



Semantic-guided adaptive dual-stream inference for marine floating litter detection

Junsen Zeng^{a,b}, Zirui Zhao^b , Honglei Jiang^b, Shuowen Huang^b, Qianyi Huang^{c,d},
Xiangrong Xu^{b,*} , Kefu Yu^b

^a School of Resources, Environment and Materials, Guangxi University, Nanning 530004, China

^b Guangxi Laboratory on the Study of Coral Reefs in the South China Sea, Coral Reef Research Center of China, School of Marine Sciences, Guangxi University, Nanning 530004, China

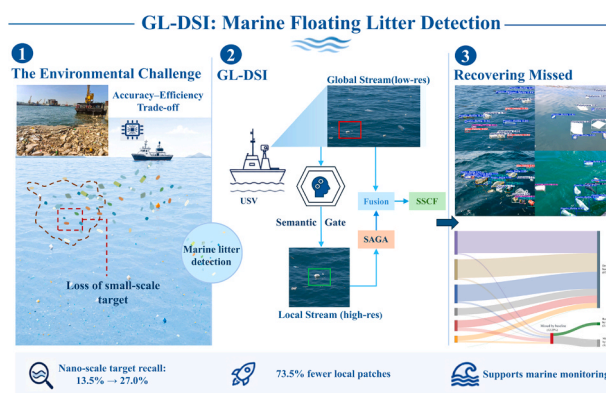
^c State Key Laboratory of Tropical Oceanography, Guangdong Provincial Key Laboratory of Applied Marine Biology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China

^d University of Chinese Academy of Sciences, Beijing 100049, China

HIGHLIGHTS

- Semantic-guided dual-stream inference selectively refines uncertain local regions.
- SAGA and SSCF suppress slice-boundary artifacts and marine pseudo-targets.
- Local patch processing is reduced by 73.5% versus exhaustive slicing.
- GL-DSI attains 93.4% F1 and an 11.0-point recall gain for targets below 256 px².
- Jetson Orin Nano tests quantify latency, power, memory, and energy trade-offs.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:

Small target detection
Semantic-guided slicing
Marine floating litter
Edge computing
Deep learning

ABSTRACT

Marine floating litter detection on autonomous edge platforms, such as Unmanned Surface Vehicles (USVs), faces a fundamental scale–compute conflict: downsampling can obscure the fine-grained features of micro-scale targets, while native high-resolution inference often exceeds the computational budgets of edge platforms. Uniform slicing recovers small-scale target recall but exhaustively processes empty sea surfaces, amplifying false positives from sun glint and sea foam and substantially increasing inference latency. We propose GL-DSI, a Global–Local Dual-Stream Inference framework that replaces blind slicing with semantic-guided adaptive extraction: a low-resolution global stream localizes suspect regions, triggering high-resolution local inference only where small-scale targets likely exist. A Size-Aware Gaussian Attenuation (SAGA) module suppresses boundary artifacts, while the Scale-Semantic Consistency Filter (SSCF) removes environment-induced pseudo-targets. On our marine floating litter dataset (MFL-Dataset), GL-DSI with YOLOv11s reaches an F1-score of 93.4% at 44.6 frames per

* Corresponding author.

E-mail address: xuxr@gxu.edu.cn (X. Xu).

<https://doi.org/10.1016/j.jhazmat.2026.143863>

Received 16 May 2026; Received in revised form 21 September 2026; Accepted 8 October 2026

Available online 8 October 2026

0304-3894/© 2026 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

second (FPS) on the desktop benchmark. Compared to its standard baseline, it improves the detection of highly degraded small-scale targets, boosting recall for targets with areas below 256 px^2 by 11.0 %age points and lifting nano-scale target recall from 13.5% to 27.0%, while remaining substantially less computationally intensive than exhaustive slicing.

1. Introduction

Marine plastic pollution has become a persistent environmental challenge, with floating debris widely distributed across coastal waters and broader ocean circulation systems [1]. A substantial portion of this pollution consists of fragmented and weathered materials whose limited visual representation, irregular morphology, and high mobility complicate reliable observation [1–3]. Accurate detection of such debris is important for pollution assessment, transport analysis, and cleanup planning. Image-based marine and water-surface monitoring has been explored across satellite, aerial, and near-surface sensing settings [4–11]. Compared with large-scale remote sensing, near-surface platforms can provide more direct visual access to floating litter in complex coastal environments.

Despite this potential, marine floating litter detection on edge platforms remains difficult because of the tension between object scale and computational budget. In real deployment scenarios, detectors are expected to operate on resource-constrained onboard hardware while maintaining sufficient sensitivity to small-scale targets. However, most lightweight detectors rely on fixed-resolution inputs and aggressive downsampling, which often suppress the fine-grained visual cues required to identify fragmented debris. In marine scenes, this issue is further aggravated by cluttered backgrounds, specular reflections, sea foam, and wave crests, all of which can obscure true targets or induce false detections. As a result, marine-oriented models still face a difficult balance among sensitivity to small-scale targets, robustness to complex backgrounds, and real-time inference performance [12,13].

A straightforward solution is to increase the input resolution or adopt slicing-based inference. High-resolution processing can preserve fine-grained visual details of small-scale targets [14–18], while slicing methods such as Slicing Aided Hyper Inference (SAHI) and task-specific variants improve their detectability by processing local patches [19–21]. However, direct high-resolution inference substantially increases memory and latency costs, whereas blind slicing exhaustively processes empty water regions, introducing redundancy [17,19–21] and potentially amplifying marine optical interference such as sun glint, foam, and wave texture [22,23]. These limitations make conventional resolution-enhancement strategies poorly suited to real-time edge monitoring.

Lightweight You Only Look Once (YOLO) architectures have been widely applied to real-time detection on resource-constrained platforms, including marine litter monitoring [12,24–27]. Recent approaches improve small-scale target representations through resolution-preserving architectures, multiscale feature fusion, context-aware representations, and hierarchical modeling [14, 16, 28], but remain constrained by onboard memory and latency budgets.

To address this trade-off, we propose the Global–Local Dual-Stream Inference (GL-DSI) framework for compute-efficient marine floating litter detection. GL-DSI employs a Semantic-Guided Dual-Stream (SGDS) mechanism in which a low-resolution global stream identifies suspect regions and selectively triggers high-resolution local processing. A Size-Aware Gaussian Attenuation (SAGA) module suppresses patch-boundary artifacts, while a Scale-Semantic Consistency Filter (SSCF) suppresses environmentally induced pseudo-targets. Experimental results on the custom marine floating litter dataset show that GL-DSI achieves strong accuracy while providing a favorable accuracy–computation trade-off across the evaluated inference settings, supporting small-scale marine debris monitoring.

Throughout this study, scale descriptors such as micro-scale and

nano-scale targets refer strictly to projected bounding-box area in the 2D image plane rather than physical particle size. For instance, nano-scale targets occupy areas below 64 px^2 under our dataset categorization, a definition distinct from physical size definitions used in environmental chemistry, where microplastics typically refer to particles smaller than 5 mm. Accordingly, our objective is to recover visually degraded floating litter in image space rather than infer physical particle dimensions or mass.

2. Methodology and dataset

2.1. Overview of the GL-DSI framework and dataset

The proposed GL-DSI framework, illustrated in Fig. 1, is designed to detect floating litter across extreme scale ranges with high computational efficiency. A low-resolution global stream establishes scene-level context, whereas a high-resolution local stream captures fine-grained details; both streams share a unified multiscale encoder to reduce parameter redundancy and memory cost.

Low-confidence global predictions serve as structural priors for the SGDS, which selectively triggers high-resolution local processing around potential small-scale targets. SAGA then penalizes boundary-affected predictions according to object pixel area, while predictions from both streams are mapped to the original coordinate space and refined by SSCF to suppress environmental noise. To characterize extreme apparent-scale variation, we further subdivide the small-scale regime into nano-scale, micro-scale, and tiny-scale categories.

The marine floating litter dataset (MFL-Dataset) comprises 3791 images (1920×1080 – 3840×2160 pixels), with over 95% originating from field imagery acquired in the Maowei Sea, Qinzhou, Guangxi, China, and less than 5% from the public IWHR AI_Lable_Floater_V1 dataset. The public subset was manually screened to remove non-debris entities, such as inland water plants or algae, and exclude sequences of video frames. All images were annotated in the Computer Vision Annotation Tool (CVAT) using a unified taxonomy and scale definitions, yielding 21,560 instances across seven categories: Plastic Bottle, Plastic Bag, Polystyrene, Metal Can, Fishing Gear, Other Plastic, and Misc Debris.

To guarantee high annotation fidelity for extremely small instances, annotators utilized digital zooming in CVAT combined with local contextual reasoning to distinguish genuine debris fragments from marine optical noise (such as sun glint and sea foam). A multi-stage quality control protocol was enforced: primary annotations were independently conducted by trained annotators and subsequently audited by a senior researcher. Ambiguous instances were excluded to eliminate false positives and maintain high inter-annotator visual agreement for small-scale target annotations.

The entire dataset was randomly partitioned into training, validation, and test sets at a ratio of 7:1:2 (specifically, 2654 images with 15,243 instances for training; 379 images with 2146 instances for validation; and 758 images with 4171 instances for testing). The field-collected images in the MFL-Dataset are discrete static photographs taken from varied handheld camera angles. Statistical analysis in Fig. 2 reveals a broad distribution of projected bounding-box areas and a nonuniform distribution of normalized object centers across the training set.

In this work, we refer to objects with limited projected pixel areas and potentially insufficient visual representation after downsampling as small-scale targets. This term serves as an umbrella category that in-

cludes nano-scale, micro-scale, and tiny-scale targets defined according to their projected bounding-box areas. To rigorously evaluate detection capabilities across ultra-small dimensions, we extended the standard Common Objects in Context (COCO) scale categorization. While COCO categorizes objects into small (area $< 1024 \text{ px}^2$), medium ($1024 \text{ px}^2 \leq \text{area} < 9216 \text{ px}^2$), and large (area $\geq 9216 \text{ px}^2$), we further subdivided the COCO small-object category into three distinct fine-grained tiers based on projected bounding box area in the image plane: nano-scale (area $< 64 \text{ px}^2$, corresponding to extremely small targets below approximately 8×8 pixels), micro-scale ($64 \text{ px}^2 \leq \text{area} < 256 \text{ px}^2$, approximately 8×8 – 16×16 pixels), and tiny-scale ($256 \text{ px}^2 \leq \text{area} < 1024 \text{ px}^2$, approximately 16×16 – 32×32 pixels). By definition, the union of nano-scale, micro-scale, and tiny-scale targets covers objects with projected bounding-box areas below 1024 px^2 in the 640×640 evaluation coordinate space. Crucially, this discrete scale categorization is strictly defined for dataset evaluation and is functionally independent of the continuous noise filtering thresholds (such as the τ_{noise} cutoff) utilized during the inference stage.

2.2. Semantic-guided dual-stream decoupling for extreme scale variation

A fundamental challenge in marine debris detection is the extreme intra-class scale variation, where substantial floating structures and heavily fragmented small-scale debris coexist within the same visual field. A physically large object captured from a significant distance presents the same extreme scale challenge as a genuinely small-scale fragment. In our training set, annotated object areas span

approximately five orders of magnitude. This extreme scale range creates an inherent representation conflict for conventional single-stream detectors [29,30], forcing an unfavorable resolution-compute trade-off that typically exceeds the capacity of edge platforms. This motivates architectures that decouple coarse contextual reasoning from fine-detail preservation and exploit complementary multiscale context fusion to strengthen representations under large scale variation.

GL-DSI addresses this trade-off through SGDS, which decouples scene-level contextual inference from selective local detail recovery rather than exhaustively processing image patches. The global stream first evaluates the downsampled image to generate candidate predictions B_{global} . Bounding boxes with confidence scores falling within a specific uncertainty interval $[C_{\text{gate}}, C_{\text{raw}}]$ are designated as suspect regions. Here, C_{gate} represents the lower activation bound for the local stream, while C_{raw} serves as the default global confidence threshold. The local stream is then exclusively triggered to extract and process high-resolution patches centered around these suspect regions. This semantic gating ensures that scarce computational resources are preferentially allocated to regions with a high probability of containing visually degraded small-scale target candidates.

To systematically determine the optimal uncertainty interval, a comprehensive two-dimensional grid search was conducted on the validation set. As illustrated in Fig. 3, we evaluated the parameter space across $C_{\text{gate}} \in [0.05, 0.25]$, and C_{raw} from 0.15 to 0.35. A fundamental architectural constraint dictates that the local trigger threshold must be lower than the global default threshold to activate local refinement, resulting in the invalid parameter space masked in the analysis.

