PAVER: PRESERVING ACROSS-SCALE VISUAL EVIDENCE WITH RISK-CONDITIONED INTERFACE FOR REMOTE-SENSING TINY-OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote-sensing detectors must recover few-pixel targets from cluttered wide-area imagery, while the reliability of candidate confidence is rarely examined separately. We identify an evidence-retention bottleneck spanning both stages: weak cues must survive feature formation and cross-scale propagation, then remain accessible for confidence conditioning at inference. Preserving Across-scale Visual Evidence with Risk-conditioned Interface (PAVER) is a framework combining an evidence-preserving detector with a separately fitted Risk-conditioned Interface. The Complementary State-Flow Mixer (CSFM), Semantic–Contour Recalibration Gate (SCRG), and High-Resolution Pyramid Extension (HRPE) maintain complementary, localization-sensitive, and high-resolution pathways. The interface makes candidate confidence more trustworthy by recalibrating each fixed detection toward its likelihood of being a true positive, using a false-positive–true-positive prototype margin from P2 and P3 region-of-interest (ROI) evidence without changing boxes or classes. Under controlled reproduction, PAVER improves YOLOv12-s by 5.7 and 4.0 points in mAP_{50} and $mAP_{50:95}$, respectively, on VisDrone, with complementary evaluation on RSOD and DIOR. On VisDrone test-dev, risk conditioning improves ECE, Brier score, NLL, and FP-AUC over raw confidence while keeping AP_{50} essentially unchanged; it also achieves the best Brier score and FP-AUC among the evaluated calibrators. PAVER thereby couples strong tiny-object detection with more trustworthy candidate confidence through a unified evidence pathway. Code is publicly available at `https://anonymous.4open.science/r/PAVER-F47C/`.

1 INTRODUCTION

Remote-sensing detection supports wide-area search and monitoring (Nikouei et al., 2025), yet high-altitude imaging can compress targets below 20×20 pixels (Cao et al., 2024; Qu et al., 2025). Dense layouts and structured clutter further obscure their limited cues (Zhou et al., 2025; Wu et al., 2025). Their loss causes missed or poorly localized targets and can also yield confident background responses. Most work is optimized and compared through aggregate AP, leaving whether individual scores reflect candidate correctness comparatively underexamined.

We view these as two stages at which detector evidence can be lost or underused. Weak cues must survive feature transformation, semantic–spatial fusion, and pyramid propagation; after detection, candidate-local evidence is often compressed into a scalar score. Multi-scale fusion (Qu et al., 2025; Cao et al., 2024), high-resolution heads (Hafeez et al., 2025; Zhou et al., 2025; Wei et al., 2025), and distance-aware assignment (Huo et al., 2023; Hu et al., 2025) improve exposure or supervision, but cannot recover evidence already erased upstream. Score-only calibration likewise cannot distinguish equally scored candidates with different ROI evidence (Alexandridis et al., 2025; Beemelmanns et al., 2026; Que et al., 2026). Existing strategies therefore largely treat feature preservation and confidence quality separately.

Our framework, Preserving Across-scale Visual Evidence with Risk-conditioned Interface (PAVER), connects both stages through one evidence pathway. The Complementary State-Flow Mixer (CSFM) retains a reference state beside evolving states, the Semantic–Contour Recalibration Gate (SCRG)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

reconciles semantic and localization-sensitive streams, and the High-Resolution Pyramid Extension (HRPE) propagates high-resolution states across P2–P5. P2 and P3 then support detection and supply ROI evidence to a validation-fitted FP–TP prototype margin. The interface makes candidate confidence more trustworthy by re-scoring each fixed detection to better represent its true-positive likelihood, while keeping boxes and classes unchanged.

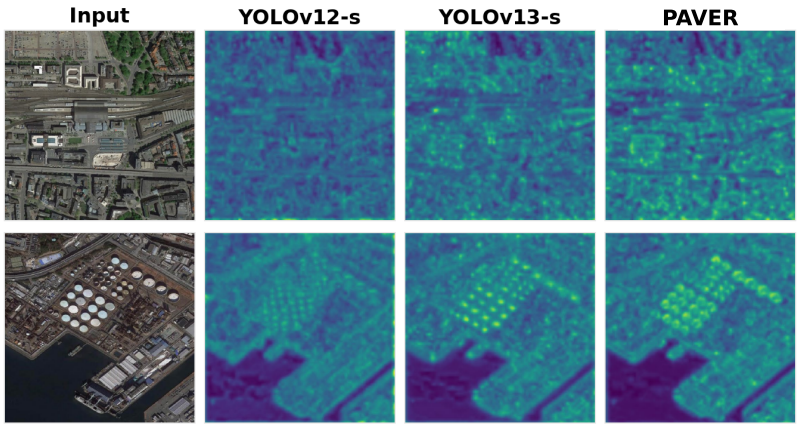


Figure 1: Qualitative activation-map comparison. In these examples, PAVER concentrates target-aligned responses and suppresses diffuse background activations; the full gallery is in Supplementary Fig. S7.

Our contributions are:

- We propose PAVER, a remote-sensing tiny-object detection framework combining an evidence-preserving detector, scale-dynamic localization supervision, and a Risk-conditioned Interface. Controlled evaluation across VisDrone, RSOD, and DIOR demonstrates strong tiny-object and strict-localization performance while reporting the computational boundary.
- CSFM, SCRG, and HRPE coordinate complementary state formation, reciprocal semantic–localization recalibration, and high-resolution propagation. Component isolation and matched replacements evaluate their contributions to detection accuracy.
- A separately validation-fitted Risk-conditioned Interface uses a signed FP–TP prototype margin from P2 and P3 ROI evidence to make fixed-candidate confidence more trustworthy without changing candidates, boxes, or classes. It improves all four reported confidence-quality metrics over raw confidence while keeping AP_{50} essentially unchanged.

2 Related Work

2.1 Remote-Sensing Tiny-Object Detection

YOLO-style and high-resolution detectors expose fine-scale features (Zhou et al., 2025; Wei et al., 2025; Liu et al., 2025a; Hafeez et al., 2025), while contrastive or distance-aware designs improve discrimination and assignment (Hu et al., 2025). Other work explores vision-language, multimodal, anisotropic, or geometric representations (Shi et al., 2026; Li et al., 2026b; Yuan et al., 2026; Zhan et al., 2026). PAVER instead coordinates feature formation, semantic–localization fusion, and pyramid propagation.

2.2 Evidence Preservation for Tiny Objects

Difference-map (Cao et al., 2024), information- and frequency-based (Bian et al., 2025; Sun et al., 2025b; Li et al., 2026a), and uncertainty-aware methods (Sun et al., 2025a) strengthen weak-target representations or optimization; boundary and geometric cues support localization (Yang et al.,

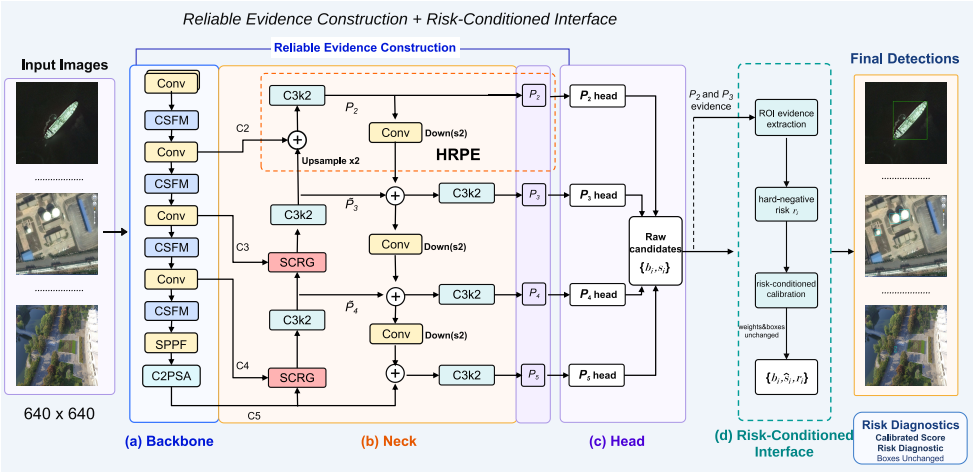


Figure 2: Overview of PAVER. Panels (a)–(c) show the CSFM backbone, SCRG–HRPE pyramid, and P2–P5 heads. Panel (d) shows the Risk-conditioned Interface, which recalibrates fixed detections using an FP–TP prototype margin over P2 and P3 ROI evidence (Eq. 11) without changing boxes or classes. Classes are omitted; $\oplus$ denotes concatenation.

2025; Guan et al., 2026; Zhan et al., 2026). PAVER treats preservation as a pathway spanning formation, fusion, and propagation.

2.3 Calibration and Risk-Aware Detection

Calibration studies separate confidence quality from accuracy under long tails and shift (Lin et al., 2025; Alexandridis et al., 2025; Beemelmanns et al., 2026; Que et al., 2026); reliability methods use consistency, ordering, or uncertainty (Yoo et al., 2025; Park et al., 2026). PAVER instead makes confidence candidate-aware through prototype-relative ROI evidence, connecting feature preservation to confidence trustworthiness while evaluating it separately from detector AP.

3 Method

3.1 Task Formulation and System Overview

Architectural lineage and design rationale. Figure 2 links an evidence-preserving detector to a separately fitted confidence stage that shares its features but does not account for detection gains. Inspired by YOLOv11-s (Khanam & Hussain, 2024), PAVER retains its compact C3k2 hierarchy and bidirectional neck as a strong, controlled scaffold, but redesigns three evidence-vulnerable sites in priority order. HRPE adds P2 and separates head refinement from bottom-up propagation to retain fine cues. SCRG replaces indiscriminate mixing at two early-neck concatenation–C3k2 sites with semantic–contour recalibration. CSFM replaces four backbone C3k2 stages, retaining weak responses in a reference state while evolving context. Because aggregate AP leaves candidate confidence quality largely unexamined, the inference-time Risk-conditioned Interface derives FP–TP prototype risk from P2 and P3 ROI evidence to make fixed-candidate confidence more trustworthy without changing boxes or classes.

For $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, the detector produces $\{P_2, P_3, P_4, P_5\}$ at strides $\{4, 8, 16, 32\}$ and candidates $d_i = (b_i, \widehat{c}_i, s_i)$, denoting box, class, and raw score $s_i \in (0, 1)$. Let $D_j = \text{channels}(P_j)$. The interface returns $\mathcal{O}_{\text{PAVER}}(\mathbf{I}) = \{b_i, \widehat{c}_i, \widehat{s}_i, r_i\}_{i=1}^{M}$, where M is the number of retained candidates and r_i is the risk proxy. After Fig. 2(c), only scores change; weights, candidates, classes, and boxes remain fixed.

Detector path and HRPE. HRPE builds P2 while separating prediction refinement from cross-scale propagation. The backbone exposes the four right-going tensors in Fig. 2(a): C_2, C_3, and C_4 follow the Conv blocks above the second, third, and fourth CSFM blocks at strides 4, 8, and 16. After

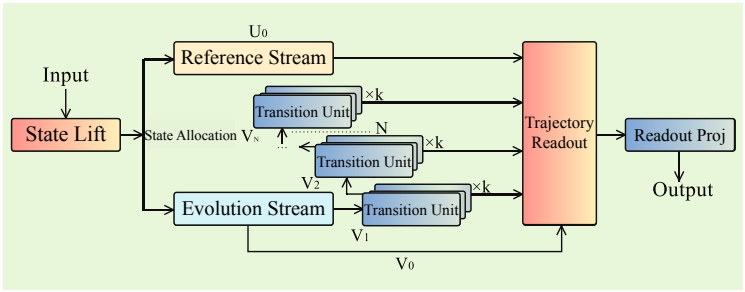


Figure 3: CSFM allocates lifted state $\mathbf{U}$ into reference $\mathbf{U}_0$ and evolution seed $\mathbf{V}_0$. The seed traverses N transitions of k bottlenecks before trajectory readout; Eq. 3 defines the operator.

the fourth CSFM, an independent stride-2 Conv–BN–SiLU precedes SPPF and C2PSA to produce C_5 at stride 32; this routine transition is omitted from the schematic. Two SCRGs consume these tensors before HRPE. Let $\mathcal{R}$, $\mathcal{U}$, $\mathcal{D}$, and $\oplus$ denote C3k2 refinement, nearest-neighbor upsampling, stride-2 Conv–BN–SiLU, and concatenation. G_2–G_5 are fusion states; stage-indexed $\mathcal{R}$ and $\mathcal{D}$ have independent parameters and destination widths:

$$\widetilde{P}_4 = \mathcal{R}(\mathrm{SCRG}(C_5, C_4)), \qquad \widetilde{P}_3 = \mathcal{R}\Big(\mathrm{SCRG}(\widetilde{P}_4, C_3)\Big), \tag{1}$$

$$\begin{aligned} G_2 &= \mathcal{U}(\widetilde{P}_3) \oplus C_2, & P_2 &= \mathcal{R}(G_2), \\ G_3 &= \mathcal{D}(P_2) \oplus \widetilde{P}_3, & P_3 &= \mathcal{R}(G_3), \\ G_4 &= \mathcal{D}(G_3) \oplus \widetilde{P}_4, & P_4 &= \mathcal{R}(G_4), \\ G_5 &= \mathcal{D}(G_4) \oplus C_5, & P_5 &= \mathcal{R}(G_5). \end{aligned} \tag{2}$$

At P3 and P4, each G_ℓ forks: $\mathcal{R}(G_\ell)$ supplies prediction, while $\mathcal{D}(G_\ell)$ propagates bottom-up. This separation avoids PAN's repeated downsampling of refined outputs (Liu et al., 2018); P2 and P3 also supply ROI evidence.

3.2 Evidence-Preserving Detection Core

Complementary State-Flow Mixer (CSFM)

To limit replacement of weak responses along one transformation chain, CSFM retains reference and evolution paths. For $\mathbf{X} \in \mathbb{R}^{C_{\text{in}} \times h \times w}$, a 1×1 Conv–BN–SiLU state lift gives $\mathbf{U} \in \mathbb{R}^{C \times h \times w}$. The reference–evolution allocation ratio $r \in (0, 1)$ partitions $\mathbf{U}$ into complementary reference state $\mathbf{U}_0$ and evolution seed $\mathbf{V}_0$, where $C_p = \lfloor rC \rfloor$, $C_t = C - C_p$, and $1 \le C_p < C$:

$$\begin{aligned} &\mathbf{U} = \phi_{\text{in}}(\mathbf{X}), \qquad [\mathbf{U}_0, \mathbf{V}_0] = \mathrm{Part}(\mathbf{U}; r), \\ &\mathbf{V}_i = \mathcal{F}_i(\mathbf{V}_{i-1}), \quad i = 1, \dots, N, \qquad \mathcal{F}_i = f_{i,k} \circ \cdots \circ f_{i,1}, \\ &\mathbf{A} = \mathbf{U}_0 \oplus \mathbf{V}_0 \oplus \mathbf{V}_1 \oplus \cdots \oplus \mathbf{V}_N, \qquad \mathbf{Y} = \phi_{\text{out}}(\mathbf{A}). \end{aligned} \tag{3}$$

Here $\mathbf{U}_0 \in \mathbb{R}^{C_p \times h \times w}$, $\mathbf{V}_i \in \mathbb{R}^{C_t \times h \times w}$, and $f_{i,j}$ is a shape-preserving RepN-CSP bottleneck. Thus N counts transition stages and k their bottlenecks. The reference stays fixed while $\mathbf{V}_0 \to \mathbf{V}_N$ evolves; C_{out} is the CSFM output width, and ϕ_{out} maps the concatenated $C_p + (N+1)C_t$ channels to this width. PAVER's detector uses $(r, N, k) = (0.5, 2, 1)$ at four stages. Supplementary Sec. S3 and Tables S12, S15 provide matched replacements and the state-flow sweep. They support the joint architecture, not a gradient-flow guarantee.

