remote sensing

MDPI

Article

SMMNet: A Plug-and-Play Lightweight Detection Framework for UAV Aerial Imagery

Minna Liu [1,2], Zhigang Luo [2,3,*], Yaowen Hu [3] and Jialang Liu [3]

[1] School of Computer Science, Xianyang Normal University, Xianyang 712000, China; liumn@xync.edu.cn
[2] Xiangjiang Laboratory, Changsha 410000, China
[3] School of Computer Science, National University of Defense Technology (NUDT), Changsha 410073, China; yaowenhu@nudt.edu.cn (Y.H.); liu_1999@nudt.edu.cn (J.L.)
* Correspondence: zgluo@nudt.edu.cn

Highlights

What are the main findings?

- We built SMMNet, a plug-and-play framework for UAV detection with three add-on modules. SDFE uses anisotropic diffusion to preserve boundary features during downsampling. MRCA scans in multiple directions with selective state-space models to capture long-range context efficiently. MBBR refines boxes using MobileSAM masks and filters false positives with an ELBO criterion.
- SMMNet improves YOLOv13-S by 3.6 mAP on VisDrone2019 while running at 87 FPS on A800 GPU and 30 FPS on Jetson Orin NX, suitable for real-time UAV deployment.

What are the implication of the main findings?

- SMMNet offers a practical way to balance accuracy and speed for UAV platforms with limited computing power. Users can turn each module on or off independently to fit their hardware budget, and the adaptive MBBR scheduling (full/selective/bypass mode) can be switched at runtime without retraining the model. This supports different mission needs, from real-time obstacle avoidance to offline high-precision mapping.
- The three modules (SDFE, MRCA, MBBR) can be plugged into various YOLO backbones (v5 to v13) and work consistently across three different UAV datasets. This suggests the design is not tied to a specific detector or scene type, and could be useful for improving other lightweight detectors facing similar problems with small objects, dense backgrounds, or imprecise bounding boxes.

Abstract

Object detection in UAV aerial imagery is challenged by dense small targets, large-scale variation, complex backgrounds, and strict onboard computation limits. To address these issues, this paper proposes SMMNet (Structured-diffusion Mamba Mixture Network), a lightweight plug-and-play detection framework evaluated with YOLO family detectors. SMMNet contains three modules. The Structured Diffusion Feature Extractor (SDFE) uses anisotropic diffusion to preserve boundary-sensitive features during downsampling. The Mamba-driven Receptive-field Context Aggregator (MRCA) performs multi-directional selective state-space scanning to capture long-range context with linear complexity. The Mask-guided Bayesian Box Refinement (MBBR) applies a MAP-inspired confidence-adaptive box update using MobileSAM mask evidence and ELBO-based false-positive filtering. Using YOLOv13-S as the main detector, SMMNet achieves 32.8% mAP@0.5:0.95 and 52.6% mAP@0.5 on VisDrone2019 at 87 FPS on an NVIDIA A800 GPU, improving the YOLOv13-S

Check for updates

Academic Editor: Riccardo Roncella

Received: 4 May 2026
Revised: 21 June 2026
Accepted: 24 June 2026
Published: 6 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

baseline by 3.6 and 4.5 points, respectively. The added modules reduce throughput compared with the detector-only baseline (168 FPS), but the resulting 87 FPS remains real-time and provides a favorable accuracy–latency trade-off. Three independent-seed runs further show a mean paired gain of 3.60 $\pm$ 0.10 mAP on VisDrone2019, 2.53 $\pm$ 0.12 mAP on DroneVehicle, and 2.77 $\pm$ 0.06 mAP on SeaDronesSee for the YOLOv13-S setting. Additional experiments on DroneVehicle and SeaDronesSee, together with cross-backbone evaluations on YOLOv5/v6/v7/v8/v10/v11/v13 across different UAV benchmarks, show aligned performance trends under matched settings. Edge deployment on an NVIDIA Jetson Orin NX reaches 30 FPS under TensorRT FP16 inference at 15 W TDP, indicating the suitability of SMMNet for resource-constrained UAV perception.

Keywords: UAV object detection; VisDrone; lightweight; plug-and-play; YOLO; anisotropic diffusion; state-space model; Bayesian refinement; MobileSAM

1. Introduction

With the rapid advancement of unmanned aerial vehicle (UAV) remote sensing technology, UAV platforms have been widely deployed in operational applications spanning precision agriculture, urban traffic surveillance, maritime search and rescue, and post-disaster damage assessment [1,2]. Unlike satellite or ground-based imaging systems, UAV remote sensing imagery presents a unique combination of perception challenges rooted in the physics of airborne acquisition: (i) fine-grained targets whose projected area is determined by the ground sampling distance (GSD), which varies continuously with flight altitude and camera focal length; (ii) drastic intra-scene scale variations caused by oblique viewing geometry and altitude fluctuations during dynamic flight; and (iii) heterogeneous spectral and textural interference from diverse land-cover types such as vegetation canopy, road markings, and building shadows, which substantially increase background complexity [3]. These inherent characteristics introduce substantial perception uncertainty and pose significant challenges for accurate and reliable object perception in aerial remote sensing scenes. In practical UAV missions, reliable perception performance is essential for ensuring stable system operation in complex and dynamic environments. However, maintaining consistent detection outputs remains difficult when objects are densely distributed, background interference from spectrally similar land-cover classes is severe, or the effective GSD changes rapidly due to altitude maneuvers [2]. Consequently, developing detection methods capable of preserving structural consistency and maintaining stable perception performance under challenging aerial conditions has become a critical requirement for improving the reliability and operational effectiveness of UAV remote sensing systems.

In recent years, object detection in optical UAV remote sensing imagery has experienced rapid development, largely following the evolution trajectory of general-purpose object detection frameworks. Early studies primarily focused on convolutional neural network (CNN)-based detectors, where improvements were achieved through enhanced feature extraction capability, optimized feature fusion strategies, and redesigned detection heads, enabling models to better handle complex backgrounds, significant scale variations, and densely distributed small targets in aerial scenes [4–6]. As the demand for robust perception in dense UAV environments has increased, subsequent research has increasingly emphasized small-object-oriented modeling, introducing mechanisms that strengthen spatial detail representation and refine multi-scale feature interaction, thereby improving object discrimination performance in crowded and cluttered scenarios [7,8]. Subsequently, more

advanced network architectures have been introduced into UAV detection pipelines. In particular, Transformer-based models have enabled stronger global dependency modeling by leveraging attention mechanisms, allowing effective interaction between global semantic context and fine-grained spatial features across large-scale scenes [9–11]. As a result, hybrid detection frameworks that integrate convolutional feature hierarchies with Transformer-based context modeling have become the dominant paradigm in UAV object detection and have continuously driven improvements in detection accuracy and robustness under complex aerial environments [6,11]. Meanwhile, with the emergence of large-scale pre-training strategies, foundation models have gradually been introduced into UAV visual perception systems. These approaches typically rely on extensive data pretraining and general-purpose feature representations to improve model generalization capability and cross-task adaptability in complex scenarios, showing promising performance in open-set perception and cross-domain transfer tasks [12–14]. However, in practical UAV deployment environments, such models often impose substantial computational requirements and face efficiency constraints during real-time inference. Moreover, their ability to preserve fine-grained structural details under complex background conditions remains limited [13,14]. Consequently, this study aims to address three critical challenges in UAV object detection scenarios, namely, structural information preservation, global context modeling, and prediction reliability.

More specifically, object detection in real-world UAV remote sensing scenarios still faces three fundamental structural limitations. First, at the feature representation level, repeated feature transformations and aggressive downsampling operations tend to weaken structured feature propagation, causing high-frequency boundary information of small and densely distributed targets (e.g., vehicles or pedestrians) to gradually dissipate. This degradation often results in blurred object contours and unstable localization performance in complex background environments [7,15]. Second, at the context modeling level, existing approaches typically rely either on limited receptive fields or on computationally expensive attention mechanisms, making it difficult to maintain consistent and stable global context aggregation in large-scale scenes characterized by long-range spatial dependencies [9,16]. Finally, at the prediction decision level, detection results remain vulnerable to environmental noise, occlusion, and appearance variations. Traditional deterministic bounding-box regression methods lack the capability to explicitly model predictive uncertainty, which may lead to unreliable confidence estimation and unstable detection outcomes under challenging operating conditions [17]. These unresolved challenges indicate the need for a perception framework capable of maintaining structural consistency during feature propagation, enabling efficient global context modeling, and improving prediction reliability, while simultaneously preserving real-time computational efficiency and practical deployment feasibility. Such a framework is essential for supporting robust object detection in complex real-world UAV remote sensing environments.

Based on the aforementioned structural limitations, we propose SMMNet (Structured-diffusion Mamba Mixture Network), a plug-and-play, architecture-agnostic perception framework for UAV object detection. SMMNet introduces three modular components, namely, SDFE, MRCA, and MBBR, which are designed for integration into multiple YOLO family detectors with minimal changes to the base detection pipeline. The baseline pipeline consists of a YOLO family detector and an external MobileSAM [18] mask generator. MobileSAM provides pixel-level mask evidence for MBBR but is not counted as a proposed SMMNet module. The core novelty of SMMNet resides in SDFE, MRCA, and MBBR. Primary experiments use YOLOv13 as the representative instantiation; the plug-and-play capability is evaluated across YOLOv8/v10/v11/v13 on VisDrone2019 and across YOLOv5/v6/v7/v8/v10/v11/v13 on additional UAV benchmarks. The framework is

designed to improve robustness in complex aerial remote sensing scenes while maintaining efficiency suitable for real-time deployment on resource-constrained UAV edge platforms. It achieves this by preserving structural consistency during feature propagation, strengthening global context awareness in large-scale environments, and improving prediction reliability under challenging imaging conditions. The main contributions are summarized as follows:

- Structured Diffusion Feature Extractor (SDFE) models feature propagation as a structure-guided anisotropic diffusion process to reduce structural information loss caused by aggressive downsampling in complex aerial scenes. The module employs a spatially adaptive conductance tensor to preserve boundary details and guide feature propagation along structural directions. This design improves contour representation and supports more stable localization for small and densely distributed targets in cluttered environments.
- Mamba-driven Receptive-field Context Aggregator (MRCA) is designed based on a selective state-space modeling mechanism to aggregate contextual features across large spatial regions using multi-directional sequence scanning. The module models long-range dependencies with linear computational complexity and helps maintain consistent semantic representation for spatially distributed targets in large-scale UAV scenes.
- By integrating the three modular components SDFE, MRCA, and MBBR, we construct a unified detection framework termed SMMNet and verify its plug-and-play capability across multiple YOLO backbones, including YOLOv5/v6/v7/v8/v10/v11/v13 on different UAV benchmarks. On the VisDrone2019 benchmark, the proposed method improves mAP@0.5:0.95 by 3.6 points over the YOLOv13-S baseline, from 29.2 to 32.8, and improves mAP@0.5 by 4.5 points, from 48.1 to 52.6, while maintaining real-time inference at 87 FPS on A800 and 30 FPS on an NVIDIA Jetson Orin NX edge device. This speed is lower than the detector-only YOLOv13-S baseline, so the contribution is positioned as an accuracy–latency trade-off rather than a pure FPS improvement. The framework shows performance gains of 2.4–3.6 mAP across seven YOLO backbones (v5, v6, v7, v8, v10, v11, and v13) and three diverse UAV benchmarks under matched settings, supporting the backbone-agnostic transferability of the proposed modules. To support reproducibility, we have supplemented the experimental settings with deployment configurations and detailed inference protocols.

2. Related Work

Object detection in UAV aerial imagery has attracted growing research interest owing to its practical importance and the technical difficulties it presents [19]. This section surveys three areas of related work: detection methods tailored for UAV and small-object scenarios, general lightweight detection architectures, and foundation segmentation models used to assist detection.

2.1. Object Detection in UAV Imagery

Object detection in UAV imagery has attracted considerable attention due to its importance in applications such as traffic monitoring, environmental inspection, and maritime search and rescue. Compared with natural scene images, UAV imagery typically contains complex backgrounds, significant scale variations, and a high proportion of small and densely distributed objects, which makes accurate detection particularly challenging. To address these issues, extensive research has focused on improving detection frameworks by enhancing feature representation and multi-scale modeling capability [7,20]. One line of research improves convolutional neural network (CNN)-based detectors by enhancing

multi-scale feature extraction, refining feature fusion structures, and redesigning detection heads to better capture small objects in UAV scenes [15,21–23]. In addition, several studies specifically address dense small-object detection by strengthening spatial detail representation and improving high-resolution object modeling [7,15,24], introducing detection heads tailored for small targets, and improving object modeling in densely distributed scenes [25,26]. Training strategies have also been explored to improve detection robustness. For example, dynamic training sample selection mechanisms have been proposed to alleviate the imbalance between positive and negative samples and improve feature learning in challenging UAV environments [27,28]. Beyond CNN-based detector improvements, recent studies have explored more advanced architectures and feature interaction mechanisms. Transformer-based detection frameworks have been introduced to enhance long-range feature interaction [9,10]. Complementary learning strategies have been proposed to model interactions among multi-scale features and improve detection under severe scale variations [29]. In addition, cross-modal detection approaches incorporate multi-modal attention mechanisms to enhance feature modeling in complex environments [30].

Despite these advances, preserving fine-grained structural information remains challenging. In UAV imagery, the edge structures of small objects are often weakened during deep feature extraction, leading to inaccurate localization or missed detections in dense scenes. While the aforementioned studies mainly focus on improving detection architectures within conventional frameworks, recent advances in large-scale vision foundation models provide new opportunities for UAV perception by enabling stronger feature representation and generalization capability.

2.2. Foundation Models in UAV Imagery

While traditional UAV detection methods mainly rely on task-specific detector design, recent advances in vision foundation models have improved representation learning and cross-task generalization. Large-scale pretrained architectures, such as Vision Transformers (ViT), self-supervised foundation models, and segmentation foundation models, have demonstrated strong performance across diverse visual tasks by leveraging massive training datasets and powerful feature representations [12,31,32]. These advantages are especially relevant to UAV imagery, where complex backgrounds, large-scale variations, and densely distributed small objects challenge conventional methods. Large vision encoders have been combined with task-specific decoders for remote sensing segmentation, while vision–language models (VLMs) have been explored for open-vocabulary scene understanding in UAV imagery [13,14]. More recently, promptable segmentation models such as the Segment Anything Model (SAM) [12] have provided flexible mask generation from prompts such as points or bounding boxes. These pixel-level priors can provide useful structural information for localization and boundary refinement. To reduce computational overhead, lightweight variants such as FastSAM [33] have been proposed for resource-constrained UAV platforms. In parallel, efforts have been made to develop foundation model architectures tailored for remote sensing imagery. Wang et al. [34] introduced a remote sensing-oriented vision transformer by incorporating rotated varied-size attention (RVSA), which adapts the attention mechanism to handle arbitrary object orientations and scale variations in aerial imagery. By enabling attention windows with flexible sizes, positions, and orientations, this approach improves spatial context modeling and enhances feature representation for downstream tasks such as detection and segmentation. Beyond architectural design, recent research has increasingly focused on strategies that improve model robustness and generalization in real-world UAV perception scenarios. To address distribution shifts encountered during deployment, test-time adaptation techniques have been actively explored. Representative entropy minimization-based approaches, such as

TENT [35], enable model adaptation during inference and have demonstrated effectiveness in maintaining stable performance under changing environmental conditions. In addition to deployment-time adaptation, multi-task pretraining strategies have been proposed to enhance feature generalization in remote sensing applications. For example, researchers introduced a multi-task pretraining (MTP) paradigm [36] that jointly optimizes semantic segmentation, instance segmentation, and rotated object detection within a unified framework. By sharing a common encoder and introducing task-specific decoders, MTP improves representation learning and reduces the gap between pretraining and downstream tasks. Notably, this framework has primarily been validated on remote sensing benchmarks, demonstrating strong performance across multiple aerial perception tasks. Although its transferability suggests potential benefits for UAV object detection, further validation in UAV-specific detection scenarios is still required. These advances indicate that foundation models provide transferable feature representations that can be further adapted to UAV object detection tasks requiring robust generalization under complex aerial conditions.