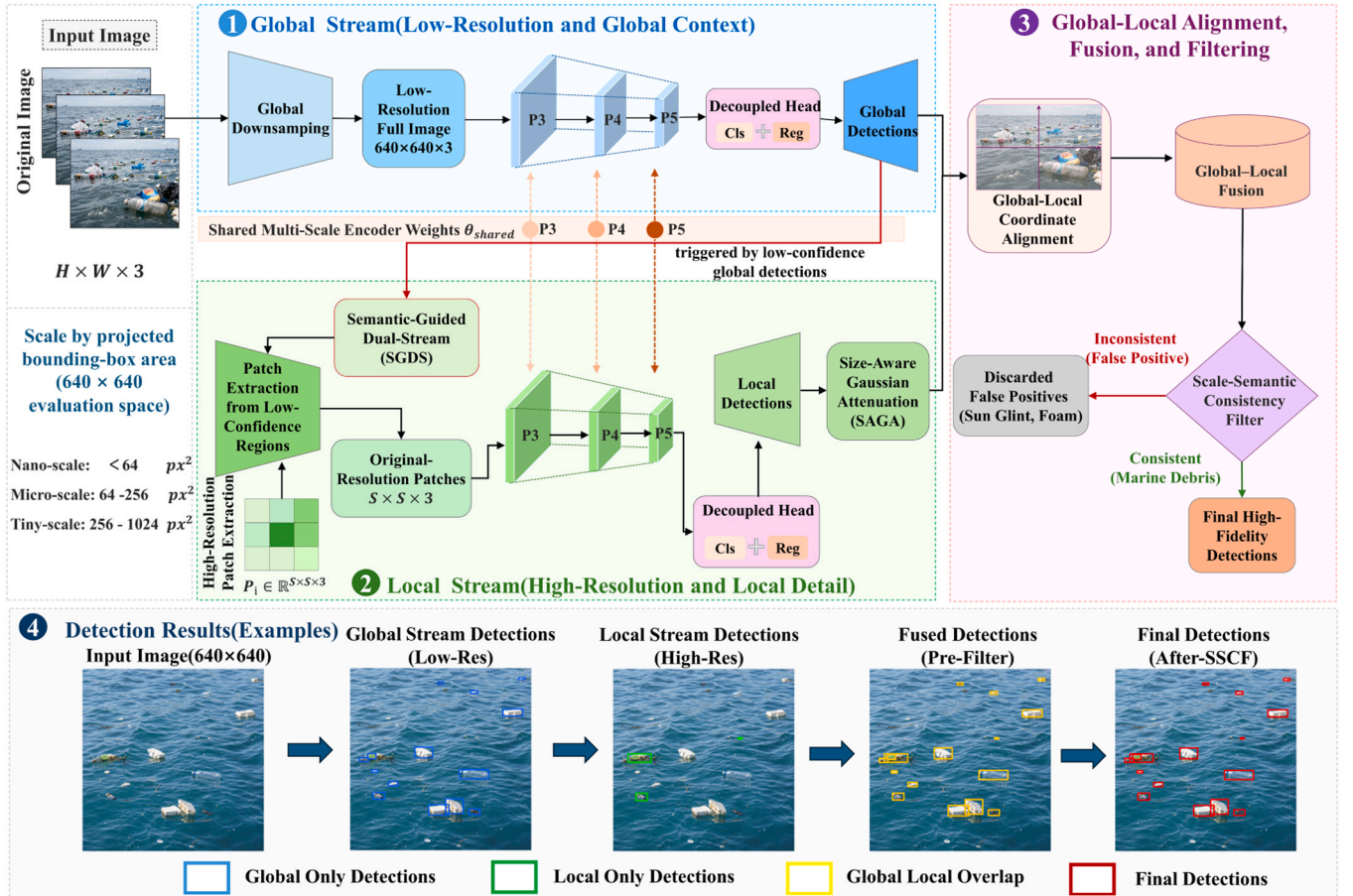


Fig. 1. Overview of the proposed GL-DSI framework. The global stream captures scene-level context and generates candidate regions for selective high-resolution local processing. The local stream recovers fine-grained visual details from adaptively extracted patches using shared encoder weights. SAGA attenuates boundary-affected predictions, after which global and local predictions are aligned and fused, and SSCF suppresses residual environmental pseudo-targets to produce final detection.

While the absolute peak F1-score of 94.33% converges near the upper boundary of 0.25 and 0.35, adopting such an aggressive gate severely throttles the activation frequency of the local stream. To maximize the recovery of highly degraded small-scale debris, the primary detection objective of this study, we deliberately prioritized micro-scale target recall over marginal precision gains. Consequently, the thresholds were established at $C_{gate} = 0.12$ and $C_{raw} = 0.25$. This configuration yields an F1-score of 93.68% while maintaining sufficient sensitivity to trigger high-resolution local patch extraction. These operating thresholds were selected exclusively using the validation set and fixed before test-set evaluation. Their post-selection sensitivity and degradation boundaries, including the effect of C_{gate} on nano-scale target recall, are further characterized on the complete test set in Section 3.4 and Fig. 7.

2.3. Size-aware Gaussian attenuation

When the local stream processes extracted patches, objects located near the boundaries of the local patch may be truncated. Standard non-maximum suppression (NMS) alone may be insufficient for handling this artifact, often leading to high-confidence false positives or duplicate partial detections. To reduce boundary artifacts while limiting confidence attenuation for small-scale targets, we introduce a SAGA module.

For a predicted local bounding box with center coordinates (c_x, c_y) inside a slice of width s_w and height s_h , its minimum distance to the patch edge is quantified as $d_{edge} = \min(c_x, s_w - c_x, c_y, s_h - c_y)$. The raw confidence S_{raw} is penalized by a dynamic weight $w \in (0, 1]$ to yield the decayed confidence $S_{decayed} = S_{raw} \times w$.

Crucially, the attenuation profile is not static; it is dynamically adjusted based on the object's mapped pixel area A_{px} . Smaller targets are less likely to be severely truncated near patch boundaries, whereas large targets are more susceptible to boundary truncation. Thus, we formulate an area-dependent linear interpolation for both the Gaussian standard deviation $\sigma(t)$ and the minimum attenuation-weight floor value $F(t)$:

$$t = \min\left(\frac{A_{px}}{A_{pivot}}, 1\right) \quad (1)$$

$$\sigma(t) = \sigma_{small} + (\sigma_{large} - \sigma_{small}) \times t \quad (2)$$

$$F(t) = F_{small} + (F_{large} - F_{small}) \times t \quad (3)$$

where A_{pivot} defines the area pivot used to continuously interpolate the attenuation parameters across the small-scale regime. The final spatial attenuation weight w is formulated as a truncated Gaussian function:

$$w = \begin{cases} 1, & \text{if } d_{edge} \geq M \\ \max\left(\exp\left[-\frac{1}{2}\left(\frac{M - d_{edge}}{\sigma(t)}\right)^2\right], F(t)\right), & \text{otherwise} \end{cases} \quad (4)$$

where M is the defined edge margin beyond which no attenuation is applied. Through empirical validation, these parameters are strictly defined in our architecture: the edge margin is set to $M = 64$ pixels, and the area pivot is $A_{pivot} = 1024 \text{ px}^2$. Consequently, the attenuation parameters vary continuously with object area: for smaller targets, $\sigma(t)$ and $F(t)$ approach 14 and 0.8, respectively, whereas for large objects they approach 25.6 and 0.25, enabling stronger suppression of truncated large-object predictions while applying a higher attenuation-weight floor to smaller targets. Fig. 4 summarizes the size-dependent boundary attenuation in SAGA and the subsequent scale-semantic filtering in SSCF.

2.4. Scale-semantic consistency filter

The prevalence of small background artifacts, such as sea foam and specular sun glint, often exhibits visual characteristics similar to small-scale debris targets. Conventional hard confidence thresholding may not reliably distinguish these artifacts from true targets, resulting in a high

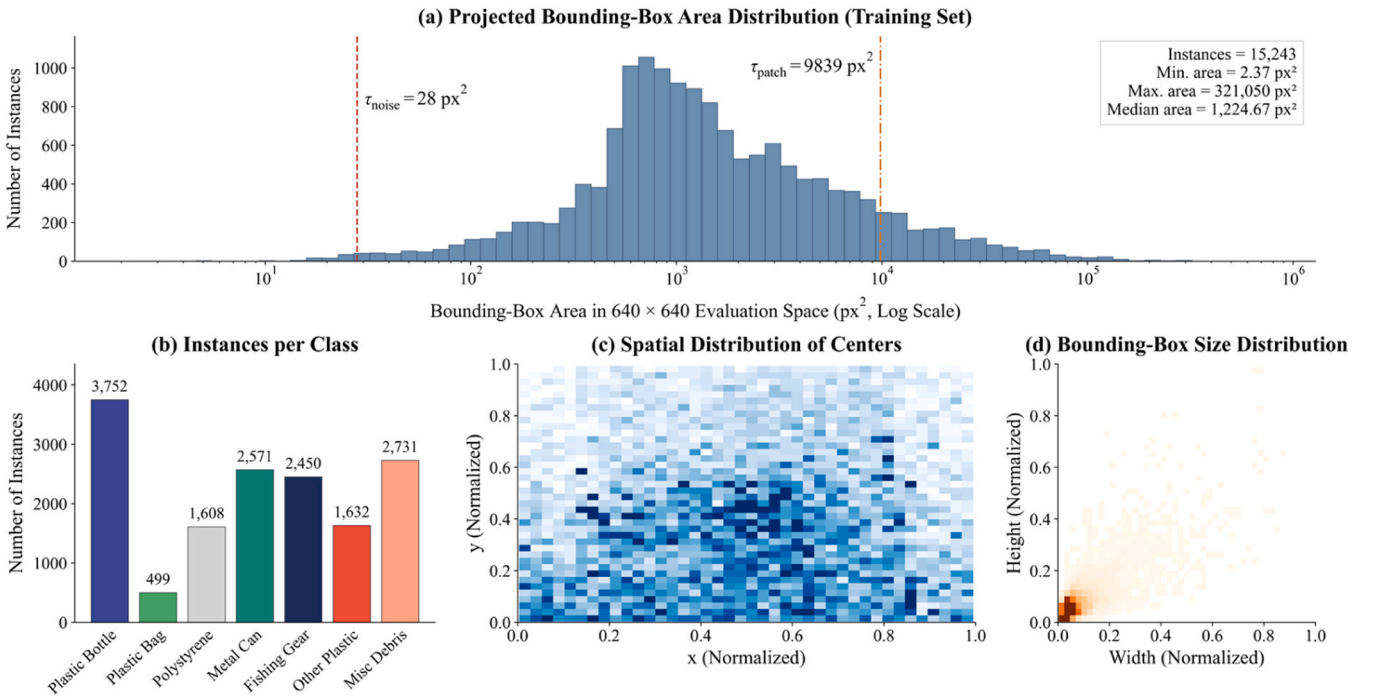


Fig. 2. Comprehensive statistical overview of the MFL-Dataset (Training Set). (a) showing the range of projected object sizes and the operating thresholds (τ_{noise} and τ_{patch}). (b) Instance count distribution across the seven annotated marine debris categories, demonstrating inherent class imbalance. (c) Normalized spatial distribution heatmap of bounding box center points (x, y) , indicating a nonuniform distribution of object centers across the training images. (d) Normalized bounding box size heatmap (width vs. height), showing that most bounding boxes are concentrated at small normalized widths and heights.

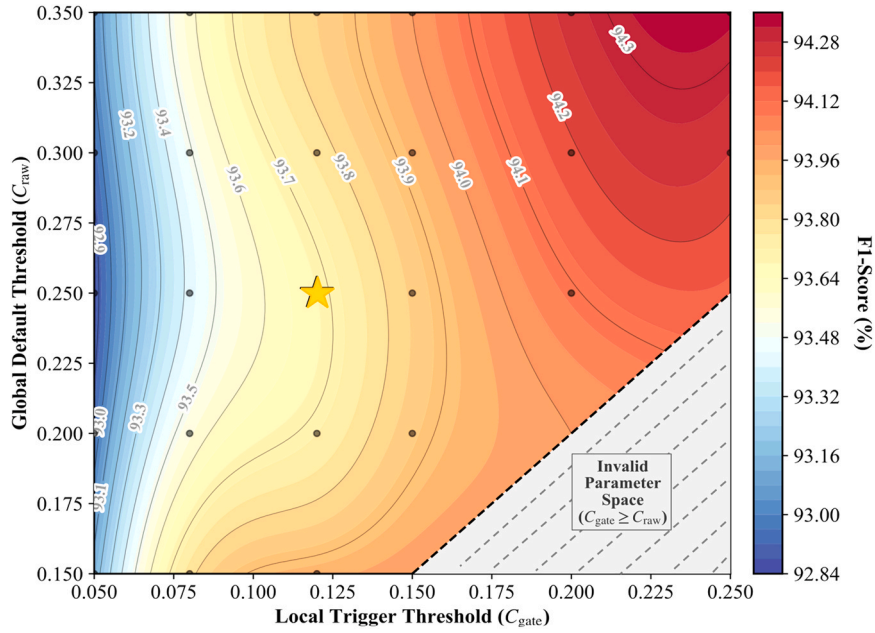


Fig. 3. Grid-search landscape for dual-stream confidence thresholds. The contour map visualizes the F1-score surface interpolated from valid parameter combinations. The grey hatched region indicates the invalid parameter space defined by the constraints ($C_{gate} \geq C_{raw}$). Black dots represent the discrete configurations evaluated during the grid search. The gold star marks the selected operating point ($C_{gate} = 0.12$, $C_{raw} = 0.25$) for the GL-DSI framework.

