ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1113-1134
DOI 10.5755/j01.itc.55.3.44693

YOLOv13-LSA: An Enhanced Target Detection Algorithm for Water Surface and Underwater Garbage with Multi-Stage Transfer Learning Framework

Received 2026/04/01 | Accepted after revision 2026/06/26

HOW TO CITE: Wang, Z., Liu, X.(2026) YOLOv13-LSA: An Enhanced Target Detection Algorithm for Water Surface and Underwater Garbage with Multi-Stage Transfer Learning Framework. *Information Technology and Control*, 55(3), 1112-1133. https://doi.org/10.5755/j01.itc.55.3.44693

YOLOv13-LSA: An Enhanced Target Detection Algorithm for Water Surface and Underwater Garbage with Multi-Stage Transfer Learning Framework

Zhihao Wang*
School of Computer Science and Technology, North University of China, Taiyuan, 030051, China;
e-mail: 18334965258@163.com

Xiaolin Liu
School of Computer Science and Technology, North University of China, Taiyuan, 030051, China;
e-mail: zhishidanggao@gmail.com

Corresponding author: 18334965258@163.com

Water-surface and underwater garbage detection remains challenging because aquatic images often contain low contrast, light attenuation, reflection, wave interference, blurred boundaries, and large scale variation. These factors cause missed detections of small debris and inaccurate localization of irregular objects. Although YOLO-based methods have improved aquatic detection, many focus on a single direction, such as lightweight design, attention enhancement, or loss optimization, and still struggle with boundary ambiguity, multi-scale conflict, and cross-domain shift. To address these issues, this study develops YOLOv13-LSA by integrating existing SlideLoss, LSKA, and ASFF components into YOLOv13. SlideLoss is adopted to strengthen boundary regression for irregular and elongated targets, LSKA is embedded to enhance weak and low-contrast features, and ASFF is incorporated for adaptive multi-scale fusion. A source-to-target aquatic transfer learning framework is further designed by pretraining on underwater garbage data and fine-tuning on water-surface garbage data. Experiments show that YOLOv13-LSA improves detection accuracy, small-target recall, and robustness in complex aquatic scenes. Transfer learning further increases mAP@50 to 83.46% and reduces training time by 7.57%, supporting intelligent aquatic garbage monitoring.

KEYWORDS: YOLOv13; Transfer Learning; Water Surface/Underwater Garbage Detection; SlideLoss Loss Function; Adaptive Feature Fusion; LSKA Module

1.Introduction

Automated detection of water-surface and underwater debris improves inspection efficiency, reduces monitoring costs, and supports ecosystem protection. Floating and submerged waste threatens biodiversity and maritime safety, making reliable localization important for pollution control [4]. Deep learning has shifted debris monitoring from manual inspection to AI-assisted detection [12]. However, refraction, scattering, sediments, dynamic backgrounds, low contrast, and blurred boundaries still reduce accuracy, speed, and cross-scene adaptability [12, 15, 24].

Aquatic waste detectors are commonly divided into two-stage and one-stage methods. Two-stage detectors [7,8,27] improve accuracy but are often too costly for real-time use. Li et al. [18] proposed PC-Net based on Faster R-CNN [27] to improve small-target representation under reflection and waves. Song et al. [29] proposed Boosting R-CNN for low-contrast, occluded, and camouflaged underwater targets. However, efficiency, boundary regression, and small low-contrast targets remain difficult.

One-stage YOLO detectors are more suitable for real-time aquatic garbage detection because they perform end-to-end localization and classification. Li et al. [19] improved YOLOv3 with scale reduction and anchor re-clustering, but small-target misses remained. Wu et al. [33] enhanced YOLOv5s with a mobile backbone [23] and attention module [10]. Jiang et al. [13] introduced PConv-ELAN [1] and adaptive fusion into YOLOv7 [31]. Zhang et al. [32] added lightweight attention [3] and multi-scale recombination to YOLOv8. Ye et al. [34] proposed YOLOv12-LSE for multi-scale features and light interference, but complexity affected edge use. Feng et al. [5] proposed Hyper-YOLO with hypergraph computation for general object detection, while YOLOv13 was introduced as a real-time detector with hypergraph-enhanced adaptive visual perception [17].