Despite these advances, several challenges remain. Lightweight detectors often struggle to preserve fine-grained structural details, particularly for small and densely distributed targets. Meanwhile, foundation model-based approaches typically rely on large model sizes and high computational cost, which limits their applicability in real-time UAV scenarios. To address these issues, we propose SMMNet, a plug-and-play lightweight detection framework that integrates edge enhancement, global semantic modeling, and segmentation-guided refinement. By leveraging structural cues from segmentation and enhancing long-range feature interactions, the proposed method achieves a balance between detection accuracy and computational efficiency for UAV aerial imagery.

2.3. Segmentation-Guided Detection Refinement

While Sections 2.1 and 2.2 review detection architectures and foundation models from the perspectives of feature representation and generalization, a parallel research direction uses segmentation masks as geometric evidence for refining detection outputs. These methods exploit pixel-level cues, including object boundaries, shape priors, and spatial extent, to improve localization accuracy and suppress false positives. Such cues are particularly useful in cluttered UAV scenes, where bounding-box-only representations may not capture the fine-grained contours of small or partially occluded objects [7,15].

SAM-assisted detection refinement. The Segment Anything Model (SAM) [12] and its lightweight variants have recently been integrated as post hoc refinement tools in detection pipelines. SAMRefiner [37] adapts SAM for universal mask refinement using noise-tolerant prompts and IoU-based self-boosting. For remote sensing imagery, SAM-RSIS [38] progressively adapts SAM with detector-predicted box prompts, while Det-SAM2 [39] combines YOLOv8 with SAM 2 in a self-prompting framework for automated image and video segmentation. In UAV agricultural applications, Peddagudreddygari et al. [40] combined YOLO-based cotton boll detection with SAM-based segmentation for yield prediction. Mayladan et al. [41] further showed that a pre-trained CNN can generate prompts for zero-shot SAM refinement. These studies demonstrate the value of segmentation-derived boundaries, but most of them use mask evidence deterministically and do not explicitly adapt the refinement strength to the confidence of each detection.

Mask-guided confidence calibration and box refinement. Beyond SAM-specific pipelines, mask quality has also been used for confidence estimation and bounding-box adjustment. Latent-MaskRCNN [42] models a distribution over plausible masks and selects high-confidence predictions using mask intersection IoU. IoU-aware calibration [43] estimates confidence from mask-box alignment and reduces duplicate detections. Mask-IoU regression in semi-supervised instance segmentation [44] uses mask quality as a pseudo-

label filtering signal, while boundary-aware mask heads [45] jointly predict boundaries and IoU scores for cluttered scenes. Salience DETR [46] refines detection queries through hierarchical salience filtering using only bounding-box supervision.

Relationship to MBBR. These segmentation-guided methods confirm that mask-derived geometric evidence can improve detection quality, but three limitations remain. First, many methods rely on deterministic mask scoring or salience filtering and do not formulate detection refinement as uncertainty-aware inference. Second, they usually apply a uniform correction strategy to all detections, rather than adapting the correction magnitude to prediction confidence. Third, several methods require detector retraining or architectural modification, which limits their use as drop-in components.

MBBR differs from the above work by treating MobileSAM as an external mask generator rather than as a proposed SMMNet module. The detector-predicted box defines a confidence-adaptive Gaussian prior: high-confidence detections receive a narrow prior that resists large mask-driven corrections, whereas low-confidence detections receive a broader prior that allows stronger geometric adjustment. The MobileSAM mask provides the likelihood evidence through mask-box IoU. A closed-form MAP-inspired convex update then balances the detector anchor against mask evidence, and an ELBO-based filter suppresses structurally inconsistent detections. This design allows MBBR to remain a plug-and-play post-processing module that requires neither detector retraining nor modification of the backbone architecture.

3. Proposed Methodology

3.1. Overall Architecture and Design Philosophy

We propose SMMNet (Structured-diffusion Mamba Mixture Network), a plug-and-play lightweight framework for UAV aerial object detection. SMMNet consists of three architecture-agnostic modules, namely, SDFE, MRCA, and MBBR, which are evaluated as drop-in components across multiple YOLO family detectors. The baseline pipeline comprises a YOLO family detector and MobileSAM [18] as an external mask generator. MobileSAM is not claimed as a proposed SMMNet module; it provides mask evidence only for the MBBR post-processing branch. The three novel contributions of this work are SDFE, MRCA, and MBBR, which can be inserted into the baseline detector without changing the underlying YOLO detection architecture. We adopt YOLOv13 as the primary instantiation, verify cross-backbone transferability on YOLOv8, YOLOv10, YOLOv11, and YOLOv13 for VisDrone2019, and further evaluate YOLOv5, YOLOv6, YOLOv7, YOLOv8, YOLOv10, YOLOv11, and YOLOv13 on DroneVehicle and SeaDronesSee (Tables 1–3).

Table 1. Cross-backbone generalizability on VisDrone2019-DET `val`. Values correspond to the one-decimal three-run means, and Δ denotes the mean mAP improvement under matched training and evaluation settings.

Backbone	Params (M)	GFLOPs	FPS	mAP	mAP_{50}	mAP_S	mAP_M	ΔmAP
YOLOv8-S (baseline)	11.2	28.6	142	25.6	43.2	14.8	33.5	–
YOLOv8-S + Ours	17.4	35.6	78	28.9	47.5	18.0	37.2	+3.3
YOLOv10-S (baseline)	7.8	24.5	160	27.2	45.5	15.9	35.0	–
YOLOv10-S + Ours	14.0	31.5	85	30.8	50.2	19.5	39.0	+3.6
YOLOv11-S (baseline)	9.5	21.5	175	28.5	47.0	16.8	36.3	–
YOLOv11-S + Ours	15.7	28.5	90	32.0	51.8	20.5	40.6	+3.5
YOLOv13-S (baseline)	9.8	22.3	168	29.2	48.1	17.5	37.0	–
YOLOv13-S + Ours (primary)	16.0	29.6	87	32.8	52.6	21.2	41.5	+3.6

Table 2. Comparison with state-of-the-art methods on the DroneVehicle validation set. "+ Ours" denotes inserting the proposed SDFE, MRCA, and MBBR modules into the corresponding YOLO baseline. Δ indicates absolute mAP improvement. Bold and underline mark accuracy metrics only; speed and complexity columns are reported to show the accuracy–latency trade-off.

Method	Params (M)	GFLOPs	FPS	mAP	mAP_{50}	ΔmAP
UAV/Vehicle Detectors						
Faster R-CNN [47]	41.4	180.5	12	35.2	58.5	–
Cascade R-CNN [48]	69.2	234.0	8	38.8	62.1	–
SSD [49]	24.1	91.2	75	28.4	49.3	–
RetinaNet [27]	36.4	97.6	45	36.5	60.2	–
FCOS [50]	32.1	92.4	52	38.1	62.0	–
TP-YOLOv5 [51]	8.6	18.2	88	42.3	67.5	–
Plug-and-Play: YOLO Baselines vs. SMMNet						
YOLOv5-S [52]	7.1	15.8	210	43.8	69.2	–
YOLOv5-S + Ours	13.3	23.2	102	46.2	71.8	+2.4
YOLOv6-S [53]	18.5	45.3	185	46.5	72.3	–
YOLOv6-S + Ours	24.7	53.0	88	49.0	74.8	+2.5
YOLOv7-S [54]	36.9	104.7	156	48.2	74.1	–
YOLOv7-S + Ours	43.1	112.4	74	50.6	76.5	+2.4
YOLOv8-S [55]	11.2	28.6	142	50.1	75.4	–
YOLOv8-S + Ours	17.4	36.3	78	52.5	77.8	+2.4
YOLOv10-S [56]	7.8	24.5	160	51.8	76.7	–
YOLOv10-S + Ours	14.0	32.2	82	54.2	79.0	+2.4
YOLOv11-S [57]	9.5	21.5	175	54.0	77.9	–
YOLOv11-S + Ours	15.7	29.2	88	<u>56.4</u>	<u>79.8</u>	+2.4
YOLOv13-S [58]	9.8	22.3	168	55.2	78.8	–
SMMNet-S (YOLOv13-S + Ours)	16.0	29.6	87	**57.8**	**81.2**	**+2.6**

Table 3. Comparison with state-of-the-art methods on the SeaDronesSee validation set. "+ Ours" denotes inserting SDFE, MRCA, and MBBR into the corresponding baseline. Δ indicates absolute mAP improvement. Bold and underline mark accuracy metrics only; FPS and complexity are reported as deployment trade-offs.

Method	Params (M)	GFLOPs	FPS	mAP	mAP_{50}	ΔmAP
General Object Detectors						
Faster R-CNN [4]	41.4	180.5	12	24.8	46.3	–
SSD [49]	24.1	91.2	75	18.6	35.2	–
RetinaNet [59]	36.4	97.6	45	26.3	49.8	–
FCOS [50]	32.1	92.4	52	28.5	52.6	–
Deformable DETR [60]	40.0	173.0	28	31.2	58.4	–
RT-DETR-R18 [61]	26.0	70.0	65	33.8	63.5	–
Plug-and-Play: YOLO Baselines vs. SMMNet						
YOLOv5-S [52]	7.1	15.8	210	30.2	57.4	–
YOLOv5-S + Ours	13.3	23.2	102	33.0	61.6	+2.8
YOLOv8-S [55]	11.2	28.6	142	35.4	64.8	–
YOLOv8-S + Ours	17.4	36.3	78	38.2	68.5	+2.8
YOLOv10-S [56]	7.8	24.5	160	37.6	66.2	–
YOLOv10-S + Ours	14.0	32.2	82	40.4	70.0	+2.8
YOLOv11-S [57]	9.5	21.5	175	41.3	71.1	–
YOLOv11-S + Ours	15.7	29.2	88	<u>44.0</u>	<u>75.0</u>	+2.7
YOLOv13-S [58]	9.8	22.3	168	42.5	72.8	–
SMMNet-S (YOLOv13-S + Ours)	16.0	29.6	87	**45.3**	**76.5**	**+2.8**

YOLOv13 follows the established YOLO design philosophy of combining accuracy with real-time efficiency [5]. Its backbone is built on cross-stage partial (CSP) networks [62] with re-parameterizable convolutional blocks, enabling deep feature extraction without excessive parameter growth. The backbone utilizes a streamlined CSP-Darknet architecture with improved C3K2 modules and incorporates a Spatial Pyramid Pooling Fast (SPPF) module, an efficient variant of the spatial pyramid pooling (SPP) strategy originally proposed by He et al. [63], to enable rapid multi-scale context aggregation. The neck employs a PAN-FPN structure [64] that fuses low-level spatial features with high-level semantic features, which is critical for UAV imagery where targets range from a few pixels to hundreds of pixels. The decoupled detection head [65] separates classification and regression into dedicated subnetworks and generates predictions at three scales ($8\times$, $16\times$, $32\times$ downsampled feature maps) [66]. We complement the detector with MobileSAM [18], a mobile-friendly adaptation of the Segment Anything Model [12], whose instance masks serve as geometric evidence for the MBBR refinement stage.

All three modules use standard primitives (convolutions, batch normalization, SiLU activations) that map efficiently onto hardware accelerators found on UAV edge platforms such as the NVIDIA Jetson series. Each module can be enabled or disabled independently, allowing practitioners to balance accuracy against inference speed according to the target platform's computational budget. The overall architecture is shown in Figure 1.

3.2. Structured Diffusion Feature Extractor (SDFE)

Motivation (Bottleneck 1). Backbone downsampling via strided convolutions attenuates all spatial frequencies uniformly. For small aerial targets occupying only a few pixels, this uniform attenuation is particularly harmful: high-frequency boundary energy dissipates at the same rate as low-frequency background structure, and edge-defining feature patterns largely disappear before the detection head is reached. The Structured Diffusion Feature Extractor (SDFE) addresses this by treating feature propagation as an anisotropic PDE, constructing the diffusion tensor from deterministic structure-tensor decomposition, and learning only the conductance and gating parameters that control the correction strength. Unlike fixed Sobel-filter injection [67], SDFE learns diffusion conductance directly from the data, yielding a differentiable mechanism for boundary preservation. As shown in Figure 2, SDFE consists of three tightly coupled components: anisotropic diffusion modeling, structure-aware tensor construction, and adaptive gating-based residual fusion. Specifically, the module first models feature propagation as an anisotropic diffusion process, where spatial feature evolution is governed by a diffusion partial differential equation that preserves structure-relevant gradients while suppressing homogeneous-region noise. A structure tensor is then computed from local feature gradients to capture directional edge information, followed by eigendecomposition to obtain principal diffusion orientations and their corresponding eigenvalues. Let λ_1 and λ_2 denote the eigenvalues of the structure tensor, while μ_1 and μ_2 denote the diffusion coefficients along the corresponding principal directions. Based on these structural cues, a diffusion tensor is constructed to enable direction-aware smoothing along edges while inhibiting diffusion across object boundaries. To ensure stable and flexible feature refinement, the continuous diffusion process is discretized using a forward Euler scheme following standard finite-difference formulations [68], producing an intermediate diffused feature representation. In parallel, an adaptive gating mechanism estimates a spatially varying gating coefficient from gradient statistics, which controls the strength of diffusion correction in different regions. Finally, the original features and the diffused features are fused through a residual gating operation, yielding the refined output feature map with enhanced boundary preservation and structural consistency.

