PRISM: Priority-Guided Scanning in the Wavelet Domain for UAV Maritime Small Object Detection

Pengqi Gao **Fan Shi*** **Mianzhao Wang** **Jiangpeng Zheng**
Xu Cheng **Shengyong Chen**
School of Computer Science and Engineering, Tianjin University of Technology
Engineering Research Center of Learning-Based Intelligent System (Ministry of Education)
Key Laboratory of Computer Vision and System (Ministry of Education)
Tianjin, China
pengqigao@stud.tjut.edu.cn shifan@email.tjut.edu.cn
wmz@email.tjut.edu.cn zjp_1997@stud.tjut.edu.cn
xu.cheng@ieee.org sy@ieee.org

Abstract

Sparse object detection remains difficult when weak target evidence occupies only a few pixels and shares local statistics with surrounding clutter. This failure is not merely due to limited model capacity, but to a representation bottleneck: spatial domain features mix target edges, clutter textures, and scene context before they can be reliably separated. We propose PRISM, a wavelet domain detector that combines frequency separated representation learning with priority conditioned feature interaction. PRISM decomposes each scene into directional high frequency evidence and low frequency context, models the two streams with dedicated encoders, and learns a continuous, image conditioned priority map that induces the ordering used for local causal refinement. Multi scale features are then reconstructed by inverse wavelet synthesis with adaptive gates and passed to a standard detection head. On the SeaDronesSee and AFO maritime benchmarks, PRISM achieves state of the art performance with its largest gains on the smallest targets. Dataset specific training on VisDrone further demonstrates applicability to aerial urban imagery characterized by distinct forms of structured clutter. Code is available at https://github.com/GPQ-phD/PRISM.

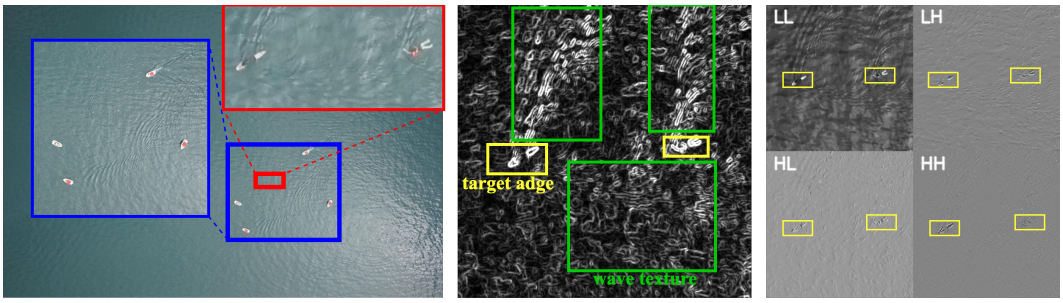
1 Introduction

Sparse object detection is difficult when weak target evidence occupies only a few pixels and shares local statistics with structured background clutter. This problem is especially acute in low altitude uncrewed aerial vehicle (UAV) imagery for maritime search and rescue, where reliable detection of small objects remains a fundamental challenge [1, 2]. Maritime targets often span only a few tens of pixels and frequently appear in close proximity, including persons in distress, life saving equipment, and clustered small boats. In such scenes, weak target boundaries are easily overwhelmed by waves, wakes, and reflections, leading to unstable localization and missed detections. This setting exemplifies a broader sparse detection problem: weak discriminative signals embedded in statistically similar structured clutter.

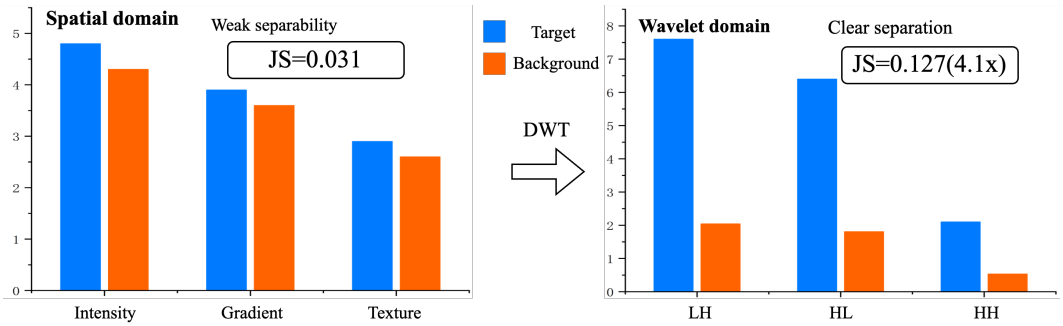
A central limitation of existing detectors is their dependence on spatial domain representations, where target sensitive edges, structured wave clutter, and scene level context are encoded together. As shown in Figure 1, the spatial gradient responses of small targets and wind driven ripples are often statistically indistinguishable, obscuring the boundary that downstream layers must learn. Existing UAV maritime detectors improve such representations through multi scale feature aggregation,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Visual motivation: target edges resemble wave textures in the spatial domain, but differ in wavelet subbands



(b) Quantitative separability over 500 images

Figure 1: Motivation of PRISM. Top: a UAV maritime image (left), its spatial domain gradient response (middle), and its one level Haar wavelet decomposition into LL, LH, HL, HH subbands (right). Target edges and wave textures are visually entangled in the spatial domain but become more separable in the wavelet domain. Bottom: target and background energy distributions over 500 sampled images in the spatial domain (left, Jensen Shannon divergence JS$=0.031$) and wavelet domain (right, JS$=0.127$).

attention based refinement, or scale adaptive convolutions [3, 4, 5]. However, enhancing an already mixed representation does not directly resolve the underlying ambiguity. **The first challenge is to disentangle target edges, background clutter, and scene context at the representation level.**

The wavelet domain provides a natural basis for this disentanglement. Multi level Haar decomposition separates an image into a low frequency subband that preserves global scene context and three high frequency subbands that capture directional edge responses. Across 500 sampled images, the Jensen Shannon divergence between target and background energy distributions increases from 0.031 in the spatial domain to 0.127 in the wavelet domain, indicating substantially improved separability. Although recent methods have introduced wavelet analysis into detection pipelines [6, 7], they typically use decomposition as an auxiliary module within a spatial backbone. As a result, frequency separated subbands are soon merged back into conventional spatial features, weakening the directional and energy cues that distinguish targets from wave textures. **A second issue therefore arises: target sensitive high frequency evidence and scene level low frequency context should be modeled separately throughout the detection pipeline.**

Separation alone does not ensure effective fusion. In maritime scenes, target responses are weak while structured wave responses are often strong. If all spatial positions interact synchronously, clutter dominated regions can suppress the weak target evidence recovered by wavelet decomposition. Existing frequency aware fusion strategies mitigate this imbalance through position wise reweighting or subband selection [8, 9]. These designs adjust how much each position contributes, but the interaction itself remains synchronous and content agnostic. **What is missing is a fusion mechanism that adapts local feature interaction to the estimated relevance of each spatial position.**

To address these challenges, we propose PRISM, a wavelet domain detector built on a divide and conquer paradigm. PRISM first performs multi level Haar decomposition to separate scene level low frequency context from target sensitive high frequency evidence. The two streams are then processed by dedicated encoders: a High Frequency Encoder (HFE) models directional edge evidence through

subband interaction, while a Low Frequency Encoder (LFE) captures contextual information with linear complexity in the number of tokens. To prevent weak target evidence from being suppressed during fusion, a Policy Network learns a continuous, image conditioned priority map from both streams. The map induces a content adaptive ordering, under which the Priority Guided Fusion (PGF) module performs local causal refinement. Finally, a Wavelet Synthesis Decoder (WSD) reconstructs multi scale detection features through inverse wavelet transform with adaptive injection gates.

Our main contributions are summarized as follows:

- **Priority induced ordering for target aware causal interaction.** We introduce a Policy Network that learns a continuous spatial priority map from high and low frequency evidence. The learned map induces an image adaptive scan order, allowing likely target regions to shape local causal feature interaction before the refined tokens are restored to their original spatial coordinates. This replaces a fixed geometric traversal with an ordering adapted to image content.
- **End to end frequency separated detection pipeline.** PRISM maintains high frequency and low frequency streams throughout encoding, fusion, and decoding. Dedicated encoders model directional target evidence and global scene context, while the Wavelet Synthesis Decoder reconstructs multi scale features through inverse wavelet transform. This persistent separation preserves frequency cues that prior wavelet aware detectors often lose after auxiliary decomposition.
- **Leading performance and broader aerial applicability.** Experiments on SeaDronesSee and AFO show that PRISM achieves leading performance, with consistent gains on small and densely distributed targets. Results on VisDrone further demonstrate its applicability to aerial urban scenes with distinct structured clutter.

2 Related Work

2.1 Spatial Domain Small Object Detection

Generic small object detection has advanced along several axes: cross scale feature propagation [10, 11], weak target response modeling [12, 13], and global dependency reasoning through Transformers [14, 15] or linear complexity sequence models [16, 17]. For UAV maritime imagery, recent works further address wave clutter and scale variation through reparameterized convolutions [18], lightweight Transformers with cross scale interaction [19], multi attention adaptive fusion [20], density aware attention [21], and adaptive zoom in for sparse foreground regions [22]. Despite their diversity, these methods mainly enhance spatial representations in which target edges, wave clutter, and scene context are encoded jointly. Feature enhancement alone therefore does not resolve the representation level ambiguity between weak targets and structurally similar clutter.

2.2 Wavelet and Directional Representation Learning

Wavelet transforms decompose visual signals into frequency subbands with distinct directional and structural characteristics, providing a principled basis for separating target edges from background clutter. WTConv [23] formulates wavelet convolution to expand the receptive field without parameter inflation, while DWTFreqNet [24] uses subband decomposition to separate structural edges from texture noise. FSUNet [25] fuses spatial and frequency features for boundary aware representations, GSFANet [26] introduces joint spatial frequency learning, and Wavelet Tree Transformer [27] applies frequency selective attention.

Directional modeling is also prominent in infrared small target detection. ISNet [28] combines edge enhancement with attention aggregation along two orientations, MSDA-Net [29] injects multiscale high frequency directional priors, and MLEDNet learns multidirectional edge features [30]. Wavelet methods typically use decomposition as an auxiliary operation and merge subbands back into spatial features after shallow processing, whereas directional infrared methods mainly strengthen local edge and shape representations. Persistent frequency separated modeling throughout encoding, fusion, and decoding, together with content adaptive spatial interaction, remains underexplored.