Semantic–Contour Recalibration Gate (SCRG)

SCRG avoids one-way fusion by symmetrically recalibrating coarse semantics and fine localization evidence. Bilinear interpolation aligns the coarse input to the finer grid. Let $\mathbf{H}, \mathbf{L} \in \mathbb{R}^{C \times h \times w}$ denote the aligned streams in Fig. 4. Its independent "Proj 1×1" Conv–BN–SiLU maps give $\mathbf{q}_H = \phi_H(\mathbf{H})$ and $\mathbf{q}_L = \phi_L(\mathbf{L})$; the pointwise gates are

$$\mathbf{g}_H = \sigma(\omega_H(\mathbf{q}_H)), \qquad \mathbf{g}_L = \sigma(\omega_L(\mathbf{q}_L)), \tag{4}$$

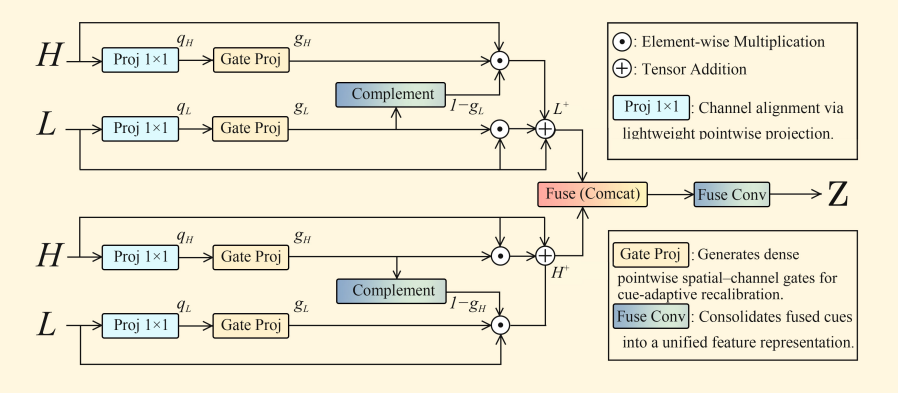


Figure 4: SCRG uses symmetric pointwise gates to recalibrate aligned semantic ($\mathbf{H}$) and localization-sensitive ($\mathbf{L}$) streams before fusion. $\odot$ and $+$ denote element-wise operations; "contour" denotes latent localization evidence, not annotation or loss.

where σ is sigmoid, ω_H, ω_L are learned 1×1 projections, and $\mathbf{g}_H, \mathbf{g}_L \in (0,1)^{C\times h\times w}$. With element-wise $\odot$ and an all-ones tensor $\mathbf{1}$, both streams update in parallel:

$$\mathbf{L}^+ = \mathbf{L} + \mathbf{g}_L \odot \mathbf{L} + (\mathbf{1} - \mathbf{g}_L) \odot (\mathbf{g}_H \odot \mathbf{H}), \tag{5}$$

$$\mathbf{H}^+ = \mathbf{H} + \mathbf{g}_H \odot \mathbf{H} + (\mathbf{1} - \mathbf{g}_H) \odot (\mathbf{g}_L \odot \mathbf{L}), \tag{6}$$

$$\mathbf{Z} = \phi_{\text{fuse}}\left(\mathbf{L}^+ \oplus \mathbf{H}^+\right). \tag{7}$$

Complement terms admit cross-stream evidence where the receiver gate is weak; residual and self-gated terms retain its evidence. Both updates use the same pre-update $(\mathbf{L}, \mathbf{H})$. The 3×3 Conv–BN–SiLU map $\phi_{\text{fuse}} : 2C \to C$ yields $\mathbf{Z} \in \mathbb{R}^{C\times h\times w}$. Tables 5 and 6 isolate SCRG at the stride-16 and stride-8 top-down refinements.

3.3 Risk-conditioned Interface

The interface uses P2 and P3 ROI evidence to calibrate each fixed candidate toward its true-positive probability under the evaluation event, capturing variation unavailable to score-only calibration. The detector is frozen, and candidates, boxes, and classes remain unchanged.

ROI evidence. For candidate b_i and $j \in \{2,3\}$, letterboxing maps the box to the P_j grid, and ROIAlign (He et al., 2017) yields $\mathbf{R}_{ij} \in \mathbb{R}^{D_j\times3\times3}$. We concatenate per-channel spatial means and maxima:

$$\mathbf{z}_i = \text{Mean}_{h,w}(\mathbf{R}_{i2}) \oplus \text{Max}_{h,w}(\mathbf{R}_{i2}) \oplus \text{Mean}_{h,w}(\mathbf{R}_{i3}) \oplus \text{Max}_{h,w}(\mathbf{R}_{i3}) \in \mathbb{R}^{384}. \tag{8}$$

Here $(D_2, D_3) = (64, 128)$. For statistics $\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z$ estimated from the relevant calibration fitting partition, set $\sigma^\dagger_{z,d} = 1$ when $\sigma_{z,d} < 10^{-6}$ and $\sigma^\dagger_{z,d} = \sigma_{z,d}$ otherwise, for descriptor coordinate $d = 1, \ldots, 384$:

$$\widetilde{\mathbf{z}}_i = \frac{\mathbf{z}_i - \boldsymbol{\mu}_z}{\boldsymbol{\sigma}^\dagger_z}, \qquad \overline{\mathbf{z}}_i = \frac{\widetilde{\mathbf{z}}_i}{\max(\|\widetilde{\mathbf{z}}_i\|_2, 10^{-6})}. \tag{9}$$

Signed prototype margin. From same-class correctness labels, spherical k-means (Dhillon & Modha, 2001) forms class-agnostic FP and TP banks with per-bank budget $K = 16$. For $q \in \{\text{FP}, \text{TP}\}$, $n_q \geq 1$ is the available descriptor count and $K_q = \min(K, n_q)$. Each prototype $\mathbf{p}^q_m \in \mathbb{R}^{384}$. Descriptors and prototypes have ℓ_2 norm at most one, with equality above 10^{-6}. Their signed margin is

$$r_i = \max_{1\leq m\leq K_{\text{FP}}} \overline{\mathbf{z}}_i^\top \mathbf{p}^{\text{FP}}_m - \max_{1\leq m\leq K_{\text{TP}}} \overline{\mathbf{z}}_i^\top \mathbf{p}^{\text{TP}}_m. \tag{10}$$

Thus $r_i \in [-2, 2]$ is a risk proxy, with larger values more FP-like, not a probability or safety certificate. Supplementary Sec. S4 details construction and degeneracy handling.

Nested score and risk calibration. Let $s_i^{\mathrm{clip}} = \mathrm{clip}(s_i, 10^{-6}, 1-10^{-6})$ and $\xi_i = \mathrm{logit}(s_i^{\mathrm{clip}})$. Clipping only stabilizes endpoint evaluation. This interface augments affine score calibration with the prototype margin:

$$\widehat{s}_i = \sigma(\eta_1 \xi_i + \eta_0 - \lambda_r r_i), \qquad \eta_1 > 0, \quad \lambda_r \geq 0. \tag{11}$$

For validation candidates $\mathcal{V}$, correctness indicators $y_i^{\mathrm{cor}} \in \{0,1\}$ (TP is 1), and image-disjoint out-of-fold (OOF) margins r_i^{OOF}, we fit

$$(\eta_1^\star, \eta_0^\star, \lambda_r^\star) = \underset{\substack{\eta_1>0,\ \eta_0\in\mathbb{R} \\ 0\leq\lambda_r\leq 20}}{\arg\min} \left\{ \frac{1}{|\mathcal{V}|} \sum_{i\in\mathcal{V}} \mathrm{BCE}\big(y_i^{\mathrm{cor}}, \sigma\big(\eta_1\xi_i + \eta_0 - \lambda_r r_i^{\mathrm{OOF}}\big)\big) + 10^{-7}\lambda_r^2 \right\}. \tag{12}$$

Here $\mathrm{BCE}(y,p) = -y\log p - (1-y)\log(1-p)$, η_1 is the positive score slope, η_0 is the intercept, and λ_r is the nonnegative risk coefficient. All calibrators share candidates and binary NLL. Temperature scaling (Guo et al., 2017) sets $\eta_1 = T^{-1}$, $\eta_0 = \lambda_r = 0$; Platt scaling (Platt, 1999) fits η_1, η_0 with $\lambda_r = 0$; PAVER fits all three and reduces to Platt when $\lambda_r = 0$. Positive η_1 preserves score-only ordering, whereas r_i may change it.

3.4 Training Objective

The detector and Risk-conditioned Interface are fitted sequentially. For a mini-batch $\mathfrak{M}$ of $n_{\mathfrak{M}}$ images, detector training uses

$$\mathcal{L}_{\mathrm{train}} = n_{\mathfrak{M}}\mathcal{L}_{\mathrm{det}}, \qquad \mathcal{L}_{\mathrm{det}} = \lambda_1\mathcal{L}_{\mathrm{reg}} + \lambda_2\mathcal{L}_{\mathrm{cls}} + \lambda_3\mathcal{L}_{\mathrm{dfl}}, \qquad (\lambda_1,\lambda_2,\lambda_3) = (7.5, 0.5, 1.5). \tag{13}$$

Here $\mathcal{L}_{\mathrm{reg}}$, $\mathcal{L}_{\mathrm{cls}}$, and $\mathcal{L}_{\mathrm{dfl}}$ denote SDIoU, logit BCE, and Distribution Focal Loss, respectively (Zheng et al., 2020; Li et al., 2020b). SDIoU modulates task-aligned assignment (Feng et al., 2021) by stop-gradient predicted area in resized-input coordinates, and regression by assigned-target area in pyramid-level coordinates. At fixed SDIoU assignment, replacing CIoU regression with SDIoU gains 0.7 mAP_{50} points with or without DFL (Supplementary Table S3). Supplementary Secs. S1–S2 define the objective; this is not a joint assignment–loss comparison.

4 Experiments

We evaluate PAVER's two stages separately: detector experiments test tiny-object localization, while calibration experiments hold detections fixed and test whether ROI evidence makes candidate confidence more trustworthy than raw confidence.

4.1 Experimental Setup and Evaluation Protocol

We evaluate horizontal-box detection on VisDrone-DET2019 (Du et al., 2019), RSOD (Long et al., 2017), and DIOR (Li et al., 2020a); Supplementary Fig. S8 contrasts their object sizes in the same order. Supplementary Sec. S12 separately reports a horizontal-box SODA-A adaptation (Cheng et al., 2023) using the official partition and released annotations.

Detection protocol. We use the 548-image VisDrone validation split for detector comparison and calibrator fitting, reserving 1,610 test-dev images for calibration evaluation. Reproduced detectors share 640×640 inputs, SGD (momentum 0.937), transformations, seed, NMS, and evaluator; the batch size is 32 for VisDrone and 16 for RSOD and DIOR. DIOR uses the configured 18,000-image training and 3,463-image evaluation split, whose evaluation partition also guides checkpoint selection; these results are controlled internal comparisons, not standard-split blind-test rankings. The shared recipe retains architecture-specific heads and non-overlap objective terms. We report mAP_{50}, $mAP_{50:95}$, AP_{75}, AP_S, parameters, and GFLOPs. mAP_{50}^{tiny} averages the four categories richest in resized-input instances below 32×32 pixels; it is not instance-filtered tiny AP. Supplementary Sec. S11 and Tables S6, S7, and S8 specify SDIoU insertion points, full protocols, size-bin intervals, and ignore rules.

Calibration protocol. Candidate metrics use fixed labels shared across calibrators. After removing VisDrone category-0 ignore-covered detections, descending-score greedy same-class matching at IoU ≥ 0.50 labels TPs; unmatched, intermediate-IoU, and duplicate candidates are FPs. Official AP

is instead recomputed per output ranking. Five image-disjoint folds generate OOF inputs to Eq. 12; components are refitted on validation and frozen for test-dev. ECE_{15}, Brier, NLL, and FP-AUC use all evaluable candidates; FP-AUC evaluates output-confidence error ranking, not r_i alone. We form percentile 95% CIs from 2,000 image-level bootstrap replicates. Supplementary Sec. S4 gives the full protocol.

4.2 Tiny-Object Detection Across Scales

We test the detection core with three reproduced benchmarks, strict-localization metrics, and category-agnostic size bins rather than AP at IoU 0.50 alone. Across main-text tables, **bold** and underlined values mark the best and second-best distinct point estimates within each comparison group; ties share rank, and shading identifies PAVER or the selected configuration.

Table 1: Reproduced comparison on VisDrone.

Method	Venue	Pedestrian↑	People↑	Bicycle↑	Motor↑	mAP_{50}^{tiny} ↑	mAP_{50} ↑	Params↓	GFLOPs↓
VAMYOLOX (Yang et al., 2023)	Sens. J'23	34.0	23.6	15.6	33.6	26.7	34.2	9.7	60.4
YOLO-v7 (Wang et al., 2023)	CVPR'23	31.1	18.6	11.3	31.6	23.2	35.8	36.5	103.3
YOLO-v8m (Jocher et al., 2023)	GitHub'23	29.4	15.3	11.5	31.2	21.9	33.8	25.9	79.1
YOLOv11-s (Khanam & Hussain, 2024)	arXiv'24	40.9	31.9	13.2	43.6	32.4	39.1	9.4	21.5
Mamba YOLO (Wang et al., 2025)	AAAI'25	30.0	16.4	11.2	32.7	22.6	34.9	21.8	49.6
Drone-YOLO (Zhang, 2023)	Drones'23	34.5	21.3	10.2	34.4	25.1	34.6	10.8	37.3
FFCA-YOLO (Zhang et al., 2024)	TGRS'24	34.3	19.8	11.2	30.0	23.8	34.8	**7.1**	51.8
LAR-YOLOv8 (Yi et al., 2024)	JSTARS'24	27.9	16.2	11.0	30.3	21.4	33.1	26.4	134.5
RT-DETR (Zhao et al., 2024)	CVPR'24	20.9	10.9	8.6	24.7	16.3	30.1	32.0	103.5
AMSFNet (Chen et al., 2025)	TGRS'25	34.8	23.4	13.1	35.7	26.8	36.6	10.5	39.8
YOLOv13-s (Lei et al., 2025)	arXiv'25	40.2	30.2	12.1	41.8	31.1	37.4	9.0	**20.8**
YOLOv12-s (Tian et al., 2025)	NeurIPS'25	45.9	34.4	13.5	47.5	35.3	40.2	9.3	21.4
Ours		**53.3**	**41.4**	**18.6**	**53.6**	**41.7**	**45.9**	10.3	40.0

Table 2: Reproduced comparison on RSOD.

Method	Venue	Aircraft↑	Oil Tank↑	Overpass↑	Playground↑	mAP_{50}^{tiny} ↑	mAP_{50} ↑	Params↓	GFLOPs↓
Sig-NMS (Dong et al., 2019)	TGRS'19	80.6	90.6	87.4	99.1	89.4	89.4	-	-
YOLO-v9c (Wang et al., 2024b)	ECCV'24	**96.8**	97.1	66.0	98.2	89.5	89.5	25.3	102.4
YOLOv11-s (Khanam & Hussain, 2024)	arXiv'24	94.9	97.7	87.9	99.4	95.0	95.0	9.4	21.5
LAR-YOLOv8 (Yi et al., 2024)	JSTARS'24	94.8	98.2	85.4	98.9	94.3	94.3	26.4	-
AGMF-Net (Gao et al., 2024)	TGRS'24	96.0	**99.0**	82.4	**99.7**	94.3	94.3	-	-
MFAE-YOLO (Liu et al., 2025b)	TGRS'25	**96.8**	97.9	85.0	99.3	94.8	94.8	**3.0**	**11.9**
YOLOv13-s (Lei et al., 2025)	arXiv'25	92.7	96.7	83.5	99.4	93.1	93.1	9.0	20.8
YOLOv12-s (Tian et al., 2025)	NeurIPS'25	94.5	97.7	84.5	99.4	94.0	94.0	9.3	21.4
Ours		95.3	96.7	**95.8**	99.1	**96.7**	**96.7**	10.3	40.0

Table 3: Controlled reproduction on the configured DIOR split.

Method	Venue	Ship↑	Storage Tank↑	Vehicle↑	Windmill↑	mAP_{50}^{tiny} ↑	mAP_{50} ↑	Params↓	GFLOPs↓
RAST-YOLO (Jiang & Wu, 2023)	Access'23	91.1	77.9	54.0	76.2	74.8	69.8	-	-
MSFC-Net (Zhang et al., 2022)	TGRS'22	89.6	76.5	55.2	**84.3**	76.4	69.8	136.1	49.6
TMAFNet (Gao et al., 2023)	TGRS'23	90.5	75.4	55.2	83.2	76.1	72.9	41.6	60.0
AGMFNet (Gao et al., 2024)	TGRS'24	91.7	75.8	58.0	83.1	77.2	73.2	-	-
YOLOv10-n (Wang et al., 2024a)	NeurIPS'24	88.1	71.7	45.6	72.2	69.4	66.2	**2.3**	**6.7**
YOLOv11-s (Khanam & Hussain, 2024)	arXiv'24	87.4	67.0	43.1	71.5	67.3	63.3	9.4	21.5
MFAE-YOLO (Liu et al., 2025b)	TGRS'25	89.4	72.6	46.4	76.1	71.1	67.0	3.0	11.9
YOLOv13-s (Lei et al., 2025)	arXiv'25	89.9	74.7	50.8	78.3	73.4	70.3	9.0	20.8
YOLOv12-s (Tian et al., 2025)	NeurIPS'25	92.2	**81.4**	58.3	81.2	78.3	**74.5**	9.3	21.4
Ours		**92.4**	81.2	**61.4**	81.0	**79.0**	**74.5**	10.3	40.0

Table 4: Strict localization. AP_S uses the COCO small-object area interval in resized-input coordinates; Supplementary Sec. S8 gives the complete size-bin and ignore protocol.