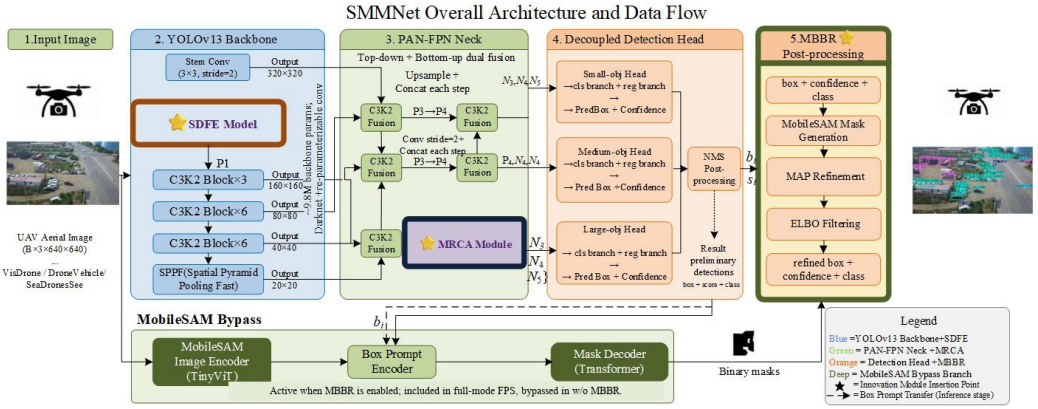


Figure 1. Overall architecture of SMMNet built on YOLOv13 and an external MobileSAM mask generator. The SDFE module is inserted after the initial downsampling stage of the backbone to inject anisotropic diffusion-corrected boundary features. The MRCA module is placed at the end of the PAN-FPN neck to aggregate global context via multi-directional selective state-space scans. The MBBR module operates as a post-detection refinement stage, using MobileSAM masks and a MAP-inspired geometric update to tighten boxes and suppress false positives. Full-mode FPS includes this MBBR/MobileSAM branch, whereas the "w/o MBBR" speed setting bypasses it. Detailed mathematical formulations are provided in Figures 2–4.

Anisotropic Diffusion Formulation. Let $\mathbf{F} : \Omega \to \mathbb{R}^C$ denote the feature map defined on the spatial domain $\Omega \subset \mathbb{R}^2$, with C channels. We model boundary-preserving feature refinement as the forward evolution of the anisotropic diffusion PDE [69]:

$$\frac{\partial \mathbf{F}}{\partial t} = \mathrm{div}(\mathbf{D}(\mathbf{F}) \nabla \mathbf{F}), \tag{1}$$

where $\mathrm{div}(\cdot)$ is the spatial divergence, ∇ the spatial gradient operator, and $\mathbf{D}(\mathbf{F}) \in \mathbb{R}^{2\times 2}$ a symmetric positive-definite diffusion tensor. The tensor $\mathbf{D}$ controls the direction and magnitude of feature spreading: strong diffusion is permitted parallel to edges (smoothing within homogeneous regions) but inhibited perpendicular to edges (preventing boundary blurring). Following structure-tensor analysis, we compute the local structure tensor:

$$\mathbf{J}_\rho(\mathbf{F}) = G_\rho * \left(\nabla \bar{\mathbf{F}} \, (\nabla \bar{\mathbf{F}})^\top \right), \tag{2}$$

where $\bar{\mathbf{F}} = \sum_{c=1}^{C} \mathbf{F}_c / C$ is a channel-averaged projection, G_ρ is a Gaussian kernel with integration scale ρ, and $*$ denotes spatial convolution. The eigendecomposition $\mathbf{J}_\rho = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^\top$ yields eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ (aligned with and orthogonal to local edges, respectively) and eigenvalues $\lambda_1 \geq \lambda_2 \geq 0$. The diffusion tensor is constructed as:

$$\mathbf{D}(\mathbf{F}) = \mu_1 \, \mathbf{v}_1 \mathbf{v}_1^\top + \mu_2(\mathbf{J}_\rho) \, \mathbf{v}_2 \mathbf{v}_2^\top, \tag{3}$$

Here, μ_1 and μ_2 denote diffusion coefficients along the principal directions defined by the eigenvectors, controlling the strength of feature smoothing parallel and perpendicular to local edges, respectively.

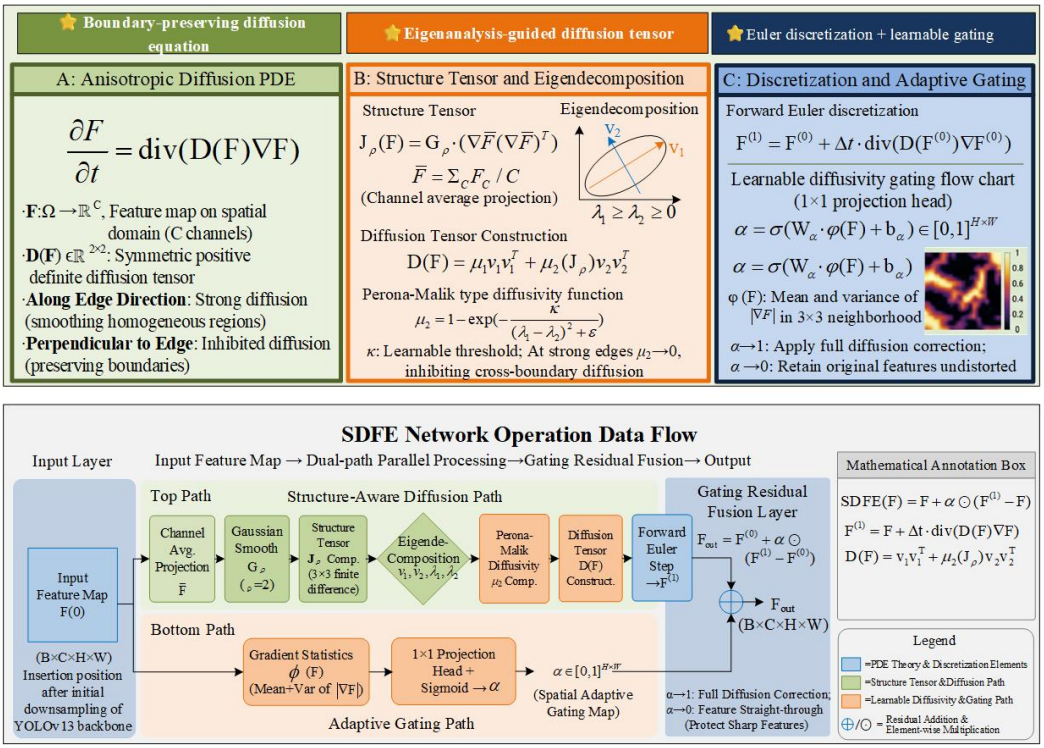


Figure 2. Overall architecture and operational workflow of the Structured Diffusion Feature Extractor (SDFE).

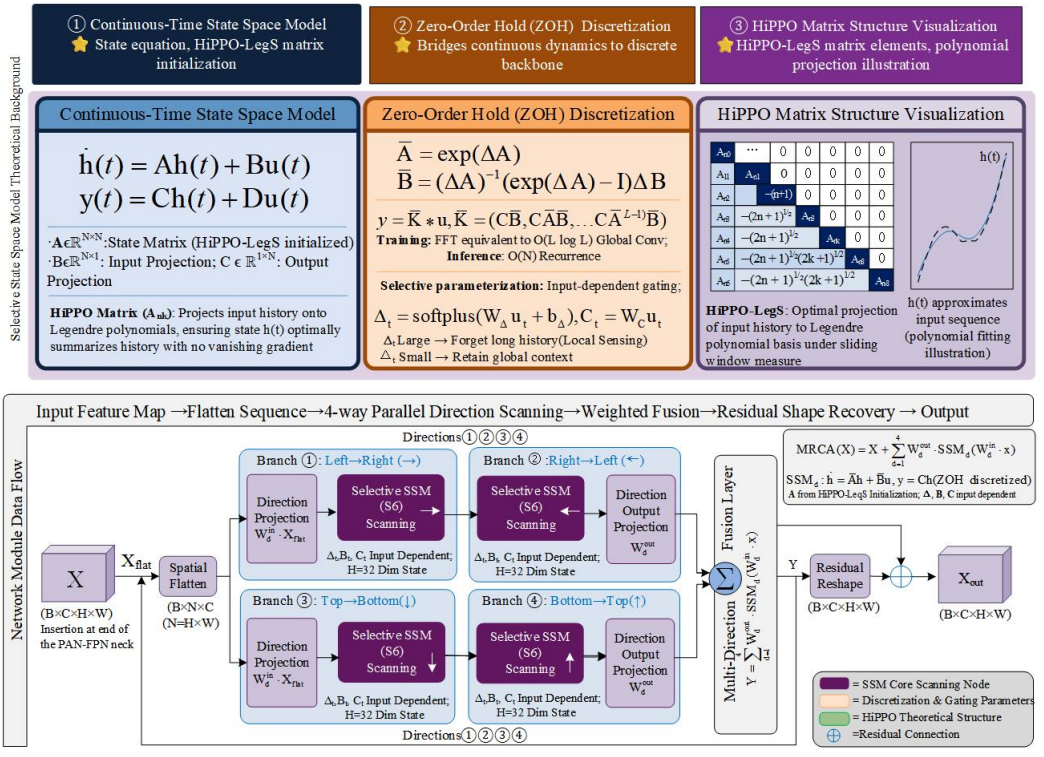


Figure 3. Overall architecture and multi-directional context aggregation process of the proposed Mamba-driven Receptive-field Context Aggregator (MRCA). Different arrow styles indicate distinct information flows: solid arrows for forward propagation, dashed arrows for residual connections, and dotted arrows for gating mechanisms. Light purple boxes denote learnable parameters.

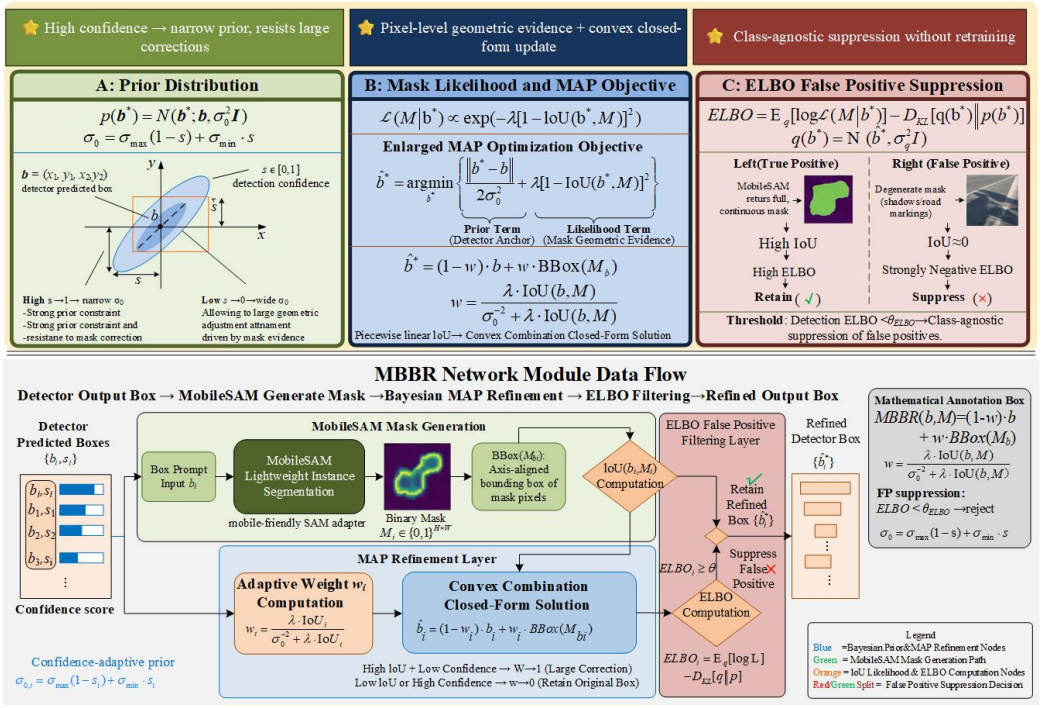


Figure 4. Mask -guided Bayesian Box Refinement (MBBR). The refinement is formulated as a Bayesian MAP inference problem: (A) a confidence-adaptive Gaussian prior is placed over the detector-predicted box, with variance inversely proportional to detection confidence; (B) the MobileSAM-generated mask serves as pixel-level geometric evidence, and a MAP objective balances the prior anchor against the mask likelihood; (C) an ELBO-based filter evaluates the evidence lower bound to suppress structurally inconsistent detections without retraining the backbone.

We set $\mu_1 = 1$ for free diffusion along edges, and define the cross-edge conductance using a learnable Perona–Malik-inspired diffusivity function [69]:

$$\mu_2(\mathbf{J}_\rho) = 1 - \exp\left(-\frac{\kappa}{(\lambda_1 - \lambda_2)^2 + \epsilon}\right), \tag{4}$$

with $\kappa > 0$ a learnable scalar threshold and $\epsilon > 0$ a numerical stability constant. In homogeneous image regions $\lambda_1 \approx \lambda_2$, so $\mu_2 \approx 1$ and diffusion is isotropic. At strong edges $\lambda_1 \gg \lambda_2$, so $\mu_2 \to 0$ and cross-edge diffusion is fully suppressed.

To clarify the learnable and computed parts of SDFE, the structure tensor $\mathbf{J}_\rho$, its eigenvectors $(\mathbf{v}_1, \mathbf{v}_2)$, and eigenvalues (λ_1, λ_2) are computed deterministically from the current feature map using local gradients and eigendecomposition. The diffusion tensor $\mathbf{D}(\mathbf{F})$ is then assembled from these computed structural directions. The learned quantities are the conductance threshold κ in Equation (4) and the parameters of the 1×1 projection head used to produce the spatial gate $\boldsymbol{\alpha}$ in Equation (6). Therefore, the 1×1 projection does not replace the tensor decomposition; it only controls how strongly the diffusion-corrected feature $\mathbf{F}^{(1)}$ is blended back into the original feature through Equation (7).

Discrete Convolutional Implementation. We discretize Equation (1) via a single Forward Euler step [68] with step size Δt:

$$\mathbf{F}^{(1)} = \mathbf{F}^{(0)} + \Delta t \, \mathrm{div}\Big(\mathbf{D}(\mathbf{F}^{(0)}) \, \nabla \mathbf{F}^{(0)}\Big), \tag{5}$$

approximated with a standard 3×3 finite-difference stencil. To maintain differentiability in training, the conductance computation is implemented as a lightweight 1×1 projection

head that maps local gradient statistics $\phi(\mathbf{F})$ (mean and variance of $|\nabla \mathbf{F}|$ over a 3×3 neighborhood) to a spatially adaptive gating map:

$$\boldsymbol{\alpha} = \sigma(\mathbf{W}_\alpha \, \phi(\mathbf{F}) + \mathbf{b}_\alpha) \in [0,1]^{H \times W}. \tag{6}$$

The SDFE output is a gated residual blend:

$$\mathbf{F}_{\text{out}} = \mathbf{F}^{(0)} + \boldsymbol{\alpha} \odot \left(\mathbf{F}^{(1)} - \mathbf{F}^{(0)} \right), \tag{7}$$

where $\odot$ is element-wise multiplication. Where $\boldsymbol{\alpha} \to 1$, the full diffusion correction is applied; where $\boldsymbol{\alpha} \to 0$, the input passes through unchanged, so already-sharp features are not degraded. SDFE introduces approximately 0.4M parameters and 0.8 GFLOPs of overhead.