2.3 Ordered Feature Interaction

The way features interact across spatial positions is a fundamental design choice in detection. Self attention [31] enables global pairwise interaction, but it does not define an explicit order in which one position should refine another. Deformable attention [32] reduces cost by restricting interaction to a

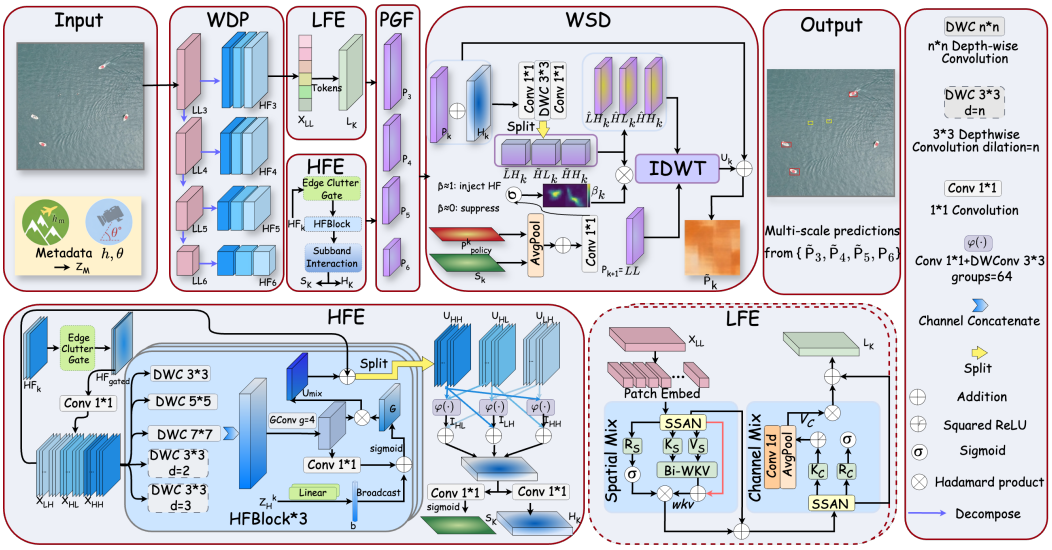


Figure 2: Architecture of PRISM. WDP produces multi scale wavelet subbands, HFE and LFE encode high frequency evidence and low frequency context, PGF performs priority guided fusion, and WSD reconstructs detection features through gated inverse DWT.

sparse subset, yet communication among selected positions remains synchronous rather than priority ordered. Causal sequence models introduce a different interaction structure, in which earlier tokens condition the representation of later ones, establishing a directional information flow [33, 34, 35]. When adapting such models to 2D inputs, the scan order becomes a central design question, typically resolved by fixed geometric rules, such as the four directional selective scans of VMamba [36] or the window based scans of LocalMamba [37]. In these methods, the scan trajectory is independent of image content, so the order in which positions influence one another is determined by geometry rather than by semantic relevance. The potential of content adaptive routing that derives local causal neighborhoods from learned spatial priorities remains largely unexplored.

3 Methodology

3.1 Overview

PRISM is a wavelet domain detector built on a divide and conquer paradigm, illustrated in Figure 2. Given an input image $\mathbf{X} \in \mathbb{R}^{H_0 \times W_0 \times 3}$ and optional flight metadata including altitude h and viewing angle θ, PRISM maintains frequency separated representations across encoding, fusion, and decoding.

In the divide stage, a Wavelet Decomposition Pyramid (WDP) applies four levels of Haar DWT, producing a low frequency subband LL_k and three directional high frequency subbands LH_k, HL_k, HH_k at each scale $k \in \{3, 4, 5, 6\}$. HFE consolidates the three high frequency subbands into a target aware feature H_k and a target prior $S_k \in [0, 1]^{h_k \times w_k}$, while LFE processes LL_k with a linear complexity sequence model to obtain a global context feature L_k. In the conquer stage, a Policy Network predicts a continuous, image conditioned priority map from the joint high and low frequency representation. Its scores induce a deterministic ordering π_k. PGF first balances the two streams through adaptive frequency mixing, then reorders the fused tokens by π_k and applies gated local causal refinement in the ordered sequence.

WSD finally reconstructs multi scale detection features $\{\tilde{P}_3, \tilde{P}_4, \tilde{P}_5, P_6\}$ through cascaded inverse DWT with adaptive injection gates. These features are passed to a shared detection head.

Throughout the network, normalization is implemented as Scene State Adaptive Normalization (SSAN), whose affine parameters are predicted from a state vector derived from low frequency channel statistics, high frequency directional energy, and flight metadata. The full formulation is provided in Appendix A.1.

3.2 Wavelet Domain Divide

The divide stage separates target sensitive high frequency evidence from scene level low frequency context. It consists of the WDP, HFE, and LFE.

Wavelet Decomposition Pyramid. WDP applies four levels of 2D Haar DWT to obtain multi scale subbands $\{LL_k, LH_k, HL_k, HH_k\}_{k=3}^{6}$. For a 640×640 input, their spatial resolutions are 320^2, 160^2, 80^2, and 40^2. LL_k preserves smooth scene structure, while LH_k, HL_k, and HH_k capture directional high frequency responses. The invertible Haar basis enables exact reconstruction in WSD with linear computational cost.

High Frequency Encoder. The high frequency subbands contain target boundary evidence, but also respond strongly to wave textures. HFE therefore aims to strengthen directionally consistent target evidence while suppressing clutter like responses. At scale k, HFE takes LH_k, HL_k, and HH_k as input and outputs a target aware feature $H_k \in \mathbb{R}^{h_k \times w_k \times C_k}$ and a target prior map $S_k \in [0,1]^{h_k \times w_k}$.

These subbands are first concatenated and refined by HFBlocks conditioned on a directional energy descriptor:

$$\mathbf{z}_H^k = \mathrm{Concat}\big(\mathrm{GAP}(|LH_k|),\, \mathrm{GAP}(|HL_k|),\, \mathrm{GAP}(|HH_k|)\big). \tag{1}$$

This descriptor summarizes the orientation energy distribution: a target boundary typically activates one or two subbands strongly, while wave clutter spreads more uniformly across the three. Conditioning HFE on $\mathbf{z}_H^k$ allows it to adapt to the current edge and clutter pattern.

After refinement, the features are split into three directional groups U_{LH}, U_{HL}, and U_{HH}. Since an oblique target boundary may respond weakly in multiple subbands but never strongly in any single one, HFE introduces cross subband interaction:

$$\hat{U}_{LH} = U_{LH} + \varphi(U_{HL} + U_{HH}), \tag{2}$$

where $\varphi = \mathrm{Conv}_{1\times1} \circ \mathrm{DWConv}_{3\times3}$. Analogous updates are applied to $\hat{U}_{HL}$ and $\hat{U}_{HH}$. The updated groups are concatenated and projected to obtain H_k. The target prior is predicted as

$$S_k = \sigma(\mathrm{Conv}_{1\times1}(\hat{U})), \tag{3}$$

where $\hat{U} = \mathrm{Concat}(\hat{U}_{LH}, \hat{U}_{HL}, \hat{U}_{HH})$. This prior provides a high frequency estimate of target relevance for the subsequent fusion stage. Detailed HFBlock design is provided in Appendix A.2.

Low Frequency Encoder. LFE models the low frequency scene context. This stream requires long range dependency aggregation, because the low frequency subband encodes global layout, water surface structure, and coarse object context. At the finest scale, the token count reaches $T_3 = 102{,}400$, making quadratic self attention impractical. LFE therefore adopts Vision-RWKV [35] as the token mixer, which captures long range context with linear complexity in the number of tokens and is well matched to the smoother spatial statistics of LL_k:

$$L_k = \mathrm{Reshape}\big(\mathrm{VRWKV}(\mathrm{Flatten}(\mathrm{Conv}_{1\times1}(LL_k)))\big), \tag{4}$$

where $L_k \in \mathbb{R}^{h_k \times w_k \times C_k}$ denotes the scene context feature at scale k. Token mixer comparisons and configuration details are in Appendix C.1 and A.3.

3.3 Priority Guided Conquer

The conquer stage integrates high frequency target evidence and low frequency scene context under a content adaptive interaction order. Figure 3 summarizes the workflow: priority scores determine which spatial positions become local neighbors in the ordered causal branch, after which inverse permutation restores the spatial layout.

3.3.1 Policy Network

High frequency features provide precise edge evidence but are sensitive to wave clutter, whereas low frequency features provide scene context but lack fine localization. The two streams should therefore contribute differently at each position, and the Policy Network estimates this contribution as a spatial priority map. Unlike the high frequency only prior S_k from HFE, it operates on the joint representation $F_{\mathrm{pol}} = [L_{k_0};\, H_{k_0}]$ at the finest scale $k_0 = 3$ to combine boundary evidence with

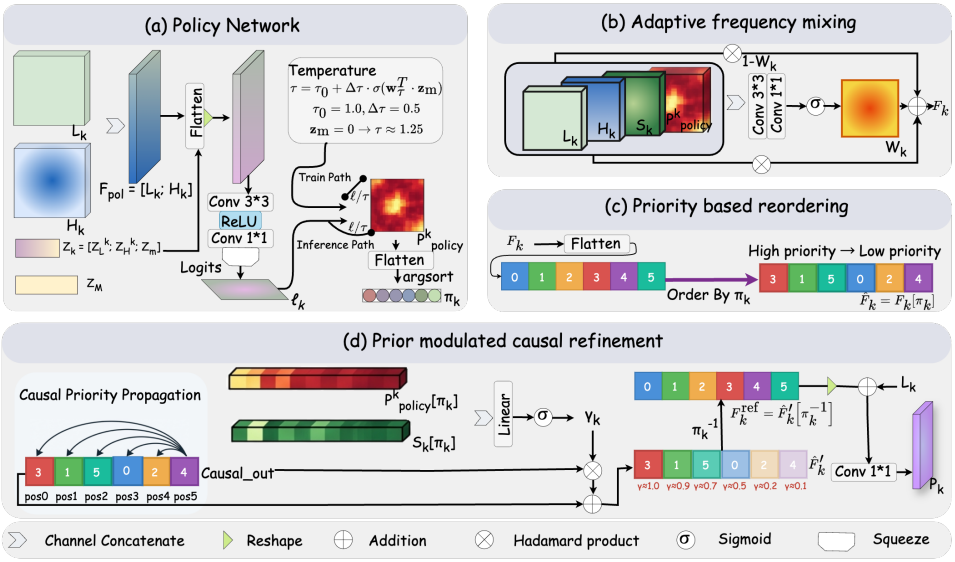


Figure 3: Priority Guided Conquer stage. The Policy Network predicts a continuous spatial priority map that induces the ordering π_k. PGF balances the two frequency streams, reorders tokens by π_k, performs priority guided causal refinement, and restores the spatial layout by inverse permutation.

scene context. A lightweight head with SSAN predicts a spatial logit map $\ell \in \mathbb{R}^{h_{k_0} \times w_{k_0}}$, scaled by a metadata aware temperature τ that adapts to flight conditions (Appendix A.4):

$$P^{k_0}_{\text{policy}} = \text{GumbelSoftmax}(\ell/\tau). \tag{5}$$

Priority maps at coarser scales are obtained by average pooling. The induced ordering at scale k is

$$\pi_k = \text{argsort}\big(\text{flatten}(P^k_{\text{policy}}), \text{descending}\big), \tag{6}$$

placing positions with stronger target priority earlier in the causal sequence. The discrete argsort receives no gradient. Instead, P^k_{policy} is optimized through its differentiable roles in the downstream mixing weight w_k and causal gate γ_k; its learned scores deterministically induce π_k. We therefore refer to the mechanism as a learned continuous priority map that induces an ordering adapted to image content.