Method	VisDrone AP_{75} ↑	VisDrone AP_S ↑	VisDrone $mAP_{50:95}$ ↑	RSOD $mAP_{50:95}$ ↑	DIOR $mAP_{50:95}$ ↑
YOLOv11-s	24.7	15.5	23.3	66.6	51.8
YOLOv13-s	23.6	14.8	22.3	65.4	50.9
YOLOv12-s	25.8	16.3	24.0	67.1	**55.0**
Ours	**31.4**	**21.6**	**28.0**	**68.9**	53.0

Accuracy. VisDrone provides the strongest result: PAVER gains 5.7 and 6.4 points over YOLOv12-s in mAP_{50} and mAP_{50}^{tiny}, respectively, and reaches 31.4% AP_{75} and 28.0% $mAP_{50:95}$. On RSOD, it reaches 96.7% mAP_{50} and 68.9% $mAP_{50:95}$. On DIOR, PAVER matches YOLOv12-s at 74.5% mAP_{50} but trails at $mAP_{50:95}$ (53.0% vs. 55.0%), delimiting a metric-dependent gain. Category-agnostic small-object AP exceeds YOLOv12-s by 5.3, 4.4, and 2.7 points on VisDrone, RSOD, and DIOR, respectively (Supplementary Table S8).

Efficiency. PAVER's detector uses 10.3M parameters and 40.0 GFLOPs, 86.9% more compute than YOLOv12-s@640. It nevertheless exceeds YOLOv12-s@1024 by 1.3 and 0.6 points in mAP_{50} and mAP_{50}^{tiny}, respectively, at 40.0 versus 54.8 GFLOPs (Supplementary Table S10). Supplementary Table S9 gives same-GPU detector-core latency and FPS, excluding the Risk-conditioned Interface; neither comparison establishes end-to-end compute parity.

4.3 Isolating the Evidence-Preserving Detection Core

We test whether gains arise from the coordinated pathway rather than only a P2 head or generic block. Tables 5–7 fix scale, objective, assignment, training, and evaluation while varying the indicated structures.

Table 5: VisDrone component isolation from YOLOv11-s; H, C, and S denote HRPE, CSFM, and SCRG, respectively.

Setting	H	C	S	mAP_{50} ↑	mAP_{50}^{tiny} ↑	$mAP_{50:95}$ ↑
YOLOv11-s base				39.1	32.4	23.3
+CSFM only		✓		40.3	34.1	23.9
+SCRG only			✓	40.6	33.7	24.3
+HRPE only	✓			43.5	38.6	26.2
+HRPE+SCRG	✓		✓	<u>44.9</u>	40.2	<u>27.2</u>
+HRPE+CSFM	✓	✓		44.5	<u>40.6</u>	27.1
PAVER full	✓	✓	✓	**45.9**	**41.7**	**28.0**

Table 6: Matched one-factor VisDrone replacements at fixed sites and widths; rankings are within each family.

Variant	mAP_{50} ↑	mAP_{50}^{tiny} ↑	GFLOPs↓
Backbone block (four matched sites; fixed SCRG+HRPE)			
C3k2	44.9	40.2	**39.1**
CSP	44.7	39.8	<u>39.6</u>
ELAN	45.3	40.8	41.7
YOLOv12 stage-wise blocks (C3k2 and A2C2f)	<u>45.4</u>	<u>41.0</u>	39.8
CSFM	**45.9**	**41.7**	40.0
Fusion gate (two matched sites; fixed CSFM+HRPE)			
SE-style	44.8	39.9	**39.4**
One-way sigmoid gate	45.1	40.6	<u>39.7</u>
Additive attention gate	45.3	40.8	40.1
CBAM	<u>45.4</u>	<u>41.0</u>	40.2
SCRG	**45.9**	**41.7**	40.0
Decoder (matched P2–P5 heads; fixed CSFM+SCRG)			
P2 head only	44.6	39.7	**37.6**
PAN-FPN (P2–P5)	45.2	40.5	<u>39.3</u>
BiFPN (P2–P5)	<u>45.5</u>	<u>40.9</u>	39.7
HRPE	**45.9**	**41.7**	40.0

Table 7: VisDrone CSFM ($N = 2, k = 1$) and SCRG ablations. "Channel" denotes spatial–channel gating.

Setting	mAP_{50} ↑	mAP_{50}^{tiny} ↑	GFLOPs↓
CSFM $r = 0.25$	45.2	40.5	41.3
CSFM $r = 0.50$	**45.9**	**41.7**	<u>40.0</u>
CSFM $r = 0.75$	<u>45.4</u>	<u>41.0</u>	**39.2**
SCRG-Spatial	45.5	41.2	<u>40.3</u>
SCRG-Deformable	<u>45.7</u>	<u>41.4</u>	41.4
SCRG-Channel	**45.9**	**41.7**	**40.0**

Component isolation. HRPE gives the largest single gain over YOLOv11-s under fixed loss and assignment. Adding CSFM and SCRG contributes 2.4, 3.1, and 1.8 points across metrics (Table 5). Because Supplementary Table S3 changes loss on the full detector, architectural and loss gains are not additive.

Controlled replacements. Matched replacements favor CSFM, SCRG, and HRPE within families (Table 6); HRPE exceeds P2–P5 PAN-FPN by 0.7 and 1.2 points on mAP_{50} and mAP_{50}^{tiny} at 0.7 additional GFLOPs. The comparisons support these selected roles but not retained information; Supplementary Tables S12–S14 and Figs. S2–S3 give full results and diagnostics.

4.4 Risk-conditioned Candidate Confidence

On 434,224 fixed non-ignore test-dev candidates, frozen validation-fitted calibrators change only scores (Table 8). Figure 5 summarizes the absolute calibration metrics and the paired PAVER-minus-Platt uncertainty reported in Supplementary Table S5. Supplementary Sec. S4 provides the full fitting, evaluation, metric, and bootstrap protocols.

Table 8: Validation-fitted calibration on fixed VisDrone test-dev candidates and labels. AP_{50} is percent; other metrics are fractional. FP-AUC ranks output-score errors. Shading identifies PAVER; rankings are point estimates, with the protocol and confidence intervals in Supplementary Sec. S4.

Variant	AP_{50} ↑	FP-AUC↑	ECE_{15} ↓	Brier↓	NLL↓
Raw confidence s_i	**33.592**	0.94456	0.01779	0.04486	0.16109
Temperature scaling	**33.592**	0.94456	**0.01244**	0.04447	**0.15973**
Platt scaling	**33.592**	0.94456	0.01545	0.04445	0.16016
PAVER interface	33.570	**0.94466**	0.01530	**0.04442**	0.16002

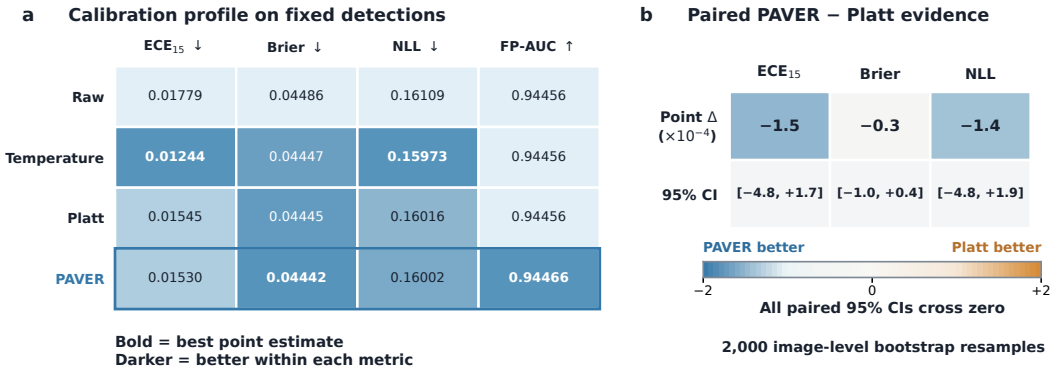


Figure 5: Candidate-confidence evaluation on fixed VisDrone test-dev detections. (a) Absolute ECE_{15}, Brier, NLL, and FP-AUC point estimates reproduce Table 8; darker cells indicate better values within each metric, and bold marks the best point estimate. (b) PAVER-minus-Platt calibration differences and paired 95% image-bootstrap CIs reproduce Supplementary Table S5 in units of 10^{-4}. Negative differences favor PAVER, and all three intervals cross zero.

Calibration quality. Across fixed candidates, PAVER improves ECE_{15}, Brier score, NLL, and FP-AUC over raw confidence while keeping AP_{50} essentially unchanged. Among all evaluated variants, PAVER attains the best Brier score and FP-AUC, whereas temperature scaling attains the best ECE_{15} and NLL. PAVER also improves all three calibration point estimates over its nested Platt submodel; the paired intervals in Figure 5 and Supplementary Table S5 quantify the uncertainty of that comparison. Together, these results show that candidate-local ROI evidence makes confidence more trustworthy without trading away detector AP.

5 Limitations and Responsible Use

Controlled ablations support the design but do not directly quantify retained information. The confidence study establishes improvements over raw scores under one VisDrone correctness protocol. Its comparison with Platt scaling remains CI-unresolved, and bootstrap CIs omit fitting and shift uncertainty. Test-dev is not externally sequestered; unseen negatives, alternative prototype budgets, ROI sizes, and P4 and P5 evidence remain untested. Consequential use requires domain validation, not a safety guarantee (Supplementary Sec. S14).

6 Conclusion

PAVER addresses a two-stage evidence bottleneck in remote-sensing tiny-object detection by coupling an evidence-preserving detector with a separately fitted Risk-conditioned Interface. CSFM, SCRG, and HRPE coordinate complementary state formation, semantic–localization recalibration, and high-resolution propagation to deliver strong tiny-object and strict-localization performance across benchmarks. On fixed VisDrone test-dev detections, the interface uses P2 and P3 ROI evidence to improve ECE_{15}, Brier score, NLL, and FP-AUC over raw confidence while keeping boxes, classes, and AP_{50} essentially unchanged. Together, PAVER connects accurate localization with more trustworthy candidate confidence through a unified formation-to-inference evidence pathway.

AI USE STATEMENT

During manuscript preparation, the authors used ChatGPT (GPT-6, OpenAI) solely for grammatical correction and English-language polishing. The tool was not used for research design, methodological development, code generation, experimentation, data analysis, or the generation of scientific results or conclusions. The authors take full responsibility for the content, accuracy, and integrity of the manuscript.

REPRODUCIBILITY STATEMENT

Sections 3 and 3.4 define the PAVER framework and its objective. The supplement specifies component operators, data splits, preprocessing, optimization, controls, candidate construction, calibration, seeds, metrics, and resampling. The anonymized repository provides configurations, training and evaluation scripts, risk calibration, and table-reproduction instructions. All evaluated datasets are public and cited; no private data are used.

REFERENCES

Konstantinos Panagiotis Alexandridis, Ismail Elezi, Jiankang Deng, Anh Nguyen, and Shan Luo. Fractal calibration for long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15139–15150, 2025.

Till Beemelmanns, Alexey Nekrasov, Stefan Vilceanu, Jonas Steinhaus, Timo Woopen, Bastian Leibe, and Lutz Eckstein. Query2uncertainty: Robust uncertainty quantification and calibration for 3d object detection under distribution shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4677–4686, 2026.

Jinghao Bian, Mingtao Feng, Weisheng Dong, Fangfang Wu, Jianqiao Luo, Yaonan Wang, and Guangming Shi. Feature information driven position gaussian distribution estimation for tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30376–30386, 2025.

Bing Cao, Haiyu Yao, Pengfei Zhu, and Qinghua Hu. Visible and clear: Finding tiny objects in difference map. In *European Conference on Computer Vision (ECCV)*, pp. 1–18. Springer, 2024.

Yan Chen, Qi Gu, Xiaofeng Wang, Xinlu Shi, Lixiang Xu, and Chen Zhang. Axis-squeeze and multi-routing scale-adaptive fusion network for remote sensing images object detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3597159.

Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488, 2023. doi: 10.1109/TPAMI.2023.3290594.

Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42:143–175, 2001. doi: 10.1023/A:1007612920971.

Ruchan Dong, Dazhuan Xu, Jin Zhao, Licheng Jiao, and Jungang An. Sig-nms-based faster r-cnn combining transfer learning for small target detection in vhr optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):8534–8545, 2019. doi: 10.1109/TGRS.2019.2921396.

Dawei Du, Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Lin, Qinghua Hu, Tao Peng, Jiayu Zheng, Xinyao Wang, Yue 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 (ICCVW)*, pp. 0–0, 2019.

Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R. Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3510–3519, 2021.

Tao Gao, Zixiang Liu, Jing Zhang, Guiping Wu, and Ting Chen. A task-balanced multiscale adaptive fusion network for object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 61:1–15, 2023.

Tao Gao, Ziqi Li, Yuanbo Wen, Ting Chen, Qianqian Niu, and Zixiang Liu. Attention-free global multiscale fusion network for remote sensing object detection. *IEEE transactions on geoscience and remote sensing (TGRS)*, 62:1–14, 2024. doi: 10.1109/TGRS.2023.3346041.

Wenchao Guan, Chuan Lin, Sihan Huang, Xiongzhen Wang, and Xintao Pang. Bdnet: Bio-inspired dual-backbone small object detection network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32724–32734, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Huma Hafeez, Matthew Garratt, Jo Plested, Sankaran Iyer, and Arcot Sowmya. Yolo11-4k: An efficient architecture for real-time small object detection in 4k panoramic images. *arXiv preprint arXiv:2512.16493*, 2025.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961–2969, 2017.

Xinkai Hu, Zhida Ren, Uzair Aslam Bhatti, Mengxing Huang, and Yirong Wu. Dcedet: Tiny object detection in remote sensing images based on dual-contrast feature enhancement and dynamic distance measurement. *Remote Sensing*, 17(16):2876, 2025.

Bihan Huo, Chenglong Li, Jianwei Zhang, Yingjian Xue, and Zhoujin Lin. Saff-ssd: Self-attention combined feature fusion-based ssd for small object detection in remote sensing. *Remote Sensing*, 15(12):3027, 2023.

Xuzhao Jiang and Yonghong Wu. Remote sensing object detection based on convolution and swin transformer. *IEEE Access*, 11:38643–38656, 2023.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. Software, version 8.0.0, 2023. URL `https://github.com/ultralytics/ultralytics`.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Mengqi Lei, Siqi Li, Yihong Wu, Han Hu, You Zhou, Xinhu Zheng, Guiguang Ding, Shaoyi Du, Zongze Wu, and Yue Gao. Yolov13: Real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv preprint arXiv:2506.17733*, 2025.

Chaolang Li, Pengwen Dai, Jingyu Li, Siyuan Yao, Yuchen Jiang, and Zhuoran Zheng. Dyfclt: Dynamic frequency-decoupled cross-modal learning transformer for multimodal tiny object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11313–11323, 2026a.

Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020a.

Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21002–21012, 2020b.

Yuxuan Li, Xiang Li, Yunheng Li, Yicheng Zhang, Yimian Dai, Qibin Hou, Ming-Ming Cheng, and Jian Yang. Sm3det: A unified model for multi-modal remote sensing object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6717–6725, 2026b. doi: 10.1609/aaai.v40i8.37603.

Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Uncertainty weighted gradients for model calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15497–15507, 2025.

Kai Liu, Zhihang Fu, Sheng Jin, Ze Chen, Fan Zhou, Rongxin Jiang, Yaowu Chen, and Jieping Ye. Esod: Efficient small object detection on high-resolution images. *IEEE Transactions on Image Processing (TIP)*, 34:183–195, 2025a.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759–8768, 2018. doi: 10.1109/CVPR.2018.00913.

Yibo Liu, Xin Cheng, Ning Xu, Luyao Wang, Xu Wang, and Xian Zhong. Mfae-yolo: Multi-feature attention-enhanced network for remote sensing images object detection. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 63:1–14, 2025b. doi: 10.1109/TGRS.2025.3583467.

Yang Long, Yiping Gong, Zhifeng Xiao, and Qing Liu. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 55(5):2486–2498, 2017.

Mahya Nikouei, Bita Baroutian, Shahabedin Nabavi, Fateme Taraghi, Atefe Aghaei, Ayoob Sajedi, and Mohsen Ebrahimi Moghaddam. Small object detection: A comprehensive survey on challenges, techniques and real-world applications. *Intelligent Systems with Applications (ISWA)*, pp. 200561, 2025.

Hyeonjeong Park, Peixi Xiong, Xiaoqian Ruan, Dian Jia, Pei Yu, and Wei Tang. Rare: Learn to rank and retrieve for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11556–11566, 2026.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Junsuo Qu, Tong Liu, Zongbing Tang, Yifei Duan, Heng Yao, and Jiyuan Hu. Remote sensing small object detection network based on multi-scale feature extraction and information fusion. *Remote Sensing*, 17(5):913, 2025.

Xiaofan Que, Dingrong Wang, Xumin Liu, and Qi Yu. Inscal: Calibrated multi-source fully test-time prompt tuning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36936–36946, 2026.