3.3. Mamba-Driven Receptive-Field Context Aggregator (MRCA)

Motivation (Bottleneck 2). In UAV scenes, objects of the same category are often distributed across large spatial extents. Local convolutional receptive fields cannot capture the inter-object relationships needed to resolve category ambiguities, and full self-attention incurs $O(N^2)$ complexity that is impractical at the resolutions required for real-time inference. The Mamba-driven Receptive-field Context Aggregator (MRCA) addresses this by using Selective State-Space Models (S6) [16], which achieve globally receptive context aggregation at $O(N)$ cost by framing the problem as a content-dependent sequence scan. This linear complexity arises from the recurrent state-space formulation, which processes the sequence in a single forward pass without explicit pairwise interactions. As illustrated in Figure 3, the proposed Mamba-driven Receptive-field Context Aggregator (MRCA) adopts a sequence-based global context modeling framework built upon selective state-space models. The module first transforms the input feature map into a flattened spatial sequence, enabling efficient long-range dependency modeling through structured sequential processing. To capture spatial interactions from multiple orientations, MRCA employs a four-way parallel directional scanning strategy, including left-to-right, right-to-left, top-to-bottom, and bottom-to-top propagation paths. Each directional branch performs selective state-space scanning with input-dependent parameterization, allowing the model to adaptively regulate memory retention and context integration according to the input content. The outputs from different directional branches are aggregated through a weighted multi-direction fusion mechanism. This operation constructs an isotropic global receptive field while preserving directional consistency. Finally, the fused sequence representation is reshaped back into the original spatial structure through a residual reconstruction operation, producing the refined feature map with enhanced global semantic consistency and efficient computational complexity.

Continuous-Time State-Space Foundation. MRCA builds on a linear time-invariant (LTI) state-space model (SSM) that maps a scalar input signal $u(t)$ to output $y(t)$ via a latent state $\mathbf{h}(t) \in \mathbb{R}^N$:

$$\dot{\mathbf{h}}(t) = \mathbf{A}\,\mathbf{h}(t) + \mathbf{B}\,u(t), \qquad y(t) = \mathbf{C}\,\mathbf{h}(t) + D\,u(t), \tag{8}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{B} \in \mathbb{R}^{N \times 1}$, $\mathbf{C} \in \mathbb{R}^{1 \times N}$, $D \in \mathbb{R}$. The state matrix $\mathbf{A}$ is initialized with the HiPPO-LegS construction [70], which projects the input history onto Legendre polynomials under a sliding-window measure:

$$A_{nk} = -\begin{cases} (2n+1)^{1/2}(2k+1)^{1/2} & n > k \\ n+1 & n = k, \\ 0 & n < k \end{cases} \tag{9}$$

ensuring that $\mathbf{h}(t)$ encodes an optimal polynomial summary of all past inputs without gradient vanishing or explosion.

Zero-Order Hold Discretization. To integrate with a discrete convolutional backbone, we discretize Equation (8) using the Zero-Order Hold (ZOH) rule with step size Δ:

$$\overline{\mathbf{A}} = e^{\Delta \mathbf{A}}, \qquad \overline{\mathbf{B}} = (\Delta \mathbf{A})^{-1}\Big(e^{\Delta \mathbf{A}} - \mathbf{I}\Big)\Delta \mathbf{B}, \tag{10}$$

yielding the discrete recurrence $\mathbf{h}_t = \overline{\mathbf{A}}\,\mathbf{h}_{t-1} + \overline{\mathbf{B}}\,u_t$, $y_t = \mathbf{C}\,\mathbf{h}_t + D\,u_t$. Because $\overline{\mathbf{A}}$ has diagonal-plus-low-rank structure, the full output sequence is equivalent to a causal convolution $\mathbf{y} = \overline{\mathbf{K}} * \mathbf{u}$ with kernel $\overline{\mathbf{K}} = (\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{A}\mathbf{B}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{L-1}\overline{\mathbf{B}})$, computed in $O(L \log L)$ via FFT during training.

Selective Input-Dependent Parameterization. To achieve content-dependent gating analogous to attention, MRCA introduces an input-selective parameterization in which Δ, $\mathbf{B}$, and $\mathbf{C}$ at each step t are functions of the current token u_t:

$$\Delta_t = \mathrm{softplus}(\mathbf{W}_\Delta\, u_t + b_\Delta), \quad \mathbf{B}_t = \mathbf{W}_B\, u_t, \quad \mathbf{C}_t = \mathbf{W}_C\, u_t, \tag{11}$$

where $\mathbf{W}_\Delta, \mathbf{W}_B, \mathbf{W}_C$ are learnable. A large Δ_t accelerates state decay, biasing the model toward local context; a small Δ_t retains long-range history. Through this mechanism, the model selectively gathers semantic evidence for ambiguous tokens and discards irrelevant background, providing attention-like context modeling capability at linear computational cost.

Multi-directional Scan Strategy. To avoid directional bias from a single raster scan, we apply four independent selective scans in four spatial traversal orders (left-to-right, right-to-left, top-to-bottom, bottom-to-top) and merge their outputs:

$$\mathbf{Y} = \sum_{d=1}^{4} \mathbf{W}_d^{\mathrm{out}}\, \mathrm{SSM}_d\Big(\mathbf{W}_d^{\mathrm{in}}\, \mathbf{X}\Big), \tag{12}$$

where $\mathbf{X} \in \mathbb{R}^{N \times C}$ is the flattened input and $\mathbf{W}_d^{\mathrm{in}}, \mathbf{W}_d^{\mathrm{out}}$ are direction-specific projections. The merged output is reshaped to the original spatial dimensions and added to $\mathbf{X}$ via a residual connection, producing isotropically receptive global context at $O(N)$ cost, consistent with the linear-time complexity of selective state-space models. MRCA introduces approximately 0.6M additional parameters and 1.1 GFLOPs of computational overhead.

3.4. Mask-Guided Bayesian Box Refinement (MBBR)

Motivation (Bottleneck 3). Detection heads minimize a deterministic regression loss with no structural prior on object shape. In densely packed UAV scenes, the resulting boxes frequently enclose neighboring objects or background, or clip object boundaries. The Mask-guided Bayesian Box Refinement (MBBR) module addresses this issue through a practical MAP-inspired refinement procedure, using MobileSAM instance masks as pixel-level geometric evidence for box correction and false-positive suppression. Figure 4 illustrates the refinement pipeline. The module begins with detector-predicted bounding boxes, which are used as prompts to generate instance masks via a lightweight MobileSAM segmentation model. These masks provide pixel-level evidence describing the spatial extent and structural consistency of detected objects. Based on the generated masks,

an axis-aligned bounding box is extracted from mask pixels, and an Intersection-over-Union (IoU) score is computed to quantify the geometric agreement between the predicted box and the mask-derived region. Because the IoU-based objective is non-smooth with respect to box coordinates, we adopt a data-adaptive convex update inspired by the MAP objective rather than relying on an exact analytic optimizer. To further improve robustness, an Evidence Lower Bound (ELBO)-based filtering mechanism is applied to identify and suppress structurally inconsistent detections without retraining the backbone network. The final output is a set of geometrically consistent bounding boxes with improved localization accuracy and structural reliability.

Prior Distribution over Box Coordinates. Let $\mathbf{b} = (x_1, y_1, x_2, y_2) \in \mathbb{R}^4$ denote the detector's predicted box. We model the prior over the true box $\mathbf{b}^*$ as a Gaussian anchored to the prediction:

$$p(\mathbf{b}^*) = \mathcal{N}\Big(\mathbf{b}^*;\, \mathbf{b},\, \sigma_0^2\, \mathbf{I}\Big), \tag{13}$$

where the prior variance is confidence-adaptive: $\sigma_0 = \sigma_{\max}(1-s) + \sigma_{\min}\, s$ with detection confidence $s \in [0,1]$. High-confidence detections carry narrow priors that resist large mask-driven corrections. Low-confidence detections receive broad priors that allow more substantial geometric adjustment.

Mask Likelihood. Given $\mathbf{b}$, MobileSAM is invoked with $\mathbf{b}$ as a box prompt to produce a binary segmentation mask $\mathbf{M} \in \{0,1\}^{H \times W}$. The likelihood of $\mathbf{M}$ under a candidate refined box $\mathbf{b}^*$ is modeled using an IoU-based exponential likelihood formulation commonly adopted in detection refinement and geometric alignment tasks [71]:

$$\mathcal{L}(\mathbf{M} \mid \mathbf{b}^*) \propto \exp\Big(-\lambda \left[1 - \mathrm{IoU}(\mathbf{b}^*, \mathbf{M})\right]^2\Big), \tag{14}$$

where $\lambda > 0$ is a temperature parameter and $\mathrm{IoU}(\mathbf{b}^*, \mathbf{M}) = |\mathbf{b}^* \cap \mathbf{M}| / |\mathbf{b}^* \cup \mathbf{M}|$ measures pixel-level agreement. Boxes that closely align with the mask boundary yield high likelihood values, whereas boxes enclosing excessive background or clipping the object receive low likelihood.

MAP Refinement. Applying Bayes' rule, the log-posterior objective is:

$$\hat{\mathbf{b}}^* = \arg\min_{\mathbf{b}^*} \underbrace{\frac{\|\mathbf{b}^* - \mathbf{b}\|^2}{2\sigma_0^2}}_{\text{prior (detector anchor)}} + \underbrace{\lambda \left[1 - \mathrm{IoU}(\mathbf{b}^*, \mathbf{M})\right]^2}_{\text{likelihood (mask evidence)}}. \tag{15}$$

The prior term keeps $\hat{\mathbf{b}}^*$ close to the detector's output; the likelihood term pulls it toward the mask. Directly optimizing Equation (15) is difficult because the IoU term is non-smooth and depends on discrete mask pixels. Therefore, MBBR uses the following practical data-adaptive convex update as a MAP-inspired approximation:

$$\hat{\mathbf{b}}^* = (1 - w)\, \mathbf{b} + w\, \mathrm{BBox}(\mathbf{M}_\mathbf{b}), \quad w = \frac{\lambda\, \mathrm{IoU}(\mathbf{b}, \mathbf{M})}{\sigma_0^{-2} + \lambda\, \mathrm{IoU}(\mathbf{b}, \mathbf{M})}, \tag{16}$$

where $\mathbf{M}_\mathbf{b} = \{(i,j) : \mathbf{M}(i,j) = 1\} \cap \mathbf{b}$ is the set of mask pixels within the predicted box and $\mathrm{BBox}(\cdot)$ is their axis-aligned bounding rectangle. When mask agreement is high and confidence is low, $w \to 1$ and the correction magnitude increases; when agreement is weak or confidence is high, $w \to 0$ and the original box is kept.

False-Positive Suppression via ELBO. To reject spurious detections, MBBR evaluates the Evidence Lower Bound (ELBO) [72] of the variational approximation $q(\mathbf{b}^*) = \mathcal{N}(\hat{\mathbf{b}}^*, \sigma_q^2 \mathbf{I})$:

$$\mathrm{ELBO} = \mathbb{E}_q[\log \mathcal{L}(\mathbf{M} | \mathbf{b}^*)] - D_{\mathrm{KL}}[q(\mathbf{b}^*) \,\|\, p(\mathbf{b}^*)], \tag{17}$$

where D_{KL} is the Kullback–Leibler divergence. For a genuine detection, the mask **M** is geometrically coherent with **b**, the likelihood is large, and the ELBO is correspondingly high. For a false positive such as a rooftop shadow or road marking, MobileSAM returns a degenerate or disjoint mask, IoU is near zero, and the ELBO is strongly negative. Detections below a threshold θ_{ELBO} are suppressed, giving a class-agnostic false-positive filter that requires no per-class confidence tuning. MBBR introduces approximately 0.2M additional parameters and 0.4 GFLOPs of computational overhead.

4. Experiments

4.1. Experimental Setup

Datasets. We use the VisDrone2019-DET dataset [3] as the primary benchmark for UAV aerial object detection. VisDrone2019 contains 6471 training images and 548 validation images. To further verify generalizability while maintaining consistent task definitions, we conduct dataset-specific training and evaluation on two complementary UAV benchmarks: (1) the DroneVehicle dataset [73], which consists of 56,878 image pairs captured across five vehicle categories; and (2) the SeaDronesSee dataset [74], which covers maritime search-and-rescue targets. This setting keeps the category definitions and imaging conditions within UAV-oriented detection scenarios.

Metrics. We report mean Average Precision at IoU 0.5:0.95 (mAP@0.5:0.95) and at IoU 0.5 (mAP@0.5), along with inference speed in FPS and model size in millions of parameters (M). For VisDrone2019, we also report mAP_S (small objects, area $< 32^2$), mAP_M (medium objects, $32^2 <$ area $< 96^2$), and mAP_L (large objects, area $> 96^2$).

Implementation Details. Input resolution is 640 × 640. We adopt the stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 5×10^{-4}. A cosine learning rate schedule is used with linear warm-up. The batch size is set to 16, and all models are trained for 300 epochs on 2× NVIDIA A800 GPUs. Standard YOLO-style data augmentations are applied, including mosaic, mixup, random horizontal flip, and color jitter. For reproducibility, all experiments follow a YOLO-style training and evaluation pipeline built upon the publicly released YOLOv13 implementation [58]. Unless otherwise specified, validation adopts a confidence threshold of 0.001, an NMS IoU threshold of 0.70, and a maximum of 300 detections per image. FPS is measured with batch size 1 after warm-up and includes image preprocessing, network forward propagation, detection decoding, NMS, and the MBBR/MobileSAM pathway when MBBR is enabled; data loading and disk I/O are excluded. In practice, MBBR is applied only after NMS and is limited to the top-K retained detections per image (we set $K = 100$ by default); all box prompts for a given image are processed in a single-batched MobileSAM call to prevent the refinement cost from scaling with the number of raw predictions. The MobileSAM image encoder output is computed once per image and reused across all box prompts. The MobileSAM branch is invoked only for MBBR refinement and is disabled in the "w/o MBBR" speed setting. The released code configuration will record the exact software versions used for the final reported runs: PyTorch 2.2.0, CUDA 12.8, cuDNN 8.9.7, Ultralytics 8.2.0, ONNX 1.16.0, and TensorRT 8.6.1. versions used for the final reported runs.

Modules SDFE, MRCA, and MBBR are progressively inserted in ablation studies to assess their individual and combined effectiveness. MobileSAM is used only for MBBR mask-guided refinement.

Reproducibility and statistical analysis. To improve reproducibility, we now report the training schedule, augmentation policy, confidence threshold, NMS IoU threshold, maximum detections per image, warm-up protocol, and MobileSAM invocation strategy in detail. In response to the reviewers' statistical concerns, the main YOLOv13-S comparisons and key VisDrone2019 ablations were rerun with three independent random seeds {0, 1, 2}.

We report mean, standard deviation, 95% confidence interval (CI), paired t-test, Wilcoxon signed-rank test, and Benjamini–Hochberg false-discovery-rate (FDR) correction where applicable. Because the recorded metrics are rounded to one decimal place, some paired differences have zero rounded variance; in those cases, the paired t-test is not estimable, and we report the paired CI and exact non-parametric p-value without overstating significance. Tables 4 and 5 and Figure 5 summarize these repeated-run results.