3.3.2 Priority Guided Fusion

PGF integrates H_k and L_k by first adapting their frequency balance and then refining the fused tokens in the priority induced order. This process is guided by P^k_{policy}, S_k, and π_k.

Adaptive frequency mixing. The relative importance of edge detail and scene context varies across positions, so the two streams are mixed under a position specific weight predicted from the policy map, target prior, and both streams:

$$w_k = \sigma\big(\Psi_w([P^k_{\text{policy}} \otimes \mathbf{1}_C;\, S_k \otimes \mathbf{1}_C;\, H_k;\, L_k])\big)\,, \tag{7}$$

where Ψ_w is a lightweight convolutional head and $\otimes\mathbf{1}_C$ broadcasts a single channel map to C channels. The initial fused feature is $F_k = w_k \odot H_k + (1 - w_k) \odot L_k$.

Priority based reordering. Mixing operates per position and does not introduce spatial interaction. To condition local interaction on content, the fused tokens are flattened and permuted by π_k, yielding $\hat{F}_k = \text{Flatten}(F_k)[\pi_k] \in \mathbb{R}^{T_k \times C}$, so that higher priority positions tend to occur earlier in the causal sequence.

Prior modulated causal refinement. The reordered sequence is refined by a causal depthwise 1D convolution with kernel size 7, stride 1, and dilation 1. Thus, the output at rank t directly depends only on $\{\max(1, t-6), \ldots, t\}$. A per token gate $\gamma_k = \sigma(\Psi_\gamma([P^k_{\text{policy}}[\pi_k];\, S_k[\pi_k]])) \in \mathbb{R}^{T_k \times 1}$ modulates the propagation strength, giving

$$\hat{F}'_k = \hat{F}_k + \gamma_k \odot \text{CausalDWConv1D}(\hat{F}_k). \tag{8}$$

A token at rank r therefore directly influences only $\{r, \ldots, \min(T_k, r+6)\}$ through this branch. The operation is applied independently at each scale, so these sequence ranges do not compound across scales. PGF is consequently a local priority conditioned refinement mechanism, with γ_k controlling the residual update at each ordered position.

The refined tokens are restored via $F_k^{\mathrm{ref}} = \mathrm{Reshape}(\hat{F}'_k[\pi_k^{-1}])$. The restored feature is concatenated with L_k, normalized by SSAN, and projected by a 1×1 convolution to produce $P_k \in \mathbb{R}^{h_k \times w_k \times 128}$. The fused features $\{P_3, P_4, P_5, P_6\}$ are passed to WSD for multi scale reconstruction.

3.4 Wavelet Synthesis Decoder

WSD reconstructs multi scale detection features from $\{P_k\}$ through coarse to fine inverse DWT. Unlike interpolation based FPN decoding, which may attenuate the high frequency cues preserved by PRISM, WSD uses the wavelet synthesis path to propagate coarse semantics while selectively injecting high frequency detail.

Starting from $Q_6 = P_6$, each step from scale $k{+}1$ to k uses Q_{k+1} as the low frequency input. A lightweight module Φ predicts the required high frequency subbands from (P_k, H_k), and a scene adaptive gate g_k suppresses clutter dominated responses before inverse DWT:

$$\tilde{P}_k = \mathrm{Conv}_{1\times1}\big([\mathrm{IDWT}(Q_{k+1},\ g_k \odot \Phi(P_k, H_k));\ P_k]\big). \tag{9}$$

Here $\tilde{P}_k$ serves as Q_k for the next step. After three steps, WSD outputs $\{\tilde{P}_3, \tilde{P}_4, \tilde{P}_5, P_6\}$ for the shared detection head. Detailed forms of Φ and g_k are provided in Appendix A.6.

3.5 Training Objective

PRISM is optimized end to end by

$$\mathcal{L} = \mathcal{L}_{\mathrm{det}} + \lambda_{\mathrm{wav}}\mathcal{L}_{\mathrm{wav}} + \lambda_{\mathrm{prior}}\mathcal{L}_{\mathrm{prior}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{det}}$ is the standard multi scale detection loss. $\mathcal{L}_{\mathrm{wav}}$ regularizes the finest reconstructed feature $\tilde{P}_3$ by encouraging smooth low frequency structure and target concentrated high frequency responses, providing direct frequency domain supervision for WSD. $\mathcal{L}_{\mathrm{prior}}$ penalizes $D_{\mathrm{KL}}(P_{\mathrm{meta}} \,\|\, P_{\mathrm{img}})$, where P_{img} and P_{meta} are policy distributions without and with metadata modulation, respectively. This anchors metadata adaptation to image driven prioritization and is omitted when metadata is unavailable. We set $\lambda_{\mathrm{wav}} = 0.1$ and $\lambda_{\mathrm{prior}} = 0.01$. Full formulations are provided in Appendix B.1.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. SeaDronesSee [38] provides 8,930 training images and 1,547 validation images across six UAV maritime categories. AFO [39] contains 3,647 images and 39,991 annotated instances across six floating object categories, with more than 99% of objects occupying less than 1% of the image area and more than 30% of images containing over 20 instances. VisDrone2019-DET [40] is used to evaluate architectural applicability after dataset specific training on aerial urban scenes with ten categories and structured background clutter.

Metrics. We follow the COCO evaluation protocol and report AP, AP_{50}, AP_{75}, AP_S, AR_1, and AR_{10}. Efficiency is measured by Params and GFLOPs on a single 640×640 forward pass.

Implementation. All models are trained for 100 epochs at 640×640 resolution with batch size 8 using AdamW. Experiments are conducted on NVIDIA RTX 4090 GPUs with a fixed random seed unless otherwise stated. For the maritime benchmarks, all baselines are reimplemented and retrained under identical settings to ensure fair comparison. Full training and model hyperparameters are provided in Appendix B.2.

4.2 Comparison with State of the Art Methods

Table 1 reports the main comparison on SeaDronesSee and AFO. PRISM achieves the best overall performance on both benchmarks. On SeaDronesSee, PRISM reaches 56.3 AP and 49.8 AP_S, leading on all six metrics. The closest maritime detector, FCDS-DETR, is behind by 5.3 AP and 4.4 AP_{75} points, while requiring 62.0M parameters and 190.0 GFLOPs, compared with 27.2M parameters and 61.7 GFLOPs for PRISM. Compared with Mamba-YOLO, the strongest general baseline, PRISM improves AP, AP_{50}, and AP_S by 5.5, 5.1, and 5.6 points, respectively.

On AFO, PRISM again achieves the highest AP, AP_{50}, AP_S, AR_1, and AR_{10}. It surpasses FCDS-DETR by 2.2 AP points while using less than half of its parameters and about one third of its GFLOPs.

Table 1: Single run comparison on SeaDronesSee and AFO validation sets. All methods are re-trained by us under identical settings to ensure fair comparison. Params and GFLOPs are measured on a single 640×640 forward pass. **Blue** and Green denote best and second best results.

Method	SeaDronesSee						AFO						Params	GFLOPs
	AP	AP_{50}	AP_{75}	AP_S	AR_1	AR_{10}	AP	AP_{50}	AP_{75}	AP_S	AR_1	AR_{10}		
General detectors														
Faster R-CNN[41]	25.2	58.9	21.5	10.2	23.6	30.4	27.5	54.4	25.0	7.5	23.8	30.3	41.1	208.0
DETR[14]	26.8	55.4	23.1	19.8	18.1	28.6	11.4	30.9	7.0	3.2	7.4	22.3	41.3	86.4
Deformable DETR[32]	39.0	74.4	35.8	28.7	28.5	44.6	34.9	59.1	35.7	8.2	20.9	42.7	41.8	179.3
RT-DETR[15]	49.4	83.7	47.8	43.5	40.8	56.3	43.5	78.2	44.8	9.5	32.5	54.2	42.3	129.6
QueryDet[12]	35.8	56.1	38.3	26.8	32.1	40.7	28.8	55.5	26.5	8.5	17.8	34.5	36.4	105.8
YOLOv8[42]	38.4	63.8	39.9	28.2	27.8	33.8	34.5	66.0	35.0	10.5	21.5	40.0	43.7	165.2
YOLOv9[43]	38.6	62.4	39.8	27.2	27.5	34.7	31.6	63.1	31.8	7.3	20.0	37.5	46.1	102.2
YOLOv10[44]	36.3	62.3	37.5	23.1	32.5	43.2	30.2	61.5	30.0	8.8	19.0	36.0	24.3	120.0
TPH-YOLOv5[45]	37.8	59.7	41.0	24.5	33.3	43.3	33.0	64.0	33.5	10.0	20.8	38.5	60.4	145.3
Mamba-YOLO[17]	50.8	83.9	47.8	44.2	40.9	56.3	44.9	80.1	45.5	10.5	26.0	55.5	19.1	45.8
SAMamba[46]	49.6	83.0	46.7	43.1	41.4	56.1	42.7	78.2	43.2	9.3	24.8	52.8	28.8	129.0
Frequency detectors														
FSUNet[25]	45.3	75.2	42.9	34.2	38.0	50.4	37.5	70.0	37.8	10.8	22.5	44.5	38.7	53.1
DWTFreqNet [24]	47.0	75.6	44.5	36.5	39.2	52.5	39.8	72.0	40.2	11.2	23.5	47.0	31.7	66.8
Maritime detectors														
YOLO-OW[18]	39.5	65.0	38.3	26.7	34.7	46.1	34.4	68.3	34.0	10.1	21.5	40.5	40.7	91.4
MARFPNet [20]	38.2	63.8	37.8	27.2	33.9	43.6	33.5	66.3	33.0	9.1	21.0	39.2	23.1	40.9
L-OWRT-DETR[19]	50.2	81.1	52.5	43.9	37.4	53.5	43.6	78.9	44.0	9.8	25.2	54.0	20.1	54.4
FCDS-DETR [21]	51.0	82.7	54.4	–	37.0	59.9	54.5	87.5	**61.6**	–	26.9	57.5	62.0	190.0
ZoomDet[22]	47.2	77.0	48.9	38.9	39.4	56.8	40.8	74.5	41.5	11.0	24.0	49.0	12.2	48.2
PRISM (Ours)	**56.3**	**89.0**	**58.8**	**49.8**	**45.2**	**64.6**	**56.7**	**91.2**	60.5	**14.3**	**38.0**	**58.2**	27.2	61.7

Table 2: Comparison on the VisDrone2019-DET validation set after dataset specific training. †: re-trained by us under identical settings. **Blue** and Green denote best and second best results.