Shuohao Shi, Qiang Fang, and Xin Xu. Vlm4rsdet: Collaborative optimization with vision-language model for enhancing remote sensing object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18450–18460, 2026.

Huixin Sun, Yanjing Li, Linlin Yang, Xianbin Cao, and Baochang Zhang. Uncertainty-aware gradient stabilization for small object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8407–8417, 2025a.

Huixin Sun, Runqi Wang, Yanjing Li, Linlin Yang, Shaohui Lin, Xianbin Cao, and Baochang Zhang. Set: Spectral enhancement for tiny object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 4713–4723, 2025b.

Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 107984–108011, 2024a.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 7464–7475, 2023.

Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information. In *European Conference on Computer Vision (ECCV)*, 2024b.

Zeyu Wang, Chen Li, Huiying Xu, Xinzhong Zhu, and Hongbo Li. Mamba yolo: A simple baseline for object detection with state space model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 8205–8213, 2025. doi: 10.1609/aaai.v39i8.32885.

Xiaotan Wei, Zhensong Li, and Yutong Wang. Sed-yolo based multi-scale attention for small object detection in remote sensing. *Scientific reports*, 15(1):3125, 2025.

Shouluan Wu, Hui Yang, Liefa Liao, Chao Song, Qiuming Liu, Jianglong Fu, and Tan Li. Dynamic small object feature enhancement and detection for remote sensing images. *Scientific Reports*, 15 (1):37225, 2025.

Lihua Yang, Yi Gu, and Hao Feng. Multi-scale feature fusion and feature calibration with edge information enhancement for remote sensing object detection. *Scientific Reports*, 15(1):15371, 2025.

Yahu Yang, Xiangzhou Gao, Yu Wang, and Shenmin Song. Vamyolox: An accurate and efficient object detection algorithm based on visual attention mechanism for uav optical sensors. *IEEE Sensors Journal*, 23(11):11139–11155, 2023.

Hao Yi, Bo Liu, Bin Zhao, and Enhai Liu. Small object detection algorithm based on improved yolov8 for remote sensing. *IEEE journal of selected topics in applied earth observations and remote sensing (JSTARS)*, 17:1734–1747, 2024. doi: 10.1109/JSTARS.2023.3339235.

Seungju Yoo, Hyuk Kwon, Joong-Won Hwang, and Kibok Lee. Automated model evaluation for object detection via prediction consistency and reliability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19764–19773, 2025.

Xinbin Yuan, Zhaohui Zheng, Yuxuan Li, Xialei Liu, Li Liu, Xiang Li, Qibin Hou, and Ming-Ming Cheng. Strip r-cnn: Large strip convolution for remote sensing object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12259–12267, 2026. doi: 10.1609/aaai.v40i15.38217.

Jialei Zhan, Li Liu, Jiehua Zhang, Yuhang Xie, Yongxiang Liu, Jiangming Chen, and Ming-Ming Cheng. Rotation invariant and symmetry aware pixel difference network for remote sensing object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13264–13274, 2026.

Tong Zhang, Yin Zhuang, Guanqun Wang, Shan Dong, He Chen, and Lianlin Li. Multiscale semantic fusion-guided fractal convolutional object detection network for optical remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing (TGRS)*, 60:1–20, 2022.

Yin Zhang, Mu Ye, Guiyi Zhu, Yong Liu, Pengyu Guo, and Junhua Yan. Ffca-yolo for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15, 2024. doi: 10.1109/TGRS.2024.3363057.

Zhengxin Zhang. Drone-yolo: An efficient neural network method for target detection in drone images. *Drones*, 7(8):526, 2023.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 16965–16974, 2024.

Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12993–13000, 2020. doi: 10.1609/aaai.v34i07.6999.

Shilong Zhou, Haijin Zhou, and Lei Qian. A multi-scale small object detection algorithm sma-yolo for uav remote sensing images. *Scientific reports*, 15(1):9255, 2025.

S1 DEFINITION OF SCALE-DYNAMIC IOU (SDIOU)

Motivation. In remote-sensing imagery, tiny objects are highly sensitive to minor localization deviations. For very tiny boxes, even a one-pixel offset can cause a large fluctuation in raw IoU, which may lead to unstable regression signals. By contrast, the normalized center-distance term is typically less sensitive to such small perturbations and therefore provides a more stable localization cue in the tiny-object regime. To better balance overlap-based supervision and center-based localization across different object scales, we adopt a scale-dynamic IoU formulation, denoted as SDIoU.

Formulation. Following CIoU regression (Zheng et al., 2020), let IoU denote the intersection-over-union between the predicted box and the ground-truth box. The aspect-ratio term is defined as

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\mathrm{gt}}}{h_{\mathrm{gt}}} - \arctan\frac{w_p}{h_p}\right)^2, \qquad \alpha_{\mathrm{CIoU}} = \frac{v}{(1-\mathrm{IoU}) + v + \epsilon}, \tag{S1}$$

where (w_p, h_p) and $(w_{\mathrm{gt}}, h_{\mathrm{gt}})$ denote the width and height of the predicted and ground-truth boxes, respectively. The term v quantifies the aspect-ratio discrepancy between the predicted and ground-truth boxes, α_{CIoU} is the adaptive CIoU coefficient controlling the aspect-ratio term, and $\epsilon = 10^{-7}$ is the numerical-stability constant. We stop gradients through α_{CIoU}; the other predicted-box terms remain differentiable in regression, while assignment is entirely stop-gradient as specified below.

Let $\rho^2(\mathbf{c}_p, \mathbf{c}_{\mathrm{gt}})$ denote the squared Euclidean distance between the predicted center $\mathbf{c}_p$ and the ground-truth center $\mathbf{c}_{\mathrm{gt}}$. If w_{enc} and h_{enc} are the width and height of the smallest enclosing box covering both boxes, define its stabilized squared diagonal as $d^2_{\mathrm{enc}} = w^2_{\mathrm{enc}} + h^2_{\mathrm{enc}} + \epsilon$. We define the scale coefficient $\beta \in [0, \delta]$ separately for assignment and regression, with $\delta = 0.5$ setting the maximum reweighting range. The two definitions reflect their different coordinate systems: task-aligned assignment operates in resized-input coordinates, whereas regression operates in the native coordinates of each pyramid level.

Given two boxes represented in a common coordinate frame and the coefficient β, the SDIoU score is

$$\mathrm{SDIoU}(\mathbf{b}_p, \mathbf{b}_{\mathrm{gt}}; \beta) = \delta - \beta + (1 - \delta + \beta)(\mathrm{IoU} - \alpha_{\mathrm{CIoU}} v) - (1 + \delta - \beta)\frac{\rho^2(\mathbf{c}_p, \mathbf{c}_{\mathrm{gt}})}{d^2_{\mathrm{enc}}}. \tag{S2}$$

For task-aligned assignment, let $\widehat{A}_i^{\mathrm{in}}$ be the area of the stop-gradient predicted box $\widehat{\mathbf{b}}_i^{\mathrm{in}}$ in resized-input pixel coordinates. The assignment coefficient and overlap are

$$\beta_i^{\mathrm{assign}} = \delta \min\left(\frac{\widehat{A}_i^{\mathrm{in}}}{81}, 1\right), \qquad o_{ig} = \mathrm{SDIoU}\left(\widehat{\mathbf{b}}_i^{\mathrm{in}}, \mathbf{b}_g^{\star,\mathrm{in}}; \beta_i^{\mathrm{assign}}\right). \tag{S3}$$

Thus assignment modulation follows the current predicted extent, consistent with prediction-dependent task-aligned matching. The target index g enters through $\mathbf{b}_g^{\star,\mathrm{in}}$ in the overlap computation, not through the area-dependent coefficient $\beta_i^{\mathrm{assign}}$.

For regression, let $\tau_i \in \{4, 8, 16, 32\}$ be the pyramid stride of foreground location i, and let $A_i^{\star,\mathrm{in}}$ be the area of its assigned ground-truth box in resized-input pixels. The symbol s_i remains reserved throughout the paper for the raw candidate score shown in Fig. 2. Both predicted and target boxes are expressed in that level's native feature-grid coordinates,

$$\widehat{\mathbf{b}}_i^{\mathrm{grid}} = \frac{\widehat{\mathbf{b}}_i^{\mathrm{in}}}{\tau_i}, \qquad \mathbf{b}_i^{\star,\mathrm{grid}} = \frac{\mathbf{b}_i^{\star,\mathrm{in}}}{\tau_i}, \qquad A_i^{\star,\mathrm{grid}} = \frac{A_i^{\star,\mathrm{in}}}{\tau_i^2}. \tag{S4}$$

The corresponding regression coefficient and per-sample loss are

$$\beta_i^{\mathrm{reg}} = \delta \min\left(\frac{A_i^{\star,\mathrm{grid}}}{81}, 1\right) = \delta \min\left(\frac{A_i^{\star,\mathrm{in}}}{81\tau_i^2}, 1\right), \qquad \ell_{\mathrm{reg}}^{(i)} = 1 - \mathrm{SDIoU}\left(\widehat{\mathbf{b}}_i^{\mathrm{grid}}, \mathbf{b}_i^{\star,\mathrm{grid}}; \beta_i^{\mathrm{reg}}\right). \tag{S5}$$

Equivalently, the regression loss can be rewritten as

$$\ell_{\mathrm{reg}}^{(i)} = (1 - \delta + \beta_i^{\mathrm{reg}}) \underbrace{(1 - \mathrm{IoU} + \alpha_{\mathrm{CIoU}} v)}_{\ell_{\mathrm{shape}}^{(i)}} + (1 + \delta - \beta_i^{\mathrm{reg}}) \underbrace{\frac{\rho^2(\mathbf{c}_p, \mathbf{c}_{\mathrm{gt}})}{d^2_{\mathrm{enc}}}}_{\ell_{\mathrm{dist}}^{(i)}}, \tag{S6}$$

which shows that SDIoU adaptively reweights the shape-and-overlap term and the center-distance term according to the assigned target scale in the native pyramid coordinates. Under either definition,

$$(1 - \delta + \beta) + (1 + \delta - \beta) = 2, \tag{S7}$$

so the sum of the two coefficients remains exactly two while their relative allocation changes with the relevant box area. This identity constrains the coefficient mass; it does not imply that the numerical loss value is invariant, because $\ell^{(i)}_{\text{shape}}$ and $\ell^{(i)}_{\text{dist}}$ generally differ.

Behavior analysis. The saturation condition depends on the definition:

- In assignment, $\beta_i^{\text{assign}} = \delta$ when $\widehat{A}_i^{\text{in}} \geq 81$; in regression, $\beta_i^{\text{reg}} = \delta$ when $A_i^{\star,\text{in}} \geq 81\tau_i^2$. In either case,
$$1 - \delta + \beta = 1, \qquad 1 + \delta - \beta = 1,$$
and the SDIoU score reduces to the standard CIoU score. At the regression site, $\ell^{(i)}_{\text{reg}}$ therefore becomes the standard CIoU loss:
$$\ell^{(i)}_{\text{reg}} = 1 - \text{IoU} + \alpha_{\text{CIoU}} v + \frac{\rho^2(\mathbf{c}_p, \mathbf{c}_{\text{gt}})}{d^2_{\text{enc}}}.$$
- Below the corresponding threshold, $\beta < \delta$ and decreases with the relevant box area. Consequently, the weight on the overlap-and-shape term becomes smaller,
$$1 - \delta + \beta < 1,$$
while the weight on the center-distance term becomes larger,
$$1 + \delta - \beta > 1.$$
Therefore, SDIoU places relatively more emphasis on stable center alignment and relatively less emphasis on the more jitter-sensitive IoU term. Assignment uses the current prediction scale, whereas regression uses the assigned target scale represented at its pyramid level.

S2 Training Loss Details

The main paper summarizes the overall objective. For completeness, we provide the per-term definitions here. All definitions operate on one mini-batch $\mathfrak{M}$ with image count $n_{\mathfrak{M}} = |\mathfrak{M}|$ and candidate-location set $\mathcal{I}_{\mathfrak{M}}$; candidate–target pairs are formed only within the same image. For candidate location $i \in \mathcal{I}_{\mathfrak{M}}$ and ground-truth instance g of class c_g, let p_{i,c_g} be the corresponding classification logit, $\pi_{ig} = \sigma(p_{i,c_g})$ its probability, let o_{ig} be the assignment-site SDIoU overlap from Eq. S3, and set $\bar{o}_{ig} = \max(o_{ig}, 0)$. The selected task-aligned assignment (Feng et al., 2021) scores each eligible pair by

$$\chi_{ig} = \pi_{ig}^{0.5}\, \bar{o}_{ig}^{6.0}. \tag{S8}$$

Let $\mathbf{x}_i$ be the anchor point of location i in resized-input coordinates and $\mathcal{Q}_g = \{i : \mathbf{x}_i \text{ lies strictly inside } \mathbf{b}_g^{\star,\text{in}}\}$ its geometrically eligible set. If $\mathcal{Q}_g$ is nonempty and $\max_{i \in \mathcal{Q}_g} \chi_{ig} > 10^{-9}$, the positive pool retains the first $n_g^{\text{sel}} = \min(10, |\mathcal{Q}_g|)$ locations after sorting by decreasing χ_{ig} and then by increasing flattened feature index. Thus an instance with fewer than ten eligible locations retains all of them, while an empty or numerically zero-scored eligible set contributes no positive. If one location is selected by multiple instances, the assignment with the largest $\bar{o}_{ig}$ is retained; equal overlaps follow the earlier ground-truth annotation order. Let $m_{ig} \in \{0, 1\}$ denote the resulting mask, so every foreground location is assigned to exactly one ground truth. For an instance with at least one assigned location, define

$$\chi_g^{\max} = \max_i m_{ig}\chi_{ig}, \qquad o_g^{\max} = \max_i m_{ig}\bar{o}_{ig}. \tag{S9}$$

The task-aligned class target is the one-hot class label scaled by the normalized alignment score:

$$y_{ic} = \begin{cases} \dfrac{\chi_{ig} o_g^{\max}}{\chi_g^{\max} + 10^{-9}}, & m_{ig} = 1 \text{ and } c = c_g, \\ 0, & \text{otherwise.} \end{cases} \tag{S10}$$

All predicted scores and boxes used by the assignment and target construction are stop-gradient quantities. Thus SDIoU participates in assignment independently of its use in the regression loss. This assignment rule is fixed in every row of Table S3. Let $\mathcal{I}^+_{\mathfrak{M}} = \{i : \sum_g m_{ig} = 1\}$ collect foreground locations across the complete mini-batch, let $w_i = \sum_c y_{ic}$, and define the shared positive-mass normalizer

$$\mathcal{N}_{\text{pos}} = \max\left(\sum_{i \in \mathcal{I}^+_{\mathfrak{M}}} w_i, 1\right). \tag{S11}$$

This is one batch-wide quantity, not a per-image denominator. If the entire mini-batch contains no assigned positive, $\mathcal{I}^+_{\mathfrak{M}} = \varnothing$ and $\mathcal{N}_{\text{pos}} = 1$; regression and DFL are then zero, while negative classification terms remain active. The aggregate regression term is

$$\mathcal{L}_{\text{reg}} = \frac{1}{\mathcal{N}_{\text{pos}}} \sum_{i \in \mathcal{I}^+_{\mathfrak{M}}} w_i \ell^{(i)}_{\text{reg}}, \tag{S12}$$

where $\ell^{(i)}_{\text{reg}}$ is the grid-coordinate SDIoU loss in Eq. S5. The classification loss is BCE on logits over all candidate locations:

$$\mathcal{L}_{\text{cls}} = \frac{1}{\mathcal{N}_{\text{pos}}} \sum_{i \in \mathcal{I}_{\mathfrak{M}}} \sum_c \text{BCE}_{\text{logit}}(p_{ic}, y_{ic}), \tag{S13}$$

where p_{ic} is the predicted class logit, $\sigma(p) = (1 + e^{-p})^{-1}$, and $\text{BCE}_{\text{logit}}(p, y) = -y \log \sigma(p) - (1 - y) \log(1 - \sigma(p))$. For Distribution Focal Loss (Li et al., 2020b), let $u \in \{l, t, r, b\}$ index the four distances from a foreground location to its assigned box sides, let t_{iu} be the corresponding continuous target distance in that pyramid level's feature-grid coordinates, and set the number of discrete bins to $R = 16$. The target is clipped and its adjacent bins are

$$t_{iu} \leftarrow \text{clip}(t_{iu}, 0, R - 1 - 0.01), \qquad t^-_{iu} = \lfloor t_{iu} \rfloor, \qquad t^+_{iu} = t^-_{iu} + 1. \tag{S14}$$

Writing $\boldsymbol{\theta}_{iu} \in \mathbb{R}^R$ for the predicted bin logits, define the integer-bin cross-entropy as

$$\text{CE}(\boldsymbol{\theta}, t) = -\log \frac{\exp(\theta_t)}{\sum_{m=0}^{R-1} \exp(\theta_m)}, \qquad t \in \{0, \ldots, R - 1\}. \tag{S15}$$

The Distribution Focal Loss is

$$\mathcal{L}_{\text{dfl}} = \frac{1}{\mathcal{N}_{\text{pos}}} \sum_{i \in \mathcal{I}^+_{\mathfrak{M}}} w_i \frac{1}{4} \sum_{u \in \{l,t,r,b\}} \left[(t^+_{iu} - t_{iu}) \,\text{CE}(\boldsymbol{\theta}_{iu}, t^-_{iu}) + (t_{iu} - t^-_{iu}) \,\text{CE}(\boldsymbol{\theta}_{iu}, t^+_{iu})\right]. \tag{S16}$$

The decoded distance is $\widehat{d}_{iu} = \sum_{m=0}^{R-1} m \,\text{softmax}(\boldsymbol{\theta}_{iu})_m$. With $\mathbf{x}^{\text{grid}}_i = (x^{\text{grid}}_i, y^{\text{grid}}_i) = \mathbf{x}_i / \tau_i$, the four distances yield $\widehat{\mathbf{b}}^{\text{grid}}_i = (x^{\text{grid}}_i - \widehat{d}_{il}, y^{\text{grid}}_i - \widehat{d}_{it}, x^{\text{grid}}_i + \widehat{d}_{ir}, y^{\text{grid}}_i + \widehat{d}_{ib})$. Thus the lower- and upper-bin cross-entropies are linearly interpolated for each side and then averaged across the four sides before foreground weighting. All three components use the same batch-wide $\mathcal{N}_{\text{pos}}$. After applying their gains, the returned optimization scalar is multiplied by $n_{\mathfrak{M}} = |\mathfrak{M}|$, exactly as specified by $\mathcal{L}_{\text{train}}$ in the main paper.