Transfer learning helps address class imbalance, limited samples, poor image quality, and domain shift [6, 24, 28]. Zhao et al. [38] proposed UWS-YOLO for underwater sonar-object detection. Lin et al. [20] developed DLFE-YOLO with deformable convolution, attention, and improved regression loss. Guo et al. [9] improved YOLOv5s using EfficientNetV2 [30], feature fusion, attention, and SIoU loss. Related attention and loss designs, including coordinate attention [11], NAM attention [22], and Focal-EIoU loss [37], also provide useful references for improving weak-feature representation and boundary regression. Yet rare samples, weak signals, and cross-scene stability remain challenging.

Despite this progress, YOLO-based transfer learning remains limited in water-surface and underwater scenes. Many methods rely on natural images or a single water-domain dataset, while reflection, waves, attenuation, scattering noise, and color shift alter target appearance and background statistics [2]. Distribution differences caused by turbidity, sea area, or imaging devices are also insufficiently analyzed [14]. Moreover, pretrained weights are often used without considering source-target domain differences, which may cause feature bias under scarce target samples [26]. Therefore, efficient, practical, and scenario-adaptive detectors are still needed.

Although YOLO-based methods perform well in aquatic object detection, several issues remain: boundary ambiguity and localization drift are insufficiently handled, fixed multi-scale fusion may cause feature conflict, and common transfer learning strategies weakly model the domain gap between underwater and water-surface images. Therefore, a task-oriented framework is needed to improve boundary localization, weak-feature representation, adaptive fusion, and cross-domain generalization.

This paper develops YOLOv13-LSA as an engineered integration of existing SlideLoss, LSKA, and ASFF components with YOLOv13 for water-surface and underwater garbage detection. These modules are integrated into a degradation-driven pipeline for feature enhancement, scale fusion, and boundary optimization. A multi-stage transfer learning framework is further introduced to improve cross-domain adaptation. The main contributions are:

1 A task-oriented YOLOv13-LSA detection framework is developed by integrating existing SlideLoss and LSKA components into YOLOv13. This engineered design improves boundary-aware regression and weak-feature representation for elongated, deformable, low-contrast, and partially occluded aquatic garbage targets.

2 ASFF is incorporated into the YOLOv13 detection head to perform spatially adaptive multi-scale feature fusion. This integration helps reduce scale conflicts and improves small-target recall and localization stability in complex aquatic scenes.

3 A source-to-target aquatic transfer learning framework is designed for the integrated YOLOv13-LSA model. The model first learns general aquatic representations from a larger underwater garbage dataset and is then fine-tuned on the target water-surface dataset, improving convergence, robustness, and cross-domain adaptability.

2. YOLOv13-LSA Algorithm

Although YOLOv13 performs well in general detection, its accuracy may decline in water-surface and underwater garbage detection due to irregular shapes, scale variation, blurred boundaries, and noisy aquatic backgrounds.

SlideLoss improves bounding-box regression for elongated or irregular targets by reweighting difficult samples with large localization errors, enhancing edge and contour perception. The ASFF-enhanced head adaptively fuses multi-scale features at each spatial location, improving cross-scale detection and localization stability without excessive complexity.

These improvements enable YOLOv13-LSA to achieve better accuracy, small-target recognition, and robustness while maintaining efficient inference in complex aquatic environments. The YOLOv13-LSA architecture is shown in Figure 1.

2.1. Traditional YOLOv13 Algorithm

The YOLO series balances speed and accuracy through its one-stage detection framework. YOLOv13 further improves feature extraction and multi-scale representation through hypergraph-enhanced adaptive visual perception while following the Backbone-Neck-Head paradigm for feature extraction, fusion, and detection [17].

However, YOLOv13 still faces challenges in water-surface and underwater garbage detection. Irregular shapes, blurred boundaries, waves, refraction, and motion blur weaken boundary localization, especially for elongated objects. Turbidity, light attenuation, suspended particles, and reflection reduce target-background contrast, causing missed detections and confidence fluctuations. In addition, large scale and pose variations make it difficult to balance local details and global context. Therefore, further optimization is needed in boundary regression, weak small-target recall, cross-scale representation, and feature fusion. The baseline YOLOv13 architecture is shown in Figure 2.

Figure 1

Architecture of YOLOv13-LSA. The backbone, neck, and ASFF-based detection head are explicitly annotated, and the FullPAD tunnel, HyperACE feature aggregation module, LSKA module, B3/B4/B5 feature paths, and multi-scale detection branches are shown in detail.