Method	AP	AP_{50}	AP_{75}	AP_S	Params	GFLOPs
Faster R-CNN [41]	24.7	40.5	26.2	15.7	41.4	208.0
ClusDet [47]	28.4	53.2	26.4	19.1	–	–
GFL [13]	26.4	42.7	27.7	17.4	52.3	283.0
TOOD [48]	27.5	44.0	28.8	18.6	32.0	199.0
QueryDet [12]	28.3	48.1	28.8	20.5	36.4	105.8
DINO [49]	34.7	**58.0**	34.9	26.5	47.0	279.0
RT-DETR† [15]	28.4	47.0	28.7	18.9	42.3	129.6
YOLOv8† [42]	30.1	48.3	30.8	20.5	43.7	165.2
YOLOv10† [44]	31.3	49.8	32.1	21.3	24.3	120.0
YOLOv11† [50]	32.0	50.4	33.0	22.1	25.3	87.3
TPH-YOLOv5† [45]	31.2	51.8	31.5	21.0	60.4	145.3
PRISM (Ours)	**36.3**	57.6	**37.1**	**29.7**	27.2	61.7

The only exception is AP_{75}, where FCDS-DETR is higher by 1.1 points. Against Mamba-YOLO, PRISM improves AP by 11.8 points and also achieves the best AP_S, indicating stronger performance on small floating objects.

PRISM consistently improves over the frequency based detectors FSUNet and DWTFreqNet on both datasets, suggesting that auxiliary frequency cues alone are insufficient in challenging maritime scenes. In contrast, PRISM preserves frequency separated representations across encoding, fusion, and wavelet synthesis, rather than merging them back into spatial features. Additional analyses of clutter strength and target density are provided in Appendices C.6 and C.7.

4.3 Applicability to Aerial Urban Scenes

To evaluate the architecture beyond the maritime setting, we train PRISM on the VisDrone2019-DET training split and evaluate it on the validation split. This aerial urban benchmark contains ten object categories and dense scene layouts, with structured clutter from building edges, road markings, vegetation, and shadows rather than ocean waves.

As shown in Table 2, PRISM reaches 36.3 AP and 29.7 AP_S, exceeding the strongest baseline DINO by 1.6 AP and 3.2 AP_S. It also uses 27.2M parameters and 61.7 GFLOPs, lower than the 47.0M

Table 3: Single run core design ablation on SeaDronesSee. Best results are shown in **bold**.

#	Config.	AP	AP_{50}	AP_{75}	AP_S	AR_{10}	Params	GFLOPs
A	w/o WDP	47.0	78.5	48.2	37.5	51.0	24.8	52.5
B	w/ WDP	49.8	82.0	51.0	41.2	54.8	22.5	48.2
C	B + PGF	54.5	87.0	56.8	47.8	61.8	25.3	54.5
D	C + WSD	55.5	88.2	58.0	49.0	63.5	26.8	59.8
E	**Full PRISM**	**56.3**	**89.0**	**58.8**	**49.8**	**64.6**	27.2	61.7

Table 4: Matched single run ablation of Priority Guided Fusion on SeaDronesSee. Best results are shown in **bold**.

#	Variant	w_k	π_k	Causal	γ_k	AP	AP_{50}	AP_{75}	AP_S	AR_{10}
a	Concat	×	×	×	×	49.8	82.0	51.0	41.2	54.8
b	Adaptive mixing	✓	×	×	×	51.3	83.5	52.5	43.0	56.0
c	Fixed order causal	✓	raster	✓	✓	52.2	84.4	53.8	44.2	57.4
d	**Full PGF**	✓	**priority**	✓	✓	**54.5**	**87.0**	**56.8**	**47.8**	**61.8**
e	Random π	✓	random	✓	✓	48.5	80.0	49.5	39.0	52.0

parameters and 279.0 GFLOPs required by DINO. PRISM trails DINO by 0.4 points on AP_{50}, but leads on AP and AP_{75}, with the largest margin on small objects.

The observed small object gains show that frequency separated representation and priority conditioned interaction remain effective under distinct structured clutter after training directly on the target dataset. This result demonstrates dataset specific architectural applicability rather than zero shot domain transfer.

4.4 Ablation Experiments

4.4.1 Core Design Ablation

Table 3 isolates the contribution of each core component by incrementally adding WDP, PGF, and WSD to a spatial baseline that retains the rest of the PRISM topology.

WDP. Adding WDP improves AP from 47.0 to 49.8 and AP_S from 37.5 to 41.2, while reducing model cost from 24.8M parameters and 52.5 GFLOPs to 22.5M and 48.2 GFLOPs. The simultaneous gain in AP_S and efficiency suggests that wavelet decomposition provides a more separable representation for weak target signals than spatial domain features.

PGF. PGF provides the largest single gain, improving AP, AP_S, and AR_{10} by 4.7, 6.6, and 7.0 points over Config B. The larger gains on AP_S and AR_{10} indicate that priority guided interaction is most beneficial for small targets and dense scenes, where weak target evidence is more likely to be suppressed by clutter. Internal PGF design choices are analyzed in Table 4.

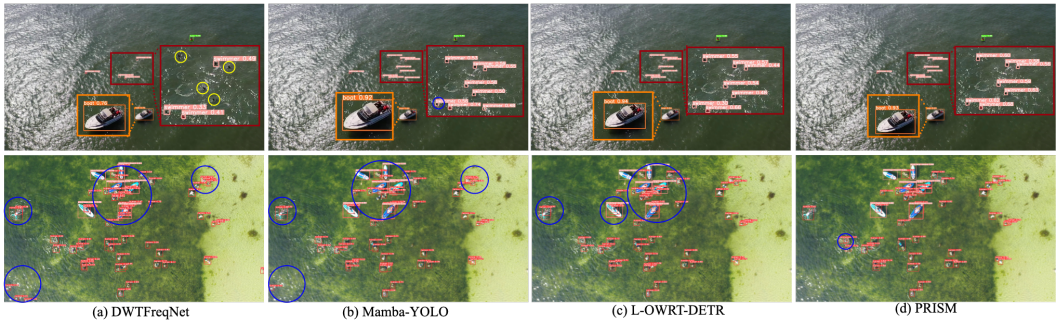
WSD. WSD yields consistent gains, including +1.0 AP, +1.2 AP_{75}, +1.2 AP_S, and +1.7 AR_{10}. The gains on AP_{75} and AR_{10} indicate improved localization quality and recovery under a larger detection budget, consistent with inverse wavelet synthesis reconstructing fine scale features.

Auxiliary components. SSAN, metadata aware temperature, and the two regularizers add 0.8 AP; individual contributions are reported in Appendix C.1. With all components enabled, PRISM reaches 56.3 AP and 49.8 AP_S with 27.2M parameters, only 2.4M more than Config A. The +9.3 AP gain is therefore unlikely to be explained by capacity alone.

4.4.2 Priority Guided Fusion Analysis

Table 4 analyzes the internal design of PGF, with particular focus on priority induced ordering. Compared with direct concatenation in row a, adaptive mixing in row b improves AP by 1.5 points and AP_S by 1.8 points, indicating that content dependent frequency balancing is more effective than directly stacking separated frequency features.

Introducing causal refinement with a fixed raster order in row c further improves AP to 52.2 and AP_S to 44.2, but the gain remains limited without content adaptive ordering. Replacing the raster order with the priority induced order in row d yields the largest improvement, reaching 54.5 AP and 47.8 AP_S. Since rows c and d share adaptive mixing, causal refinement, and gating, the additional gain is associated with priority induced ordering. Row e provides a negative control: random permutation



(a) DWTFreqNet (b) Mamba-YOLO (c) L-OWRT-DETR (d) PRISM

Figure 4: Qualitative comparison on SeaDronesSee (top) and AFO (bottom). Yellow circles: targets missed by baselines but detected by PRISM. Blue circles: false positives of baselines. Additional cases in Appendix D.2.

drops AP to 48.5 and AP_S to 39.0, below the concatenation baseline, showing that arbitrary reordering weakens local refinement; further ordering analyses are provided in Appendix C.2.

4.5 Qualitative Detection Comparison

Figure 4 compares PRISM with representative frequency based, SSM based, and maritime Transformer detectors. On SeaDronesSee, DWTFreqNet produces false positives on wave textured regions, while Mamba-YOLO and L-OWRT-DETR miss several peripheral targets in the dense cluster. These errors reflect two common failure modes in maritime scenes: structured wave responses are confused with target boundaries, and weak targets near dense groups are suppressed. PRISM reduces both errors, producing fewer clutter induced false alarms and recovering more peripheral targets. On AFO, the dominant error is missed detection rather than false alarm. All three baselines miss multiple closely packed small objects, whereas PRISM provides more complete coverage across the dense layout. This qualitative trend is consistent with the quantitative gains on AP_S and AR_{10}, indicating improved small object sensitivity and multi target recovery under structured clutter.

5 Conclusion

We presented PRISM, a wavelet domain detector for sparse target detection in cluttered UAV imagery. PRISM preserves high frequency target evidence and low frequency context throughout the pipeline and fuses them using an order induced by a learned priority map, placing likely target regions earlier in the sequence for local causal refinement. Experiments on SeaDronesSee, AFO, and VisDrone show consistent gains in detection accuracy, particularly for small objects, while maintaining practical efficiency. Ablations indicate that these gains come from persistent frequency separation, priority guided fusion, and wavelet synthesis decoding, rather than increased model capacity. These results suggest the broader value of priority ordered feature interaction for sparse visual detection under structured clutter.

Acknowledgments and Disclosures

This work was supported by the National Natural Science Foundation of China under Grants 62272342, T2422015, and 62603497; the Tianjin Natural Science Foundation Project under Grant 23JCJQJC00070; the China Postdoctoral Science Foundation under Grant 2026M791667; and the Postdoctoral Fellowship Program of CPSF under Grant GZC20260889.