Table S1: Complete HRPE architecture. The table reports both top-down SCRG outputs, every prediction feature, and the channel width of each independent bottom-up downsampling output. Each P_ℓ is subsequently passed to its corresponding detection head.

Node	Input	Operation	Output	Output width
P4 top-down feature	C_5, C_4	SCRG + C3k2	$\widetilde{P}_4$	256
P3 top-down feature	$\widetilde{P}_4$, C_3	SCRG + C3k2	$\widetilde{P}_3$	128
P2 fusion state	C_2, $\text{Up}_2(\widetilde{P}_3)$	Concatenate	G_2	256
P2 prediction feature	G_2	C3k2	P_2	64
P3 propagation branch	P_2	3×3 Conv, s2	$\mathcal{D}(P_2)$	64
P3 fusion state	$\mathcal{D}(P_2)$, $\widetilde{P}_3$	Concatenate	G_3	192
P3 prediction feature	G_3	C3k2	P_3	128
P4 propagation branch	G_3	3×3 Conv, s2	$\mathcal{D}(G_3)$	128
P4 fusion state	$\mathcal{D}(G_3)$, $\widetilde{P}_4$	Concatenate	G_4	384
P4 prediction feature	G_4	C3k2	P_4	256
P5 propagation branch	G_4	3×3 Conv, s2	$\mathcal{D}(G_4)$	256
P5 fusion state	$\mathcal{D}(G_4)$, C_5	Concatenate	G_5	768
P5 prediction feature	G_5	C3k2	P_5	512

Table S2: Complementary State-Flow Mixer operator details corresponding to Eq. 3.

Component	Kernel	Dilation	Norm. and act.	Role
State lift ϕ_{in}	1×1	1	BN + SiLU	Initialize $\mathbf{U}$
State allocation $\text{Part}(\cdot; r)$	–	–	–	Form complementary $\mathbf{U}_0, \mathbf{V}_0$
Transition stage $\mathcal{F}_i$	1×1 and 3×3	1	BN + SiLU	Advance $\mathbf{V}_{i-1}$ to $\mathbf{V}_i$
Readout projection ϕ_{out}	1×1	1	BN + SiLU	Mix reference and trajectory states

Table S3: Regression-loss attribution on VisDrone. Task-aligned assignment is fixed to the selected SDIoU-based formulation in every row; only the regression overlap term and the presence of DFL are varied. The complete PAVER detector architecture and all remaining training settings are unchanged. **Bold** and underlined mark best and second-best.

Loss setting	mAP_{50} ↑	mAP_{50}^{tiny} ↑	$mAP_{50:95}$ ↑
CIoU regression without DFL	44.1	38.7	26.0
CIoU regression + DFL	<u>45.2</u>	<u>40.6</u>	<u>27.4</u>
SDIoU regression without DFL	44.8	39.9	26.8
PAVER (SDIoU regression + DFL)	**45.9**	**41.7**	**28.0**

S3 Architectural Specification of PAVER Components

HRPE architecture. The backbone labels in Fig. 2 identify specific architecture features rather than generic stage outputs: C_2, C_3, and C_4 are the right-going outputs of the Conv blocks immediately above the second, third, and fourth CSFM blocks, at strides 4, 8, and 16. After the fourth CSFM, an independent stride-2 Conv–BN–SiLU precedes SPPF and C2PSA to produce C_5 at stride 32; this routine transition is omitted from the schematic. These are exactly the tensors denoted by C_2–C_5 in Eqs. 1–2. HRPE forms a compact P2 extension with separate prediction and propagation branches. It first fuses the stride-4 backbone feature with the upsampled stride-8 feature and refines the result into P_2. At the next two levels, the pre-refinement fusion states G_3 and G_4 each fork: a C3k2 produces the feature used by the corresponding detection head, while an independent 3×3, stride-2 Conv–BN–SiLU block propagates the unrefined fusion state to the next coarser level. Every C3k2 refinement and downsampling block in Eqs. 1–2 has its own parameters; the shared symbols specify operator type rather than weight sharing. This differs from the conventional iterative PAN path, in which each newly refined bottom-up feature becomes the input to the next downsampling block (Liu et al., 2018). The fork shortens the P2-to-P5 route by bypassing intermediate prediction refinements; it does not remove the C3k2 used at any prediction level. For PAVER-s, $(\widetilde{P}_3, \widetilde{P}_4)$ have widths $(128, 256)$, the three bottom-up outputs $(\mathcal{D}(P_2), \mathcal{D}(G_3), \mathcal{D}(G_4))$ have widths $(64, 128, 256)$, and the final prediction widths are $(D_2, D_3, D_4, D_5) = (64, 128, 256, 512)$. The fusion states (G_2, G_3, G_4, G_5) have widths $(256, 192, 384, 768)$. Table S1 records this complete data flow. HRPE denotes this branching topology rather than an additional standalone detection head.

CSFM architecture. The paper-level instantiation follows Eq. 3. A 1×1 state lift first produces $\mathbf{U} \in \mathbb{R}^{C \times h \times w}$. State allocation assigns $C_p = \lfloor rC \rfloor$ channels to reference state $\mathbf{U}_0$ and the remaining $C_t = C - C_p$ channels to evolution seed $\mathbf{V}_0$. The reference stream carries $\mathbf{U}_0$ directly to trajectory readout. The evolution stream applies N transition stages, where stage i contains k ordered, shape-preserving RepN-CSP bottlenecks and produces $\mathbf{V}_i \in \mathbb{R}^{C_t \times h \times w}$. To define each bottleneck $f_{i,j}$ uniquely, let $C_e = \lfloor 0.5 C_t \rfloor$ and let $\psi_a, \psi_b : C_t \to C_e$ and $\psi_o : 2C_e \to C_t$ be independent 1×1 Conv–BN–SiLU projections. For input $\mathbf{x} \in \mathbb{R}^{C_t \times h \times w}$,

$$\begin{aligned} \mathbf{x}_a &= \psi_a(\mathbf{x}), \qquad \mathbf{x}_b = \psi_b(\mathbf{x}), \qquad \mathbf{h}^{(0)} = \mathbf{x}_a, \\ \mathbf{h}^{(q)} &= \mathbf{h}^{(q-1)} + \psi_{q,2}\Big(\text{RepConv}_{q,1}(\mathbf{h}^{(q-1)})\Big), \quad q = 1, \ldots, \nu, \\ f_{i,j}(\mathbf{x}) &= \psi_o\Big(\mathbf{h}^{(\nu)} \oplus \mathbf{x}_b\Big). \end{aligned} \tag{S17}$$

Here $\nu = 1$ is the internal repeat count of each evaluated RepN-CSP bottleneck and is distinct from the figure-level k, which counts bottlenecks per transition stage. The transition and bottleneck indices (i, j) on the ψ and RepConv parameters are suppressed in Eq. S17; different $f_{i,j}$ instances do

not share weights. $\mathrm{RepConv}_{q,1}$ and $\psi_{q,2}$ are 3×3 operators with unit stride, unit dilation, group count 1, and BN–SiLU; their hidden expansion is 1. The displayed residual addition is enabled because their input and output widths agree. This specifies the split–transform–concatenate topology, outer expansion 0.5, internal repeat, and shortcut used by every $f_{i,j}$. The ordered states $\mathbf{V}_0, \ldots, \mathbf{V}_N$ form the discrete state flow. Trajectory readout concatenates $\mathbf{U}_0$ with every evolution state before the final 1×1 readout projection. Here "state flow" denotes an ordered feature-update sequence, not optical flow, a continuous-time model, or an SSM. The four CSFM sites in the PAVER-s detector use $(C_{\mathrm{in}}, C, C_{\mathrm{out}}) = (64, 64, 128), (128, 128, 256), (256, 128, 256), (512, 256, 512)$. At $r = 0.5$, their (C_p, C_t) pairs are therefore $(32, 32), (64, 64), (64, 64), (128, 128)$. The evaluated setting is $(r, N, k) = (0.5, 2, 1)$, which is highlighted in Table S15. Table S2 records the corresponding operator roles. The reference state remains unchanged only within its CSFM block; the architecture does not assert invariant downstream representations.

SCRG architecture. Each SCRG receives a coarse high-level source and a finer low-level source. The coarse source is resized to the finer destination grid by bilinear interpolation under the standard non-corner-aligned convention; at both evaluated sites, the two incoming sources have the same width C. We denote the resulting same-shaped residual streams by $\mathbf{H}$ and $\mathbf{L}$. The diagram's two "Proj 1×1" blocks are exactly the independent $C \to C$ Conv–BN–SiLU maps ϕ_H and ϕ_L that produce $\mathbf{q}_H$ and $\mathbf{q}_L$; no additional unreported channel adapter is used inside SCRG. Independent linear 1×1 gate projections ω_H and ω_L, followed by sigmoid, generate dense spatial–channel masks $\mathbf{g}_H$ and $\mathbf{g}_L$ without global pooling. No projection or gate weights are shared between streams or between the two SCRG sites. Equations 5 and 6 compute the two complementary corrections symmetrically from the same pre-update aligned pair $(\mathbf{H}, \mathbf{L})$; neither branch consumes the other branch's updated tensor. Their concatenation is consolidated by the 3×3 Conv–BN–SiLU map $\phi_{\mathrm{fuse}} : 2C \to C$ in Eq. 7, producing $\mathbf{Z}$ at the common input width. Here "contour" denotes latent high-resolution localization-sensitive evidence, not an external contour label or a contour loss.

S4 Validation-Fitted Risk and Calibration Protocol

Candidate construction and ROI evidence extraction. Single-view inference uses 640×640 inputs, confidence threshold 0.001, NMS IoU 0.70, and at most 300 candidates per image. Candidate boxes, scores, classes, and the final P_2 and P_3 feature tensors are obtained from the same forward pass. Boxes are retained in original-image corner coordinates (x_1, y_1, x_2, y_2) after inverse letterbox mapping. The corresponding gain and rounded offsets return them to detector-input coordinates. For original size (H_0, W_0) and detector-input size $(H_{\mathrm{in}}, W_{\mathrm{in}})$, this round-trip convention is

$$\begin{aligned} \gamma &= \min\left(\frac{H_{\mathrm{in}}}{H_0}, \frac{W_{\mathrm{in}}}{W_0}\right), \\ \Delta_x^{\mathrm{rt}} &= \mathrm{round}\left(\frac{W_{\mathrm{in}} - \gamma W_0}{2} - 0.1\right), \qquad \Delta_y^{\mathrm{rt}} = \mathrm{round}\left(\frac{H_{\mathrm{in}} - \gamma H_0}{2} - 0.1\right). \end{aligned} \tag{S18}$$

Here Δ_x^{rt} and Δ_y^{rt} are round-trip coordinate offsets, not a reconstruction of the exact left and top raster padding used to create the letterboxed image after rounding its resized width and height. The forward map is $(x', y') = (\gamma x + \Delta_x^{\mathrm{rt}}, \gamma y + \Delta_y^{\mathrm{rt}})$, followed by clipping to the valid detector-input extent. Because the detector's output inverse and the ROI forward map use the same convention, this operation recovers the detector-input candidate geometry up to boundary clipping. For an actual feature map $P_j \in \mathbb{R}^{D_j \times h_j \times w_j}$, where $D_j = \mathrm{channels}(P_j)$ as defined in the main paper, each horizontal coordinate is then multiplied by $\frac{w_j}{W_{\mathrm{in}}}$ and each vertical coordinate by $\frac{h_j}{H_{\mathrm{in}}}$. Because the boxes are already expressed in feature-grid coordinates, ROIAlign (He et al., 2017) uses unit spatial scaling, a 3×3 output grid, two samples per bin, and the half-pixel alignment convention. Per-channel spatial means and maxima over the two ROI-grid axes form the 384-dimensional vector in Eq. 8, ordered as $[P_2^{\mathrm{mean}}, P_2^{\mathrm{max}}, P_3^{\mathrm{mean}}, P_3^{\mathrm{max}}]$. Detector inference is not repeated when fitting the three scalar calibration variants.

Fold-specific normalization and prototype banks. For each out-of-fold split, the per-dimension mean and population standard deviation are estimated only from evaluable candidates whose images belong to that fold's training partition. A standard-deviation entry below 10^{-6} is replaced by 1 before z-score standardization, and each resulting row is divided by $\max(\|\widetilde{\mathbf{z}}_i\|_2, 10^{-6})$, as in Eq. 9. The FP and TP banks are class-agnostic, although their labels come from same-class matching. Each bank

preferentially uses evaluable examples with $s_i \geq 0.10$; this is a prototype-construction score floor, not a rule for excluding candidates from evaluation. If the selection is empty, the bank falls back to all evaluable examples with the corresponding label.

For bank type $q \in \{\mathrm{FP}, \mathrm{TP}\}$, let n_q be the descriptor count under the prototype-eligibility rule and empty-selection convention above. A fitted fold is admitted only when $n_{\mathrm{FP}}, n_{\mathrm{TP}} \geq 1$; this condition holds for every OOF fold and the final validation refit. With prototype budget $K = 16$, spherical k-means (Dhillon & Modha, 2001) uses the effective bank size $K_q = \min(K, n_q)$ and runs for at most 30 updates. The first center is sampled uniformly. Each subsequent center is sampled with probability proportional to its nearest nonnegative cosine distance $\max(0, 1 - \overline{\mathbf{z}}^{\top}\mathbf{p})$ from the centers already selected. If these distances sum to zero, the next center is sampled uniformly from descriptors not yet selected as centers. After assignment, a nonempty cluster center is the ℓ_2-normalized mean of its members. A zero-norm mean or an empty cluster is replaced by a uniformly sampled bank descriptor, followed by ℓ_2 normalization. Iterations stop when the largest absolute center-coordinate change is below 10^{-5}. With base seed 2027, OOF fold $f \in \{0, \ldots, 4\}$ uses seeds $2027 + 101f + 1$ for the FP bank and $2027 + 101f + 2$ for the TP bank. The final all-validation refit uses seeds $2027 + 9001$ and $2027 + 9002$, respectively. Both descriptors and centers are normalized to norm at most one (unit norm outside the degenerate zero-vector case), so the FP–TP inner-product margin remains bounded by $r_i \in [-2, 2]$.

Ignore-aware candidate labels. Let Ω_{ign} be the union of VisDrone category-0 regions. Positive-sized category-1–10 ground truths enter same-class matching; category-11 "others" annotations enter neither matching nor Ω_{ign}. Released $xywh$ boxes are converted by $x_2 = x_1 + w$ and $y_2 = y_1 + h$, clipped to image bounds, and compared in floating-point coordinates without intermediate rounding. Candidate b_i is removed from our correctness event when its continuous-geometry coverage

$$\mathrm{cov}_i^{\mathrm{ign}} = \frac{|b_i \cap \Omega_{\mathrm{ign}}|}{|b_i|} \tag{S19}$$

is at least 0.50. Filtering precedes TP and FP matching. Candidates are processed by descending raw score, with ties preserving the detector's original order. Let $\mathcal{G}_{\widehat{c}_i}^{\mathrm{unmatched}}$ be the still-unassigned ground truths of predicted class $\widehat{c}_i$. If nonempty, matching selects

$$g_i^{\star} = \underset{g \in \mathcal{G}_{\widehat{c}_i}^{\mathrm{unmatched}}}{\arg\max} \ \mathrm{IoU}(b_i, g). \tag{S20}$$

IoU ties follow the earlier ground-truth annotation order. Candidate i is a TP when $\mathrm{IoU}(b_i, g_i^{\star}) \geq 0.50$; $g_i^{\star}$ then becomes unavailable. All other candidates are FPs, including intermediate-IoU and duplicate detections. Unlike this continuous-geometry event, the official VisDrone AP matcher uses rounded raster geometry and filters ignore-covered ground truths and detections.