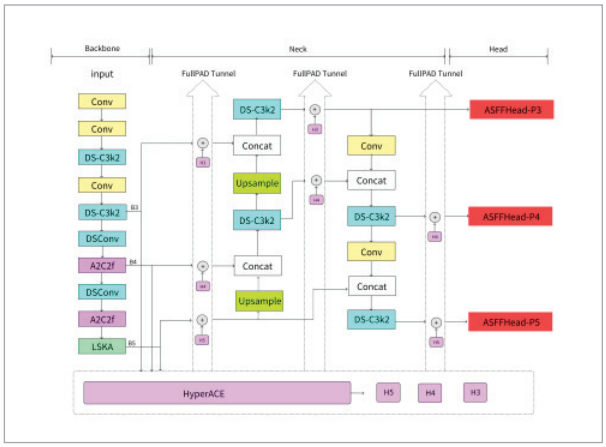
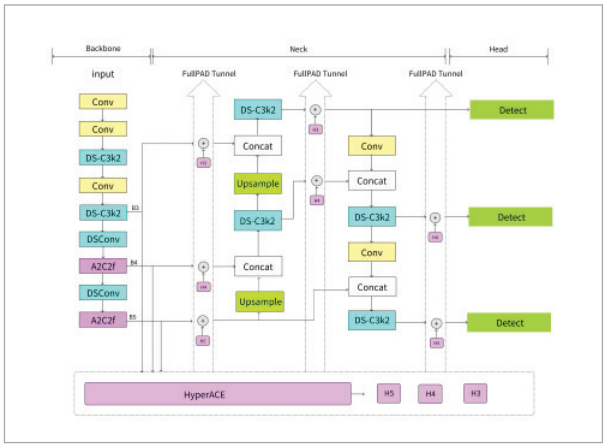


Figure 2

Architecture of YOLOv13. The baseline detector is redrawn as a high-resolution diagram with clearer annotations of the backbone, neck, detection head, FullPAD tunnel, HyperACE module, B3/B4/B5 feature transmission paths, and multi-scale detection outputs.



2.2. SlideLoss

Bounding box regression directly affects localization accuracy in object detection, and this issue is more critical in water-surface and underwater garbage detection. Aquatic imaging conditions such as light scattering, refraction, turbidity, surface reflection, waves, and slight object motion often weaken edge gradients and blur object contours. As a result, a predicted box may obtain a relatively high IoU while its boundaries still slide away from the actual object, especially for elongated, deformable, or irregular targets such as plastic strips, ropes, floating films, and partially submerged bottles.

SlideLoss [35] is introduced to address this boundary ambiguity. Conventional IoU-based losses mainly measure area overlap, so they may be insensitive when two boxes have similar IoU values but different boundary alignments. In aquatic scenes, small horizontal or vertical offsets can include reflections, ripples, or background noise in the predicted box, reducing localization reliability. Therefore, a loss function that emphasizes boundary misalignment and sliding offsets is needed. This work adopts SlideLoss to enhance the regression loss function, tackling issues such as blurred target boundaries, box drift, and localization errors in underwater and water-surface garbage detection. By strengthening horizontal and vertical sliding offset constraints between predicted and ground-truth boxes, SlideLoss reduces reliance on IoU alone, enhancing boundary alignment and directional error optimization for more precise detection in challenging aquatic environments.

The Slide weighting function is as follows:

$$f(x)=\begin{cases}1, & x\le\mu-\delta\\ exp\left(\frac{1}{1-\mu}\right), & \mu-\delta<x<\mu\\ exp\left(\frac{1}{1-x}\right), & x\ge\mu\end{cases}, \quad (1)$$

where μ represents the mean IoU threshold, and δ denotes the boundary bandwidth near the threshold, which is used to emphasize hard-to-classify samples.

During training, the SlideLoss weighting function is multiplied by the classification/target confidence loss (e.g., BCE/CE) or the overall sample loss, as expressed by:

$$L_{Silde}=\frac{1}{N}\sum_{i=1}^{N}f(x_i)\,L_i. \quad (2)$$

This method effectively suppresses the "masking effect" of easy samples on the optimization direction, allowing YOLOv13 to achieve higher stability in detecting small targets, low-contrast targets, and in scenarios with reflection interference in water surface/underwater garbage detection. During the optimization process, SlideLoss introduces a directional loss function for the localization error of each target box, expressed as:

$$L_{Slide\ Direction}=\sum_{i=1}^{N}(\alpha\cdot\|\Delta x_i\|+\beta\cdot\|\Delta y_i\|), \quad (3)$$

where Δx_i and Δy_i represent the horizontal and vertical offsets, while α and β are the directional weights.