Competing interests. The authors declare no competing interests.

References

[1] J. Wu, L. Cheng, S. Chu, and Y. Song, "An autonomous coverage path planning algorithm for maritime search and rescue of persons-in-water based on deep reinforcement learning," *Ocean Engineering*, 2024.

[2] P. Gao, F. Shi, Y. Yin, X. Cheng, M. Wang, M. Zhao, and S. Chen, "A lightweight global-local multiscale fusion network for small ship detection in remote sensing images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 19, pp. 760–777, 2025.

[3] Y. Yin, X. Cheng, F. Shi, X. Liu, H. Huo, and S. Chen, "High-order spatial interactions enhanced lightweight model for optical remote sensing image-based small ship detection," *IEEE Trans. Geosci. Remote Sens.*, 2024.

[4] J. Song, M. Zhou, J. Luo, H. Pu, Y. Feng, X. Wei, and W. Jia, "Boundary-aware feature fusion with dual-stream attention for remote sensing small object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–13, 2025.

[5] W. Zou, H. Gao, W. Yang, and T. Liu, "Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement," in *Proceedings of the 32nd ACM international conference on multimedia*, 2024, pp. 1534–1543.

[6] Y. Yang, G. Yuan, and J. Li, "Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–17, 2024.

[7] N. Wang, S. Mei, Y. Wang, Y. Zhang, and D. Zhan, "Whanet: Wavelet-based hybrid asymmetric network for spectral super-resolution from rgb inputs," *IEEE Transactions on Multimedia*, vol. 27, pp. 414–428, 2024.

[8] Q. Zhu, K. Li, G. Zhang, X. Wang, J. Huang, and X. Li, "Gdsr: global-detail integration through dual-branch network with wavelet losses for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[9] Z.-H. Cao, Y.-J. Liang, L.-J. Deng, and G. Vivone, "An efficient image fusion network exploiting unifying language and mask guidance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[11] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[12] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 668–13 677.

[13] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 21 002–21 012.

[14] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020*. Springer, 2020, pp. 213–229.

[15] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 965–16 974.

[16] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[17] Z. Wang, C. Li, H. Xu, X. Zhu, and H. Li, "Mamba yolo: A simple baseline for object detection with state space model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025.

[18] J. Xu, X. Fan, H. Jian, C. Xu, W. Bei, Q. Ge, and T. Zhao, "Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from uav aerial imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[19] S. Ma, Y. Zhang, L. Peng, C. Sun, L. Ding, and Y. Zhu, "Owrt-detr: A novel real-time transformer network for small object detection in open water search and rescue from uav aerial imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[20] Q. Ge, W. Da, and M. Wang, "Marfpnet: Multiattention and adaptive reparameterized feature pyramid network for small target detection on water surfaces," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–17, 2024.

[21] Z. Jiao, M. Wang, S. Qiao, Y. Zhang, and Z. Huang, "Transformer-based object detection in low-altitude maritime uav remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[22] T. Wang, C. Lin, C. Tang, J. Zhou, D. Xiong, J. Li, J. Zhao, and J. Lv, "Adaptive image zoom-in with bounding box transformation for uav object detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 233, pp. 452–466, 2026.

[23] S. E. Finder, R. Amoyal, E. Treister, and O. Freifeld, "Wavelet convolutions for large receptive fields," in *Proceedings of the European Conference on Computer Vision*. Springer, 2024, pp. 363–380.

[24] Q. Ma, S. Deng, B. Li, Z. Zhu, Z. Song, X. Li, and H. Hu, "Dwtfreqnet: Infrared small target detection via wavelet-driven frequency matching and saliency-difference optimization," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[25] W. Guo, F. Cai, and W. Quan, "What is in the frequency: Wavelet-guided semantic understanding for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 63, pp. 1–14, 2025.

[26] C. Deng, Z. Zhao, X. Xu, Y. Xia, J. Li, and A. Plaza, "Gsfanet: Global spatial-frequency attention network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[27] J. Pan, C. He, W. Huang, J. Cao, and M. Tong, "Wavelet tree transformer: Multihead attention with frequency-selective representation and interaction for remote sensing object detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–23, 2024.

[28] M. Zhang, R. Zhang, Y. Yang, H. Bai, J. Zhang, and J. Guo, "Isnet: Shape matters for infrared small target detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 867–876.

[29] J. Zhao, Z. Shi, C. Yu, Y. Liu, X. Ying, and Y. Dai, "Multi-scale direction-aware network for infrared small target detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[30] Y. Li, W. Kang, W. Zhao, and X. Liu, "Mlednet: Multi-directional learnable edge information-assisted dense nested network for infrared small target detection," *Electronics*, vol. 14, no. 17, p. 3547, 2025.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[32] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[33] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *Conference on Language Modeling*, 2024.

[34] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, L. Derczynski *et al.*, "Rwkv: Reinventing rnns for the transformer era," in *Findings of the association for computational linguistics: EMNLP 2023*, 2023, pp. 14 048–14 077.

[35] Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, "Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures," *arXiv preprint arXiv:2403.02308*, 2024.

[36] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu, "Vmamba: Visual state space model," *ArXiv*, vol. abs/2401.10166, 2024.

[37] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," *arXiv preprint arXiv:2403.09338*, 2024.

[38] L. A. Varga, B. Kiefer, M. Messmer, and A. Zell, "Seadronessee: A maritime benchmark for detecting humans in open water," *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3686–3696, 2021.

[39] J. Gasienica-Józkowy, M. Knapik, and B. Cyganek, "An ensemble deep learning method with optimized weights for drone-based water rescue and surveillance," *Integrated Computer-Aided Engineering*, vol. 28, pp. 221 – 235, 2021.

[40] D. Du, P. Zhu, L. Wen, X. Bian, H. Ling, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 213–226.

[41] S. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[42] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics yolo," 2023, version 8.0.0, accessed 2026-04-29.

[43] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. Springer, 2024, pp. 1–21.

[44] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "Yolov10: Real-time end-to-end object detection," *ArXiv*, vol. abs/2405.14458, 2024.

[45] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "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)*, Montreal, QC, Canada, 2021, pp. 2778–2788.

[46] W. Xu, S. Zheng, C. Wang, Z. Zhang, C. Ren, R. Xu, and S. Xu, "Samamba: Adaptive state space modeling with hierarchical vision for infrared small target detection," *Information Fusion*, vol. 124, p. 103338, 2025.

[47] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8311–8320.

[48] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "Tood: Task-aligned one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3510–3519.

[49] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J.-J. Zhu, L. M. shuan Ni, and H. yeung Shum, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *ArXiv*, vol. abs/2203.03605, 2022.

[50] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024, version 11.0.0, software.

A Architecture Details

A.1 Scene State Adaptive Normalization

All normalization layers in PRISM are implemented as SSAN, a conditional normalization scheme that adapts affine parameters to the current scene state. At each scale k, SSAN is conditioned on a state vector

$$\mathbf{z}^k = [\mathbf{z}_L^k;\, \mathbf{z}_H^k;\, \mathbf{z}_m] \in \mathbb{R}^{44}, \tag{11}$$

composed of three components.

Low frequency channel statistics. $\mathbf{z}_L^k \in \mathbb{R}^3$ is obtained by global average pooling over the three channels of LL_k, capturing the dominant chromatic and luminance characteristics of the scene.

Directional energy descriptor. $\mathbf{z}_H^k \in \mathbb{R}^9$ is computed by global average pooling over $|LH_k|$, $|HL_k|$, and $|HH_k|$, capturing the per direction high frequency energy distribution that helps distinguish target edges from wave textures.

Metadata embedding. $\mathbf{z}_m = \mathrm{MLP}([\log h,\, \sin\theta,\, \cos\theta]) \in \mathbb{R}^{32}$ encodes flight altitude h and viewing angle θ. When metadata is unavailable, $\mathbf{z}_m$ is replaced by a learned default embedding.

Given an input feature map $\mathbf{U}$, SSAN applies conditional affine modulation after LayerNorm:

$$\mathrm{SSAN}(\mathbf{U};\, \mathbf{z}^k) = \hat{\mathbf{U}} \odot \gamma(\mathbf{z}^k) + \beta(\mathbf{z}^k), \tag{12}$$

where

$$\gamma(\mathbf{z}^k) = 1 + \alpha_s \tanh(W_\gamma\, \mathbf{z}^k), \quad \beta(\mathbf{z}^k) = \alpha_s \tanh(W_\beta\, \mathbf{z}^k). \tag{13}$$

The modulation scale α_s is initialized to 0.01, so that SSAN behaves as standard LayerNorm at the beginning of training and progressively learns scene adaptive modulation during optimization.

A.2 High Frequency Encoder Details

The High Frequency Encoder consists of three internal components in addition to the cross subband interaction described in the main text.

Residual sigmoid gating. The three high frequency subbands are first concatenated along the channel dimension to form HF_k. Since both target edges and wave textures can produce strong responses, a residual sigmoid gate adaptively enhances informative responses while preserving the original signal:

$$HF_k^{\text{gated}} = HF_k \odot \big(1 + \sigma(\mathrm{Conv}_{3\times3}(HF_k))\big). \tag{14}$$

HFBlock structure. The gated features are projected to $3C_k$ channels and refined by $R{=}3$ cascaded HFBlocks. Each HFBlock consists of a five branch parallel depthwise convolution with kernel sizes $\{3, 5, 7\}$ and dilated $3{\times}3$ kernels at rates $\{2, 3\}$, capturing edge patterns at multiple spatial scales; a grouped 1×1 convolution that mixes features across branches; and a directional energy gate conditioned on $\mathbf{z}_H^k$, which adapts the feature response to the current orientation statistics through channel wise affine modulation.

A.3 Low Frequency Encoder Details

For each scale k, the low frequency subband LL_k is first projected to C_k channels by a $1{\times}1$ convolution and flattened into a sequence of T_k tokens. The sequence is then processed by $N{=}2$ Vision-RWKV blocks, each comprising a Spatial Mix module for context aggregation along bidirectional scan paths and a Channel Mix module for token wise feature refinement. The output sequence is finally reshaped back to the spatial domain to obtain the context feature $L_k \in \mathbb{R}^{h_k \times w_k \times C_k}$. The linear complexity in token count makes this choice suitable for the finest scale, where $T_3 = 102{,}400$.