Candidate populations. Table S4 reports the image and evaluable-candidate counts used for fitting and evaluation. All evaluable candidates contribute to the calibration metrics, including non-ignore duplicates and intermediate-IoU negatives.

Table S4: VisDrone image and evaluable-candidate counts under the validation-fitted protocol. Evaluable candidates are fixed after ignore filtering and shared by all calibration variants.

Split	Images	Evaluable candidates
Validation	548	139,384
Test-dev	1,610	434,224

OOF fitting and nested calibrators. The 548 validation image identifiers are shuffled once with seed 2027 and divided into five folds. Each fold fits its own standardizer and FP and TP banks on the other four folds with budget $K = 16$, so every validation candidate's OOF margin is obtained without its image entering the corresponding bank. Folds are formed at image level, so no image contributes to both fitting and held-out partitions. We do not claim scene-disjointness because no independent scene-group metadata are used. Equation 12 then fits $(\eta_1, \eta_0, \lambda_r)$ by binary NLL with $0 \leq \lambda_r \leq 20$ and the fixed $10^{-7}\lambda_r^2$ penalty. Temperature scaling uses $(\eta_1, \eta_0, \lambda_r) = (T^{-1}, 0, 0)$ with $T > 0$, while Platt scaling fits (η_1, η_0) with $\eta_1 > 0$ and $\lambda_r = 0$ on the same labels. The

PAVER interface fits all three terms and reduces to the Platt submodel when $\lambda_r = 0$. The normalizer, prototype banks, and scalar parameters are finally refitted on all evaluable validation candidates and held fixed for test-dev.

Validation-fitted test-dev evaluation. For the nested comparisons in main-paper Table 8, the normalizer, prototype banks, and scalar parameters are selected and fitted using validation candidates only. During test-dev evaluation, these fixed quantities produce raw, temperature-scaled, Platt-scaled, and PAVER-interface scores without access to test-dev annotations; annotations are introduced only for metric computation. Candidate boxes, predicted classes, and candidate membership are identical across these variants; only the PAVER interface may change their score ordering. For the custom candidate-level metrics, all variants share the ignore mask and TP and FP labels determined in raw-score order; AP_{50} is recomputed using each variant's output scores. Applying the interface requires only the fitted normalizer, prototype banks, and scalar parameters in addition to detector evidence; candidate-correctness labels are not required. This ordering prevents label-dependent rescoring during evaluation, but it does not constitute an externally sequestered blind test. The margin r_i remains an explicit diagnostic output rather than being interpreted as a calibrated probability.

Uncertainty control. Two thousand bootstrap replicates (seed 2027) resample the 1,610 test-dev images with replacement. Every candidate, ground-truth instance, and ignore region inherits its image multiplicity; candidates are never resampled independently. The validation-fitted calibrators are not refitted inside a replicate, so the intervals quantify test-image sampling uncertainty conditional on the validation fit. Calibration differences are computed on the resampled evaluable candidates, whereas AP_{50} is recomputed from the corresponding weighted precision–recall curve. Table S5 reports percentile 95% CIs for the nested comparisons.

Figure summary. Main-paper Fig. 5 visualizes only the absolute point estimates in Table 8 and the paired differences and intervals in Table S5; it introduces no additional candidate stratification or case selection.

S4.1 Candidate-Level Reliability Metrics

Let $\mathcal{E}$ denote the candidate set on which a reliability metric is evaluated (the test-dev evaluable set for the reported results). For candidate $i \in \mathcal{E}$, let $\zeta_i \in [0, 1]$ be its reported output confidence and let $y_i^{\mathrm{cor}} \in \{0, 1\}$ be its correctness indicator (TP is 1). This notation distinguishes the metric-evaluation set $\mathcal{E}$ from the calibration-fitting set $\mathcal{V}$ in Eq. 12. ECE uses 15 equal-width confidence bins, with the last bin right-closed so that $\zeta_i = 1$ is included:

$$\mathcal{B}_m = \begin{cases} \left\{ i \in \mathcal{E} : \frac{m-1}{15} \le \zeta_i < \frac{m}{15} \right\}, & m = 1, \ldots, 14, \\ \left\{ i \in \mathcal{E} : \frac{14}{15} \le \zeta_i \le 1 \right\}, & m = 15. \end{cases} \tag{S21}$$

Let $\mathcal{M}_{\mathcal{E}} = \{m \in \{1, \ldots, 15\} : |\mathcal{B}_m| > 0\}$ be the indices of nonempty bins. Then

$$\mathrm{ECE}_{15} = \sum_{m \in \mathcal{M}_{\mathcal{E}}} \frac{|\mathcal{B}_m|}{|\mathcal{E}|} \left| \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} y_i^{\mathrm{cor}} - \frac{1}{|\mathcal{B}_m|} \sum_{i \in \mathcal{B}_m} \zeta_i \right|. \tag{S22}$$

The Brier score and binary negative log likelihood are

$$\mathrm{Brier} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} (\zeta_i - y_i^{\mathrm{cor}})^2, \tag{S23}$$

$$\mathrm{NLL} = -\frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \left[y_i^{\mathrm{cor}} \log \zeta_i^{\mathrm{clip}} + (1 - y_i^{\mathrm{cor}}) \log(1 - \zeta_i^{\mathrm{clip}}) \right], \tag{S24}$$

where $\zeta_i^{\mathrm{clip}} = \mathrm{clip}(\zeta_i, 10^{-6}, 1 - 10^{-6})$. FP-AUC treats the false-positive indicator $1 - y_i^{\mathrm{cor}}$ as the positive label and $e_i = 1 - \zeta_i$ as the output-score error signal:

$$\text{FP-AUC}(\zeta) = \Pr(e_{\mathrm{FP}} > e_{\mathrm{TP}}) + \tfrac{1}{2} \Pr(e_{\mathrm{FP}} = e_{\mathrm{TP}}). \tag{S25}$$

Here e_{FP} and e_{TP} are output-score error signals drawn from the evaluable FP and TP sets, respectively, and the half-weight handles ties. ECE, Brier score, NLL, and FP-AUC use the same non-ignore set $\mathcal{E}$. FP-AUC evaluates whether the final reported confidence separates FPs from TPs;

it does not directly evaluate r_i as a standalone predictor. AP_{50} uses the official detection evaluator. Because temperature and positive-slope Platt scaling are monotone in the raw score, they preserve raw AP and FP-AUC ranking; the PAVER interface can change both through its signed prototype margin r_i.

Table S5: Paired image-level bootstrap comparisons on VisDrone test-dev (2,000 replicates; seed 2027). Calibration cells give the point difference above the percentile 95% CI; the AP_{50} cell summarizes the essentially unchanged point estimate and reports its paired interval. Calibration metrics use fractional units and AP_{50} uses percentage points (pp). Negative calibration differences favor the PAVER interface. Dashes denote metrics not reported for that contrast.

Comparison	ΔECE_{15} [95% CI]	ΔBrier [95% CI]	ΔNLL [95% CI]	ΔAP_{50} [95% CI]
PAVER − Platt	-0.00015 $[-0.00048, +0.00017]$	-0.00003 $[-0.00010, +0.00004]$	-0.00014 $[-0.00048, +0.00019]$	–
PAVER − raw confidence	–	–	–	Essentially unchanged $[-0.082, +0.037]$

The PAVER–Platt row isolates prototype conditioning against the $\lambda_r = 0$ submodel of Eq. 11. PAVER improves all three calibration point estimates over Platt, with 95% CIs including zero. Against raw confidence, it improves ECE_{15}, Brier score, NLL, and FP-AUC while keeping AP_{50} essentially unchanged, with a paired interval of $[-0.082, +0.037]$ pp. It also attains the best Brier score and FP-AUC among the evaluated variants, while temperature scaling retains the best ECE_{15} and NLL. These results show that candidate-local prototype conditioning makes confidence more trustworthy while preserving detector AP.

S5 Regression-Loss Attribution under Fixed Assignment

The main paper uses BCE for classification, while DFL and the selected overlap term supervise box regression. To isolate these regression choices from positive-sample selection, Table S3 fixes the complete PAVER detector architecture and the SDIoU-based task-aligned assignment in all four rows, changing only CIoU versus SDIoU inside the regression loss and switching DFL off or on. Replacing CIoU regression with SDIoU regression increases mAP_{50} by 0.7 points both without and with DFL. Adding DFL increases mAP_{50} by 1.1 points under either regression formulation.

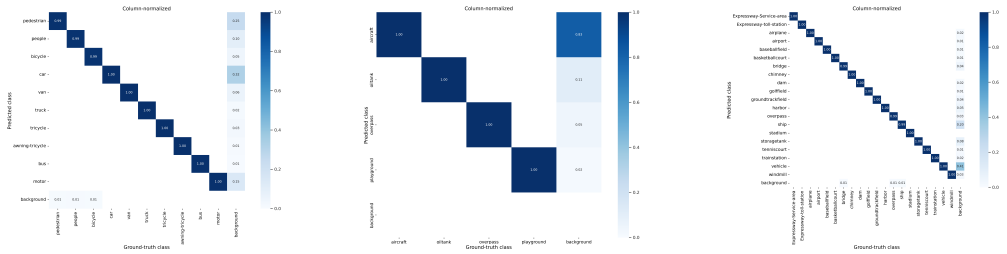
S6 More Quantitative Results

The main paper summarizes overall detection, strict localization, and selected-category metrics in the order VisDrone, RSOD, and DIOR. Tables S6 and S7 provide the complete VisDrone and DIOR per-category AP_{50} results. RSOD has only four categories, all reported in main-paper Table 2.

Comprehensive Results on VisDrone. Table S6 reports the complete VisDrone category comparison. The PAVER detector reaches 53.3% Pedestrian AP and 18.6% Bicycle AP. These results use the 548-image validation split, distinct from the validation-fitted test-dev calibration results.

Table S6: Full per-category comparison on VisDrone. **Bold** and underlined mark best and second-best.

Method	Venue	Pedestrian↑	People↑	Bicycle↑	Car↑	Van↑	Truck↑	Tricycle↑	Aw-Tricycle↑	Bus↑	Motor↑	mAP_{50} ↑
VAMYOLOX(Yang et al., 2023)	Sens. J'23	34.0	23.6	15.6	72.2	36.7	35.7	20.4	17.4	52.9	33.6	34.2
YOLOv7(Wang et al., 2023)	CVPR'23	31.1	18.6	11.3	76.1	40.0	**45.8**	24.3	21.9	57.6	31.6	35.8
YOLOv8(Jocher et al., 2023)	GitHub'23	29.4	15.3	11.5	72.0	39.8	43.9	16.0	19.0	59.6	31.2	33.8
Mamba YOLO(Wang et al., 2025)	AAAI'25	30.0	16.4	11.2	74.4	40.8	43.1	18.7	21.5	60.0	32.7	34.9
Drone-YOLO(Zhang, 2023)	Drones'23	34.5	21.3	10.2	74.7	38.1	37.6	19.1	19.2	57.2	34.4	34.6
YOLOv11-s(Khanam & Hussain, 2024)	arXiv'24	40.9	31.9	13.2	78.9	45.1	37.6	27.8	16.3	55.3	43.6	39.1
FFCA-YOLO(Zhang et al., 2024)	TGRS'24	34.3	19.8	11.2	76.4	41.3	38.6	20.0	19.0	57.2	30.0	34.8
LAR-YOLOv8(Yi et al., 2024)	JSTARS'24	27.9	16.2	11.0	71.8	39.3	40.1	16.8	19.3	58.2	30.3	33.1
RT-DETR(Zhao et al., 2024)	CVPR'24	20.9	10.9	8.6	66.0	31.6	45.3	19.7	18.1	55.4	24.7	30.1
AMSFNet(Chen et al., 2025)	TGRS'25	34.8	23.4	13.1	76.5	41.4	38.9	20.2	**23.1**	58.8	35.7	36.6
YOLOv13-s(Lei et al., 2025)	arXiv'25	40.2	30.2	12.1	78.6	44.1	35.2	25.8	12.4	53.5	41.8	37.4
YOLOv12-s(Tian et al., 2025)	NeurIPS'25	45.9	34.4	13.5	81.2	45.4	36.1	28.7	15.4	54.3	47.5	40.2
Ours		**53.3**	**41.4**	**18.6**	**84.6**	**49.5**	41.2	**32.4**	20.3	**64.4**	**53.6**	**45.9**



(a) VisDrone (b) RSOD (c) DIOR

Figure S1: Column-normalized confusion matrices on the VisDrone validation split (548 images), RSOD evaluation split (187 images), and configured DIOR evaluation split (3,463 images). The matrices use post-NMS detections under the 640×640 evaluation protocol: candidate confidence floor 0.001, NMS IoU 0.70, displayed-matrix confidence threshold 0.25, and one-to-one matching at IoU 0.45. Columns are ground-truth classes and rows are predicted classes; each column sums to one up to rounding under the common $[0, 1]$ color scale. The BG column records unmatched false positives by predicted class, whereas the BG row records unmatched ground-truth instances by class.

Complete Results on RSOD. RSOD contains only Aircraft, Oil Tank, Overpass, and Playground. Main-paper Table 2 therefore already reports every category, with PAVER reaching 96.7% mAP_{50} and the strongest Overpass AP at 95.8%.

Comprehensive Results on DIOR. Table S7 reports all 20 DIOR categories. Vehicle AP is 61.4% for the PAVER detector, compared with 58.3% for YOLOv12-s and 58.0% for AGMFNet; the full table shows the accompanying category-level trade-offs.

Table S7: Full per-category comparison on DIOR. AL: Airplane; AT: Airport; BF: Baseball Field; BC: Basketball Court; B: Bridge; C: Chimney; D: Dam; ESA: Expressway Service Area; ETS: Expressway Toll Station; GF: Golf Field; GB: Ground Track Field; HB: Harbor; OP: Overpass; S: Ship; SD: Stadium; ST: Storage Tank; TC: Tennis Court; TS: Train Station; V: Vehicle; WM: Windmill. **Bold** and underlined mark best and second-best.

Method	Venue	AL↑	AT↑	BF↑	BC↑	B↑	C↑	D↑	ESA↑	ETS↑	GF↑	GB↑	HB↑	OP↑	S↑	SD↑	ST↑	TC↑	TS↑	V↑	WM↑	mAP_{50} ↑
RAST-YOLO(Jiang & Wu, 2023)	Access'23	84.3	76.4	78.7	85.9	40.2	76.8	50.2	62.6	56.5	77.1	73.7	61.1	56.6	91.1	74.3	77.9	89.3	53.3	54.0	76.2	69.8
MSFC-Net(Zhang et al., 2022)	TGRS'22	85.8	76.2	74.4	91.1	44.2	78.1	55.5	60.9	59.5	76.9	73.7	49.6	57.2	89.6	69.2	76.5	86.7	51.8	55.2	**84.3**	69.8
TMAFNet(Gao et al., 2023)	TGRS'23	**92.2**	77.7	75.0	**91.3**	**47.1**	78.6	53.6	67.1	**66.2**	78.5	76.3	64.9	61.4	90.5	72.2	75.4	90.7	62.1	55.2	83.2	72.9
AGMFNet(Gao et al., 2024)	TGRS'24	90.9	72.8	79.3	89.7	44.7	**81.4**	59.3	66.0	62.7	73.8	**79.2**	65.0	**61.7**	91.7	**78.6**	75.8	90.7	60.0	58.0	83.1	73.2
YOLOv10-n(Wang et al., 2024a)	NeurIPS'24	72.4	74.1	77.0	84.3	34.3	74.1	53.6	59.5	52.0	72.2	67.5	61.0	53.8	88.1	69.5	71.7	85.9	54.7	45.6	72.2	66.2
YOLOv11-s(Khanam & Hussain, 2024)	arXiv'24	76.2	67.4	76.9	85.9	27.9	73.6	47.8	59.3	51.2	72.7	63.9	51.7	50.1	87.4	60.8	67.0	86.1	45.5	43.1	71.5	63.3
MFAE-YOLO(Liu et al., 2025b)	TGRS'25	80.9	72.8	76.5	86.1	33.8	74.8	53.9	62.3	55.2	72.7	65.1	61.2	53.8	89.4	69.5	72.6	87.4	49.0	46.4	76.1	67.0
YOLOv13-s(Lei et al., 2025)	arXiv'25	81.3	77.7	76.9	88.1	40.9	77.2	63.6	58.8	53.4	79.6	73.1	64.5	57.6	89.9	70.9	74.7	88.8	59.8	50.8	78.3	70.3
YOLOv12-s(Tian et al., 2025)	NeurIPS'25	84.5	83.1	**81.5**	91.1	46.8	80.6	64.7	**71.3**	61.5	**82.8**	74.6	65.4	61.0	92.2	74.3	**81.4**	**91.8**	**62.9**	58.3	81.2	**74.5**
Ours		85.7	**83.8**	80.1	90.9	46.6	79.8	**67.1**	65.9	64.9	81.6	74.1	**66.9**	60.9	**92.4**	73.5	81.2	91.6	61.5	**61.4**	81.0	**74.5**

Strict localization metrics. The main paper reports the compact $mAP_{50:95}$ comparison and VisDrone AP_{75} and AP_S. We retain the full per-category AP_{50} tables here for detailed inspection.