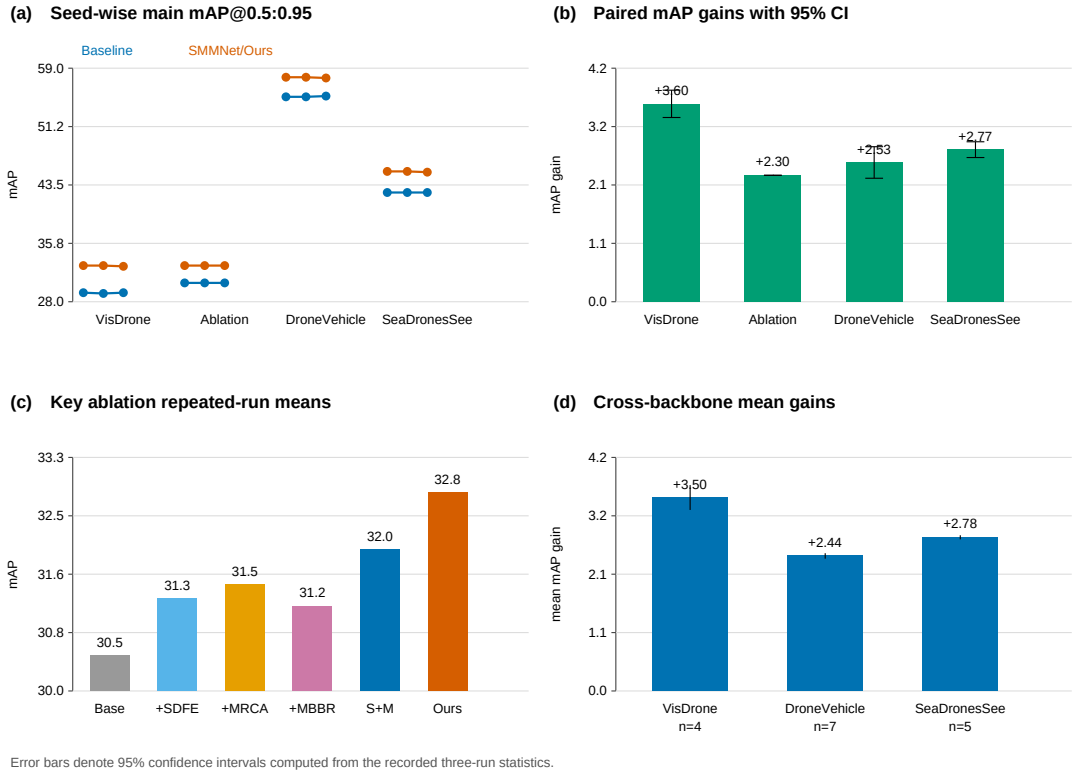


Figure 5. Repeated-run statistical analysis. (**a**) Seed-wise mAP@0.5:0.95 values for the main YOLOv13-S comparisons. (**b**) Paired mAP gains with 95% CIs. (**c**) Three-run means for key VisDrone2019 ablation configurations. (**d**) Cross-backbone mean gains, where each paired unit is one YOLO backbone and the value for each backbone is the mean gain over seeds {0, 1, 2}.

Edge Device Deployment. To assess the practical deployability on resource-constrained UAV platforms, we additionally benchmark SMMNet-S on three representative NVIDIA Jetson edge devices. Models are exported from PyTorch to ONNX and then optimized using TensorRT 8.6 with FP16 precision. Inference latency is measured as the average over 1000 frames at 640×640 resolution after a 100-frame warm-up, with batch size 1. The TensorRT path includes the detector backbone, SDFE, MRCA, decoding, NMS, and the complete MBBR post-processing path when full SMMNet is reported. For the "w/o MBBR" setting, only the MobileSAM-based refinement path is disabled; the SDFE and MRCA modules remain enabled. Jetson measurements are performed under the listed TDP modes with the same confidence and NMS thresholds as GPU evaluation. Table 6 summarizes the results.

Table 4. Three-run statistical summary for the main SMMNet comparisons. Values are mean ± standard deviation over seeds {0, 1, 2}. Δ is paired treatment–baseline mAP@0.5:0.95; CI is the 95% confidence interval of the paired difference. "n.e." denotes a paired *t*-test not estimable because the rounded paired differences have zero variance.

Comparison	Dataset	Baseline mAP	SMMNet mAP	ΔmAP	95% CI	p_t/p_W
VisDrone	VisDrone2019-DET	29.17 ± 0.06	32.77 ± 0.06	3.60 ± 0.10	[3.35, 3.85]	<0.001/0.250
Ablation	VisDrone2019-DET	30.50 ± 0.00	32.80 ± 0.00	2.30 ± 0.00	[2.30, 2.30]	n.e./0.250
DroneVehicle	DroneVehicle	55.23 ± 0.06	57.77 ± 0.06	2.53 ± 0.12	[2.25, 2.82]	<0.001/0.250
SeaDronesSee	SeaDronesSee	42.50 ± 0.00	45.27 ± 0.06	2.77 ± 0.06	[2.62, 2.91]	<0.001/0.250

Table 5. Repeated-run summary of the key VisDrone2019 ablation rows. All rows were retrained with seeds {0, 1, 2}. Δ is computed relative to the YOLOv13-S + MobileSAM baseline. "n.e." = not estimable: paired *t*-test undefined due to zero rounded variance.

Configuration	mAP	mAP_{50}	mAP_S	mAP_M	ΔmAP	p_t/p_W
Baseline (YOLOv13-S + MobileSAM)	30.50 ± 0.00	49.80 ± 0.00	18.63 ± 0.06	38.23 ± 0.06	–	–
+SDFE	31.30 ± 0.00	50.73 ± 0.06	19.50 ± 0.00	39.00 ± 0.00	0.80 ± 0.00	n.e./0.250
+MRCA	31.50 ± 0.00	51.00 ± 0.00	19.80 ± 0.00	39.30 ± 0.00	1.00 ± 0.00	n.e./0.250
+MBBR	31.20 ± 0.00	50.50 ± 0.00	19.33 ± 0.06	38.77 ± 0.06	0.70 ± 0.00	n.e./0.250
SDFE + MRCA	32.00 ± 0.00	51.60 ± 0.00	20.30 ± 0.00	40.10 ± 0.00	1.50 ± 0.00	n.e./0.250
SDFE + MBBR	31.80 ± 0.00	51.27 ± 0.06	20.00 ± 0.00	39.80 ± 0.00	1.30 ± 0.00	n.e./0.250
MRCA + MBBR	31.90 ± 0.00	51.50 ± 0.00	20.10 ± 0.00	39.90 ± 0.00	1.40 ± 0.00	n.e./0.250
Ours (SDFE + MRCA + MBBR)	32.80 ± 0.00	52.60 ± 0.00	21.20 ± 0.00	41.50 ± 0.00	2.30 ± 0.00	n.e./0.250

Table 6. Edge-device inference benchmark of SMMNet-S (TensorRT FP16, 640 × 640, batch 1). "w/o MBBR" denotes disabling the MobileSAM-based refinement pathway.

Device (Manufacturer; City, Country)	TDP (W)	FPS	FPS w/o MBBR	Latency (ms)
Jetson Orin NX 8 GB (NVIDIA; Santa Clara, CA, USA)	15	30	41	33.3
Jetson Xavier NX 8 GB (NVIDIA; Santa Clara, CA, USA)	15	20	27	50.0
Jetson Orin Nano 4 GB (NVIDIA; Santa Clara, CA, USA)	7	13	18	76.9
A800 GPU (reference) (NVIDIA; Santa Clara, CA, USA)	300	87	108	11.5

On the Jetson Orin NX, a representative compute module for industrial-grade UAV platforms, SMMNet-S achieves 30 FPS (33.3 ms latency) with all modules enabled, comfortably exceeding the 25 FPS threshold commonly adopted for real-time UAV perception. Disabling the MBBR refinement pathway further increases throughput to 41 FPS, providing a lower-latency configuration suitable for time-sensitive tasks such as obstacle avoidance or real-time tracking. Even on the ultra-low-power Jetson Orin Nano (7 W TDP), the model sustains 13 FPS with full MBBR and 18 FPS without it, which remains appropriate for offline high-precision mapping and surveying tasks where frame-level real-time constraints are relaxed. These results indicate that the proposed modules, despite incorporating PDE-based diffusion and state-space sequence modeling, introduce computational overhead that remains manageable on edge-class accelerators and does not hinder practical UAV deployment.

Table 7 provides a coarse latency decomposition on the A800 GPU. The MobileSAM/MBBR overhead is computed from the difference between the full SMMNet setting and the MBBR-bypass setting. This breakdown clarifies that SMMNet improves accuracy at the cost of reduced throughput relative to the detector-only YOLOv13-S baseline; the intended claim is a practical accuracy–latency trade-off, not a higher FPS than the unmodified detector.

Table 7. Coarse A800 latency breakdown at 640 × 640 resolution and batch size 1. Detector-only latency is derived from YOLOv13-S FPS, and the MBBR/MobileSAM overhead is derived from the difference between the full and bypass modes.

Configuration/Stage	Latency (ms)	FPS
YOLOv13-S detector only	5.95	168
YOLOv13-S + MobileSAM baseline	8.33	120
SDFE + MRCA with MBBR bypass	9.26	108
MobileSAM/MBBR refinement overhead	+2.23	–
Full SMMNet-S	11.49	87

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

To evaluate the effectiveness of SMMNet, we conduct comparisons across three UAV aerial detection benchmarks that differ in scene domain, target category, altitude range, and background complexity: VisDrone2019 (dense urban pedestrian and vehicle detection), DroneVehicle (vehicle-centric multi-illumination aerial imagery), and SeaDronesSee (maritime search-and-rescue). This multi-dataset evaluation is now supported by three-seed statistics for the main YOLOv13-S setting and by cross-backbone paired analyses using the three-run mean of each backbone as the paired unit.

VisDrone2019 Dataset

VisDrone2019 serves as the primary benchmark, featuring multi-category dense crowds and extreme small-object distributions. We compare against: (i) UAV-specific detectors (TPH-YOLOv5, ClusDet, GLSAN, DMNet, QueryDet, UFPMP-Det, CEASC); (ii) general-purpose YOLO variants (YOLOv8-S, YOLOv10-S, YOLOv11-S, YOLOv13-S); and (iii) transformer-based RT-DETR-R18. Full COCO-style metrics including mAP@0.5:0.95, mAP@0.5, mAP@0.75, mAP_S, mAP_M, and mAP_L are reported in Table 8.

Table 8. Comparison with state-of-the-art methods on the VisDrone2019-DET val set. Bold are used only for accuracy metrics, because Params, GFLOPs, and FPS represent deployment trade-offs rather than accuracy rankings.

Method	Params (M)	GFLOPs	FPS	mAP	mAP_{50}	mAP_{75}	mAP_S	mAP_M	mAP_L
UAV-Specific Detectors									
TPH-YOLOv5-S [51]	8.1	16.8	95	23.4	39.6	24.1	13.2	31.8	38.5
ClusDet [7]	41.2	112.0	18	26.7	44.3	27.5	16.9	35.1	41.3
GLSAN [75]	36.5	95.0	22	25.8	43.0	26.2	15.6	34.2	40.8
DMNet [76]	33.8	88.0	25	27.1	45.0	28.0	17.3	35.6	42.0
QueryDet [24]	44.0	125.0	15	28.3	46.5	29.2	18.5	36.8	43.2
UFPMP-Det [77]	38.0	101.0	20	27.8	45.8	28.7	17.8	36.2	42.5
CEASC [78]	29.3	68.0	35	28.0	46.2	28.9	18.1	36.5	42.8
General-Purpose YOLO Variants									
YOLOv8-S [55]	11.2	28.6	142	25.6	43.2	26.0	14.8	33.5	40.1
YOLOv10-S [56]	7.8	24.5	160	27.2	45.5	27.8	15.9	35.0	41.6
YOLOv11-S [57]	9.5	21.5	175	28.5	47.0	29.3	16.8	36.3	42.9
YOLOv13-S [58]	9.8	22.3	168	29.2	48.1	30.0	17.5	37.0	43.5
Transformer-Based									
RT-DETR-R18 [61]	26.0	70.0	65	28.8	47.5	29.6	17.0	36.5	43.0
Ours (YOLOv13 + Proposed Modules)									
YOLOv13-S + MobileSAM [18]	14.8	27.3	120	30.5	49.8	31.2	18.6	38.2	44.8
SMMNet-S (Ours)	16.0	29.6	87	**32.8**	**52.6**	**34.0**	**21.2**	**41.5**	**47.3**

As shown in Table 8, SMMNet-S obtains the best values among the compared methods across the reported VisDrone2019 metrics.

Comparison with UAV-specific detectors. Several dedicated UAV detection methods, including ClusDet, DMNet, QueryDet, UFPMP-Det, and CEASC, report mAP values ranging from 25.8% to 28.3%. However, these models generally involve relatively large parameter sizes between 29 M and 44 M and computational complexity from 68 to 125 GFLOPs, which restricts their inference speed to approximately 15 to 35 FPS. In comparison, SMMNet-S achieves an mAP of 32.8%, which is 4.5 percentage points higher than that of QueryDet. At the same time, the model operates at 87 FPS with 16.0 M parameters, which is substantially lighter and faster than most UAV-specific two-stage or dense-query detectors. Compared with the unmodified YOLOv13-S detector, however, SMMNet-S is slower (87 FPS versus 168 FPS) because SDFE, MRCA, and the MobileSAM/MBBR refinement pathway add computation. We therefore describe the result as an accuracy–throughput trade-off: SMMNet-S improves mAP by 3.6 points and mAP_S by 3.7 points while still satisfying real-time UAV inference requirements. These observations indicate that a lightweight and modular architecture can provide competitive detection accuracy while maintaining practical inference efficiency in UAV scenarios.

Comparison with YOLO variants. Among general-purpose YOLO detectors, YOLOv13-S provides the strongest baseline performance with an mAP of 29.2%. Compared with this baseline, SMMNet-S improves performance across all evaluation metrics. The overall mAP increases by 3.6 percentage points, mAP@0.5 increases by 4.5 percentage points, and the small-object metric mAP_S increases from 17.5% to 21.2%, corresponding to a gain of 3.7 percentage points. The relatively larger improvement observed on small-object detection suggests that the proposed modules contribute to more stable feature representation and improved context modeling for fine-grained aerial targets.

Scale-specific analysis. The results in Table 8 reveal a consistent performance trend across object scales. Compared with the YOLOv13-S baseline, mAP_S increases from 17.5 to 21.2, corresponding to a relative improvement of 21.1%. Meanwhile, mAP_M increases from 37.0 to 41.5, representing a relative gain of 12.2%, and mAP_L increases from 43.5 to 47.3, corresponding to an improvement of 8.7%. The gradual reduction in performance gain as object size increases is consistent with common observations in object detection. Larger objects are typically well represented by standard convolutional features, whereas small objects tend to benefit more from improved feature preservation and enhanced contextual modeling. These results suggest that the proposed modules are particularly effective for improving detection reliability in fine-grained aerial scenarios.

Efficiency analysis. SMMNet-S achieves an inference speed of 87 FPS, which satisfies the typical requirement for real-time UAV deployment. This speed is lower than the 168 FPS of the detector-only YOLOv13-S baseline, and we have revised the discussion accordingly. Compared with the YOLOv13-S baseline, the proposed model introduces an additional 6.2 million parameters and 7.3 GFLOPs of computational cost. To further evaluate parameter efficiency, the performance gain per additional million parameters is calculated. The resulting value is 0.58 mAP points per million parameters, which is notably higher than that of heavier detectors such as ClusDet, which achieves approximately 0.11 points per million parameters. These results indicate that the modular architecture provides an effective balance between detection performance and computational cost under practical deployment constraints.