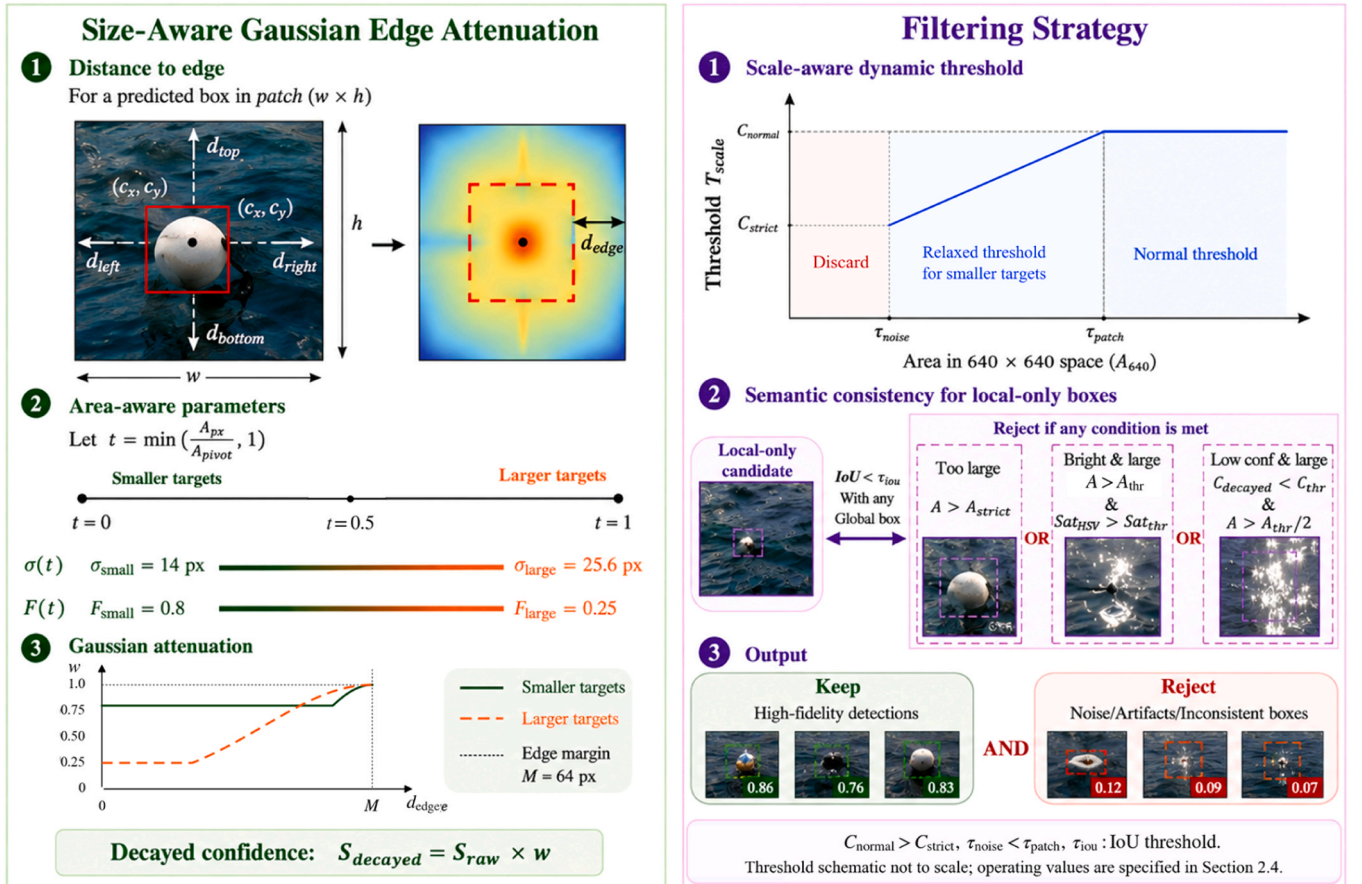


Fig. 4. Operational mechanisms of SAGA and SSCF. The left panel illustrates the size-aware Gaussian edge attenuation process, highlighting the dynamic penalty applied to boundary predictions based on the projected bounding-box area. The right panel details the scale-semantic consistency filter, which combines a continuous dynamic confidence boundary with semantic verification steps to effectively suppress environmentally induced pseudo-targets.

false-positive rate [22,23]. To combat this, GL-DSI incorporates the

SSCF as a unified, single-frame post-processing module.

Operating purely in the spatial domain to maintain edge efficiency, the first stage of this filter establishes an area-aware confidence lower bound instead of a step-function threshold. For a fused prediction mapped to the 640×640 normalized space with area A_{640} , the required confidence threshold T_{scale} linearly relaxes as the target becomes smaller, acknowledging the inherently lower confidence of smaller targets:

$$T_{scale}(A_{640}) = \begin{cases} C_{strict} + (C_{normal} - C_{strict}) \times \frac{A_{640}}{\tau_{patch}}, & A_{640} < \tau_{patch} \\ C_{normal}, & otherwise \end{cases} \quad (5)$$

where C_{strict} and C_{normal} represent the lower and upper confidence bounds, respectively. The parameter τ_{patch} denotes the area pivot governing the transition of the scale-aware confidence threshold, while τ_{noise} defines the lower area cutoff below which predicted boxes are rejected as likely environmental artifacts. In our implementation, $C_{strict} = 0.24$, $C_{normal} = 0.30$, $\tau_{patch} = 9839 \text{ px}^2$ and $\tau_{noise} = 28 \text{ px}^2$ were empirically determined using the validation data and fixed before final test-set evaluation. It is worth emphasizing that τ_{patch} serves as an area pivot for scale-aware confidence thresholding rather than an object-scale classification boundary.

Even with scale-aware thresholding, certain local stream predictions may falsely identify distinct background noise such as highly saturated colorful reflections as debris. The second stage of the filter explicitly quantifies semantic consistency to suppress these pseudo-targets. First, the algorithm isolates predictions exclusive to the local stream, defined as local boxes whose maximum intersection over union (IoU) with any global prediction is less than τ_{iou} . Such a box is semantically suppressed if it exhibits structurally implausible characteristics based on three empirical criteria. First, predictions with an extremely large area exceeding A_{strict} are directly discarded, as such objects should inherently be captured by the global stream rather than appearing exclusively in local patches. Second, to filter environmental artifacts such as specular reflections, boxes with a large area exceeding A_{thr} coupled with abnormally high mean saturation in the hue, saturation, and value (HSV) color space exceeding Sat_{thr} are rejected. Finally, boxes with decayed confidence below C_{thr} and an area larger than half of A_{thr} are also removed to reduce false positive. In our implementation, A_{strict} and A_{thr} were determined from statistics of target area, whereas Sat_{thr} and C_{thr} were empirically selected on the validation set. All thresholds were fixed before final test-set evaluation.

SSCF uses scalar operations and localized color-space conversions on the boxes remaining after initial filtering, typically fewer than 50 per frame, resulting in minimal computational overhead. While the scale-aware thresholding ($C_{strict} = 0.24$) is parameterized to mitigate false rejections, the filter operates under inherent technical trade-offs. True small-scale targets risk inadvertent suppression under three specific conditional overlaps: (1) Boundary Attenuation: A degraded micro-scale target located near slice margins ($d_{edge} < 64 \text{ px}$) with an initial confidence close to C_{strict} may be attenuated below the rejection cutoff by SAGA; (2) Optical Ambiguity: Highly specular plastic fragments exhibiting localized high HSV saturation might be classified as optical artifacts if omitted by the global detection stream; and (3) Noise Floor Truncation: Candidates with an apparent area below $T_{noise} = 28 \text{ px}^2$ are excluded by design to control false positives caused by ocean foam. Because genuine nano-scale targets may occupy the neighboring low-area regime, increasing this cutoff too aggressively can suppress genuine detections; the corresponding trade-off between false-positive suppression and nano-scale target recall and its observed degradation boundary are quantified in Section 3.4 and Fig. 7(c).

3. Experiments and results

3.1. Experimental configuration and evaluation metrics

The detector models were trained using PyTorch 2.8.0, CUDA 12.9, and Ultralytics 8.4.36 on a desktop platform equipped with a single NVIDIA RTX 5060 GPU with 8 GB of memory. The training utilized the Stochastic Gradient Descent optimizer with an initial learning rate of 0.01, momentum of 0.937, and weight decay of 0.0005. Each detector was trained using the standard loss formulation of its corresponding YOLO implementation. Because GL-DSI modifies the inference procedure rather than the detector training objective, each YOLO backbone uses its corresponding trained detector weights consistently across the evaluated inference strategies, without additional task-specific retraining.

To assess detection performance and operational feasibility, we employ a multi-dimensional metric system. For detection accuracy, we utilize precision (P), recall (R), F1-score, and mAP@50. Precision, recall, and F1-score are defined as follows:

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$R = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (8)$$

Here, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Because the primary operational objective is floating-litter detection rather than subtype classification, all seven debris categories are treated as a unified foreground class in the main detection benchmarks. Predictions are matched to ground-truth boxes using class-agnostic one-to-one matching at $\text{IoU} \geq 0.5$. Under this unified-foreground evaluation, mAP@50 corresponds to the area under the interpolated precision-recall curve at an IoU threshold of 0.5. For scale-stratified precision and recall, TP and FN are assigned by ground-truth area, whereas unmatched FP predictions are assigned by predicted-box area, with areas measured in the 640×640 evaluation space.

Crucially, GL-DSI adopts an Asymmetric Training-Inference Resolution Strategy. The detector was trained exclusively at 640×640 pixels without additional high-resolution patch fine-tuning. During inference, the global branch operates on downsampled views, whereas the local branch analyzes patches extracted from the original images. This asymmetry may introduce a potential feature distribution shift (domain shift) between the global and local representations. To quantify this representational discrepancy, we evaluated the Centered Kernel Alignment (CKA) similarity between the deep feature maps (15th layer) of the resized global images and their corresponding high-resolution patches across the test set. The average CKA similarity score was 0.0978, indicating limited representational similarity between the paired global and local features under the adopted layer-wise feature comparison protocol.

To further characterize scale-dependent detection behavior under asymmetric global-local inference, we stratified the locked YOLOv11s baseline and GL-DSI predictions using the mutually exclusive target-size intervals defined in Section 2.1. Under the asymmetric GL-DSI inference setting, nano-scale targets ($A < 64 \text{ px}^2$) showed increases in precision from 41.67% to 45.45% and recall from 13.51% to 27.03%; micro-scale targets ($64 \leq A < 256 \text{ px}^2$) increased from 73.03% to 77.82% in precision and from 63.31% to 74.03% in recall; and tiny-scale targets ($256 \leq A < 1024 \text{ px}^2$) increased from 86.74% to 92.88% in precision and from 95.17% to 95.89% in recall. For the remaining targets with $A \geq 1024 \text{ px}^2$, precision increased from 91.47% to 94.05%, whereas recall decreased slightly from 98.11% to 97.33% (− 0.78 %age points).

These results indicate that the performance change under asymmetric global–local inference is scale dependent rather than a uniform degradation, with the largest recall improvements concentrated in the nano- and micro-scale regimes. Because CKA quantifies representational similarity rather than causal influence, these scale-wise performance changes are interpreted as downstream detection behavior observed under the asymmetric inference setting, rather than as effects caused solely by the CKA discrepancy.

Furthermore, in a supplementary YOLOv8s patch fine-tuning experiment, directly fine-tuning the baseline on high-resolution patches reduced mAP@50 from 0.9583 to 0.8983 on the 758-image full-image test set evaluated at 640×640 pixels. These values were obtained using standard seven-class evaluation and are reported separately from the unified-foreground benchmarks. This experiment indicates that direct patch fine-tuning did not improve full-image detection performance under the tested setting.

In addition to the desktop evaluation, we conducted an embedded-hardware benchmark on an NVIDIA Jetson Orin Nano Engineering Reference Developer Kit Super as a representative edge-computing platform operating in MAXN_SUPER power mode. The deployment environment used TensorRT 10.3.0 with 16-bit floating-point (FP16) inference, an input size of 640×640 , and a batch size of 1. Both YOLOv8s- and YOLOv11s-based implementations were evaluated using the same trained weights as in the main experiments. Before timing, 50 warm-up frames were processed to stabilize the execution environment. CUDA synchronization was applied immediately before and after each timed inference to ensure that asynchronous GPU execution did not bias latency measurements. Each method was then evaluated over the complete 758-image test set for five consecutive cycles, resulting in 3790 timed frames per configuration.

Pipeline latency includes model preprocessing, detector inference, prediction post-processing, and the additional GL-DSI inference and filtering operations, while image loading from disk was measured separately and excluded from the reported inference latency. Board-level power consumption, system memory usage, device temperature, GPU utilization, and GPU operating frequency were sampled using NVIDIA tegrastats at 500-ms intervals. Average and peak board power were recorded, and per-frame inference energy was estimated from the measured average board power and mean pipeline latency. Peak process memory and system random-access memory usage were also monitored. Model floating-point operation counts for a single detector forward pass were obtained from the Ultralytics model analysis utility, whereas the

effective detector computation per frame for GL-DSI was estimated based on the average number of activated detector inference operations per image; therefore, this value should be interpreted as an estimate of computation for detector forward passes rather than the complete cost of all auxiliary post-processing operations. Thermal behavior was assessed from the recorded temperature and GPU-frequency traces over the repeated benchmark.

Alongside conventional detection metrics, we report three nested small-scale target recall metrics in Table 1, based on bounding-box pixel areas in the 640×640 evaluation coordinate space: nested tiny-scale target recall for $A < 1024 \text{ px}^2$, nested micro-scale target recall for $A < 256 \text{ px}^2$, and nested nano-scale target recall for $A < 64 \text{ px}^2$. These nested metrics should be distinguished from the mutually exclusive size intervals used for the detailed scale-stratified analysis in Fig. 8(b). These task-specific metrics provide a fine-grained assessment of difficult small-scale target recovery and should not be interpreted as official COCO size-specific metrics.

To ensure a fair and rigorous cross-paradigm comparison, the configurations of the competing inference strategies were fixed before final test-set evaluation. For high-resolution inference (1280×1280), standard letterbox padding was applied without slicing to maintain aspect ratios during a single forward pass. Multiscale Test Time Augmentation (TTA) performed full-image inference at input sizes of 480, 640, 800, and 1024 pixels without horizontal flipping. Predictions from all input scales were mapped to the original image coordinates and fused using NMS. Crucially, for the SAHI paradigm (implemented via the SAHI Python library, version 0.11.36), the framework was configured with a slice window of 320×320 pixels, an overlap ratio of 0.25 for both height and width, and a full-image forward pass. The confidence and NMS matching thresholds were selected using the validation set and subsequently fixed for test-set evaluation. The resulting operating point used a confidence threshold of 0.25 and an NMS intersection threshold of 0.30.

To compare foreground- and density-guided adaptive slicing strategies, we implemented two controlled baselines: Feature-Energy Adaptive Slicing (FE-AS), inspired by Efficient Small Object Detection (ESOD) [17] and Density-Guided Quadtree Slicing (DG-QS). To examine whether the comparative behavior depends on the detector backbone, both FE-AS and DG-QS were evaluated with YOLOv8s and YOLOv11s using the corresponding trained detector weights and the same 758-image test split. The method-specific slicing parameters were kept fixed across the two backbones to avoid backbone-specific retuning on

Table 1
Cross-paradigm performance comparison of inference strategies on the MFL-Dataset.