S7 Confusion Analysis

Fig. S1 reports column-normalized confusion matrices using the splits and post-NMS thresholds specified in its caption. Matching is class-agnostic and IoU-based before the matched prediction and ground-truth class labels determine the matrix cell. A diagonal entry is therefore the fraction of a ground-truth class assigned to the matching prediction class. Off-diagonal object rows represent inter-class confusion; the BG row represents unmatched ground-truth objects, and the BG column summarizes unmatched false positives. Across all three benchmarks, the dominant object-class diagonal responses indicate that most represented ground-truth instances are assigned to the matching class, while the visible residual mass is concentrated in a small number of inter-class or background errors.

S8 CATEGORY-AGNOSTIC SIZE BINS AND EFFICIENCY DIAGNOSTICS

Parameter and computation reporting. The checkpoint used for the detector comparisons and the validation-fitted Risk-conditioned Interface contains 10,261,048 trainable parameters. We therefore report 10.3M parameters throughout, using the standard decimal-million convention. Its computational cost is reported consistently as 40.0 GFLOPs at 640 pixels.

Category-agnostic size-bin evaluation. To complement the selected-category mAP_{50}^{tiny} metric, Table S8 reports category-agnostic COCO-style AP averaged over IoU thresholds 0.50:0.05:0.95 within three size strata. Let A be box area in resized detector-input pixel coordinates after the common isotropic gain; letterbox padding changes position but not A. The intervals are $0 < A < 32^2$ for AP_S, $32^2 \leq A < 96^2$ for AP_M, and $A \geq 96^2$ for AP_L. For each stratum, ground-truth instances outside its interval are marked ignored. A detection matched to an ignored ground truth is also ignored; an unmatched detection is ignored when its own area lies outside the active interval and is counted as a false positive otherwise. The same coordinate and ignore rules are applied to both compared detectors on all three datasets.

Table S8: Category-agnostic size-bin evaluation. Rankings are computed within each dataset. **Bold** and underlined mark best and second-best.

Dataset	Method	$\text{AP}_S \uparrow$	$\text{AP}_M \uparrow$	$\text{AP}_L \uparrow$
VisDrone	YOLOv12-s	16.3	33.9	45.8
VisDrone	PAVER	**21.6**	**38.7**	**49.6**
RSOD	YOLOv12-s	51.4	68.7	75.9
RSOD	PAVER	**55.8**	**70.9**	**76.8**
DIOR	YOLOv12-s	38.6	**59.4**	**66.2**
DIOR	PAVER	**41.3**	56.8	63.5

Detector-core latency. The main paper reports Params and GFLOPs for accuracy–efficiency comparison. Table S9 reports batch-size-1 TensorRT detector-core inference time and reciprocal throughput. The timing excludes image decoding, preprocessing, NMS, P2 and P3 feature capture for the external Risk-conditioned Interface, ROIAlign, prototype comparison, and confidence calibration; the evaluation batch size used for metric computation should therefore not be interpreted as deployment latency.

Table S9: Detector-core TensorRT FP16 efficiency under batch size 1 on an NVIDIA T4. Latency measures detector-core execution only; it is not an end-to-end timing of the external Risk-conditioned Interface. Bold and underlined values denote best and second-best results, respectively.

Model	Params↓	GFLOPs↓	Detector core (ms per image)↓	FPS↑	Peak mem. (GB)↓
YOLOv13-s	**9.0**	**20.8**	2.78	359.7	**0.76**
YOLOv12-s	9.3	21.4	**2.61**	**383.1**	0.82
PAVER	10.3	40.0	4.89	204.5	1.31

High-resolution diagnostic.

Table S10: Inference-resolution sensitivity on VisDrone using checkpoints trained at 640×640. Rankings are computed across all method–resolution rows. **Bold** and underlined mark best and second-best.

Method	Input	$mAP_{50} \uparrow$	$mAP_{50}^{tiny} \uparrow$	GFLOPs↓
YOLOv12-s	640×640	40.2	35.3	**21.4**
PAVER	640×640	45.9	41.7	40.0
YOLOv12-s	1024×1024	44.6	41.1	54.8
PAVER	1024×1024	49.0	46.8	102.4
YOLOv12-s	1280×1280	46.9	44.0	85.6
PAVER	1280×1280	**50.6**	**49.1**	160.0

Table S11: Progressive component ablation from the YOLOv11-s structural baseline on VisDrone. The SDIoU+DFL objective, SDIoU-based assignment, model scale, and common training and evaluation protocol are fixed. **Bold** and underlined mark best and second-best.

Model stage	CSFM	SCRG	HRPE	mAP_{50} ↑	Ped.↑	Bicycle↑	Bus↑
YOLOv11-s base				39.1	40.9	13.2	55.3
+ HRPE only			✓	43.5	50.9	16.8	<u>58.5</u>
+ HRPE + CSFM	✓		✓	<u>44.5</u>	<u>51.9</u>	<u>17.9</u>	56.8
PAVER (Full)	✓	✓	✓	**45.9**	**53.3**	**18.6**	**64.4**

S9 Extended Ablation Studies

S9.1 Class-Wise Progressive Ablation

The main paper reports the component-isolation table with strict metrics. Table S11 preserves the corresponding class-wise progressive ablation for Pedestrian, Bicycle, and Bus on VisDrone. Its first row is the YOLOv11-s structural parent, and every row retains the selected SDIoU+DFL objective, SDIoU-based task-aligned assignment, model scale, and common training and evaluation protocol.

S9.2 Controlled Replacement Details

The main paper summarizes the strongest controlled alternatives. Tables S12, S13, and S14 provide the full one-factor replacement details. In every study, the alternative occupies the same designated sites and preserves the corresponding feature interfaces; the remaining modules, SDIoU+DFL objective, SDIoU-based assignment, input size, training schedule, and evaluation protocol are fixed.

Backbone replacements. Each alternative replaces CSFM at all four backbone stages with matched input and output widths and the same stage-wise repeat schedule. The YOLOv12 alternative follows its stage-wise block schedule, using C3k2 in shallower stages and A2C2f in deeper stages. SCRG, HRPE, the detection heads, and all optimization settings remain unchanged.

Fusion replacements. Each alternative replaces SCRG at the same two early-neck sites and receives the same aligned high-level and low-level tensors. The one-way sigmoid gate predicts a sigmoid mask from the coarse stream and applies it to the fine stream; the additive attention gate jointly transforms the aligned coarse and fine streams before predicting a spatial attention mask. SE-style and CBAM provide standard channel and channel–spatial recalibration controls, respectively. CSFM, HRPE, and all remaining settings are fixed.

Decoder replacements. All alternatives retain prediction at P2–P5. The P2-head-only variant adds P2 prediction without reinjecting P2 into the P3–P5 bottom-up path. PAN-FPN follows the conventional iterative bottom-up pattern (Liu et al., 2018): each fused feature is refined into P_ℓ^{PAN}, and that refined output is downsampled for the next level. HRPE instead downsamples the pre-refinement states G_3 and G_4 while sending their parallel C3k2 outputs to the P3 and P4 heads. BiFPN uses normalized weighted bidirectional fusion over the same levels. CSFM, SCRG, head count, feature interfaces, and all remaining settings are fixed.

Table S12: Backbone block replacement on VisDrone. Alternatives occupy all four CSFM sites with matched widths and repeats; SCRG, HRPE, the SDIoU+DFL objective, SDIoU-based assignment, and common training and evaluation settings are fixed. **Bold** and underlined mark best and second-best.

Block	mAP_{50} ↑	mAP_{50}^{tiny} ↑	$mAP_{50:95}$ ↑	Params↓	GFLOPs↓
C3k2	44.9	40.2	27.2	**9.5**	**39.1**
CSP	44.7	39.8	27.0	9.6	39.6
ELAN	45.3	40.8	27.6	10.2	41.7
YOLOv12 stage-wise blocks (C3k2 and A2C2f)	45.4	41.0	27.7	9.7	39.8
CSFM	**45.9**	**41.7**	**28.0**	10.3	40.0

Table S13: Fusion-gate replacement at the two SCRG sites on VisDrone. All alternatives receive the same aligned tensors; CSFM, HRPE, the SDIoU+DFL objective, SDIoU-based assignment, and common training and evaluation settings are fixed. **Bold** and underlined mark best and second-best.

Fusion gate	mAP_{50} ↑	mAP_{50}^{tiny} ↑	$mAP_{50:95}$ ↑	GFLOPs↓
SE-style	44.8	39.9	27.1	**39.4**
One-way sigmoid gate	45.1	40.6	27.4	39.7
Additive attention gate	45.3	40.8	27.6	40.1
CBAM	45.4	41.0	27.7	40.2
SCRG	**45.9**	**41.7**	**28.0**	40.0

Table S14: P2–P5 decoder replacement on VisDrone. CSFM, SCRG, four prediction levels, the SDIoU+DFL objective, SDIoU-based assignment, and common training and evaluation settings are fixed. **Bold** and underlined mark best and second-best.

Decoder	mAP_{50} ↑	mAP_{50}^{tiny} ↑	$mAP_{50:95}$ ↑	GFLOPs↓
P2 head only	44.6	39.7	26.9	**37.6**
PAN-FPN (P2–P5)	45.2	40.5	27.4	39.3
BiFPN (P2–P5)	45.5	40.9	27.7	39.7
HRPE	**45.9**	**41.7**	**28.0**	40.0

S9.3 Full CSFM State-Flow Grid

Table S15 gives the complete design sweep over the three variables in Eq. 3: reference–evolution allocation ratio r, state-flow horizon N, and per-transition depth k. These VisDrone measurements select $(r, N, k) = (0.5, 2, 1)$ for the paper-level PAVER detector. The validation-fitted Risk-conditioned Interface consumes features from this selected detector and does not use the grid to choose a test-time configuration.

Table S15: CSFM state-flow ablation on VisDrone. The sweep varies the reference–evolution allocation ratio r, state-flow horizon N, and per-transition depth k in Eq. 3. Metrics are mAP_{50} ↑, mAP_{50}^{tiny} ↑, and GFLOPs↓ (@640). **Bold** and underlined mark best and second-best.

$N=1$					$N=2$					$N=3$				
r	k	mAP_{50} ↑	mAP_{50}^{tiny} ↑	GFLOPs↓	r	k	mAP_{50} ↑	mAP_{50}^{tiny} ↑	GFLOPs↓	r	k	mAP_{50} ↑	mAP_{50}^{tiny} ↑	GFLOPs↓
0.25	1	44.5	39.5	39.7	0.25	1	45.2	40.5	41.3	0.25	1	45.0	40.2	42.6
0.25	2	44.3	39.2	40.6	0.25	2	44.9	40.2	42.0	0.25	2	44.7	39.8	44.0
0.25	3	44.0	38.8	41.4	0.25	3	44.6	39.8	42.6	0.25	3	44.2	39.2	45.4
0.50	1	45.3	41.0	38.9	0.50	1	**45.9**	**41.7**	40.0	0.50	1	45.5	41.2	41.0
0.50	2	45.1	40.7	39.8	0.50	2	45.6	41.3	41.2	0.50	2	45.1	40.7	42.5
0.50	3	44.8	40.3	40.6	0.50	3	45.2	40.8	42.4	0.50	3	44.7	40.2	44.0
0.75	1	45.0	40.6	**38.1**	0.75	1	45.4	41.0	39.2	0.75	1	45.1	40.7	40.0
0.75	2	44.8	40.3	38.8	0.75	2	45.0	40.6	40.0	0.75	2	44.8	40.3	41.0
0.75	3	44.5	39.9	39.5	0.75	3	44.7	40.2	40.8	0.75	3	44.4	39.8	42.0

Scaling guidance. The selected $(r, N, k) = (0.5, 2, 1)$ configuration gives the highest mAP_{50} and mAP_{50}^{tiny} in the grid. Extending the state-flow horizon N to 3 or increasing per-transition depth k beyond 1 does not improve either metric for $r = 0.5$ and increases computation. Changing r instead reallocates channel width between the reference and evolution streams. These observations directly motivate the CSFM instantiation specified in Eq. 3.

S9.4 SCRG Variant Details

The main paper reports the compact SCRG variant comparison. Table S16 records the corresponding controlled results with all other PAVER detector components fixed. The label "SCRG-Channel" follows the ablation nomenclature and denotes the selected variant; its gate remains spatially varying across channels. The operator applies sigmoid pointwise to projected feature maps and therefore uses dense spatial–channel masks, as specified by Eqs. 5–7. This comparison supports the selected variant relative to the alternatives, but it does not isolate spatial from channel dependence within that gate.

Table S16: SCRG variants on VisDrone. "SCRG-Channel" follows the ablation nomenclature and denotes the selected variant, whose operator uses dense spatial–channel gates. All remaining components and evaluation settings are fixed. **Bold** and underlined mark best and second-best.

Variant	$mAP_{50}\uparrow$	$mAP_{50}^{tiny}\uparrow$	GFLOPs↓
SCRG-Spatial	45.5	41.2	<u>40.3</u>
SCRG-Deformable	<u>45.7</u>	<u>41.4</u>	41.4
SCRG-Channel (ours)	**45.9**	**41.7**	**40.0**

S10 Additional Qualitative Results

We present additional qualitative results of the PAVER detector on VisDrone, RSOD, and DIOR in Fig. S4, Fig. S5, and Fig. S6, respectively. These examples illustrate its detection quality under diverse object scales, dense layouts, and cluttered backgrounds.

Fig. S3 presents the CSFM state-progression illustration referenced in the main paper. It compares an S2 feature without CSFM with the subsequent S3 feature with CSFM and is therefore a qualitative progression rather than a same-stage, module-only ablation; the controlled evidence for CSFM is provided by Tables 5 and 6.

Fig. S2 visualizes SCRG responses for ships, aircraft, and cars. Because these per-image normalized maps are qualitative diagnostics, the controlled evidence for SCRG is provided by Tables 5 and 6.

Fig. S7 shows additional feature-map comparisons with recent YOLO-style baselines. In the yellow-boxed regions, the PAVER detector produces clearer local structures and sharper responses, suggesting stronger preservation of tiny-object cues and more discriminative feature representations.

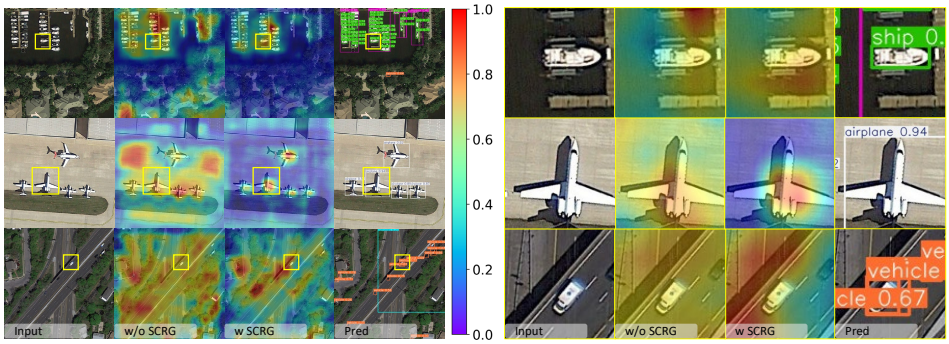
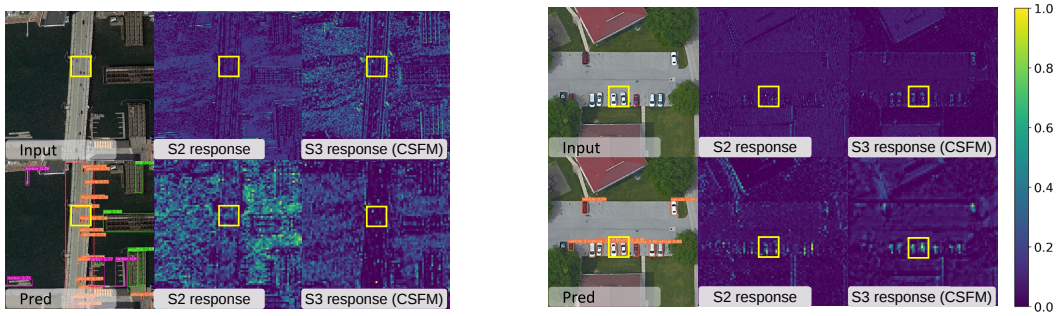


Figure S2: Qualitative SCRG responses (absolute-channel sum; per-image min–max) for ships, aircraft, and cars. In these examples, yellow boxes show sharper boundaries and lower background activation.