A.4 Policy Network Details

The spatial logit map of the Policy Network is computed as

$$\ell = \mathrm{Conv}_{1\times1}\big(\mathrm{ReLU}(\mathrm{Conv}_{3\times3}(\mathrm{SSAN}(F_{\text{pol}})))\big) \in \mathbb{R}^{h_3 \times w_3}. \tag{15}$$

Metadata aware temperature. Since flight conditions affect the spatial distribution of target evidence, the priority logits are scaled by a metadata aware temperature:

$$\tau = \tau_0 + \Delta\tau \cdot \sigma(\mathbf{w}_\tau^\top \mathbf{z}_m), \tag{16}$$

where τ_0 and $\Delta\tau$ are learnable scalars initialized to 1.0 and 0.5, respectively. Higher altitudes typically lead to sparser target distributions and may benefit from sharper priority peaks, while lower altitudes often have more distributed evidence and may benefit from softer priorities.

Inference behavior. During inference, the Gumbel Softmax in Eq. (5) is replaced by standard Softmax, providing deterministic priority maps. Priority maps at coarser scales P^4_{policy}, P^5_{policy}, and P^6_{policy} are obtained by average pooling P^3_{policy} to the corresponding resolutions.

A.5 Priority Guided Fusion Details

Adaptive mixing weight. The channel wise spatial mixing weight w_k in Eq. (7) is generated from the priority map, target prior, and both frequency streams:

$$w_k = \sigma\big(\text{Conv}_{1\times1}(\text{Conv}_{3\times3}([P^k_{\text{policy}} \otimes \mathbf{1}_C;\, S_k \otimes \mathbf{1}_C;\, H_k;\, L_k]))\big), \tag{17}$$

where $\otimes\mathbf{1}_C$ denotes broadcasting of a single channel map to C channels.

Causal gate. The per token causal gate in Eq. (8) is predicted from the permuted priority and target prior signals:

$$\gamma_k = \sigma(\text{Linear}([P^k_{\text{policy}}[\pi_k];\, S_k[\pi_k]])) \in \mathbb{R}^{T_k\times1}. \tag{18}$$

Inverse permutation and output. After causal refinement, tokens are restored to spatial order through the inverse permutation, concatenated with the low frequency stream, normalized by SSAN, and projected by a 1×1 convolution:

$$P_k = \text{Conv}_{1\times1}(\text{SSAN}([\text{Reshape}(\hat{F}'_k[\pi_k^{-1}]);\, L_k])). \tag{19}$$

A.6 Wavelet Synthesis Decoder Details

Subband prediction module. At each reconstruction step from scale $k{+}1$ to scale k, the high frequency subbands are predicted from the concatenation of P_k and H_k:

$$(\tilde{LH}_k,\, \tilde{HL}_k,\, \tilde{HH}_k) = \Phi\big(\text{AvgPool}([P_k;\, H_k])\big), \tag{20}$$

where Φ is a lightweight convolutional module producing three subbands of dimension $\mathbb{R}^{h_{k+1}\times w_{k+1}\times128}$.

Scene adaptive injection gate. The injection gate g_k controls how much high frequency detail is reintroduced at each position:

$$g_k = \sigma\big(\text{Conv}_{1\times1}(\text{AvgPool}([H_k;\, P^k_{\text{policy}} \otimes \mathbf{1};\, S_k \otimes \mathbf{1}_C]))\big). \tag{21}$$

The predicted subbands are modulated as $g_k \odot (\tilde{LH}_k,\, \tilde{HL}_k,\, \tilde{HH}_k)$. At likely target positions, larger values of g_k preserve edge detail; at background positions, smaller values suppress wave artifacts.

IDWT integration. The gated subbands and the coarser scale feature are combined through inverse DWT:

$$U_k = \text{IDWT}(LL{=}Q_{k+1},\, LH{=}\hat{LH}_k,\, HL{=}\hat{HL}_k,\, HH{=}\hat{HH}_k), \tag{22}$$

where $Q_{k+1} = \tilde{P}_{k+1}$ for $k+1 \in \{4, 5\}$, and $Q_6 = P_6$ at the coarsest scale. The restored feature is fused with the current scale feature:

$$\tilde{P}_k = \text{Conv}_{1\times1}([U_k;\, P_k]) \in \mathbb{R}^{h_k\times w_k\times128}. \tag{23}$$

B Training Details

B.1 Loss Formulation

PRISM is optimized with the standard detection loss augmented by two auxiliary regularizers:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{wav}}\mathcal{L}_{\text{wav}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}. \tag{24}$$

Table 5: Hyperparameters used in all experiments.

Parameter	Value
Architecture	
Channel dimensions C_3, C_4, C_5, C_6	96, 192, 384, 768
HFBlock repetitions R	3
LFE Vision-RWKV blocks N per scale	2
Final fused channel dimension	128
CausalDWConv1D kernel size	7
Loss weights	
λ_{wav}	0.1
λ_{prior}	0.01
Initialization	
SSAN scale α_s	0.01
Policy temperature τ_0	1.0 (learnable)
Policy temperature $\Delta\tau$	0.5 (learnable)
Optimization	
Optimizer	AdamW
Initial LR	1×10^{-4}
LR schedule	Decay by 0.1 after epoch 80
Weight decay	5×10^{-4}
Batch size	8
Epochs	100
Input resolution	640×640

Detection loss. The detection loss is computed over all prediction scales:

$$\mathcal{L}_{\text{det}} = \sum_{k=3}^{6} \big(\mathcal{L}_{\text{cls}}^{k} + \mathcal{L}_{\text{obj}}^{k} + 5\, \mathcal{L}_{\text{box}}^{k} \big), \tag{25}$$

where $\mathcal{L}_{\text{cls}}$ is implemented by Focal Loss, $\mathcal{L}_{\text{obj}}$ by binary cross entropy, and $\mathcal{L}_{\text{box}}$ by the sum of CIoU and ℓ_1 regression losses.

Wavelet domain regularization. A one level Haar DWT is applied to the finest reconstructed feature $\tilde{P}_3$, producing one low frequency component and three high frequency components:

$$\mathcal{L}_{\text{wav}} = \mathcal{L}_{\text{lf}} + 0.5\, \mathcal{L}_{\text{hf}}, \tag{26}$$

where $\mathcal{L}_{\text{lf}}$ is an ℓ_1 penalty on the gradient magnitude of the reconstructed low frequency component, encouraging smooth scene structure, and $\mathcal{L}_{\text{hf}}$ is an ℓ_1 sparsity penalty applied to the high frequency components outside ground truth bounding boxes, suppressing clutter dominated responses.

Policy regularization. Let $P_{\text{img}} = \text{Softmax}(\ell/\tau_0)$ denote the image driven policy and $P_{\text{meta}} = \text{Softmax}(\ell/\tau(h, \theta))$ denote the metadata modulated policy. The regularization aligns the two distributions through KL divergence:

$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}\big(P_{\text{meta}} \,\|\, P_{\text{img}}\big). \tag{27}$$

This term is omitted when metadata is unavailable.

B.2 Hyperparameters

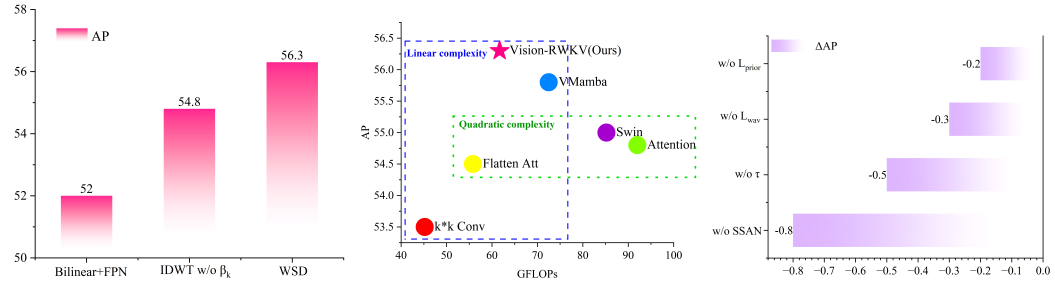
B.3 Compute Resources

All experiments are conducted on a workstation with two NVIDIA RTX 4090 GPUs, each with 24 GB memory. Params and GFLOPs are measured on a single 640×640 forward pass. All reported training runs use the same input resolution, batch size, and epoch schedule described in Appendix B.2. Beyond the reported experiments, the project also included preliminary architecture exploration and intermediate ablation runs, which required additional compute not detailed here.

C Additional Experimental Analysis

C.1 Component Analysis

We analyze three groups of component design choices in PRISM (Figure 5).



(a) Decoder ablation (b) LFE token mixer (c) Auxiliary designs

Figure 5: Component analysis of PRISM. (a) Decoder ablation. (b) LFE token mixer comparison in the accuracy efficiency space. (c) Auxiliary design ablation, where each bar indicates the AP drop after removing one component from the full model.

Table 6: Extended matched single run ordering analysis on SeaDronesSee. Rows (a)-(e) reproduce the main paper PGF ablation (Table 4); rows (f)–(h) further analyze the priority induced ordering.

#	Variant	w_k	π_k	Causal	γ_k	AP	AP_{50}	AP_{75}	AP_S	AR_{10}
a	Concat	×	×	×	×	49.8	82.0	51.0	41.2	54.8
b	Adaptive mixing	✓	×	×	×	51.3	83.5	52.5	43.0	56.0
c	Fixed order causal	✓	raster	✓	✓	52.2	84.4	53.8	44.2	57.4
d	**Full PGF**	✓	**priority**	✓	✓	**54.5**	**87.0**	**56.8**	**47.8**	**61.8**
e	Random π	✓	random	✓	✓	48.5	80.0	49.5	39.0	52.0
f	Reversed π	✓	reversed	✓	✓	46.0	77.5	46.8	36.4	49.8
g	HFE prior only	✓	S_k	✓	✓	53.1	85.5	55.0	45.8	60.0
h	Image shared π	✓	shared	✓	✓	52.6	84.8	54.4	45.0	58.5

Decoder design (Figure 5(a)). Replacing the proposed WSD with a conventional Bilinear+FPN decoder reduces AP from 56.3 to 52.0, indicating that direct spatial domain upsampling is less suitable for the wavelet domain pipeline. Retaining IDWT while removing the injection gate g_k achieves 54.8 AP, showing that inverse wavelet reconstruction itself provides a clear benefit. With the full WSD, AP further increases to 56.3, suggesting that g_k helps regulate high frequency reconstruction by suppressing clutter dominated wave responses while preserving target relevant structures.

LFE token mixer (Figure 5(b)). Six token mixing strategies are compared. The $k \times k$ convolution obtains the lowest AP of 53.5, indicating that local receptive fields alone are insufficient for low frequency encoding. Self attention and Swin improve AP to 54.8 and 55.0, respectively, but require higher computational cost. Among the more efficient alternatives, Flatten Attention reaches 54.5 AP at 55.8 GFLOPs, while VMamba obtains 55.8 AP at 72.5 GFLOPs. Vision-RWKV achieves the best AP of 56.3 at 61.7 GFLOPs, providing the best accuracy efficiency trade off.