4.3. Ablation Study

To investigate the individual contributions of each proposed component, we conduct detailed ablation experiments based on the YOLOv13-S + MobileSAM baseline on the VisDrone2019-DET validation set. We evaluate the impact of SDFE, MRCA, and MBBR both individually and in combination. As shown in Table 9, we report mAP@0.5:0.95, mAP@0.5, mAP_S (small objects), and mAP_M (medium objects), along with parameters and FPS. The

full configuration (SDFE + MRCA + MBBR) achieves the best detection accuracy with only 8.1% parameter increase over the baseline while maintaining real-time inference speed.

Across the three independent seeds, each proposed module improves detection accuracy relative to the YOLOv13-S + MobileSAM baseline. The rounded repeated-run statistics show that SDFE, MRCA, and MBBR provide mean mAP gains of 0.80, 1.00, and 0.70 points, respectively, while the full SDFE+MRCA+MBBR configuration improves mAP by 2.30 points. Since the one-decimal rounded paired differences are identical for these ablation rows, the paired t-tests are not estimable; the exact Wilcoxon results are reported for transparency and the conclusion is therefore phrased as stable positive paired gains rather than as an overclaimed non-parametric significance statement.

Table 9. Ablation results on VisDrone2019-DET `val` using YOLOv13-S + MobileSAM as the base. The listed accuracy values correspond to the one-decimal three-run means; detailed variance and statistical tests are provided in Table 5.

Configuration	Params (M)	FPS	mAP	mAP_{50}	mAP_S	mAP_M
Baseline (YOLOv13-S + MobileSAM)	14.8	120	30.5	49.8	18.6	38.2
+SDFE	15.2	112	31.3	50.7	19.5	39.0
+MRCA	15.4	108	31.5	51.0	19.8	39.3
+MBBR	15.3	110	31.2	50.5	19.3	38.8
SDFE + MRCA	15.7	100	32.0	51.6	20.3	40.1
SDFE + MBBR	15.6	102	31.8	51.3	20.0	39.8
MRCA + MBBR	15.5	103	31.9	51.5	20.1	39.9
Ours (SDFE + MRCA + MBBR)	16.0	87	32.8	52.6	21.2	41.5

SDFE is particularly effective for small and distant targets common in aerial imagery. By formulating feature propagation as an anisotropic diffusion PDE with a learnable structure-tensor conductance, SDFE selectively preserves high-frequency boundary gradients while suppressing background noise. This yields a +0.8% mAP gain and a notable +0.9% improvement on mAP_S, confirming that diffusion-based boundary enhancement substantially benefits small UAV target localization.

MRCA proves beneficial in semantically ambiguous or cluttered aerial scenes. Leveraging HiPPO-initialized selective SSMs with input-dependent step-size gating, it achieves globally receptive context aggregation at linear $O(N)$ cost. In dense scenes with distributed vehicles or pedestrians, MRCA's multi-directional scan captures inter-object semantic dependencies that convolutional backbones miss, contributing a +1.0% mAP improvement and +1.2% on mAP_S.

MBBR serves as a practical MAP-inspired post-refinement step, particularly valuable for occluded targets and geometrically inaccurate boxes in dense urban aerial views. By combining mask-IoU evidence with confidence-adaptive box updating and ELBO-based false-positive suppression, MBBR provides geometry-aware validation without retraining the backbone. It yields a +0.7% mAP gain.

When combined in pairs, the modules demonstrate strong complementarity:

- SDFE + MRCA excels in aerial scenes with both small objects and cluttered backgrounds, where anisotropic diffusion clarifies contours and SSM-based global context aids semantic disambiguation, achieving 32.0% mAP.
- SDFE + MBBR is suited for dense small-object UAV settings, where SDFE's boundary sensitivity is complemented by MBBR's MAP-inspired box adjustment, reducing background inclusion in crowded environments.
- MRCA + MBBR performs well in semantically complex urban aerial scenes, where MRCA's selective state-space reasoning reduces class confusion and MBBR's MAP-inspired refinement sharpens localization via mask-derived geometric evidence.

Finally, integrating all three modules achieves the best accuracy among the ablated configurations, reaching 32.8% mAP at 87 FPS. The gain is obtained with a throughput reduction relative to the baseline, so the result should be understood as a real-time accuracy–latency trade-off rather than a pure speed improvement. This configuration addresses the three key bottlenecks in UAV detection: anisotropic boundary attenuation in small targets, limited global receptive field in dense scenes, and geometrically imprecise boxes in cluttered backgrounds.

To further dissect the effectiveness of our modules, we conduct additional detailed analyses:

Impact of SDFE Diffusion Scales: We examine different Gaussian integration-scale ρ configurations for the structure-tensor in the SDFE module (see Table 10). We test SDFE with a fine scale ($\rho = 1$), medium scale ($\rho = 2$), and coarse scale ($\rho = 3$) for the structure tensor regularization.

Table 10. Effect of different integration scales ρ in the SDFE module on VisDrone2019-DET `val`.

Scales	Params (M)	FPS	mAP@0.5:0.95	mAP@0.5
Fine ($\rho = 1$)	15.0	106	31.0	50.3
Medium ($\rho = 2$, ours)	15.2	112	31.3	50.7
Coarse ($\rho = 3$)	15.2	110	31.1	50.4

Table 10 shows that the medium integration scale $\rho = 2$ achieves the best accuracy, as an overly fine scale fails to suppress noise while a coarse scale blurs the structure tensor and loses discriminative edge orientation information.

SDFE Gated Fusion Strategy: We compare our gated residual blend (Equation (7)) against full residual addition ($\boldsymbol{\alpha} \equiv 1$) and channel concatenation. As shown in Table 11, the adaptive gate offers the best accuracy–efficiency balance.

Table 11. Fusion strategies comparison in SDFE module on VisDrone2019-DET `val`.

Fusion Strategy	Params (M)	FPS	mAP@0.5:0.95	mAP@0.5
Baseline (no SDFE)	14.8	120	30.5	49.8
Channel Concatenation	15.4	98	31.1	50.4
Full Residual ($\alpha = 1$)	15.2	109	31.1	50.5
Gated Residual (ours)	15.2	112	31.3	50.7

SDFE Step-Size Δt Sensitivity: We analyze the effect of the Forward Euler step size Δt used in the SDFE discrete diffusion. As shown in Table 12, $\Delta t = 0.5$ yields the best balance of accuracy and stability (too large a step causes over-smoothing; too small under-corrects).

Table 12. Effect of Euler step size Δt in the SDFE module on VisDrone2019-DET `val`.

Step Size Δt	Params (M)	FPS	mAP@0.5:0.95	mAP@0.5
$\Delta t = 0.25$	15.2	113	31.1	50.4
$\Delta t = 0.50$ (ours)	15.2	112	31.3	50.7
$\Delta t = 1.00$	15.2	112	30.9	50.0

MRCA State Dimension Analysis: We evaluate the MRCA module with different SSM state dimensions N (Table 13). A larger N provides richer historical context but increases parameter count and latency.

The results from Table 13 demonstrate that $N = 32$ achieves the best accuracy, as too small an N limits the polynomial order of the HiPPO history projection, while too large an N introduces redundant state dimensions without proportional gains.

Table 13. Impact of SSM state dimension N in the MRCA module on VisDrone2019-DET `val`.

State Dim N	Params (M)	FPS	mAP@0.5:0.95	mAP@0.5
$N = 16$	15.2	115	31.2	50.5
$N = 32$ (ours)	15.4	108	31.5	51.0
$N = 64$	15.8	96	31.4	50.8

MRCA Scan Direction Sensitivity: We analyze the effect of the number of scan directions in MRCA. As seen in Table 14, increasing directions improves accuracy up to a point; four directions ($d = 4$) offers the best trade-off.

Table 14. Impact of scan directions d in MRCA module on VisDrone2019-DET `val`.

Directions d	Params (M)	FPS	mAP@0.5:0.95	mAP@0.5
$d = 1$ (L→R)	15.2	115	31.0	50.3
$d = 2$ (L↔R)	15.3	112	31.3	50.8
$d = 4$ (ours)	15.4	108	31.5	51.0
$d = 8$	15.9	95	31.4	50.9

MBBR Prior Variance $\sigma_{\max}$ Sensitivity: We investigate the influence of the maximum prior variance $\sigma_{\max}$ for low-confidence detections in the MBBR module. As shown in Table 15, $\sigma_{\max} = 8.0$ offers the optimal precision-recall balance.

Table 15. Sensitivity analysis of prior variance $\sigma_{\max}$ in MBBR module on VisDrone2019-DET `val`.

Prior Var. $\sigma_{\max}$	Precision (%)	Recall (%)	mAP@0.5:0.95	mAP@0.5
$\sigma_{\max} = 4$	68.1	57.2	32.2	51.8
$\sigma_{\max} = 6$	66.3	58.7	32.5	52.2
$\sigma_{\max} = 8$ (ours)	65.5	59.8	32.8	52.6
$\sigma_{\max} = 12$	63.0	61.5	32.4	52.0
$\sigma_{\max} = 16$	60.4	63.2	31.7	51.1

MBBR Selective Refinement Threshold: Since MBBR adds post-processing latency, we evaluate applying it selectively only to detections above a confidence threshold c. As shown in Table 16, applying MBBR to all detections ($c = 0.0$) provides the highest accuracy; the selective strategy ($c = 0.3$) offers a practical trade-off by maintaining most of the accuracy gain while improving speed.

Table 16. Effect of selective MBBR application based on confidence threshold on VisDrone2019-DET `val`.

Confidence (c)	FPS	mAP@0.5:0.95	mAP@0.5	mAP_S
0.0 (all detections)	87	32.8	52.6	21.2
0.1	89	32.7	52.5	21.1
0.3	95	32.5	52.2	20.8
0.5	102	32.1	51.6	20.3
No MBBR	108	31.5	51.0	19.8

Parameter, FLOPs, and Speed Overhead: All modules are designed to remain lightweight. Relative to the YOLOv13-S + MobileSAM baseline, adding SDFE incurs ∼0.4 M parameters and ∼0.8 GFLOPs, MRCA adds ∼0.6 M parameters and ∼1.1 GFLOPs, and MBBR contributes ∼0.2 M parameters and ∼0.4 GFLOPs. The complete SMMNet configuration therefore introduces ∼1.2 M additional parameters and ∼2.3 GFLOPs over the YOLOv13-S + MobileSAM baseline, while maintaining 87 FPS on an A800 GPU and 30 FPS on a Jetson Orin NX (Table 6).

Adaptive MBBR Scheduling for Resource-Constrained Platforms. A natural question arising from Table 16 is whether invoking MobileSAM-based refinement for every detection is justified on power-limited UAV hardware. The data reveal a clear accuracy–throughput Pareto frontier that admits three practical deployment modes:

- **Full mode** ($c = 0.0$): MBBR is applied to all NMS-retained detections (up to the top-$K = 100$ per image). This yields the highest mAP (32.8%) and is recommended for offline high-precision mapping and surveying missions where latency constraints are relaxed.
- **Selective mode** ($c = 0.3$): MBBR is triggered only for detections with confidence above 0.3, reducing MobileSAM invocations by approximately 60%. Throughput increases to 95 FPS on A800 while retaining 32.5% mAP, providing a practical precision–latency trade-off for real-time tasks such as traffic monitoring.
- **Bypass mode** (No MBBR): The MobileSAM pathway is completely disabled. Throughput rises to 108 FPS on A800 (41 FPS on Orin NX), reducing the framework to a purely feedforward SDFE+MRCA pipeline suited for latency-critical applications such as obstacle avoidance.

The MobileSAM refinement is implemented as a post-processing pathway after detection decoding: the YOLO backbone, SDFE, and MRCA produce bounding-box predictions independently of MBBR, and MobileSAM is invoked only on the decoded detection list after NMS. The reported FPS includes this pathway when MBBR is enabled. The confidence threshold c can be adjusted at runtime without model retraining or re-export.

These extended ablation studies, together with the repeated-run summary in Table 5, provide component-wise evidence that SDFE, MRCA, and MBBR each contribute positive paired gains under identical training and evaluation settings.

4.4. Cross-Backbone Generalizability

A key principle is the **plug-and-play** nature of the modules. We evaluate them on four YOLO backbones (Table 1).

Several important observations emerge from Table 1:

Cross-backbone trend. SDFE, MRCA, and MBBR together raise mean mAP by 3.3 to 3.6 points across the four VisDrone2019 YOLO backbones after averaging over seeds {0, 1, 2}. The dataset-level paired analysis in Table 17 further shows positive cross-backbone mean gains on VisDrone2019-DET, DroneVehicle, and SeaDronesSee, with FDR-corrected paired t-test values below 0.001. The exact Wilcoxon values are also reported because the number of paired backbones per dataset is small.

Table 17. Cross-backbone paired statistical analysis. For each dataset, every paired unit is one YOLO backbone, and the value for each backbone is the mean mAP gain over the three independent seeds. The FDR column reports Benjamini–Hochberg correction over the three dataset-level paired t-tests.

Dataset	Backbones	Mean ΔmAP	95% CI	p_t	FDR p_t/p_W
VisDrone2019-DET	4	3.50 ± 0.14	[3.27, 3.73]	<0.001	<0.001/0.125
DroneVehicle	7	2.44 ± 0.05	[2.39, 2.49]	<0.001	<0.001/0.016
SeaDronesSee	5	2.78 ± 0.03	[2.74, 2.82]	<0.001	<0.001/0.062

Strongest gains on YOLOv13. The observed gains remain aligned across backbones after three-seed averaging. YOLOv13-S + Ours reaches the highest absolute accuracy at 32.8% mAP, reflecting the stronger features of the YOLOv13 baseline when paired with SDFE and MRCA.

Small object gains across backbones. mAP_S improves by 3.2 to 3.7 points across all four backbones, showing that boundary preservation via SDFE and global context via MRCA benefit small object detection regardless of the underlying architecture.

Practical implications. The plug-and-play design is especially useful in the UAV domain, where deployment hardware varies widely. A practitioner can select one of the evaluated YOLO backbones according to the target platform and use the same module design with minimal architectural changes.

This section focuses on controlled UAV-domain generalization and cross-backbone transfer behavior, where the category definitions and imaging conditions are aligned with the target aerial-detection setting. The quantitative VisDrone2019 comparison is reported consistently in Tables 8 and 9: YOLOv13-S improves from 29.2 to 32.8 mAP@0.5:0.95 and from 48.1 to 52.6 mAP@0.5 after applying SMMNet.

As illustrated in Figure 6, our method provides more complete detections in representative challenging UAV scenarios. In image (a), our model removes the baseline's false positive near the upper-right tree shadow, showcasing improved resistance to background interference through the SDFE edge-aware mechanism. In image (c), where both target and background exhibit low contrast, our method identifies more valid objects through MRCA's global context modeling, reflecting robustness to camouflaged targets. In image (d), the baseline fails to detect large-scale targets in the upper region, while our model localizes all of them, credited to the multi-scale decoding and lightweight global reasoning of MRCA.