(a) (b)

Figure S3: Qualitative CSFM state-progression visualization for two scenes (a,b). The embedded labels identify an S2 feature without CSFM and the subsequent S3 feature with CSFM; S2 and S3 are consecutive backbone-stage labels used only in this visualization, not the P_2 and P_3 pyramid levels. Each response map is min–max normalized independently, so color intensity is comparable only within a map and indicates spatial concentration rather than absolute activation magnitude. Yellow boxes mark tiny objects. Because stage and module presence change together, this figure is an illustrative progression rather than a controlled module-only comparison.



Figure S4: Qualitative detection results of the PAVER detector on VisDrone.

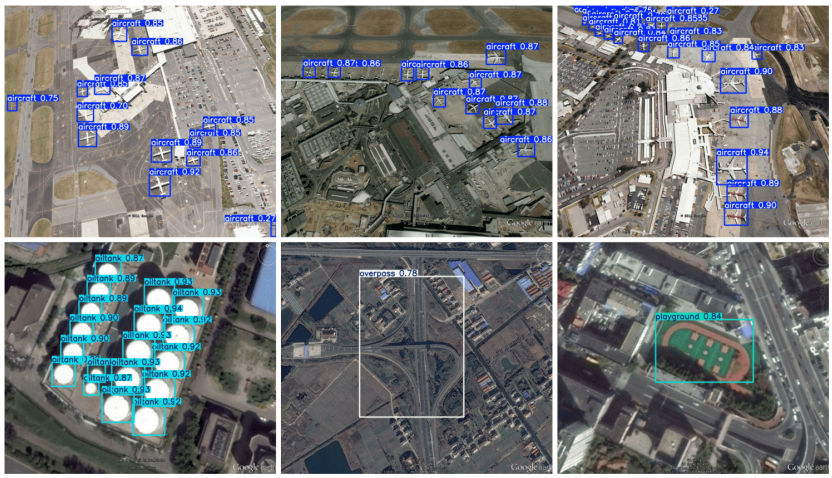


Figure S5: Qualitative detection results of the PAVER detector on RSOD.

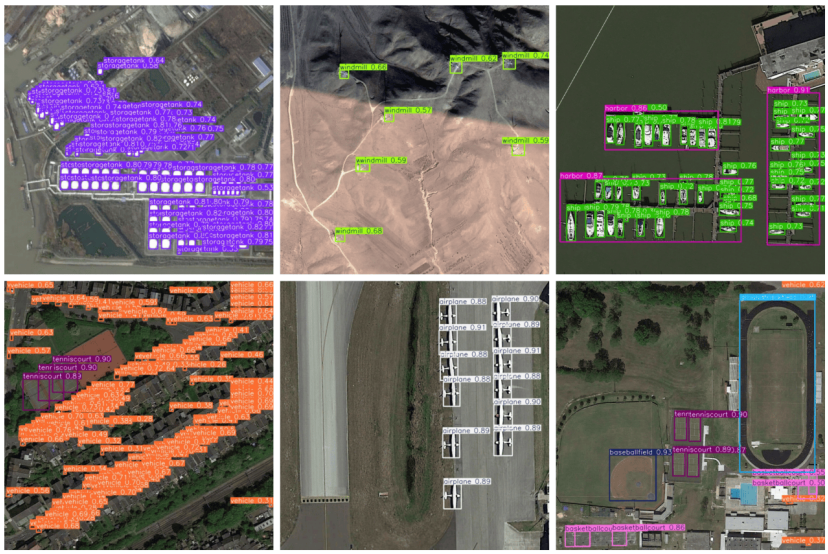


Figure S6: Qualitative detection results of the PAVER detector on DIOR.

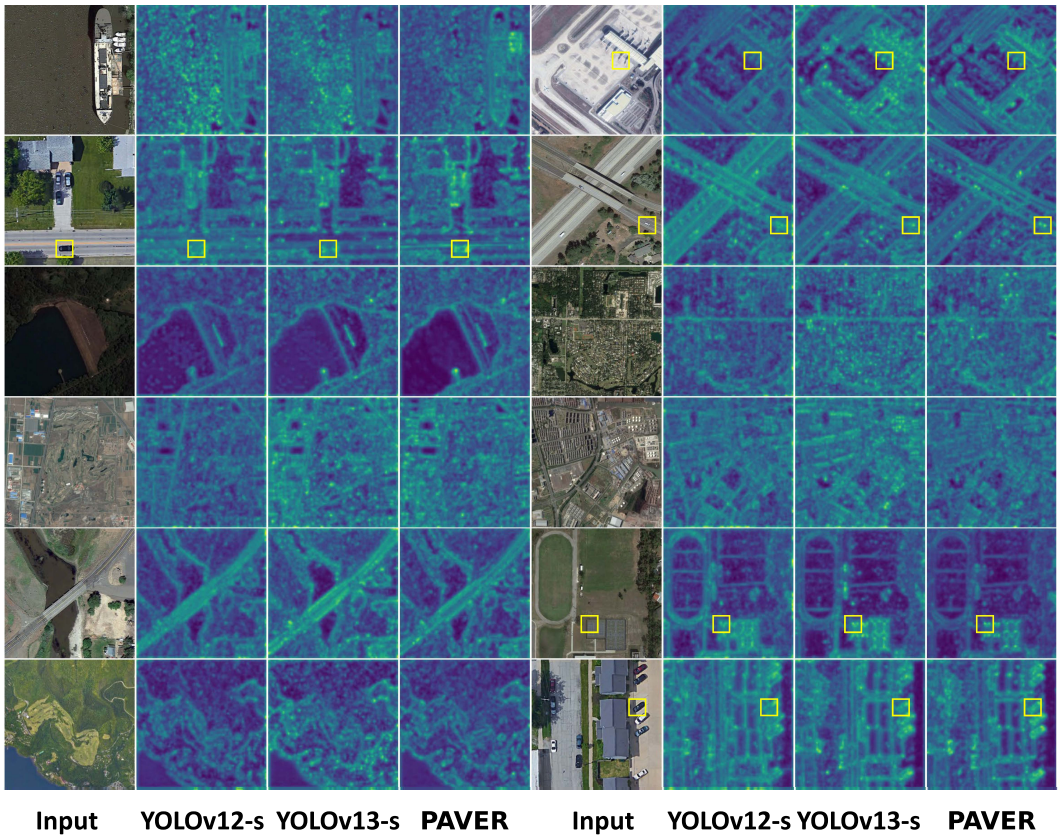
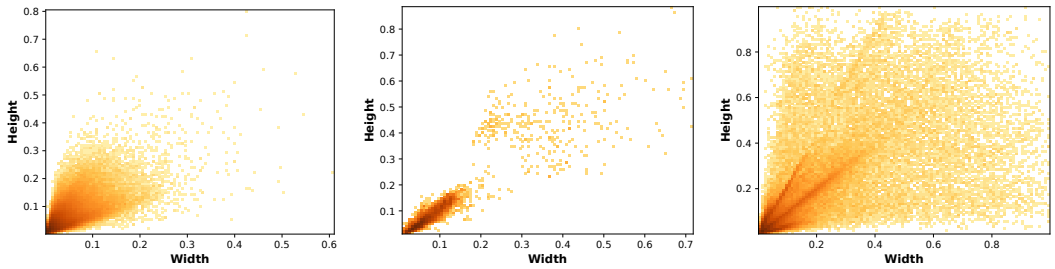


Figure S7: Additional feature-map comparison. The yellow-boxed regions highlight clearer local structures and stronger tiny-object responses in the PAVER detector than in recent YOLO-style baselines.

S11 Fairness Note for Reproduced Baselines



(a) VisDrone (b) RSOD (c) DIOR

Figure S8: Distribution of object sizes across VisDrone, RSOD, and DIOR. Box width and height are divided by the corresponding image width and height, respectively.

Dataset coverage and splits. VisDrone-DET2019 provides 6,471 training images, 548 validation images, a 1,610-image annotated test-dev split, and a separate 1,580-image test-challenge split. Detector comparisons report the validation result; calibrators are fitted on validation and evaluated on test-dev without further fitting. RSOD covers aircraft, oil tank, overpass, and playground. Its 936 images are divided by a deterministic class-stratified 80:20 split (seed 2027) into 749 training and 187 evaluation images; the evaluation partition also supplies validation fitness, and no separate blind test

Table S17: Controlled protocol used for the VisDrone, RSOD, and DIOR reproductions.

Component	Controlled setting
Input resolution	640×640
Training duration	Up to 300 epochs, with an early-stopping patience of 100 epochs
Training batch size	VisDrone: 32; RSOD and DIOR: 16
Optimizer	SGD (momentum 0.937, weight decay 5×10^{-4})
Learning-rate policy	Linear decay from 0.01 to 10^{-4}, with 3 warmup epochs
Initialization	From scratch; no external detector weights
Checkpoint selection	EMA checkpoint maximizing $0.1mAP_{50} + 0.9mAP_{50:95}$
Precision and reproducibility	Mixed precision, seed 0, deterministic computation
Color augmentation	HSV perturbation $(0.015,\ 0.7,\ 0.4)$
Geometric augmentation	Translation 0.1, scaling 0.5, horizontal flip 0.5
Composite augmentation	Mosaic disabled in the last 10 epochs
Unused composite transforms	MixUp and Copy-Paste disabled
Inference post-processing	NMS IoU threshold 0.7, maximum 300 detections

split is used. DIOR contains 23,463 optical remote-sensing images and 192,472 instances across 20 categories. We use 18,000 images for training and the configured 3,463-image evaluation partition for validation fitness, checkpoint selection, and the reported comparison; the auxiliary 2,000-image partition is not used in that comparison. We therefore refer to the configured partition as the evaluation split rather than an externally sequestered test set.

Compute environment and shared recipe. The detector checkpoints use Ultralytics 8.3.35 on a single GPU. Models are instantiated from their cited architecture definitions without loading external detector weights. All inputs are letterboxed to 640×640; the dataset-specific training batch sizes are 32 for VisDrone and 16 for RSOD and DIOR. Reproduced models share the dataset split, input size, optimizer, executed detection transforms, seed, NMS threshold, evaluator, and scale-aware overlap convention, but the common recipe does not force them into a common prediction head. For dense detectors with task-aligned assignment, the architecture-specific head is retained; SDIoU replaces the overlap used for stop-gradient assignment and positive-box regression, while classification, objectness when defined, and DFL when defined remain unchanged. For RT-DETR, the query head and Hungarian assignment are retained; SDIoU replaces the GIoU term in the matching cost and the GIoU regression term, while the classification and ℓ_1 box terms remain unchanged. Its normalized boxes are mapped to the 640×640 input frame for SDIoU: matching uses predicted-query area for β^{assign}, regression uses matched-target area for β^{reg}, and neither uses a pyramid stride. The PAVER detector follows the dense route with its P2–P5 head, BCE, and DFL. Thus the reported rows are controlled architectural reproductions under a shared overlap convention, rather than reproductions of each architecture's individually optimized official recipe. Table S17 records the common settings, and Table S18 records the architecture-specific substitutions for the principal comparison models.

With epoch index e and maximum duration $E = 300$, the post-warmup base learning rate follows

$$\mathrm{LR}(e) = \mathrm{LR}_0 \left[\left(1 - \frac{e}{E}\right)(1-\kappa) + \kappa \right], \qquad \mathrm{LR}_0 = 0.01, \quad \kappa = 0.01. \tag{S26}$$

The planned terminal learning rate is therefore 10^{-4}; an early-stopped run terminates at the rate corresponding to its final epoch. Early stopping and checkpoint selection use the same validation fitness reported in Table S17.

Augmentation check. Training includes Mosaic, affine translation and scaling, HSV perturbation, and horizontal flipping; MixUp and Copy-Paste are disabled. Disabling Mosaic reduces VisDrone mAP_{50}, mAP_{50}^{tiny}, and $mAP_{50:95}$ from 45.9, 41.7, and 28.0 to 44.4, 39.6, and 27.0, respectively.

The testing batch size is used for batched metric evaluation, not single-image latency measurement; the latter is reported separately in Sec. S8.

Controlled architecture adaptations. Our main comparisons use a controlled 640×640 protocol to isolate architectural effects under matched training and evaluation. They are not intended to reproduce the individually optimized headline result of every baseline. The principal architecture definitions follow YOLOv11 (Khanam & Hussain, 2024), YOLOv12 (Tian et al., 2025), YOLOv13 (Lei

Table S18: Architecture-specific adaptation of the principal VisDrone models under the controlled protocol. "Retained pathway" lists the head and objective components left unchanged from the cited architecture; all rows are initialized from architecture definitions without external detector weights. **Bold** and underlined mark best and second-best.

Method	Retained pathway	SDIoU substitution	mAP_{50} ↑
YOLOv11-s	Dense head; task-aligned assignment; classification and DFL	Assignment overlap and positive-box regression overlap	39.1
YOLOv13-s	Dense head; task-aligned assignment; classification and DFL	Assignment overlap and positive-box regression overlap	37.4
YOLOv12-s	Dense head; task-aligned assignment; classification and DFL	Assignment overlap and positive-box regression overlap	<u>40.2</u>
RT-DETR	Query head; Hungarian assignment; classification and ℓ_1 box loss	GIoU matching-cost and regression-overlap terms	30.1
PAVER	P2–P5 dense head; task-aligned assignment; BCE and DFL	Assignment overlap and positive-box regression overlap	**45.9**

et al., 2025), and RT-DETR (Zhao et al., 2024). Table S18 makes the retained training pathway and exact overlap substitution explicit.

S12 Supplementary Horizontal-Box Evaluation on SODA-A

The official SODA benchmark defines SODA-A as 2,513 high-resolution aerial images with 872,069 oriented-box instances from nine classes (Cheng et al., 2023). We retain the official training and validation partitions and derive each horizontal target as the tight axis-aligned enclosure of its released oriented annotation. The released image units are letterboxed to 640×640, and all compared models use the same controlled detector recipe without further image slicing or test-time tiling. Accordingly, this experiment is a supplementary horizontal-box adaptation rather than an official oriented-box leaderboard submission. Table S19 reports the resulting validation comparison.

Table S19: Supplementary SODA-A horizontal-box validation using the official partition and a common annotation-conversion and evaluation protocol. These results are not official oriented-box leaderboard entries. **Bold** and underlined mark best and second-best.

Method	mAP_{50} ↑	$mAP_{50:95}$ ↑	Params↓	GFLOPs↓
YOLOv13-s (Lei et al., 2025)	54.6	16.7	**9.0**	**20.8**
YOLOv12-s (Tian et al., 2025)	<u>57.2</u>	<u>17.8</u>	<u>9.3</u>	<u>21.4</u>
PAVER	**61.6**	**20.7**	10.3	40.0

S13 Potential Applications

The PAVER framework targets tiny-object detection in remote sensing under extreme scale variation, dense layouts, and cluttered backgrounds, which are common in practical aerial monitoring.

Environmental Monitoring. Improved sensitivity to tiny instances can benefit wide-area monitoring of tiny vessels, debris, or localized ecological changes under low contrast and complex textures.

Disaster Response and Damage Assessment. After floods, earthquakes, or wildfires, detecting tiny man-made objects (e.g., vehicles or temporary structures) and fine-grained changes can help prioritize regions for inspection across varying altitudes and viewing angles.

Infrastructure Inspection. Accurate localization of tiny objects around roads, bridges, and industrial facilities can support large-scale asset inventory and anomaly screening, such as identifying vehicles on overpasses or tiny objects near facilities.

Traffic and Mobility Analytics. In urban aerial imagery, crowded pedestrians, bicycles, and tiny vehicles are highly sensitive to localization jitter. More stable detections can therefore improve downstream counting, density estimation, and flow analysis.

Operational Considerations. Remote-sensing detection may also be used in sensitive surveillance settings. Responsible deployment should consider domain shifts, potential biases, and privacy-related constraints, and should report evaluation conditions under representative shifts.

S14 Scope and Future Directions

PAVER is evaluated in the common RGB, horizontal-box, closed-set setting. This controlled scope isolates the formation–fusion–decoding pathway and the Risk-conditioned Interface while leaving natural extensions to multispectral inputs, rotated heads, and open-vocabulary detection. For deployment-oriented variants, quantization, hardware-aware scheduling, and end-to-end profiling of feature capture and ROI processing provide direct routes to reduce runtime and memory overhead.

The confidence study fits all calibration components on VisDrone validation and evaluates them on fixed test-dev candidates under the stated correctness protocol. Broader evaluation can examine transfer to RSOD, DIOR, SODA-A, and distribution shifts, repeat calibration across additional seeds, and vary prototype budgets, ROI scales, and pyramid evidence. Incorporating P4 and P5 or temporal evidence is also compatible with the current interface. These extensions broaden the evaluated operating regimes while preserving PAVER's detector–interface separation.