Auxiliary designs (Figure 5(c)). The auxiliary components provide complementary improvements on top of the main architecture. Removing SSAN causes the largest AP drop of 0.8, indicating the importance of scene adaptive normalization. Removing the metadata aware temperature leads to a 0.5 AP drop, suggesting that adaptive priority sharpness helps the Policy Network adjust to different flight and scene conditions. The wavelet domain loss $\mathcal{L}_{\text{wav}}$ and prior loss $\mathcal{L}_{\text{prior}}$ cause AP drops of 0.3 and 0.2, respectively, providing stable regularization.

C.2 Extended Scan Order Analysis

Rows (f) to (h) further analyze the role of priority induced ordering. Reversing the priority scores reduces AP to 46.0, below the random ordering result of 48.5, showing that the direction of the induced order is performance critical. Ordering directly by the HFE prior S_k reaches 53.1 AP, 1.4 points below full PGF, demonstrating the additional benefit of jointly conditioning on H_k and L_k. Sharing one ordering across all images reaches 52.6 AP, supporting the value of per image adaptation.

C.3 Priority Alignment and Multi Run Stability

We evaluate whether the induced order aligns with detection targets at the finest implemented policy scale, $k = 3$. Each ground truth instance is mapped to the token containing the center of its bounding

Table 7: Priority target alignment on the SeaDronesSee validation set. Lower mean rank is better.

Ranking source	Mean GT center rank (%) ↓	GT center Recall@5% ↑	GT center Recall@10% ↑
Random ordering (expected)	50.0	5.0	10.0
HFE derived prior S_3	35.8	14.6	25.7
Learned policy P^3_{policy}	**23.4**	**27.9**	**44.6**

Table 8: Three seed results on SeaDronesSee, reported as mean ± sample standard deviation over complete training runs.

Method	AP	AP_{50}	AP_{75}	AP_S	AR_{10}
Mamba-YOLO	50.8 ± 0.4	83.9 ± 0.3	47.8 ± 0.6	44.2 ± 0.5	56.3 ± 0.7
Fixed order causal PGF	52.2 ± 0.5	84.4 ± 0.3	53.8 ± 0.6	44.2 ± 0.5	57.4 ± 0.6
Full PGF (priority induced)	54.5 ± 0.4	86.9 ± 0.3	56.7 ± 0.5	47.8 ± 0.4	61.8 ± 0.5
Full PRISM	$\mathbf{56.3 \pm 0.3}$	$\mathbf{88.9 \pm 0.2}$	$\mathbf{58.7 \pm 0.4}$	$\mathbf{49.8 \pm 0.4}$	$\mathbf{64.6 \pm 0.5}$

box, and its percentile in the priority ranking is measured. Ground truth annotations are used only for this post hoc analysis and not for priority learning.

Table 7 shows that the learned policy places tokens containing target centers substantially earlier than both random ordering and the image derived HFE prior. This establishes measurable target alignment without using target locations to supervise the priority map.

We additionally repeat four key configurations with seeds 0, 1, and 2. Each architecture uses its corresponding official pretrained backbone checkpoint; parameters not covered by the checkpoint are initialized under the stated seed. All runs use the same data split, training budget, optimization schedule, augmentation, and evaluation protocol. Full PGF and fixed order causal PGF differ only in the ordering mechanism.

Full PRISM exceeds Mamba-YOLO by 5.5 AP, while priority induced PGF exceeds the matched fixed order variant by 2.3 AP. These margins are substantially larger than the observed variation across runs.

C.4 Metadata Robustness

To isolate the core architecture from flight information, we train a strictly metadata free variant from scratch. It replaces flight metadata with a learned default embedding shared across images, uses a learned metadata independent temperature τ_0, and removes $\mathcal{L}_{\text{prior}}$; all other settings remain unchanged.

The fully metadata free model remains within 0.9 AP of Full PRISM and exceeds Mamba-YOLO by 4.6 AP. We also perturb the metadata of Full PRISM at test time without retraining. Altitude uses multiplicative Gaussian noise with standard deviation 0.1 or 0.2, and viewing angle uses additive Gaussian noise with standard deviation 5° or 10° for the moderate and severe settings, respectively.

C.5 Wavelet Basis and Runtime Analysis

Performance varies by at most 0.4 AP across the three bases, indicating that the gains are not restricted to Haar. Haar gives the highest measured result and remains the default.

Latency is averaged over 500 CUDA synchronized iterations after 50 warmup iterations and excludes data loading, preprocessing, and NMS. Dynamic ordering adds 0.7 ms, or 3.9%, relative to the matched fixed raster implementation. Embedded edge hardware is not evaluated.

C.6 Performance Stratified by Background Clutter Strength

To understand under what conditions priority guided fusion contributes most, we stratify the SeaDronesSee validation set by background clutter strength and reevaluate PRISM against three strong baselines on each stratum. This analysis complements the density wise stratification reported in Appendix C.7, examining a different axis of scene difficulty.

Table 9: Matched single run metadata ablation on SeaDronesSee.

Variant	AP	AP_{50}	AP_{75}	AP_S	AR_{10}
Full PRISM	**56.3**	**89.0**	**58.8**	**49.8**	**64.6**
Without metadata aware temperature	55.8	88.6	58.2	49.3	63.9
Fully metadata free	55.4	88.1	57.8	48.6	63.3

Table 10: Robustness to test time metadata noise. Noisy results are mean $\pm$ sample standard deviation over five noise realizations.

Metadata setting	AP	AP_S	AR_{10}
Clean metadata	56.3	49.8	64.6
Moderate noise	56.0 ± 0.1	49.4 ± 0.2	64.1 ± 0.2
Severe noise	55.6 ± 0.2	48.9 ± 0.3	63.5 ± 0.3

Clutter metric. For each image, a clutter score is computed as the mean high frequency energy outside ground truth bounding boxes:

$$c(I) = \frac{1}{|\Omega_{\mathrm{bg}}|} \sum_{(i,j)\in\Omega_{\mathrm{bg}}} \big(|LH_1| + |HL_1| + |HH_1|\big)_{i,j}, \tag{28}$$

where Ω_{bg} denotes the background region, namely image positions outside any ground truth box, and LH_1, HL_1, and HH_1 are the level one Haar high frequency subbands. Images are sorted by $c(I)$ and divided into three equal sized strata: low, medium, and high clutter.

Table 13 shows that the advantage of PRISM grows with clutter strength, from $+2.4$ AP under low clutter to $+5.1$ under medium clutter and $+7.8$ under high clutter. From low to high clutter, AP decreases by 10.2 to 11.9 points for the baselines, compared with only 5.6 points for PRISM. This pattern is consistent with the intended role of priority guided fusion: target relevant positions are processed earlier, reducing the chance that weak target evidence is suppressed by strong background responses. The margin becomes larger under stronger clutter.

C.7 Density Stratified Analysis

Complementing the clutter strength stratification in Appendix C.6, we further analyze PRISM along a second axis of scene difficulty: target density. We stratify the SeaDronesSee validation set into sparse, medium, and dense subsets based on the number of target instances per image, and reevaluate the model and three strong baselines on each subset.

External comparison. All methods exhibit performance degradation as scene density increases. PRISM consistently achieves the highest AP across all density levels, obtaining 54.8, 53.2, and 49.1 AP in sparse, medium, and dense scenes, respectively. In comparison, L-OWRT-DETR decreases from 52.0 to 44.8 AP and Mamba-YOLO decreases from 52.4 to 45.4 AP from sparse to dense scenes. PRISM therefore maintains stronger robustness when target crowding and maritime background interference increase.

Internal comparison. Full PGF improves AP by 4.5 points over concat and 2.8 points over fixed order fusion in dense scenes, indicating that learned priority guided scanning is most beneficial when target responses are densely distributed.

D Additional Visualizations

D.1 Priority Map Visualization

The priority maps learned by the Policy Network exhibit consistent behavior across scenes of varying complexity. In sparse scenes, the priority map produces sharp, isolated high priority activations that align with the target positions, while the calm background remains suppressed. As the number of targets increases in medium density scenes, the Policy Network maintains distinct high priority peaks for individual targets without merging nearby responses.

Table 11: Matched single run wavelet basis comparison on SeaDronesSee.

Wavelet basis	AP	AP_S	AR_{10}
Haar	**56.3**	**49.8**	**64.6**
Daubechies-2	56.1	49.6	64.3
Coiflet-1	55.9	49.3	64.0

Table 12: Model forward latency on one RTX 4090 at 640×640, batch size 1, and FP32.

Method	Latency (ms) ↓	FPS ↑	Params (M)	GFLOPs
Mamba-YOLO	14.2	70.4	19.1	45.8
Full PRISM, fixed raster	18.1	55.2	27.2	61.7
Full PRISM, priority induced	18.8	53.2	27.2	61.7

In dense scenes, the priority map shifts from isolated peaks to sustained high activation across the target cluster. This cluster level prioritization allows PGF to process targets in the group before attending to background positions, consistent with the density stratified analysis in Appendix C.7.

The mixed scale scene is particularly informative. Despite the simultaneous presence of a large vessel, small targets, and strong wave texture, the priority map assigns high activation to targets at multiple scales while suppressing wave dominated regions. Minor residual responses are visible in wave textured areas but remain weaker than the target activations, suggesting that the Policy Network integrates low frequency context from LFE to distinguish target edges from background clutter.

D.2 Additional Qualitative Comparisons

Figure 8 extends the qualitative comparison in the main text to four representative scenarios.

Row 1: Dense maritime scene with strong wave interference. DWTFreqNet generates false positives in wave textured regions (blue circles), indicating that auxiliary wavelet processing is less effective at suppressing background high frequency responses after the features are merged back into the spatial domain. Mamba-YOLO and L-OWRT-DETR miss several targets at the cluster periphery (yellow circles), where weak target signals are overwhelmed by adjacent wave responses during synchronous fusion. PRISM reduces both error types, producing fewer clutter induced false positives and recovering more peripheral targets.

Row 2: Tiny targets on a calm sea surface. All three baselines miss the smallest targets (yellow circles), whose edge signals occupy only a few pixels. PRISM recovers these targets, consistent with the role of wavelet decomposition in separating faint edge responses from low frequency background structure and the role of the Policy Network in assigning priority to weak target evidence.

Row 3: AFO scene with strong wave reflections. Both DWTFreqNet and Mamba-YOLO produce false positives in reflective regions (blue circles), while L-OWRT-DETR misses low contrast targets. PRISM produces more complete detections with fewer false alarms, consistent with full pipeline frequency separation under reflective water surface clutter.