By introducing a direction-sensitive penalty, SlideLoss further optimizes the translation and rotation errors of the bounding box, expressed as:

$$L_{Translation\ \&\ Rotation}=\sum_{i=1}^{N}(\lambda_{trans}\|\Delta x_i\|+\lambda_{rot}\|\Delta\theta_i\|), \quad (4)$$

where, $\Delta\theta_i$ represents the rotational deviation of the target box, while λ_{trans} and λ_{rot} are the penalty coefficients for translation and rotation, respectively.

The final loss function is expressed as:

$$L_{Total}=L_{Slide}+\lambda_{Slide\ Direction}+\lambda_{\gamma}L_{Translation\ \&\ Rotation}. \quad (5)$$

Through the above method, SlideLoss enhances YOLOv13's regression accuracy in complex environments, improving the localization stability and precision of garbage target detection.

In YOLOv13-LSA, SlideLoss is used only during training and does not change the inference structure. It calculates IoU between predicted and ground-truth boxes, classifies samples by localization quality, and assigns higher weights to hard samples with blurred boundaries or localization ambiguity. The weighted loss then guides more accurate boundary regression. Thus, SlideLoss improves localization stability for irregular and weak-boundary garbage targets without increasing inference cost.

2.3. LSKA

Feature extraction and spatial context modeling are important for water-surface and underwater garbage detection. Aquatic images often suffer from low contrast, light attenuation, scattering noise, color distortion, suspended particles, ripples, and specular reflection. These factors weaken texture and edge responses, making small fragments, thin plastic films, ropes, and partially occluded objects difficult to separate from the background. Therefore, local convolutional responses alone may be insufficient, because useful target cues are often distributed over a wider spatial region.

LSKA [16] is introduced to address feature attenuation in aquatic images. By expanding the receptive field, it helps the network capture long-range spatial dependencies, including contour continuity, spatial contrast, texture differences, and target-background relationships. Unlike standard large-kernel convolutions, LSKA decomposes large-kernel modeling into separable one-dimensional depthwise convolutions in horizontal and vertical directions, reducing parameters and computational cost. The generated spatial attention map enhances informative target regions and suppresses background interference from ripples, reflections, turbidity, and texture noise, improving weak and boundary-blurred target representation while maintaining inference efficiency.

Given the input feature map, with a given separable large kernel size k (and an optional dilation rate d), the spatial attention computation performed by LSKA can be expressed as: $X \in R^{H\times W\times C}$

$$Y=LSKA(X)=X \odot \sigma(A), \tag{6}$$

where $\odot$ denotes element-wise multiplication, $\sigma(\cdot)$is the Sigmoid activation, and A is the spatial attention map. The attention map A is composed of separable large kernel depthwise convolutions and point convolutions:

$$A=PWConv\left(DWConv_{1\times k}^{(d)}\left(DWConv_{k\times 1}^{(d)}(X)\right)\right). \tag{7}$$

For convenience of expression, the two one-dimensional large kernel depthwise convolutions are denoted as:

$$Z=DWConv_{k\times 1}^{(d)}(X), \quad A=PWConv\left(DWConv_{1\times k}^{(d)}(Z)\right). \tag{8}$$

The resulting enhanced feature map Y is then used as input to subsequent convolution/fusion modules, aiming to improve the model's capacity to capture weakly prominent, small-scale, and boundary-blurred garbage targets:

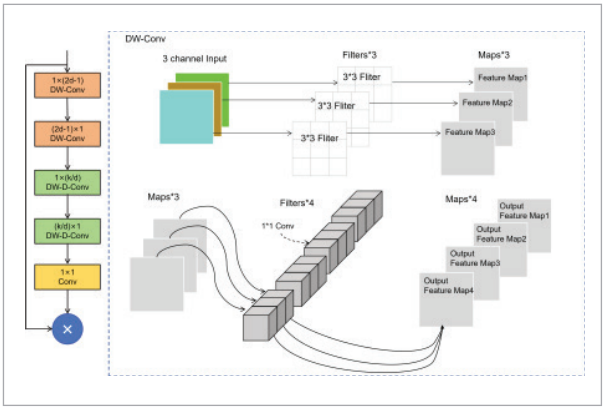
$$Fout=\Phi(Y). \tag{9}$$

Through this design, LSKA establishes large-scale spatial dependencies with limited overhead, reducing weak-feature suppression caused by reflections, ripples, and turbidity in aquatic scenes. It improves small-target recall, localization stability, and robustness in complex underwater and water-surface garbage detection. The module structure is shown in Figure 3.