Method	Prec%	Rec%	F1	mAP@50	TinyR%	MicroR%	NanoR%	FPS	Lat(ms)
YOLOv5s (640)	86.0	93.6	89.7	92.7	87.5	58.0	5.4	135.4	7.4
YOLOv8s (640)	88.1	94.1	91.0	93.3	88.2	57.7	10.8	138.7	7.2
YOLOv11s (640)	88.6	93.8	91.1	93.2	87.8	58.0	13.5	128.8	7.8
YOLOv8s (1280)	88.9	98.1	93.2	97.6	97.2	91.9	73.0	58.5	17.1
YOLOv11s (1280)	88.5	97.6	92.8	97.0	96.6	89.3	59.5	54.9	18.2
YOLOv8s + SAHI	41.5	90.5	56.9	74.5	94.2	85.1	62.2	3.3	305.6
YOLOv11s + SAHI	43.6	90.6	58.9	79.8	94.7	87.0	64.9	2.9	350.3
YOLOv8s + Multiscale TTA	89.4	97.6	93.3	97.1	96.0	86.7	56.8	29.3	34.1
YOLOv11s + Multiscale TTA	89.5	97.3	93.2	96.8	95.8	84.6	46.0	27.1	36.9
YOLOv8s + FE-AS	78.5	96.4	86.5	93.0	94.2	81.2	46.0	21.9	45.6
YOLOv11s + FE-AS	78.3	96.6	86.5	93.8	94.3	82.6	46.0	20.1	49.8
YOLOv8s + DG-QS	73.5	94.9	82.9	87.6	92.5	77.4	51.4	31.4	31.9
YOLOv11s + DG-QS	73.4	95.0	82.8	88.3	92.4	78.3	43.2	29.3	34.2
YOLOv8s + GL-DSI	92.0	94.8	93.3	93.7	90.8	69.6	29.7	49.6	20.2
YOLOv11s + GL-DSI	92.3	94.5	93.4	93.8	90.5	69.0	27.0	44.6	22.4

Note: Object areas are measured after mapping the ground-truth boxes to the 640×640 evaluation coordinate space. Nano (Area $< 64 \text{ px}^2$), Micro (Area $< 256 \text{ px}^2$), and Tiny (Area $< 1024 \text{ px}^2$) are nested fine-grained categories, where Nano is a subset of Micro and Micro is a subset of Tiny. These categories are derived from, but are not identical to, the standard COCO Small regime (Area $< 1024 \text{ px}^2$). SAHI results are reported using confidence and NMS settings selected on the validation set and fixed before test-set evaluation. FE-AS denotes our controlled implementation of an ESOD-inspired feature-energy-guided adaptive slicing strategy rather than an official code reproduction of ESOD. DG-QS denotes our implementation of a density-guided quadtree slicing baseline. For each backbone, FE-AS and DG-QS use the corresponding trained detector weights, the same 758-image test split, and the same evaluation protocol as the corresponding standard and GL-DSI configurations. The method-specific slicing parameters were kept fixed across YOLOv8s and YOLOv11s.

the test set. FE-AS isolates regions above the 85th percentile of the feature-energy response distribution computed from the selected intermediate feature maps to trigger local patch extraction. DG-QS recursively subdivides the image via a quadtree strategy driven by candidate-density estimates under a predefined threshold. For both baselines, local patches were processed in constrained batches, coordinates were explicitly mapped, and global-local predictions were fused using standard NMS (IoU = 0.35).

3.2. Comparative evaluation against baseline strategies

To comprehensively evaluate the performance characteristics of the proposed GL-DSI framework, we benchmarked it against prevalent advanced inference paradigms, including brute-force High-Resolution inference (1280×1280), SAHI [19], and multiscale TTA, and contemporary adaptive slicing strategies (FE-AS and DG-QS). The benchmark included YOLOv5s, YOLOv8s, and YOLOv11s standard detectors, while GL-DSI was evaluated with YOLOv8s and YOLOv11s to examine its applicability across different detector generations. The comprehensive performance metrics are summarized in Table 1.

This cross-paradigm comparison evaluates accuracy–computation trade-offs across standard inference, resolution enhancement, and slicing strategies rather than performance at a common input resolution. It assesses the benefits and costs of selective local processing without implying a speed advantage over standard low-resolution inference.

While standard inference at 640×640 achieves excellent operational speeds (>128 FPS), it demonstrates a significant degradation in resolving small-scale targets, with nano-scale target recall ranging from 5.4% to 13.5%. Conventional strategies attempt to solve this by increasing computational burden. For instance, scaling the input resolution to 1280×1280 or employing multiscale TTA improves recall but significantly penalizes the frame rate. More critically, the prevalent SAHI strategy exhibits a pronounced precision–latency trade-off in the tested marine scenes. Even with its confidence and NMS thresholds selected before final test-set evaluation, its exhaustive blind slicing mechanism necessarily processes artifact-rich regions containing sun glint and sea foam at high resolutions. Under the YOLOv11s configuration, SAHI yielded an F1-score of 58.9% and a precision of 43.6%, substantially below the corresponding standard detector and GL-DSI configurations. Its computational cost was also markedly higher, with an inference speed of 2.9 FPS (350.3 ms latency) on the RTX 5060 desktop platform. These results indicate that exhaustive slicing incurs substantial computational and false-positive costs in the present marine-scene benchmark, rather than establishing its general unsuitability for all edge platforms.

Furthermore, our evaluation reveals that the tested feature-energy and density-guided adaptive strategies (FE-AS and DG-QS) also exhibit limitations in simultaneously maintaining high precision and micro-scale target sensitivity in complex marine backgrounds. While methods such as ESOD-inspired feature-energy slicing reuse detector features to reduce unnecessary background processing, activation intensity alone may respond strongly to visually salient environmental structures such as waves and reflections. Consequently, as shown in Table 1, while FE-AS and DG-QS mitigate the extreme latency of SAHI, they still exhibit elevated false-positive rates. This behavior is consistent across the YOLOv8s and YOLOv11s backbones. For FE-AS, precision/F1 change only from 78.5%/86.5% with YOLOv8s to 78.3%/86.5% with YOLOv11s; for DG-QS, the corresponding values change from 73.5%/82.9–73.4%/82.8%. Although both adaptive baselines achieve higher small-scale target recall than GL-DSI, this gain is accompanied by substantially lower precision and F1-score. GL-DSI instead uses low-confidence detector outputs as semantic cues for triggering local inference, thereby basing patch selection on detector uncertainty rather than feature intensity alone.

Within this desktop comparison, the proposed GL-DSI framework achieves a favorable engineering balance between precision, recall, and

computational cost. The controlled YOLOv11s comparison further shows that this trade-off advantage is not attributable solely to the choice of backbone. When integrated with YOLOv11s, GL-DSI achieves the highest precision (92.3%) and F1-score (93.4%) among the evaluated configurations, while high-resolution inference and TTA retain advantages in recall-oriented evaluation settings. By replacing exhaustive cropping with the SGDS mechanism and post-inference SSCF filtering, GL-DSI increases the nano-scale target recall from 13.5% under standard YOLOv11s inference to 27.0%, while maintaining an inference speed of 44.6 FPS on the RTX 5060 desktop platform. This precision-preserving enhancement is explicitly visualized in the scale-stratified delta matrix (Fig. 5).

Fig. 6 illustrates the spatial difference between uniform slicing and semantic-guided local processing, together with the corresponding detection outputs in two example scenes. Across the test set, semantic-guided triggering reduced the number of processed local patches by 73.5% relative to exhaustive slicing.

3.3. Ablation study

To systematically assess the individual contributions of the proposed components, a progressive ablation study was conducted. YOLOv11s was selected as the representative baseline for this ablation study and subsequent visual analysis because it exhibited strong baseline detection performance in the comparative evaluation, providing a consistent reference for isolating the contribution of each component. As summarized in Table 2, five configurations were sequentially evaluated to examine their effects on micro-scale target recall and background noise suppression. All configurations were evaluated on the same 758-image test set using the same YOLOv11s weights and matching protocol. Configs A and E reproduced the per-image detection counts and matched ground-truth identities of the locked baseline and GL-DSI results underlying Table 1. Config A represents the baseline detector using only the global stream. Config B adds uniform blind slicing. Its full-coverage grid uses half-image-width and half-image-height patches with 45% overlap and differs from the SAHI configuration in Table 1. Config C replaces blind slicing with the proposed SGDS mechanism. Config D introduces the SAGA module. Config E incorporates the SSCF. The local-output confidence threshold was fixed at 0.40 across Configs B–E, with thresholding applied after confidence attenuation in Configs D and E.

The transition from Config A to B shows that uniform local slicing substantially increases recall across small-scale targets. Specifically, the nano-scale target recall improves from 13.51% to 48.65%. However, this recall gain is accompanied by a marked reduction in precision from 88.59% to 76.34%. Compared with Config B, Config C restores precision to 90.24% and increases the overall F1-score from 85.39% to 92.61%, while retaining a nano-scale target recall of 29.73%. These results indicate a trade-off between small-target recovery and false-positive suppression.

Following the introduction of SAGA in Config D, precision increases from 90.24% to 91.64%, while the F1-score increases from 92.61% to 93.10%, although nano-scale target recall decreases from 29.73% to 27.03%. Adding SSCF in Config E further suppresses residual environmental pseudo-targets, increasing precision from 91.64% to 92.30% and F1-score from 93.10% to 93.39%. False positives decrease from 360 to 329, while true positives decrease from 3946 to 3942, reducing overall recall from 94.61% to 94.51%. Nested tiny-, micro-, and nano-scale recall remain unchanged at 90.54%, 68.99%, and 27.03%, respectively.

3.4. Sensitivity analysis

After fixing the operating thresholds on the validation set, we conducted a post-selection sensitivity analysis of the local-stream triggering threshold, scale-aware confidence lower bound, and noise-area cutoff on all 758 test images. These sweeps characterize sensitivity and degradation boundaries without reselecting the operating thresholds.

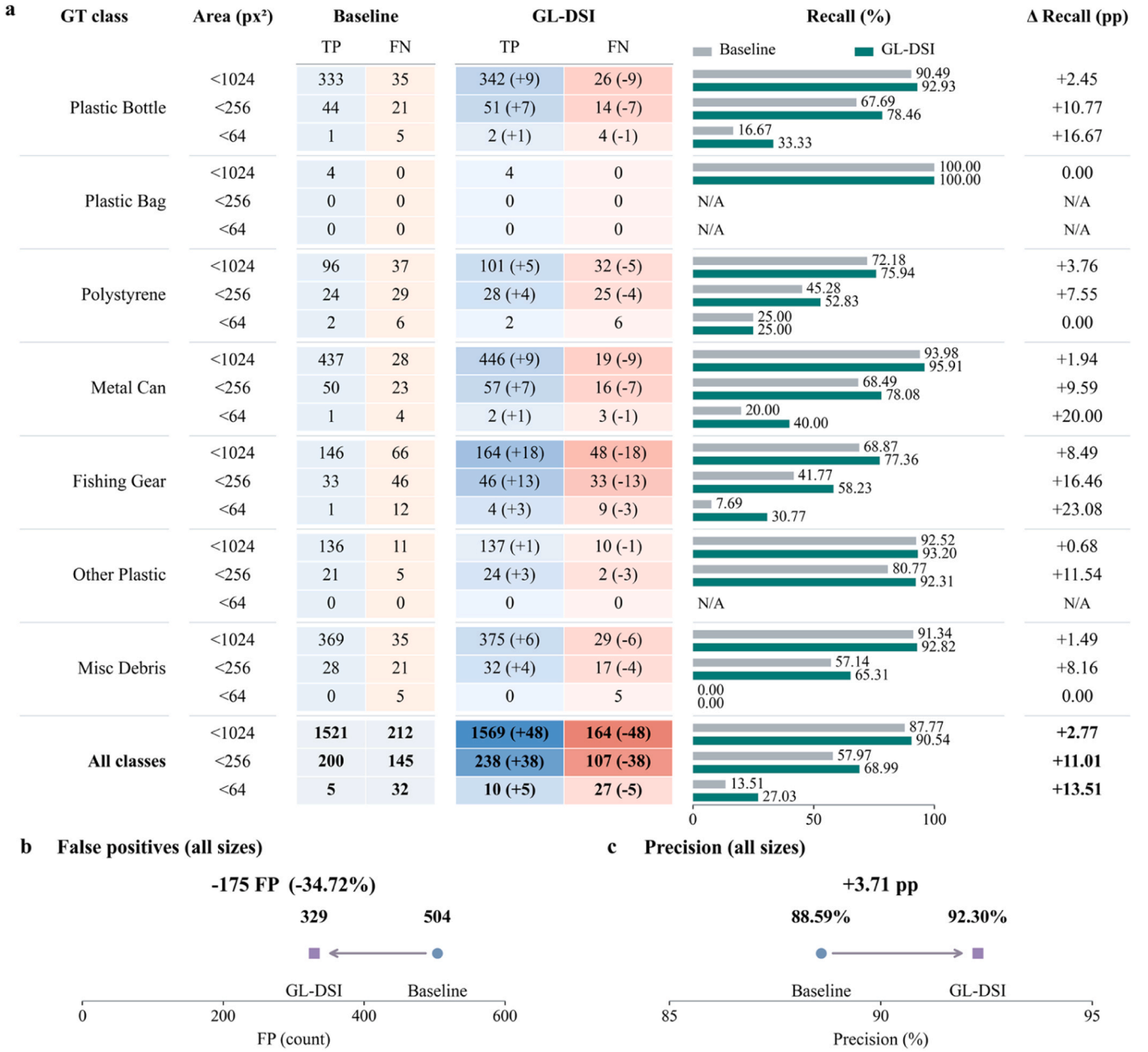


Fig. 5. Scale-stratified target recovery and overall detection changes for YOLOv11s. (a) TP, FN, and recall by debris class under the nested thresholds of $A < 1024$, $A < 256$, and $A < 64$ px² in the 640×640 evaluation space. Parentheses in the GL-DSI matrix indicate signed count changes relative to the baseline, while color intensity reflects the magnitude of TP/FN changes. Δ Recall denotes the net recall difference in percentage points (pp); N/A indicates no ground-truth instances. All statistics are derived from the same locked predictions using class-agnostic one-to-one matching at IoU ≥ 0.50 . (b) Overall false positives decrease from 504 to 329. (c) Overall precision increases from 88.59% to 92.30%.