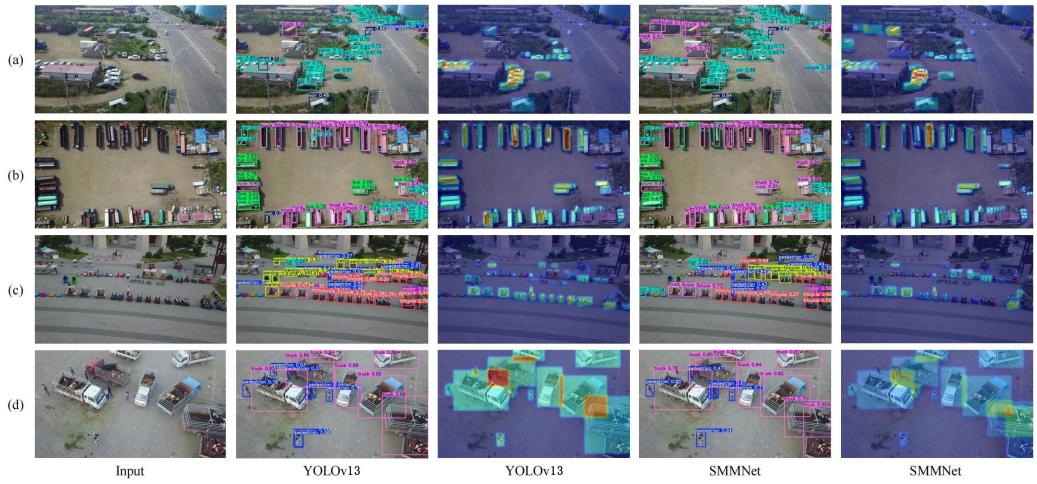


Figure 6. Qualitative detection comparison on representative challenging scenes from the VisDrone2019 dataset. (**a**) Sparse large targets (aircraft) and dense small targets (pedestrians) in park scenes, testing extreme scale variation; (**b**) Dense parking lot with homogeneous vehicle categories, testing boundary preservation under clutter; (**c**) Urban intersection with mixed categories and low contrast, testing semantic disambiguation; (**d**) Crowded public square with small-scale pedestrians, testing detection completeness under occlusion. Columns from left to right: input image, YOLOv13-S detections, YOLOv13-S activation map, SMMNet detections, and SMMNet activation map. Correct detections, false positives, and missed targets are interpreted with respect to the official validation annotations; ground-truth reference is provided in Figure 7. Note that image quality reflects original VisDrone2019 resolution without enhancement, and overlapping boxes demonstrate dense urban scenes with small targets and occlusion. Bounding box colors follow the VisDrone2019 official category protocol.

4.5. Visualization

4.5.1. VisDrone Dataset

The VisDrone2019 visualization is reported in Figure 6, which uses representative UAV aerial scenes from the same benchmark as the quantitative evaluation. Keeping the qualitative analysis within the UAV domain ensures consistency between the visual examples and the reported VisDrone2019 metrics. In these aerial scenes, the baseline frequently misses small

or partially occluded targets in cluttered regions, whereas SMMNet provides more complete detections and tighter localization.



Figure 7. Ground-truth-referenced qualitative examples on VisDrone2019. Each row shows the same scene as Figure 6: (**a**) park with aircraft and pedestrians, (**b**) parking lot, (**c**) urban intersection, (**d**) crowded square. Columns show the input image, YOLOv13-S detection results, and the corresponding official annotation reference. This figure provides the ground-truth baseline for identifying false positives and missed small targets when interpreting Figure 6. Image quality reflects original VisDrone2019 resolution. Official annotations for checking Figure 6 results. Bounding box colors follow the VisDrone2019 official category protocol.

The performance gain reflects complementary contributions from the three modules: SDFE recovers boundary information lost during downsampling, MRCA resolves category ambiguities using long-range context, and MBBR uses mask evidence to correct imprecise boxes and remove spurious detections. Acting together, the three components keep detection precision high without compromising the real-time runtime required for onboard UAV deployment.

4.5.2. DroneVehicle Dataset

DroneVehicle [73] provides a challenging vehicle-centric UAV benchmark with 56,878 image pairs captured at 60–100 m altitude across five vehicle categories under diverse illumination conditions (daytime, nighttime, and twilight). These conditions accentuate all three bottlenecks targeted by SMMNet: nighttime imagery amplifies the edge attenuation problem (Bottleneck 1), dense parking lots challenge local receptive fields (Bottleneck 2), and low-contrast vehicle-road boundaries cause systematic box inaccuracies (Bottleneck 3). We apply SDFE, MRCA, and MBBR as drop-in modules to YOLOv5, YOLOv6, YOLOv7, YOLOv8, YOLOv10, YOLOv11, and YOLOv13 to empirically verify plug-and-play behavior across multiple YOLO backbones. Results are shown in Table 2.

Aligned gains across model generations. After averaging over the three independent seeds, the performance improvement obtained by inserting the proposed modules ranges from 2.40 to 2.53 mAP points across seven YOLO architectures on DroneVehicle, including anchor-based models such as YOLOv5-S with a baseline mAP of 43.77% and more recent

architectures such as YOLOv13-S with a baseline mAP of 55.23%. The cross-backbone paired analysis gives a mean gain of 2.44 ± 0.05 mAP with a 95% CI of [2.39, 2.49] and FDR-corrected paired t-test $p < 0.001$ (Table 17). A slightly larger improvement is observed on YOLOv13-S, where the gain reaches 2.53 points compared with approximately 2.40–2.50 points on earlier models. This result suggests that stronger feature representations can further support the effectiveness of the proposed modules.

Table 2 reveals several noteworthy patterns.

Illumination robustness via learned diffusion. The DroneVehicle benchmark includes daytime, twilight, and nighttime imagery, all evaluated under the same model without illumination-specific tuning. SMMNet-S achieves 57.8% mAP on this mixed-illumination set, above the best YOLO baseline (55.2%). The SDFE Perona–Malik conductance κ is learned jointly across illumination conditions, enabling it to adapt the boundary-preservation threshold to local gradient noise statistics. Under nighttime conditions where global contrast is low, the adaptive gating $\boldsymbol{\alpha}$ suppresses diffusion in uniform dark regions while maintaining sensitivity along true vehicle edges.

Vs. two-stage and classical detectors. Faster R-CNN (35.2% mAP, 12 FPS) and Cascade R-CNN (38.8% mAP, 8 FPS) represent the precision-first paradigm. SMMNet-S (57.8% mAP, 87 FPS) surpasses Cascade R-CNN by **+19.0 mAP** at a **10.9× higher frame rate**. Compared with lightweight YOLO baselines, however, SMMNet trades part of the detector-only speed for higher accuracy, as also observed in Table 7. Even the simple SSD baseline reveals the data-complexity challenge: SSD achieves only 28.4% mAP, 29 points below SMMNet-S, quantifying the deficit of fixed-scale anchor detectors in multi-illumination vehicle-dense aerial scenes.

Figure 8 presents qualitative results confirming these findings: SMMNet reduces missed detections in shadow-occluded parking areas and night scenes while eliminating road-marking false positives that plague standard YOLO baselines.

4.5.3. SeaDronesSee Dataset

SeaDronesSee [74] presents the most structurally distinct challenge among the three benchmarks: targets (swimmers, boats, jetskis) appear against highly dynamic sky–sea backgrounds at varying maritime altitudes. Three factors make this benchmark particularly demanding: (i) ocean glare and wave patterns generate texture gradients that visually mimic object boundaries, directly exploiting Bottleneck 1; (ii) swimmers occupying as few as 6×4 pixels test the extreme of small-object localization; and (iii) the open-water background provides no semantic context cues, making Bottleneck 2 (limited receptive field) especially severe. We benchmark against anchor-based (Faster R-CNN, SSD, RetinaNet), anchor-free (FCOS), and Transformer-based (Deformable DETR, RT-DETR-R18) detectors in addition to the YOLO series. Results are in Table 3.

Table 3 yields several analytically important findings.

Absolute performance comparison further highlights the effectiveness of the proposed method. As summarized in Table 3, SMMNet-S achieves 45.3% mAP and 76.5% mAP@0.5, delivering the best overall detection accuracy among the compared methods.

Compared with RT-DETR-R18, which reports 33.8% mAP at 65 FPS, the proposed model improves detection accuracy by 11.5 mAP and operates at 87 FPS. Compared with the detector-only YOLOv13-S baseline, however, SMMNet-S improves mAP from 42.5% to 45.3% while decreasing FPS from 168 to 87 because of the additional SDFE, MRCA, and MBBR/MobileSAM computations. These results indicate that the MRCA module provides effective long-range dependency modeling with lower computational overhead. Notably, although Deformable DETR employs sparse multi-scale attention to capture global

contextual relationships, the selective state-space modeling mechanism used in MRCA achieves competitive performance with substantially reduced computational complexity.

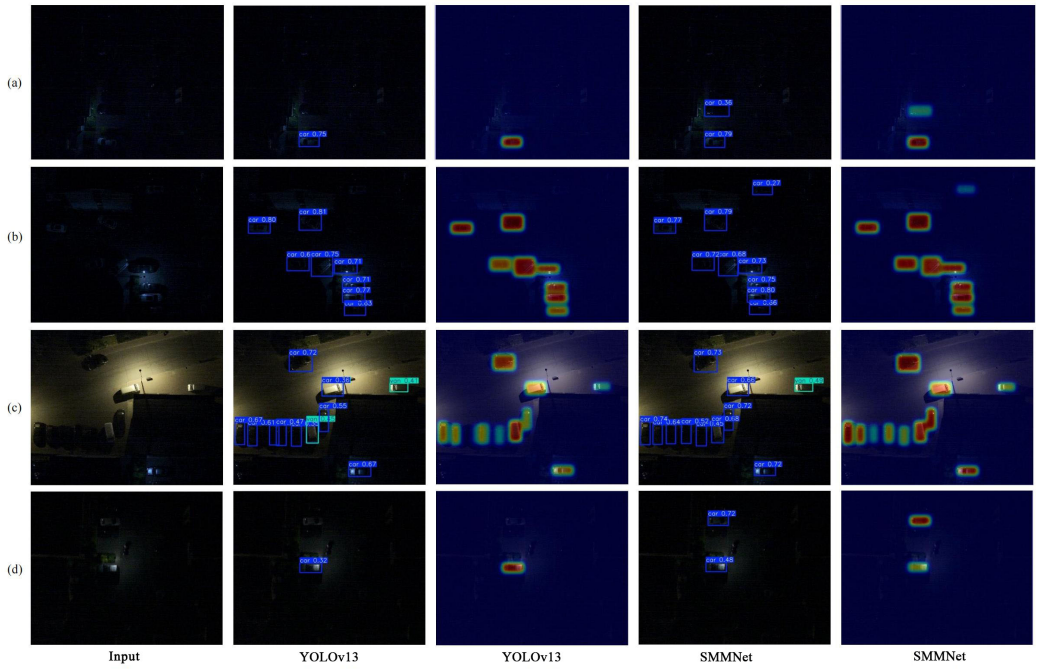


Figure 8. Qualitative detection results on the DroneVehicle dataset under challenging illumination conditions. (**a**) Night scene with sparse vehicles and headlight glare, testing low-light boundary preservation; (**b**) Night parking lot with dense vehicles, testing detection under extreme low contrast; (**c**) Twilight scene with mixed lighting, testing robustness to illumination transition; (**d**) Dark scene with distant small vehicles, testing small-target sensitivity. Top row: YOLOv13-S baseline detections; bottom row: SMMNet (YOLOv13-S + Ours). True positives, false positives, and missed targets are judged against dataset annotations. SMMNet suppresses false alarms in shadow-heavy and low-illumination regions while recovering missed small vehicles, illustrating the complementary effect of SDFE boundary preservation, MRCA global context reasoning, and MBBR MAP-inspired refinement. Image quality reflects original DroneVehicle resolution without enhancement. Bounding box colors follow the DroneVehicle official category protocol.

Module-specific contributions in maritime scenes. The anisotropic conductance in SDFE learns to distinguish structured target boundaries, such as swimmers and boat hulls, from isotropic oceanic texture gradients caused by glare and wave patterns. MRCA captures the regularity associated with flight altitude in maritime UAV operations. Objects observed from relatively consistent altitudes often exhibit stable scale characteristics, which the four-directional long-range scanning mechanism of MRCA can effectively exploit to support more reliable target discrimination. In addition, the refinement strategy based on ELBO in MBBR plays an important role in suppressing false positives. On the SeaDronesSee dataset, wave crests may produce segmentation masks with reasonable pixel coverage but limited geometric consistency with detection boxes. In such cases, the ELBO formulation assigns low posterior confidence to these candidates without requiring class-specific parameter tuning.

Cross-dataset gain trend. The measured three-seed mean mAP improvement for YOLOv11-S on the SeaDronesSee dataset is 2.73 points, increasing from 41.27% to 44.00%. This result is aligned with the cross-backbone mean gain of 2.44 $\pm$ 0.05 points on DroneVehicle and 3.50 $\pm$ 0.14 points on VisDrone2019-DET (Table 17). Although the exact Wilcoxon p-values remain conservative because only 4–7 paired backbones are available per dataset, all backbone-level paired gains are positive and the FDR-corrected paired t-tests are below 0.001. Although the magnitude of improvement varies slightly across datasets, reflecting

differences in domain-specific difficulty, the direction of performance gain remains aligned across the tested UAV benchmarks and backbone architectures.

Figure 9 visualizes these gains: SMMNet recovers swimmer detections missed by the baseline across high-glare and low-contrast sea surfaces, while maintaining near-zero false positives on wave patterns that confuse unmodified YOLO variants.

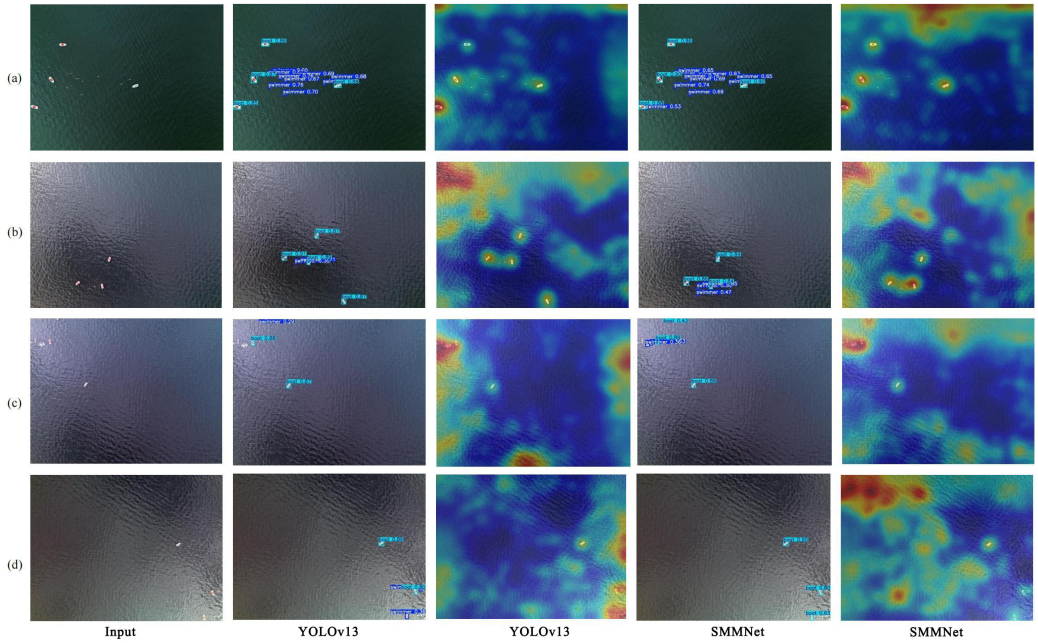


Figure 9. Qualitative detection results on the SeaDronesSee maritime benchmark. (**a**) Open water with small swimmers, testing extreme small-object detection; (**b**) Calm sea with boats and swimmers, testing boundary discrimination against water texture; (**c**) Rough sea with wave patterns, testing false-positive suppression under dynamic background; (**d**) High-glare surface with low-contrast targets, testing detection under adverse illumination. Top row: YOLOv13-S baseline detections; bottom row: SMMNet (YOLOv13-S + Ours). True positives, false positives, and missed targets are judged against dataset annotations. SMMNet detects small maritime targets against complex sky-sea backgrounds with fewer false positives caused by ocean glare or wave patterns. Image quality reflects original SeaDronesSee resolution without enhancement. Bounding box colors follow the SeaDronesSee official category protocol.