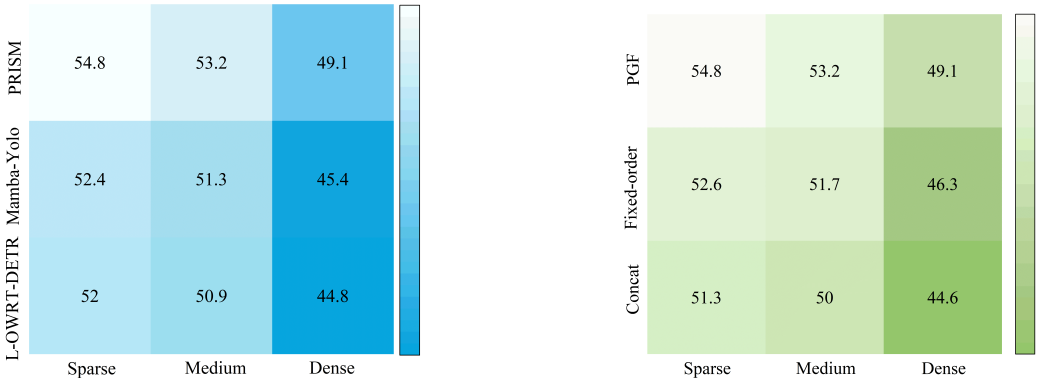
Row 4: AFO dense scene with numerous small objects. All three baselines exhibit scattered missed detections throughout the scene (yellow circles), particularly for closely packed targets. PRISM achieves more complete coverage, indicating that learned priority guided scanning is beneficial when target responses are densely distributed.

E Dataset and Asset Usage

All datasets used in this paper are public research benchmarks. We cite the original sources of SeaDronesSee, AFO, and VisDrone2019-DET in the main paper and follow the access terms provided by the respective dataset creators. Baseline implementations follow official repositories when available or standard public implementations otherwise. The official training and evaluation code used for the reported PRISM experiments is available at `https://github.com/GPQ-phD/PRISM`. The repository provides environment setup, dataset preparation, training, evaluation, inference, and independent seed instructions, together with a manuscript to code component mapping.

Table 13: AP on SeaDronesSee stratified by background clutter strength. Each stratum contains one third of the validation images.

Method	Low	Medium	High	Overall
FCDS-DETR	55.2	51.8	45.0	51.0
Mamba-YOLO	56.0	51.2	44.1	50.8
L-OWRT-DETR	54.8	50.3	43.9	50.2
PRISM	**58.4**	**56.9**	**52.8**	**56.3**
PRISM gain	+2.4	+5.1	+7.8	+5.3



(a) External comparison (b) Internal comparison

Figure 6: Density wise performance analysis on SeaDronesSee. (a) External comparison with three strong baselines across sparse, medium, and dense subsets. (b) Internal comparison among PGF variants under the same density stratification.

F Limitations

PRISM has several limitations that suggest directions for future work. First, although Haar, Daubechies-2, and Coiflet-1 give stable results, the current model uses a fixed basis rather than learning filters for the target domain. Second, the continuous priority map induces a discrete, non differentiable ordering. Sorting changes adjacency inside the gated local causal branch, whose direct kernel-7 range is limited to six preceding positions; it does not provide direct propagation to all later tokens. The HFE/LFE representations and residual path retain 2D features, and inverse permutation restores spatial coordinates, but a fully 2D priority conditioned interaction remains unexplored. Third, metadata provides a complementary gain. The metadata free and corruption experiments show retained performance when it is absent or noisy, though deployment conditions may differ from these controlled tests. Fourth, VisDrone is evaluated after dataset specific training and therefore supports architectural applicability rather than zero shot transfer. Finally, latency is measured on an RTX 4090; embedded edge hardware, extreme weather, low visibility, motion blur, and non RGB sensors remain to be evaluated.

G Broader Impact

PRISM may support maritime search and rescue by improving detection of small objects in cluttered UAV imagery. However, UAV based detection systems may also raise privacy and surveillance concerns if deployed without appropriate governance. Practical deployment should therefore include human oversight, clear operational constraints, and compliance with local regulations.

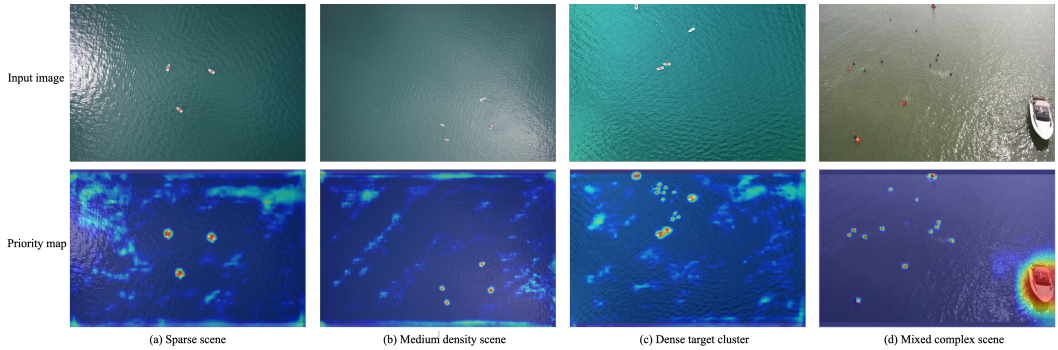


Figure 7: Visualization of the learned priority maps P_{policy} on SeaDronesSee. Top row: input images. Bottom row: corresponding priority maps, where warm colors indicate high target likelihood. From left to right: sparse targets, medium density, dense cluster, and mixed scale scene with wave interference.

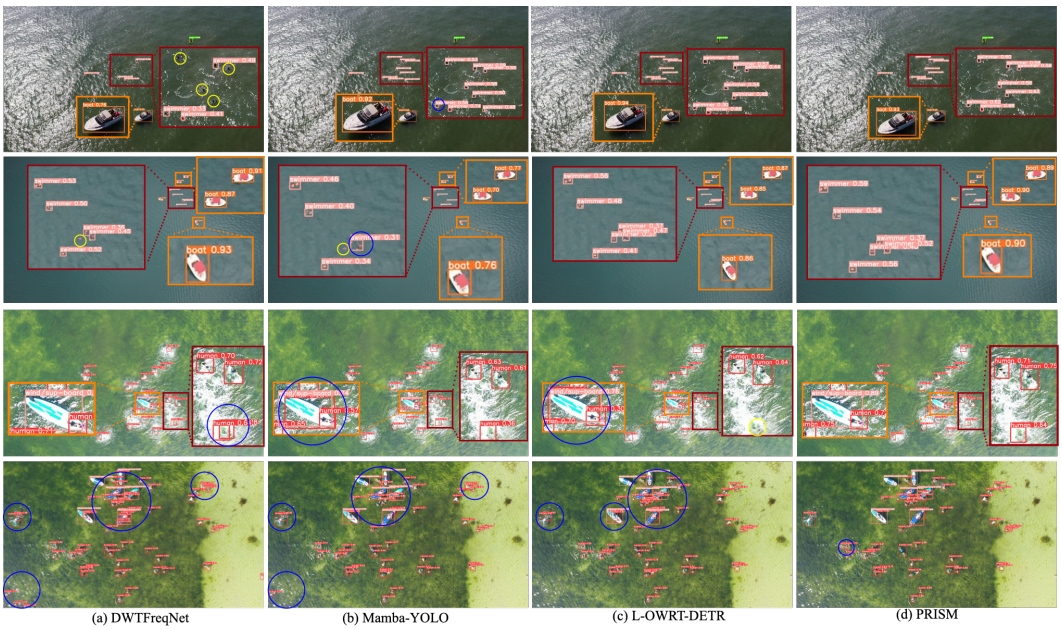


Figure 8: Qualitative detection comparison on SeaDronesSee (rows 1–2) and AFO (rows 3–4). Columns from left to right: DWTFreqNet, Mamba-YOLO, L-OWRT-DETR, and PRISM. Yellow circles highlight targets missed by baseline methods but detected by PRISM. Blue circles indicate false positives produced by baseline methods.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that PRISM targets sparse object detection under structured clutter through frequency separated representation and priority conditioned local interaction. The experimental claims are supported by results on SeaDronesSee and AFO, with dataset specific evaluation on VisDrone; the scope is limited to UAV maritime and aerial detection benchmarks.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix F, including fixed rather than learned wavelet filters, the non differentiable induced ordering and local causal range, the remaining 2D locality tradeoff, metadata availability, dataset specific VisDrone training, the lack of embedded edge hardware measurements, and evaluation scope under adverse conditions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper proposes an empirical detection architecture and does not include formal theoretical results or theorem proof claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the architecture, datasets, evaluation metrics, training protocol, implementation settings, and ablation configurations in the main text and Appendix B.2. Appendix E also identifies the public implementation used for the reported experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: The experiments use public datasets cited in the paper, and the official training and evaluation code used for the reported PRISM experiments is publicly available. Appendix E provides the repository location, whose documentation covers environment setup, dataset preparation, training, evaluation, inference, and independent seed runs.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup specifies the datasets, splits, metrics, optimizer, input resolution, training schedule, batch size, hardware, and fairness protocol for retraining baselines. Full hyperparameters are provided in Appendix B.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix C.3 reports mean and sample standard deviation over three complete training runs with seeds 0, 1, and 2 for the main baseline, the matched ordering control, Full PGF, and Full PRISM. Appendix C.4 also reports sample standard deviations over five independent metadata noise realizations. The factors of variability and protocols are stated with the corresponding tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute environment in Appendix B.3, including the use of a workstation with two NVIDIA RTX 4090 GPUs with 24 GB memory each. Params and GFLOPs are measured on a single 640×640 forward pass, and the training schedule is specified in Appendix B.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study uses publicly available datasets and focuses on object detection for UAV imagery. No private data, personally collected human subject data, or deceptive data collection procedures are used.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix G discusses potential positive impacts for maritime search and rescue, as well as possible risks related to UAV based detection systems, privacy, and surveillance if deployed without appropriate governance.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a high risk pretrained generative model, language model, or scraped dataset. The work uses existing public detection benchmarks and proposes an object detection architecture.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original dataset papers and baseline methods used in the experiments. Dataset access and usage follow the terms provided by the original dataset creators.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper releases the official PRISM implementation. The repository documents installation, dataset preparation, training, evaluation, inference, independent seed runs, and the correspondence between manuscript components and source files. The paper does not introduce a new dataset or benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use public detection benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not conduct new research with human subjects or collect new human subject data. It evaluates models on existing public datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or scientific claims. Any use for writing or formatting assistance does not affect the core methodology or results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.