The local stream triggering threshold determines the fraction of uncertain global detections routed to the local branch. As shown in Fig. 7a, decreasing C_{gate} increases small-scale target recovery at the cost of additional local inference, whereas increasing C_{gate} improves computational efficiency but progressively reduces small-scale target recall. At $C_{gate} = 0.02$, small-scale target recall ($A < 1024$ px²) reaches 92.09% at 23.50 FPS, whereas at $C_{gate} = 0.30$, the inference speed increases to 54.95 FPS while small-scale target recall decreases to 87.36%. At the operating point selected on the validation set ($C_{gate} = 0.12$), GL-DSI achieves an F1-score of 93.39%, a small-scale target recall of 90.54%, and an inference speed of 42.02 FPS in the sensitivity analysis on the complete test set. These results demonstrate a clear recall–efficiency trade-off across the complete test population.

The nano-scale response reveals a more distinct degradation

boundary. Nano-scale target recall remains at 27.03% over $C_{gate} = 0.12 \sim 0.16$, but decreases to 21.62% at $C_{gate} = 0.18$. It continues to decline with increasing gate values and reaches 13.51% at $C_{gate} = 0.26$, equal to the standard YOLOv11s baseline nano-scale target recall. Thus, within the evaluated grid, the onset of noticeable nano-scale degradation is observed above approximately $C_{gate} = 0.16$, while local refinement is disabled for $C_{gate} \geq C_{raw} = 0.25$. Fusion NMS and SSCF remain active in this regime. These values are interpreted as diagnostic degradation boundaries rather than parameter-selection criteria.

Similarly, the scale-aware confidence lower bound controls the trade-off between false-positive suppression and target retention within the post-processing filter. As shown in Fig. 7b, increasing C_{strict} progressively improves overall precision while reducing overall recall. At the validation-selected value of $C_{strict} = 0.24$, the framework achieves an

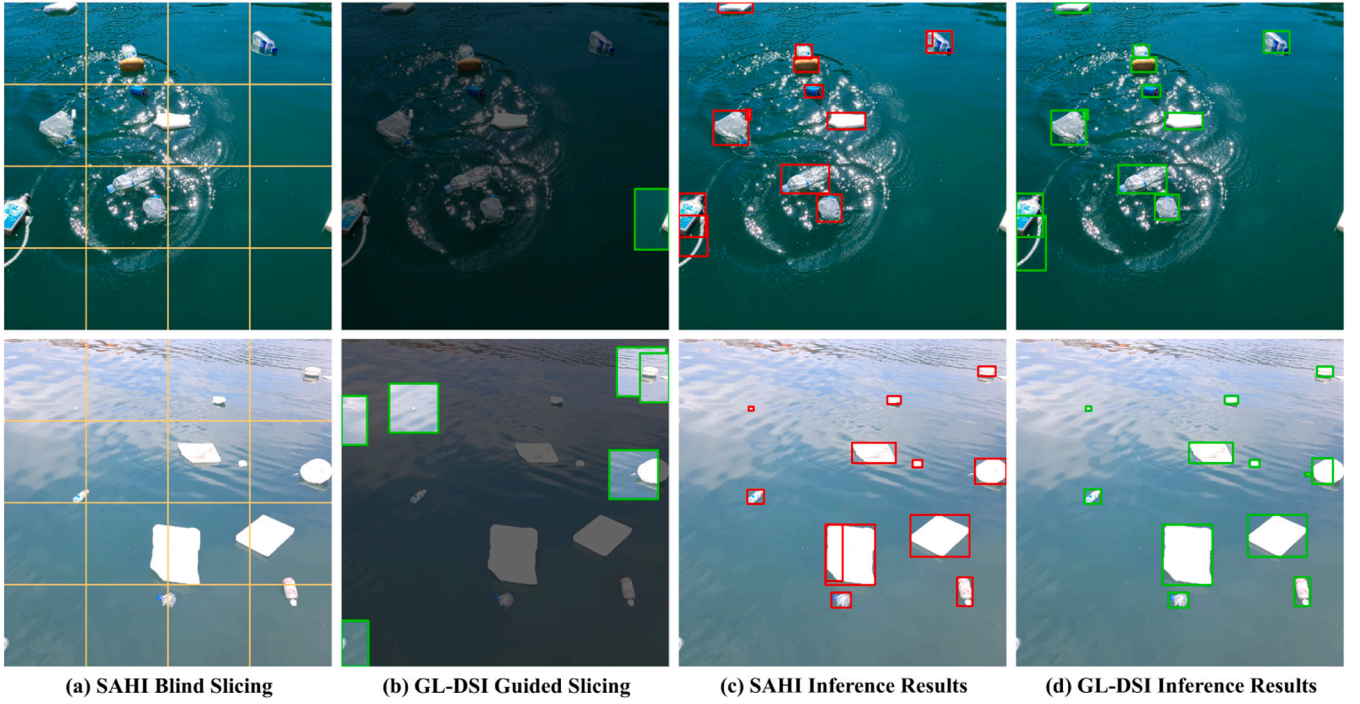


Fig. 6. Qualitative comparison of slicing layouts and detection outputs. Two example scenes are shown. (a) Uniform slicing layout used by SAHI. (b) Regions selected for local processing by GL-DSI. (c) SAHI detection outputs. (d) GL-DSI detection outputs. Red and green boxes distinguish the predictions of the two methods and do not indicate ground-truth correctness.

Table 2

Ablation study of the proposed GL-DSI framework.

Config	Global	Blind Slicing	SGDS	SAGA	SSCF	Prec%	Rec%	F1	TinyR%	MicroR%	NanoR%
A	✓	×	×	×	×	88.59	93.81	91.13	87.77	57.97	13.51
B	✓	✓	×	×	×	76.34	96.86	85.39	95.79	86.67	48.65
C	✓	×	✓	×	×	90.24	95.11	92.61	91.81	73.91	29.73
D	✓	×	✓	✓	×	91.64	94.61	93.10	90.54	68.99	27.03
E	✓	×	✓	✓	✓	92.30	94.51	93.39	90.54	68.99	27.03

Note: Tiny, Micro, and Nano recall use the nested area thresholds $A < 1024$, $A < 256$, and $A < 64 \text{ px}^2$, respectively.

overall precision of 92.30%, recall of 94.51%, and F1-score of 93.39%. Although the precision and recall curves approach and cross at a higher threshold, this intersection was not used as an operating-point criterion. Increasing C_{strict} toward this region reduces small-scale sensitivity; for example, at $C_{strict} = 0.32$, the small-scale target recall decreases to 88.81% and the nested micro-scale target recall decreases to 63.48%.

The noise-area cutoff τ_{noise} exhibits a distinct nano-scale target degradation pattern (Fig. 7c). Around the validation-selected operating point of 28 px^2 , nano-scale target recall remains stable at 27.03% (10/37) throughout the evaluated range from 0 to 32 px^2 . The first genuine nano-scale true positive is lost at $\tau_{noise} = 36 \text{ px}^2$, reducing nano-scale target recall to 24.32% (9/37). Increasing the cutoff to 48 px^2 reduces nano-scale target recall to 21.62% (8/37), corresponding to the loss of two genuine nano-scale target detections relative to the selected operating point. At 64 px^2 , only 3 of 37 nano-scale targets remain detected (8.11%), which is below the YOLOv11s baseline nano-scale target recall of 13.51%; at 80 px^2 , nano-scale target recall falls to 0%. In contrast, increasing τ_{noise} from 28 to 64 px^2 improves overall precision only from 92.30% to 92.54% and reduces false positives from 329 to 317. These results indicate that the onset of nano-scale target degradation is first observed at 36 px^2 within the evaluated grid, while aggressive increases in τ_{noise} provide only limited false-positive suppression at the cost of substantial genuine nano-scale target loss.

3.5. Analysis of small-scale target recovery

Standard metrics do not directly characterize changes in the apparent spatial coverage of fragmented debris detections. By applying the GL-DSI framework, the recall for micro-scale targets occupying $64\text{--}256 \text{ px}^2$ in the evaluation coordinate space improved from 63.3% to 74.0% (Fig. 8b). This interval follows the micro-scale definition introduced in Section 2.1 and is used here for interval-specific recovery analysis; it should be distinguished from the nested micro-scale target recall metric (area $< 256 \text{ px}^2$) defined in Section 3.1. Table 3 reports the GT, TP, FN, and recall values for both the mutually exclusive and nested size groups under the same matching protocol. GT denotes the number of ground-truth instances. To quantitatively characterize the relative change in detected bounding-box coverage between different models, we introduce the Relative Bounding Box Coverage Ratio (R_{cov}). Given a test image, R_{cov} is calculated as follows:

$$R_{cov} = \frac{\sum_{i \in \mathcal{D}} \text{Area}(b_i)}{\sum_{j \in \mathcal{D}_{base}} \text{Area}(b_j) + \epsilon} \quad (9)$$

where $\mathcal{D}$ and $\mathcal{D}_{base}$ denote the sets of bounding boxes predicted by GL-DSI and the baseline model, respectively. $\text{Area}(b) = w_b \times h_b$ represents the 2D pixel area of bounding box b , and $\epsilon = 10^{-6}$ is a small numerical stabilization constant. For the distributional and correlation

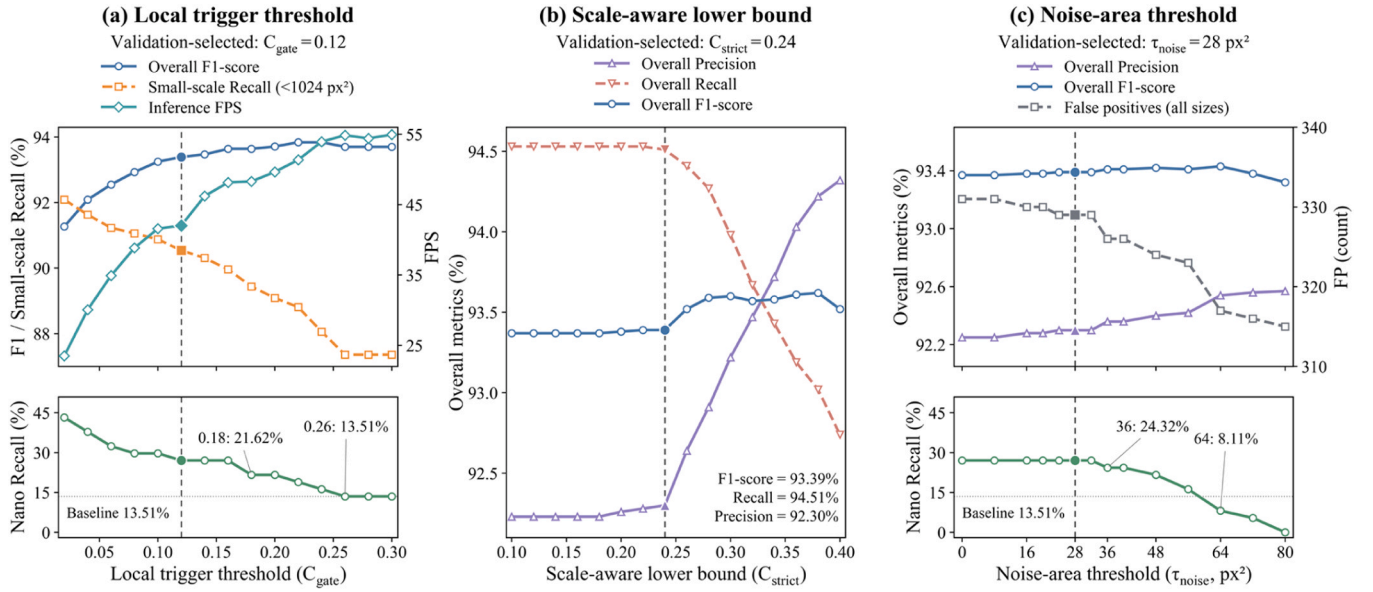


Fig. 7. Post-selection sensitivity analysis of key GL-DSI thresholds on the complete 758-image test set. (a) Sensitivity to the local-stream triggering threshold C_{gate} , showing the overall F1-score, nested small-scale target recall ($A < 1024 \text{ px}^2$), inference speed, and nano-scale target recall. (b) Sensitivity to the scale-aware lower confidence bound C_{strict} , showing overall precision, recall, and F1-score. (c) Sensitivity to the noise-area cutoff τ_{noise} , showing overall precision, F1-score, false-positive count, and nano-scale target recall. The dashed vertical lines indicate the operating thresholds selected using the validation set and fixed before test-set evaluation. The horizontal dotted lines in panels (a) and (c) indicate the YOLOv11s baseline nano-scale target recall of 13.51%. These sweeps were used only to characterize sensitivity and degradation boundaries. In panel (a), $C_{gate} \geq 0.25$ yields an empty triggering interval, so no local patches are processed.

analyses reported below, the single image with zero baseline predicted-box area was excluded to avoid an undefined or artificially extreme R_{cov} value.

Across the 757 valid images used for this analysis, the mean and median R_{cov} values are 0.971 and 1.000, respectively, with the first and third quartiles of 0.976 and 1.000 (interquartile range = 0.024; Fig. 8d). The central 50% of the observations is therefore concentrated close to unity. To examine whether changes in predicted bounding-box coverage are associated with detection recovery, we further performed a per-image Spearman correlation analysis between R_{cov} and the change in recall. For all targets, a weak positive association was observed ($\rho = 0.186$, $p < 0.001$, $n = 757$). For small-scale targets ($A < 1024 \text{ px}^2$), however, the association was negligible and not statistically significant ($\rho = 0.054$, $p = 0.242$, $n = 477$). These results indicate that R_{cov} measures relative changes in predicted bounding-box coverage and should not be interpreted as a surrogate for recall improvement.