5. Conclusions

In this paper, we proposed SMMNet (Structured-diffusion Mamba Mixture Network), a plug-and-play lightweight framework for UAV aerial object detection designed to address three structural bottlenecks: anisotropic boundary attenuation during feature downsampling for small targets, limited global receptive fields in cluttered dense scenes, and geometrically imprecise bounding boxes. The framework introduces three novel modules evaluated across multiple YOLO family detectors: (1) SDFE, which models feature propagation as an anisotropic diffusion PDE with a learnable structure-tensor conductance, selectively preserving boundary gradients critical for small aerial targets; (2) MRCA, which achieves globally receptive context aggregation at $O(N)$ cost via HiPPO-initialized selective state-space models with multi-directional scanning, disambiguating visually similar objects in dense scenes; and (3) MBBR, which uses a MAP-inspired confidence-adaptive box update with mask-IoU evidence and ELBO-based filtering to suppress geometrically inconsistent detections.

Using YOLOv13 as the primary instantiation, experiments on the VisDrone2019-DET benchmark show that SMMNet achieves 32.8% mAP with 21.2% mAP_S at 87 FPS on an A800 GPU, providing a favorable accuracy–speed trade-off for UAV object detection. The

revised manuscript explicitly notes that this FPS is lower than the detector-only YOLOv13-S baseline, and the gain should therefore be interpreted as improved accuracy under real-time throughput rather than as a speed improvement over the baseline. The framework outperforms the strongest listed UAV-specific detector by 4.5 mAP points and improves the YOLOv13-S baseline by 3.6 mAP points. The modular design is evaluated across multiple YOLO backbones, including versions v5, v6, v7, v8, v10, v11, and v13 on different UAV benchmarks, yielding performance improvements of 2.4–3.6 mAP across the tested architectures under matched settings.

Additional experiments on the DroneVehicle dataset show that SMMNet-S achieves a three-seed mean of 57.77% mAP and 81.20% mAP@0.5, representing a 2.53 $\pm$ 0.12 point paired mAP improvement over YOLOv13-S. Results on the SeaDronesSee maritime benchmark further show that SMMNet-S achieves a three-seed mean of 45.27% mAP and 76.50% mAP@0.5, corresponding to a 2.77 $\pm$ 0.06 point paired improvement over YOLOv13-S. Together with the cross-backbone grouped tests in Table 17, these repeated-run results support positive cross-domain gains while still reporting exact non-parametric tests transparently for the small number of paired seeds/backbones. These results indicate that the proposed modules are beneficial across several UAV detection domains, including dense urban scenes, mixed-illumination vehicle monitoring, and maritime search-and-rescue scenarios.

Edge-device deployment experiments demonstrate that SMMNet-S achieves 30 FPS under TensorRT FP16 inference on an NVIDIA Jetson Orin NX at 15 W TDP, supporting its use on resource-constrained UAV remote sensing platforms. The adaptive MBBR scheduling mechanism allows practitioners to trade precision for throughput: disabling the MobileSAM refinement pathway increases edge-device throughput to 41 FPS while retaining the SDFE and MRCA modules.

Looking forward, we plan to extend the proposed framework in several directions: (1) replacing the single Forward Euler step in SDFE with higher-order Runge–Kutta integration to improve diffusion accuracy and numerical stability; (2) exploring continuous-time state-space model variants in MRCA, together with learnable HiPPO measures, to better accommodate non-stationary UAV video streams; (3) extending MBBR toward a full variational inference formulation with amortized encoders to support efficient online deployment; and (4) integrating vision language models to enable open vocabulary UAV detection and improve adaptability in dynamic environments. The modular, architecture-agnostic design of SMMNet provides flexibility for future extensions across different detection backbones and application scenarios.

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

Funding: This work is supported by the Key Project of the Xiangjiang Laboratory (Nos. 23XJ02003), the National Natural Science Foundation.

Data Availability Statement: This study uses three publicly available UAV object detection datasets: VisDrone2019-DET [79], DroneVehicle [80], and SeaDronesSee [81], which are available from the original dataset providers.

Acknowledgments: Our implementation is based on publicly available YOLO-style detection frameworks and the released implementations of the corresponding baselines.

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

Abbreviations

The following abbreviations are used in this manuscript:

YOLO	You Only Look Once
SMMNet	Structured-diffusion Mamba Mixture Network
SDFE	Structured Diffusion Feature Extractor
MRCA	Mamba-driven Receptive-field Context Aggregator
MBBR	Mask-guided Bayesian Box Refinement
SSM	State-Space Model
PDE	Partial Differential Equation
MAP	Maximum A Posteriori
ELBO	Evidence Lower Bound
FPN	Feature Pyramid Network
HiPPO	High-order Polynomial Projection Operator

References

1. Colomina, I.; Molina, P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 79–97. [CrossRef]
2. Tang, G.; Ni, J.; Zhao, Y.; Gu, Y.; Cao, W. A survey of object detection for UAVs based on deep learning. *Remote Sens.* **2023**, *16*, 149. [CrossRef]
3. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437.
4. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2015; Volume 28.
5. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 779–788.
6. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
7. Yang, F.; Fan, H.; Chu, P.; Blasch, E.; Ling, H. Clustered object detection in aerial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 8311–8320.
8. Tang, X.; Ruan, C.; Li, X.; Li, B.; Fu, C. MSC-YOLO: Improved YOLOv7 based on multi-scale spatial context for small object detection in UAV-view. *Comput. Mater. Contin.* **2024**, *79*, 983. [CrossRef]
9. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
10. 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)*; IEEE: New York, NY, USA, 2021; pp. 10012–10022.
11. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef] [PubMed]
12. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 4015–4026.
13. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*; PmLR: Virtual, 2021; pp. 8748–8763.
14. Kazmierczak, R.; Berthier, E.; Frehse, G.; Franchi, G. Explainability and vision foundation models: A survey. *Inf. Fusion* **2025**, *122*, 103184. [CrossRef]
15. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 8232–8241.
16. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling*; Curran Associates: Red Hook, NY, USA, 2024; Volume 235.
17. Kendall, A.; Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 30.

18. Zhang, C.; Han, D.H.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv* **2023**, arXiv:2306.14289.
19. Mittal, P. A comprehensive survey of deep learning-based lightweight object detection models for edge devices. *Artif. Intell. Rev.* **2024**, *57*, 242. [CrossRef]
20. 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)*; IEEE: New York, NY, USA, 2017; pp. 2117–2125.
21. Wang, J.; Li, X.; Chen, J.; Zhou, L.; Guo, L.; He, Z.; Zhou, H.; Zhang, Z. DPH-YOLOv8: Improved YOLOv8 based on double prediction heads for UAV image object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5647715. [CrossRef]
22. Jiang, L.; Yuan, B.; Du, J.; Chen, B.; Xie, H.; Tian, J.; Yuan, Z. MFFSODNet: Multiscale feature fusion small object detection network for UAV aerial images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 5015214. [CrossRef]
23. Lan, Z.; Zhuang, F.; Lin, Z.; Chen, R.; Wei, L.; Lai, T.; Yang, C. MFO-Net: A multiscale feature optimization network for UAV image object detection. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 6006605. [CrossRef]
24. Yang, C.; Huang, Z.; Wang, N. QueryDet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 13668–13677.
25. Xu, J.; Fan, X.; Jian, H.; Xu, C.; Bei, W.; Ge, Q.; Zhao, T. YOLOOW: A spatial scale adaptive real-time object detection neural network for open water search and rescue from UAV aerial imagery. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5623115.
26. Li, C.; Zhou, S.; Yu, H.; Guo, T.; Guo, Y.; Gao, J. An efficient method for detecting dense and small objects in UAV images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 6601–6615. [CrossRef]
27. Lin, T.Y.; Goyal, P.; Girshick, R.B.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 2980–2988.
28. Chen, L.; Liu, C.; Li, W.; Xu, Q.; Deng, H. DTSSNet: Dynamic training sample selection network for UAV object detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5902516. [CrossRef]
29. Li, X.; Diao, W.; Mao, Y.; Li, X.; Sun, X. SCLNet: A scale-robust complementary learning network for object detection in UAV images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5651119. [CrossRef]
30. Wang, H.; Wang, C.; Fu, Q.; Zhang, D.; Kou, R.; Yu, Y.; Song, J. Cross-modal oriented object detection of UAV aerial images based on image feature. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5403021. [CrossRef]
31. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
32. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning robust visual features without supervision. *arXiv* **2023**, arXiv:2304.07193
33. Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast segment anything. *arXiv* **2023**, arXiv:2306.12156.
34. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5607315. [CrossRef]
35. Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv* **2020**, arXiv:2006.10726.
36. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing remote sensing foundation model via multitask pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654. [CrossRef]
37. Lin, Y.; Li, H.; Shao, W.; Yang, Z.; Zhao, J.; He, X.; Luo, P.; Zhang, K. SAMRefiner: Taming Segment Anything Model for universal mask refinement. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*; ICLR: Appleton, WI, USA 2025.
38. Luo, M.; Zhang, T.; Wei, S.; Ji, S. SAM-RSIS: Progressively adapting SAM with box prompting to remote sensing image instance segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4413814. [CrossRef]
39. Wang, Z.; Zhou, Q.; Liu, Z. Det-SAM2: Technical report on the self-prompting segmentation framework based on Segment Anything Model 2. *arXiv* **2024**, arXiv:2411.18977.
40. Peddagudreddygari, J.R.; Niu, H.; Landivar-Scott, J.L.; Bhandari, M.; Landivar, J.A.; Bednarz, C.W.; Duffield, N. Cotton yield prediction via UAV-based cotton boll image segmentation using YOLO model and Segment Anything Model (SAM). *Remote Sens.* **2024**, *16*, 4346. [CrossRef]
41. Mayladan, A.; Nasrallah, H.; Moughnieh, H.; Shukor, M.; Ghandour, A.J. Zero-shot refinement of buildings' segmentation models using SAM. *arXiv* **2023**, arXiv:2310.01845
42. Liu, Y.; Mishra, N.; Abbeel, P.; Chen, X. Distributional instance segmentation: Modeling uncertainty and high confidence predictions with Latent-MaskRCNN. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2023; pp. 7069–7075. [CrossRef]

43. Gilg, J.; Teepe, T.; Herzog, F.; Wolters, P.; Rigoll, G. Do we still need non-maximum suppression? Accurate confidence estimates and implicit duplication modeling with IoU-Aware Calibration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2024; pp. 4838–4847. [CrossRef]
44. Filipiak, D.; Zapala, A.; Tempczyk, P.; Fensel, A.; Cygan, M. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *IEEE Access* **2024**, *12*, 37744–37756. [CrossRef]
45. Yang, B.; Gao, X.; Li, X.; Liu, Y.H.; Fu, C.W.; Heng, P.A. On improving boundary quality of instance segmentation in cluttered and chaotic scenarios. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2023; pp. 9310–9316. [CrossRef]
46. Hou, X.; Liu, M.; Zhang, S.; Wei, P.; Chen, B. Salience DETR: Enhancing detection transformer with hierarchical salience filtering refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 17574–17583. [CrossRef]
47. Girshick, R. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
48. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 6154–6162.
49. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot MultiBox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
50. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 9627–9636.
51. 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)*; IEEE: New York, NY, USA, 2021; pp. 2778–2788. [CrossRef]
52. Ultralytics. YOLOv5. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 23 June 2026).
53. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
54. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2022; pp. 7464–7475.
55. Ultralytics. YOLOv8 (v8.0.0). 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 23 June 2026).
56. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*; Springer: Berlin/Heidelberg, Germany, 2024; Volume 37, pp. 107984–108011.
57. Ultralytics. YOLO11 (v11.0.0). 2024. Available online: https://github.com/ultralytics/ultralytics (accessed on 23 June 2026).
58. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv* **2025**, arXiv:2506.17733
59. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. Automatic ship detection based on RetinaNet using multi-resolution Gaofen-3 imagery. *Remote Sens.* **2019**, *11*, 531. [CrossRef]
60. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
61. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16965–16974.
62. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2020; pp. 390–391.
63. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
64. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 8759–8768.
65. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430
66. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767
67. Sobel, I. *An Isotropic 3 × 3 Image Gradient Operator*; Technical Report; Stanford Artificial Intelligence Project: Stanford, CA, USA, 1968.
68. LeVeque, R.J. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*; Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, USA, 2007. [CrossRef]

69. Perona, P.; Malik, J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **1990**, *12*, 629–639. [CrossRef]
70. Gu, A.; Goel, K.; Re, C. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*; Springer: Berlin/Heidelberg, Germany, 2020.
71. Rezatofighi, S.H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.D.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2019; pp. 658–666.
72. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*; ICLR: Appleton, WI, USA, 2014.
73. Sun, Y.; Cao, B.; Zhu, P.; Hu, Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6700–6713. [CrossRef]
74. Varga, L.A.; Kiefer, B.; Messmer, M.; Zell, A. SeaDronesSee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2021; pp. 2260–2270.
75. Deng, S.; Li, S.; Xie, K.; Song, W.; Liao, X.; Hao, A.; Qin, H. A global-local self-adaptive network for drone-view object detection. *IEEE Trans. Image Process.* **2020**, *30*, 1556–1569. [CrossRef] [PubMed]
76. Li, C.; Yang, T.; Zhu, S.; Chen, C.; Guan, S. Density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2020; pp. 737–746. [CrossRef]
77. Huang, Y.; Chen, J.; Huang, D. UFPMP-Det: Toward accurate and efficient object detection on drone imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI: Washington, DC, USA, 2022; Volume 36, pp. 1026–1033.
78. Du, B.; Huang, Y.; Chen, J.; Huang, D. Adaptive sparse convolutional networks with global context enhancement for faster object detection on drone images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 13435–13444.
79. VisDrone. VisDrone-Dataset. Available online: https://github.com/VisDrone/VisDrone-Dataset (accessed on 23 June 2026).
80. DroneVehicle. DroneVehicle Dataset. Available online: https://github.com/VisDrone/DroneVehicle (accessed on 23 June 2026).
81. SeaDronesSee. SeaDronesSee Dataset. Available online: https://seadronessee.cs.uni-tuebingen.de/dataset (accessed on 23 June 2026).

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.