The multi-dimensional analysis in Fig. 8 further characterizes this detection-level change. These pixel sums represent the spatial extents of bounding boxes, including water background, rather than litter segmentation areas. Without camera geometric calibration, temporal tracking, and complementary field measurements, they support relative cross-model comparison but cannot be directly converted into physical litter area, volume, or mass. The area decomposition (Fig. 8c) shows 212.3 million px^2 of baseline predicted bounding-box area. After GL-DSI processing, the pre-filter candidate area is 208.3 million px^2 , consisting of 204.7 million px^2 retained detections and 3.6 million px^2 removed candidate area after SSCF filtering; this quantity therefore describes the behavior of the filtering procedure rather than a direct measurement of physical environmental noise.

To further examine the fate of challenging debris, we constrained the small-scale target distribution (Fig. 9) to instances with areas $< 1024 \text{ px}^2$. The 1733 instances comprise 37 nano-scale targets (2.14%; $A < 64 \text{ px}^2$), 308 micro-scale targets (17.77%; $64 \leq A < 256 \text{ px}^2$), and 1388 tiny-scale targets (80.09%; $256 \leq A < 1024 \text{ px}^2$). Despite the numerical dominance of the tiny-scale group, the recall gains were concentrated primarily in the smaller-scale regimes: nano-scale target recall increased from 13.51% to 27.03% (+13.51 %age points), micro-scale target recall

from 63.31% to 74.03% (+10.71 %age points), whereas tiny-scale target recall increased only from 95.17% to 95.89% (+0.72 %age points). The corresponding net TP gains were +5, +33, and +10, respectively. Out of the 1733 small-scale target instances, the baseline missed 212 (12.23%). Among these missed instances, GL-DSI recovered 55, while 157 remained missed by both methods. Because seven instances detected by the baseline were missed by GL-DSI, the net true-positive gain was 48, increasing the nested tiny-scale target recall ($A < 1024 \text{ px}^2$) from 87.77% to 90.54% (+2.77 %age points), as summarized in Fig. 9.

The net improvement was also distributed across multiple debris classes rather than being attributable to a single target category: Fishing Gear contributed +18 net TPs, Plastic Bottle +9, Metal Can +9, Misc Debris +6, Polystyrene +5, and Other Plastic +1. No individual category accounted for more than half of the total net TP improvement. These results provide additional evidence that semantic-guided local refinement can improve the recovery of difficult small-scale targets, while the associated embedded-device computational cost is evaluated separately in Section 3.6.

3.6. Edge deployment evaluation on NVIDIA Jetson Orin Nano

To assess the deployment feasibility of GL-DSI under resource-constrained edge computing scenarios, we conducted a hardware evaluation on an NVIDIA Jetson Orin Nano. The benchmark followed the TensorRT FP16 protocol described in Section 3.1, with 3790 timed frames per configuration.

The edge benchmark results demonstrate the computational characteristics of the proposed dual-stream strategy. Compared with the baseline detector, GL-DSI introduces additional local refinement operations triggered by semantic uncertainty regions. Consequently, the average latency increased from 46.83 ms to 130.37 ms, while the throughput decreased from 21.35 FPS to 7.67 FPS. This increase reflects the additional high-resolution local inference required for recovering small-scale fragmented debris targets rather than an optimization toward maximum inference speed.

Although the additional refinement stage increases computational demand, the memory footprint remained comparable with the baseline,

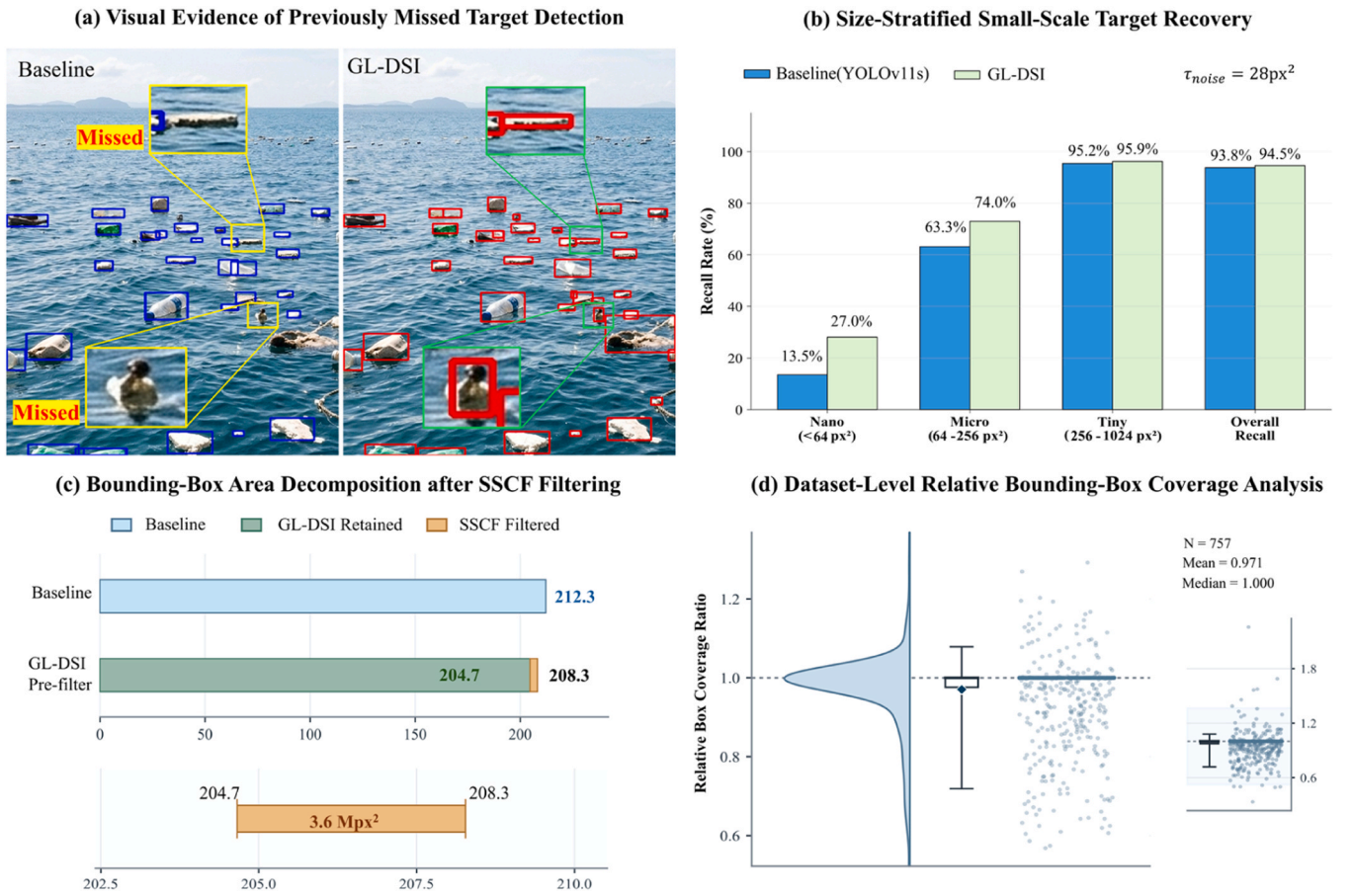


Fig. 8. Comprehensive analysis of small-scale target recovery and relative bounding-box coverage. **(a)** Visual comparison showing small-scale targets missed by the baseline and recovered by GL-DSI. **(b)** Scale-stratified recall of the baseline (YOLOv11s) and GL-DSI across nano-, micro-, and tiny-scale target bins; the SSCF noise cutoff $\tau_{noise} = 28px^2$ is used as a filtering reference and does not define the target-scale boundaries. **(c)** Decomposition of the predicted bounding-box coverage, including baseline detections, GL-DSI retained detections, and candidate areas removed by SSCF filtering. **(d)** Distribution of R_{cov} across the test set (757 valid images after excluding one image for which the total area of bounding boxes predicted by the baseline was zero); the mean, median, Q1, and Q3 values are 0.971, 1.000, 0.976, and 1.000, respectively.

Table 3

Scale-stratified numerical detection results for the YOLOv11s baseline and GL-DSI.

Group type	Size range	GT	Baseline			GL-DSI		
			TP	FN	Recall (%)	TP	FN	Recall (%)
Exclusive	A < 64	37	5	32	13.51	10	27	27.03
	64 ≤ A < 256	308	195	113	63.31	228	80	74.03
	256 ≤ A < 1024	1388	1321	67	95.17	1331	57	95.89
Nested	A < 64	37	5	32	13.51	10	27	27.03
	A < 256	345	200	145	57.97	238	107	68.99
	A < 1024	1733	1521	212	87.77	1569	164	90.54

Note: All counts were computed from the same locked prediction file using class-agnostic one-to-one matching at $IoU \geq 0.50$ in the 640×640 evaluation coordinate space.

with peak memory consumption of approximately 110 MB. The average power consumption increased moderately from 8.25 W to 9.18 W, corresponding to an increase in energy consumption per frame from 0.386 J to 1.197 J. During the repeated benchmark process, no obvious thermal throttling was detected, indicating stable execution under the tested hardware condition.

When combined with the accuracy validation results, these measurements demonstrate that GL-DSI provides a practical trade-off between detection capability and computational cost. Rather than replacing lightweight detectors for maximum throughput, GL-DSI is designed for scenarios where recovering small-scale fragmented debris is prioritized under edge-computing constraints. Under TensorRT FP16

inference on Jetson, YOLOv11s + GL-DSI achieved an F1-score of 93.56% under the complete GL-DSI configuration, and increased nano-scale target recall from 13.51% to 27.03% relative to the deployed baseline. These results indicate that deployment preserved the expected detection performance.

4. Discussion

4.1. Interpretability and cross-architectural applicability

To better understand the internal feature representation of our framework and examine its applicability across different detector

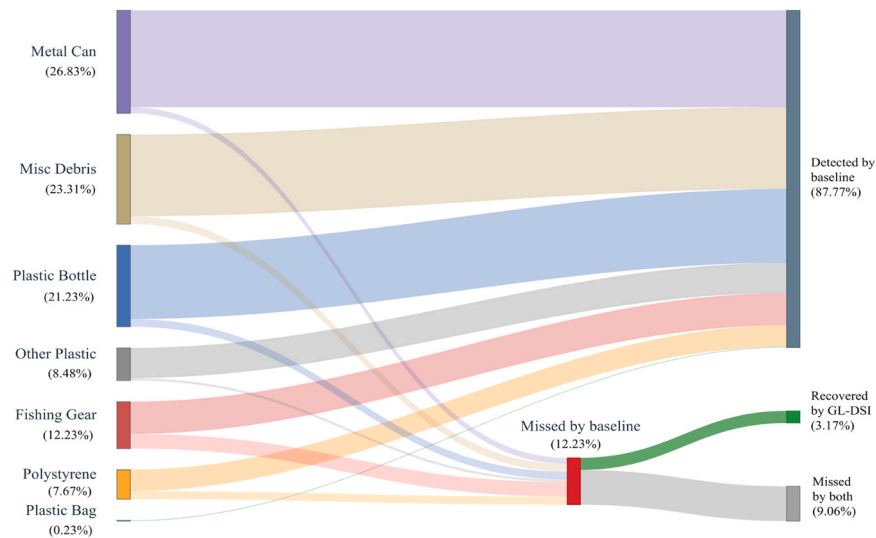


Fig. 9. Small-scale target distribution and detection outcomes. The Sankey diagram traces the detection fate of 1733 small-scale instances ($A < 1024 \text{ px}^2$) in the 758-image test set using the locked prediction results. Scale-specific counts and recall values are reported in Table 3. The baseline detects 1521 instances (87.77%) and misses 212 (12.23%). Among the instances missed by the baseline, GL-DSI recovers 55 (3.17% of all small-scale targets), while 157 (9.06%) remain missed by both methods. Seven instances detected by the baseline are missed by GL-DSI and are not shown in the Sankey diagram. Accounting for these losses, the nested tiny-scale target recall ($A < 1024 \text{ px}^2$) increases from 87.77% to 90.54%, corresponding to a net gain of 2.77 %age points.

generations, we visualized the multiscale Class Activation Maps for both YOLOv8s and YOLOv11s. In the examples shown in Fig. 10, standard inference produces relatively weak and diffuse activation around small-

scale targets, including plastic bottles and fragments against the marine background. With selective high-resolution local refinement, GL-DSI produces more concentrated activation around difficult small-scale

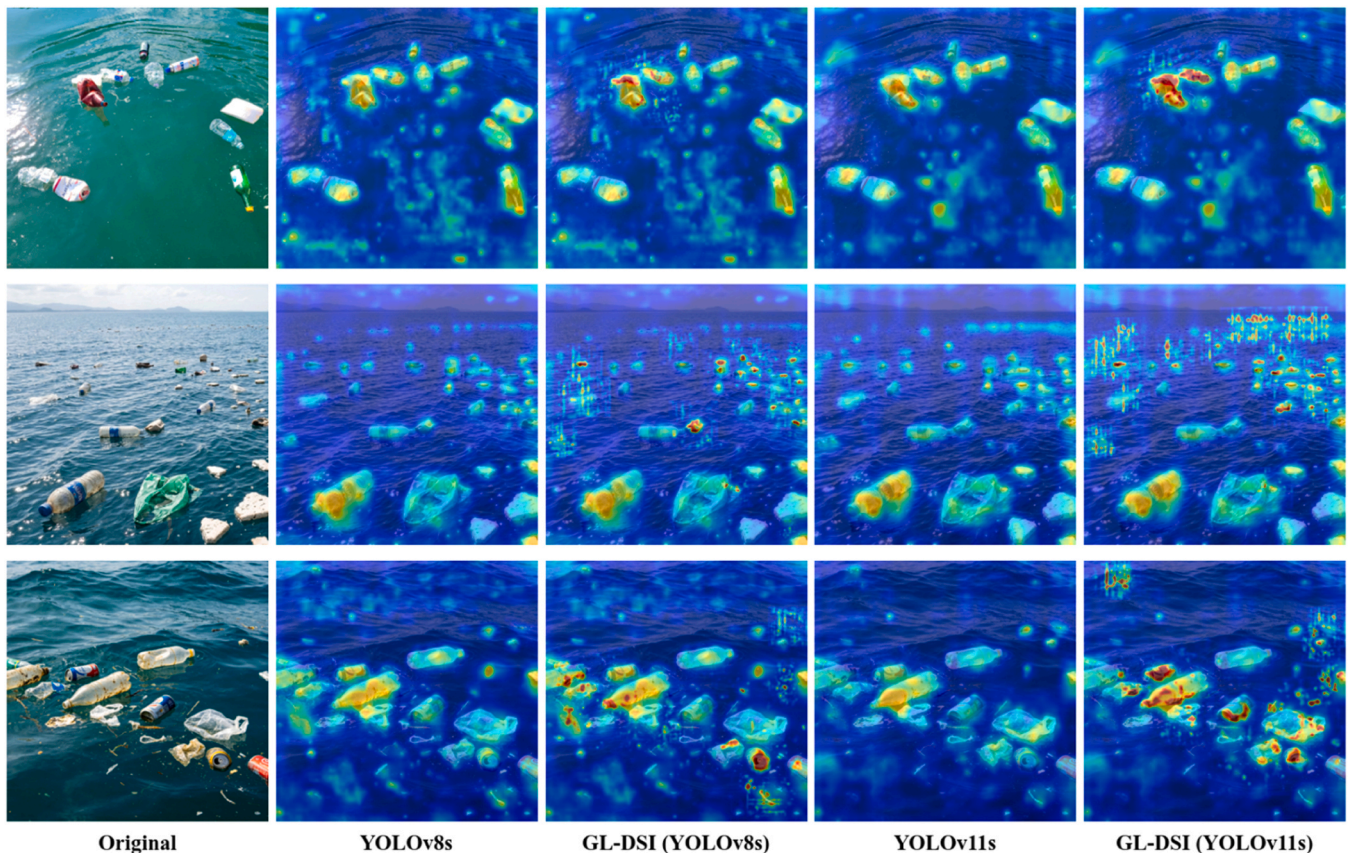


Fig. 10. Cross-architecture comparison of visual activation patterns. The visualization compares feature attention maps between standard inference and the proposed GL-DSI framework using YOLOv8s and YOLOv11s. The standard models (columns 2 and 4) exhibit relatively diffuse activation around small-scale targets under standard inference. GL-DSI (columns 3 and 5) shows more concentrated target-related activation following selective high-resolution local refinement and global-local integration.

targets while retaining scene-level context. The consistent trend observed across YOLOv8s and YOLOv11s suggests that GL-DSI can be integrated across different YOLO generations as an inference-time enhancement strategy for improving attention to difficult small-scale targets.

4.2. Operational trade-offs and paradigm comparison

The advantages of the GL-DSI framework extend beyond raw detection metrics to actual operational viability. As demonstrated by the operational trade-off analysis (Fig. 11), standard lightweight inference at 640×640 operates rapidly but suffers in overall detection reliability. While brute-force resolution scaling to 1280×1280 improves overall detection capability, its practical computational characteristics are strongly hardware-dependent. Conversely, in the tested marine scenes, SAHI exhibited reduced robustness; despite optimized threshold settings, its exhaustive blind slicing resulted in substantial false positives and computational costs.

Furthermore, when benchmarked against feature-guided (e.g., ESOD-inspired FE-AS) and density-guided (DG-QS) adaptive slicing paradigms, GL-DSI exhibits a more favorable precision–recall trade-off under both YOLOv8s and YOLOv11s backbones. The controlled comparison across YOLOv8s and YOLOv11s shows that the relative behavior of FE-AS and DG-QS is broadly consistent across the two backbones, indicating that the observed trade-off is not explained solely by backbone capacity. Feature- and density-guided slicing may respond to strong background activations or false candidate clusters caused by sun glint, wave crests, and foam, triggering unnecessary high-resolution patches and additional false positives. In contrast, GL-DSI replaces unguided feature activations with explicit low-confidence global detection cues (SGDS) and incorporates a scale-aware post-inference filter (SSCF), thereby concentrating high-resolution refinement on semantically relevant regions while suppressing background-induced false positives. Crucially, GL-DSI operates at 22.4 ms per frame (44.6 FPS) on the desktop benchmark, demonstrating a higher F1-score and lower latency than SAHI, FE-AS, and DG-QS with the same YOLOv11s backbone. Compared with multiscale TTA using YOLOv11s, GL-DSI achieves a

similar overall F1-score (93.4% versus 93.2%) at lower latency (22.4 versus 36.9 ms), whereas TTA retains higher small-scale target recall. However, the actual sensing coverage rate depends on platform-specific factors, including camera configuration, field of view, and platform motion, and requires further field validation. The influence of input resolution on inference latency and detection performance under the tested configurations is summarized in Table 4.

On the Jetson platform, however, selective local refinement introduces additional computation relative to the standard detector in exchange for improved small-scale target recovery, while maintaining a comparable memory footprint and stable execution. Its edge-device performance should therefore be interpreted as an accuracy–computation trade-off rather than an energy-saving modification of the baseline. At the other extreme, brute-force 1280×1280 inference achieves higher overall recall and mAP@50 and on the tested desktop GPU, also exhibits lower latency than GL-DSI. GL-DSI therefore should not be interpreted as dominating high-resolution inference in either accuracy or latency. Instead, the two strategies represent different operating points: high-resolution inference prioritizes recall and mAP@50, whereas GL-DSI achieves higher precision and F1-score through selective local refinement and background-noise suppression.

4.3. Implications for image-based marine litter monitoring

The small-scale target analysis indicates that standard baseline inference can underrepresent difficult debris instances in image space. In the analyzed small-scale subset, comparatively high miss rates were

Table 4

Performance gradient across different local-stream input resolutions.

Resolution	Latency (ms)	FPS	Recall (%)	Precision (%)
1080 P (FHD)	24.5	40.8	88.5	89.5
1440 P (2 K)	26.5	37.8	84.6	86.4
2160 P (4 K)	34.3	29.1	70.8	79.3

Note: All metrics are computed against ground-truth annotations at an IoU threshold of 0.5 using the same test set and evaluation protocol.

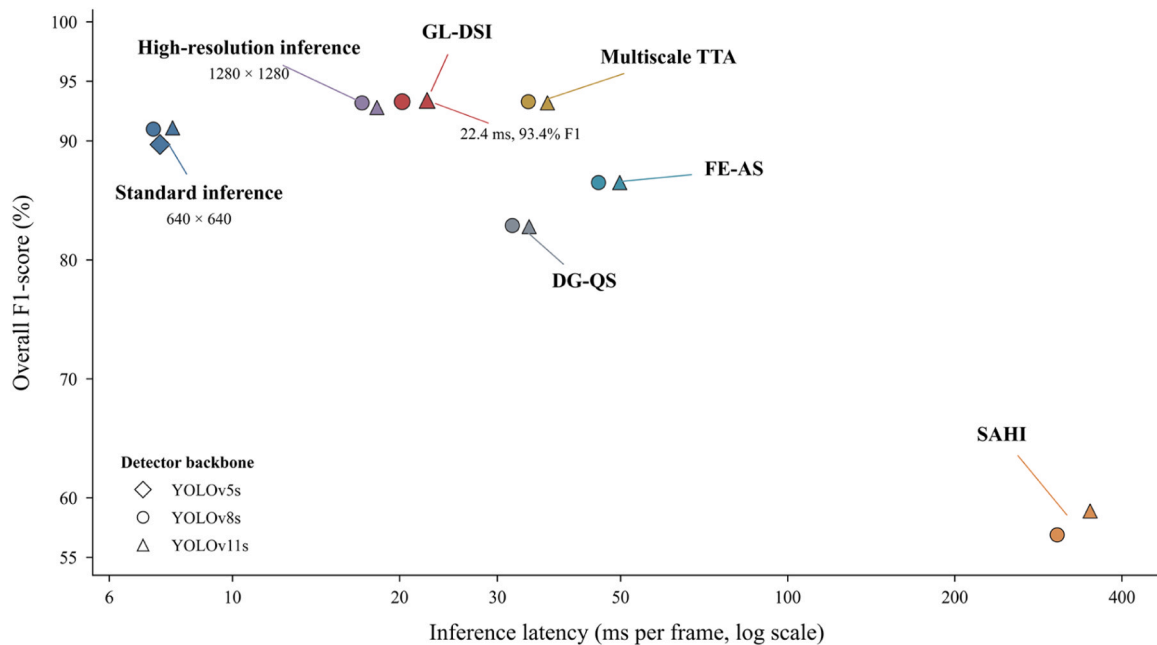


Fig. 11. Accuracy latency trade-off among the evaluated inference paradigms on the desktop benchmark. The x-axis represents inference latency on a logarithmic scale, while the y-axis represents the overall F1-score. All 15 configurations in Table 1 are shown. Colors distinguish inference strategies, while marker shapes indicate detector backbones. GL-DSI achieves a higher F1-score with additional latency relative to standard inference, while remaining substantially faster than SAHI. YOLOv11s + GL-DSI reaches an F1-score of 93.4% at 22.4 ms per frame (44.6 FPS).

observed for polystyrene and fishing gear, while GL-DSI recovered a portion of instances missed by the baseline. Recovering some of these omitted instances may reduce omissions in image-based debris records. This enhanced image-space detection capability may support waste traceability and spatial hotspot analysis [31], while providing improved visual information for future cleanup planning.

Our framework defines these targets based solely on projected 2D bounding-box area in image space. As camera distance, altitude, lens geometry, object orientation, occlusion, and image resolution dynamically affect a target's pixel footprint, bounding box geometries cannot be directly extrapolated to physical particle size or mass estimates. Objects appearing as micro-scale or nano-scale targets in the image plane may represent physically small fragments, or they may simply be distant or semi-submerged large debris. Furthermore, absent direct physical and chemical validation, these visual detections do not, by themselves, quantify chemical composition, toxicity, biological availability, or ecological risk. Consequently, GL-DSI functions strictly as an advanced image-space object detection tool, serving to complement rather than replace downstream physical and environmental assessments.

4.4. Limitations and broader impacts

While the proposed framework significantly enhances near-surface small-scale target detection, certain limitations remain. First, the evaluation using static images lacks camera geometric calibration and temporal tracking. Physical quantification requires calibration and complementary field measurements, while continuous monitoring requires tracking to suppress repeated observations. Second, cross-domain scenarios, such as underwater environments featuring severe optical attenuation [32,33] may introduce additional degradation in detection precision. Specific seabed textures, including coral rubble, can mimic the visual features of weathered plastic, indicating that texture-aware domain adaptation may be required for broader cross-domain deployment. Third, turbulent sea conditions introduce motion blur that may degrade local refinement performance by smearing high-frequency details. Future work will investigate lightweight object trackers to leverage temporal consistency across motion-blurred frames and reduce duplicate observations [28], together with multi-spectrum fusion [34] and underwater-specific feature enhancement strategies [32,33] to improve robustness under optical degradation. The Jetson benchmark demonstrates on-device execution, but further field deployment and system-level evaluation on active USVs are needed.

Despite these constraints, the broader impact of this work is relevant to resource-aware edge sensing, where computational efficiency remains an important system-level consideration [35]. By reducing redundant local processing relative to exhaustive slicing, the proposed dual-stream paradigm provides a selective inference strategy that remains deployable on embedded hardware, although the Jetson results show that it does not reduce energy consumption relative to the standard detector. The resulting image-based observations may support future assessments of coastal plastic pollution distribution [36]. Future integration with marine mechanical energy-harvesting and renewable-power systems may further improve the autonomy of long-duration sensing platforms [37,38].

The current dataset annotations distinguish genuine debris from background artifacts during quality control but do not assign mutually exclusive target-level labels for specific environmental nuisance types, such as specular glint, sea foam, wave crests, or turbulent water textures. Because these nuisance factors can overlap within the same scene, the present analysis does not attribute the observed recovery gain to individual noise categories; establishing a dedicated benchmark with annotations for environmental nuisance types is left for future work.

5. Conclusions

In response to the severe scale-compute conflict inherent in marine

floating litter detection on resource-constrained edge platforms, this study proposes the GL-DSI framework to shift the paradigm from exhaustive blind slicing to semantic-guided adaptive extraction, achieving an average of 73.5% reduction in processed local patches relative to exhaustive slicing across the test set. By structurally decoupling scene-level context from fine-grained visual details, GL-DSI utilizes a low-resolution global stream to selectively trigger high-resolution local patches around potential regions containing small-scale target candidates, while synergistically employing SAGA and the SSCF to reduce boundary artifacts and environmentally induced false positives such as sun glint and sea foam. Extensive evaluations on the custom MFL-Dataset demonstrate that GL-DSI achieves a favorable trade-off between detection efficacy and computational efficiency, delivering a robust F1-score of 93.4% and a real-time inference speed of 44.6 FPS (22.4 ms) on the desktop benchmark. Compared with the YOLOv11s baseline, GL-DSI increases recall for targets with areas below 256 px^2 by 11.0 %age points and nano-scale target recall from 13.5% to 27.0%. Ultimately, this precision-preserving architecture offers a practical accuracy-computation trade-off under edge-device constraints, reducing redundant processing relative to exhaustive slicing and providing a foundation for future AI-based marine sensing networks.

Environmental implication

Accurate detection of highly weathered and fragmented floating plastics remains challenging because their limited projected image footprints and complex marine backgrounds often result in systematic visual omission during image-based monitoring. In this study, we develop a resource-aware edge-AI framework that improves the recovery of visually fragmented floating litter by combining semantic-guided local refinement with adaptive noise suppression. By enabling more complete image-based observations on resource-constrained platforms, such as autonomous surface vehicles, the proposed framework provides a practical tool for improving marine debris monitoring and reducing missed detections of small-scale floating-litter targets. These enhanced observation capabilities may further support future assessments of coastal plastic pollution distribution and the development of targeted monitoring strategies.

CRedit authorship contribution statement

Junsen Zeng: Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Zirui Zhao:** Visualization, Validation, Investigation, Data curation. **Honglei Jiang:** Writing – review & editing, Formal analysis. **Shuowen Huang:** Writing – review & editing. **Qianyi Huang:** Resources, Investigation. **Xiangrong Xu:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Kefu Yu:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Guangxi Key Research and Development Program Project (No. AB25069113 & 2025FN9619929), Guangxi Science and Technology Program (No. AD25069075), and Hainan Province Science and Technology Special Fund (No. ZDYF2022SHFZ317).

Statement of significance

Accurate quantification of floating plastics under realistic marine

conditions is critical for evaluating their physicochemical fate and ecological risks. In the context of edge-vision monitoring, scale descriptors refer strictly to the extremely small apparent pixel area projected within the two-dimensional image plane, rather than the absolute physical dimensions commonly used in traditional environmental chemistry. This visual characterization introduces a severe resolution-compute conflict that systematically biases existing observation data. Lightweight detectors relying on fixed low-resolution inputs and aggressive downsampling inevitably obscure the fine-grained visual cues required to identify small-pixel fragments. Conversely, native high-resolution processing imposes prohibitive computational burdens that exceed edge hardware budgets, whereas conventional uniform slicing exhaustively inspects empty sea surfaces, which amplifies false positives from background noise and collapses processing speeds.

To structurally overcome this dilemma, we propose a global-local dual-stream inference framework that replaces blind slicing with semantic-guided adaptive extraction. This architecture uses a low-resolution global stream to capture scene-level context and localize suspect regions, selectively triggering high-resolution local inference exclusively where micro-scale instances likely exist to minimize parameter and computational redundancy. By incorporating a size-aware Gaussian edge attenuation module and a scale-semantic consistency filter to suppress boundary artifacts and environmentally induced pseudo-targets, the proposed framework achieves an optimal Pareto boundary. It delivers a robust real-time inference speed of 40.5 frames per second on resource-constrained settings while effectively doubling the absolute recall rate for targets occupying extremely limited pixel areas.

Consequently, this study provides a structurally robust and eco-efficient analytical tool that fundamentally corrects existing mass-balance underestimations of hidden marine pollution. By reliably reclaiming visually fragmented hazards from complex backgrounds without imposing excessive power demands, our approach establishes a precise, scalable technological baseline for advanced waste traceability, global plastic emission inventories, and targeted hazardous waste management.

Data availability

Data will be made available on request.

References

- [1] Galappaththi, E.K., Russell, J.D., Dolby, M., Newsome, T., Jayasekara, S.M., 2025. Status of global accumulation of marine debris. *Glob Environ Change* 95, 103058. <https://doi.org/10.1016/j.gloenvcha.2025.103058>.
- [2] Starcevic, L.A., Burgess, H.K., Uhrin, A.V., 2025. Optimizing marine debris monitoring: balancing statistical power and test size in survey design. *Ecol Indic* 178, 113807. <https://doi.org/10.1016/j.ecolind.2025.113807>.
- [3] Soon, Z.Y., Cheng, M.M.-C., Ouyang, L., Tamburri, M.N., 2026. A multimodal Raman – XRF approach for identifying marine coating microplastics in environmental samples. *J Hazard Mater* 505, 141482. <https://doi.org/10.1016/j.jhazmat.2026.141482>.
- [4] Wang, S., Jiao, W., Zhao, W., Xu, C., Lang, S., Jia, Y., 2026. Detecting multiple types of marine floating debris via satellite remote sensing informed by hyperspectral features. *Ecol Indic* 182, 114529. <https://doi.org/10.1016/j.ecolind.2025.114529>.
- [5] Zhang, Y., Chen, J., Wang, J., Shi, D., Han, S., Deng, L., 2025. C2DFF-net for object detection in multimodal remote sensing images. *IEEE Trans Geosci Remote Sens* 63, 1–16. <https://doi.org/10.1109/TGRS.2025.3614295>.
- [6] Cózar, A., Arias, M., Suaria, G., Viejo, J., Aliani, S., Koutroulis, A., et al., 2024. Proof of concept for a new sensor to monitor marine litter from space. *Nat Commun* 15, 4637. <https://doi.org/10.1038/s41467-024-48674-7>.
- [7] Li, Z., An, W., Guo, G., Wang, L., Wang, Y., Lin, Z., 2025. SpecDETR: a transformer-based hyperspectral point object detection network. *ISPRS J Photogramm Remote Sens* 226, 221–246. <https://doi.org/10.1016/j.isprsjprs.2025.05.005>.
- [8] Mäyrä, J., Virtanen, E.A., Jokinen, A.-P., Koskikala, J., Väkevä, S., Attila, J., 2025. Mapping recreational marine traffic from Sentinel-2 imagery using YOLO object detection models. *Remote Sens Environ* 326, 114791. <https://doi.org/10.1016/j.rse.2025.114791>.
- [9] Zheng, Y., Nong, X., Chen, L., Long, D., 2025. A novel deep learning-based floating garbage detection approach and its effectiveness evaluation in environmentally sustainable development. *J Environ Manag* 381, 125154. <https://doi.org/10.1016/j.jenvman.2025.125154>.
- [10] Dang, L.M., Sagar, A.S., Bui, N.D., Nguyen, L.V., Nguyen, T.H., 2025. Attention-guided marine debris detection with an enhanced transformer framework using drone imagery. *Process Saf Environ Prot* 197, 107089. <https://doi.org/10.1016/j.psep.2025.107089>.
- [11] Qiao, G., Yang, M., Wang, H., 2025. An annotated dataset and benchmark for detecting floating Debris in inland waters. *Sci Data* 12, 385. <https://doi.org/10.1038/s41597-025-04594-9>.
- [12] Wang, Y., Liu, S., He, Y., Zhang, Y., 2025. YOLO11-YX: an efficient algorithm for marine debris target detection. *Mar Pollut Bull* 221, 118511. <https://doi.org/10.1016/j.marpolbul.2025.118511>.
- [13] Chen, L.-J., Zou, J.-M., Zhou, Y., Hao, G.-Z., Wang, Y., 2025. Real-Time Semantic Segmentation of Maritime Navigation Scene Based on Improved Fast-SCNN. In: The Proceedings of 2024 International Conference on Artificial Intelligence and Autonomous Transportation (AIAT 2024), Lecture Notes in Electrical Engineering 1394, 346–362. https://doi.org/10.1007/978-981-96-3977-9_38.
- [14] Xu, W., Zheng, S., Wang, C., Zhang, Z., Ren, C., Xu, R., et al., 2025. SAMamba: Adaptive state space modeling with hierarchical vision for infrared small target detection. *Inf Fusion* 124, 103338. <https://doi.org/10.1016/j.inffus.2025.103338>.
- [15] Sun, Y., Zhang, C., Li, X., Jing, X., Kong, H., Wang, Q.-G., 2025. MDSF-YOLO: advancing object detection with a multiscale dilated sequence fusion network. *IEEE Trans Neural Netw Learn Syst* 1–12. <https://doi.org/10.1109/TNNLS.2025.3617122>.
- [16] Liu, W., Li, Q., Lin, X., Yang, W., He, S., Yu, Y., 2024. Ultra-high resolution image segmentation via locality-aware context fusion and alternating local enhancement. *Int J Comput Vision* 132, 5030–5047. <https://doi.org/10.1007/s11263-024-02045-3>.
- [17] Liu, K., Fu, Z., Jin, S., Chen, Z., Zhou, F., Jiang, R., et al., 2025. ESOD: efficient small object detection on high-resolution images. *IEEE Trans Image Process* 34, 183–195. <https://doi.org/10.1109/TIP.2024.3501853>.
- [18] Zhang, T., Ren, L., Sun, X., Gao, L., Zhang, B., 2026. Resolution preserving and utilization network for tiny object detection in large-size remote sensing imagery. *IEEE Trans Image Process* 35, 4208–4222. <https://doi.org/10.1109/TIP.2026.3678372>.
- [19] Akyon, F.C., Altinuc, S.O., Temizel, A., 2022. Slicing aided hyper inference and fine-tuning for small object detection. 2022 IEEE International Conference on Image Processing (ICIP), 966–970. <https://doi.org/10.1109/ICIP46576.2022.9897990>.
- [20] Turečková, A., Tureček, T., Komínková Oplatková, Z., 2024. Artificial size slicing aided fine tuning (ASSAFT) and hyper inference (ASSAHI) in tomato detection. *Comput Electron Agric* 225, 109280. <https://doi.org/10.1016/j.compag.2024.109280>.
- [21] Liu, J., Chen, Y., Si, Y., Chen, T., 2025. An approach to microalgae identification based on joint optimization model of slicing and detection. *Expert Syst Appl* 280, 127565. <https://doi.org/10.1016/j.eswa.2025.127565>.
- [22] Hu, C., 2025. On the logic of remote detection of plastic litter in the aquatic environments: A revisit. *Remote Sens Environ* 329, 114911. <https://doi.org/10.1016/j.rse.2025.114911>.
- [23] Papageorgiou, D., Topouzelis, K., 2024. Experimental observations of marginally detectable floating plastic targets in Sentinel-2 and Planet SuperDove imagery. *Int J Appl Earth Obs Geoinf* 134, 104245. <https://doi.org/10.1016/j.jag.2024.104245>.
- [24] Xu, B., Zhu, Z., Qiu, X., Wang, S., Chen, Z., Zhang, H., et al., 2024. A lightweight object detection algorithm for autonomous driving based on edge computing. *Expert Syst Appl* 238, 121752. <https://doi.org/10.1016/j.eswa.2023.121752>.
- [25] Yang, S., Wang, W., Gao, S., Deng, Z., 2023. Strawberry ripeness detection based on YOLOv8 algorithm fused with LW-Swin Transformer. *Comput Electron Agric* 215, 108360. <https://doi.org/10.1016/j.compag.2023.108360>.
- [26] Yang, Z., Liang, W.-Z., Lawrence, N., Qiao, X., Riggan, B., Harveson, R., et al., 2026. WeedCAM: An edge-computing camera system for multi-species weed detection in sugar beet production fields. *Comput Electron Agric* 244, 111498. <https://doi.org/10.1016/j.compag.2026.111498>.
- [27] Zhu, Y., Sui, S., Du, W., Li, X., Liu, P., 2025. Picking point localization method of table grape picking robot based on you only look once version 8 nano. *Eng Appl Artif Intell* 146, 110266. <https://doi.org/10.1016/j.engappai.2025.110266>.
- [28] Kos, A., Majek, K., Belter, D., 2025. Enhanced lightweight detection of small and tiny objects in high-resolution images using object tracking-based region of interest proposal. *Eng Appl Artif Intell* 153, 110852. <https://doi.org/10.1016/j.engappai.2025.110852>.
- [29] Yao, R., Li, W., Zhou, Y., Sun, J., Yin, Z., Zhao, J., 2024. Dual-stream edge-target learning network for infrared small target detection. *IEEE Trans Geosci Remote Sens* 62, 1–14. <https://doi.org/10.1109/TGRS.2024.3488054>.
- [30] Huang, Z., Qiu, H., Hou, M., Yu, Z., Wang, S., Li, X., et al., 2025. MCECF: a multiscale complementary enhanced context fusion network for remote sensing change detection. *IEEE Trans Geosci Remote Sens* 63, 1–14. <https://doi.org/10.1109/TGRS.2025.3556237>.
- [31] Liang, W., Fan, L., Yang, Z., Yang, J., Zou, Q., Huang, C., Huang, Q., et al., 2025. How to track the dumped waste? From data collection to advanced source identification. *J Hazard Mater* 499, 140052. <https://doi.org/10.1016/j.jhazmat.2025.140052>.
- [32] Chen, L., Huang, Y., Dong, J., Xu, Q., Kwong, S., Lu, H., et al., 2026. Underwater optical object detection in the era of artificial intelligence: current, challenge, and future. *ACM Comput Surv* 58, 3759243. <https://doi.org/10.1145/3759243>.
- [33] Zhou, J., He, Z., Zhang, D., Liu, S., Fu, X., Li, X., 2025. Spatial residual for underwater object detection. *IEEE Trans Pattern Anal Mach Intell* 47, 4996–5013. <https://doi.org/10.1109/TPAMI.2025.3548652>.

- [34] Nan, Y., Zhang, Q., Kou, T., He, T., Huang, C., Lu, C., Shen, J., 2025. ISP-free multi-spectrum fused imaging for extremely low-light enhanced photography. *Inf Fusion* 124, 103400. <https://doi.org/10.1016/j.inffus.2025.103400>.
- [35] Thakur, D., Guzzo, A., Fortino, G., Piccialli, F., 2025. Green federated learning: a new era of green aware AI. *ACM Comput Surv* 57, 1–36. <https://doi.org/10.1145/3718363>.
- [36] Cottom, J.W., Cook, E., Velis, C.A., 2024. A local-to-global emissions inventory of macroplastic pollution. *Nature* 633, 114–123. <https://doi.org/10.1038/s41586-024-07758-6>.
- [37] Sun, Y., Tan, T., Yan, Z., 2025. Smart Ocean powering and sensing via mechanical energy harvesting: methods, advances, and challenges. *Nano Energy* 141, 111128. <https://doi.org/10.1016/j.nanoen.2025.111128>.
- [38] Wang, G., Xu, T., Yang, M., Niu, W., 2025. Review on renewable energy systems of unmanned marine vehicles and guidance for energy selection. *Renew Sustain Energy Rev* 224, 116085. <https://doi.org/10.1016/j.rser.2025.116085>.