In YOLOv13-LSA, LSKA is embedded in the high-level feature extraction stage of the backbone. The input feature map is first processed by depthwise separable convolution to reduce redundancy, then large-kernel spatial attention captures long-range context. The generated attention weights enhance informative regions and suppress background interference before the refined features enter subsequent fusion layers. This improves representation of small, low-contrast, and partially occluded garbage targets.

Figure 3

Architecture of LSKA and its embedding position in the high-level backbone feature extraction stage of YOLOv13-LSA.



2.4. ASFF

In water-surface and underwater garbage detection, target scales vary greatly and background interference is strong. Distant debris often appears as small, weak-textured targets, while nearby plastic bags or foam blocks have large, irregular boundaries. Ripples, reflections, and turbidity introduce false textures and weaken edge gradients, causing semantic and detail differences across feature layers. Although YOLOv13 supports multi-scale fusion, fixed fusion methods such as concat or add may cause cross-scale conflicts and noise interference, leading to missed small targets, false positives, and unstable localization. Therefore, ASFF [21] is introduced to perform scale alignment and spatially adaptive weighting, improving cross-scale coordination and suppressing noise for more stable detection.

Let the input features from the three-level feature pyramid be denoted as multi-scale feature maps. ASFF uses a target scale as the fusion output scale. First, the features from the other scales are mapped to the same resolution and channel number as the l-th layer through upsampling/downsampling and channel alignment, resulting in the aligned features: $\{X^{(1)}, X^{(2)}, X^{(3)}\} l \in \{1,2,3\}$

$$\widetilde{X}^{(k\to l)} = \phi_{k\to l}(X^{(k)}), \quad k \in \{1,2,3\}, \tag{10}$$

where represents the mapping function composed of scale transformations such as interpolation, pooling, strided convolutions, and 1 x 1 channel alignment. Subsequently, ASFF generates spatial weight logits for each aligned scale feature: $\phi_{k\text{-}l}(\cdot)$

$$\lambda_k^{(l)} = Conv\left(\widetilde{X}^{(k\to l)}\right), \quad k \in \{1,2,3\}. \tag{11}$$

At each spatial position (i, j), the three-scale logits are normalized using Softmax, resulting in spatial weights that are non-negative and sum to 1:

$$\alpha_k^{(l)}(i,j) = \frac{exp\left(\lambda_k^{(l)}(i,j)\right)}{\sum_{t=1}^{3} exp\left(\lambda_t^{(l)}(i,j)\right)}, \quad \sum_{k=1}^{3} \alpha_k^{(l)}(i,j) = 1. \tag{12}$$

Finally, ASFF performs position-wise weighted fusion at the target scale l, resulting in the output feature:

$$Y^{(l)} = \sum_{k=1}^{3} \alpha_k^{(l)} \odot \widetilde{X}^{(k\to l)}. \tag{13}$$

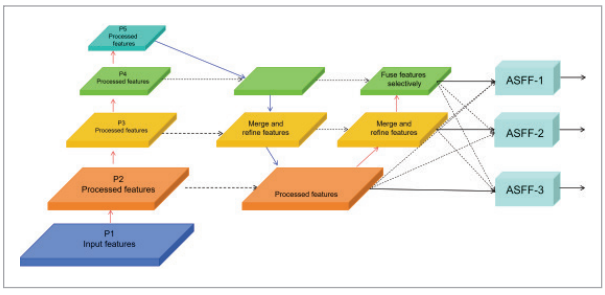
Through spatially correlated fusion, ASFF adaptively selects scale contributions at different locations. In regions with strong reflection or wave disturbance, it relies more on deeper semantic features for noise suppression, while for small floating debris, it emphasizes shallow detailed features to improve recall. This reduces cross-scale conflict and redundancy caused by fixed fusion.

ASFF performs scale alignment, adaptive weighting, and dynamic fusion, allowing the network to select suitable scale information for each location. With lightweight convolutions and Softmax-based weight learning, it keeps lower overhead than complex global attention mechanisms. Therefore, ASFF is integrated into YOLOv13's fusion stage to improve small-target recall, reflection/wave robustness, and localization stability. The module structure is shown in Figure 4.

In the proposed framework, ASFF is introduced into the multi-scale detection head to replace the fixed feature fusion strategy. Its workflow is as follows: features from different scales are first resized to the same spatial resolution; then, a lightweight convolutional layer learns spatially adaptive fusion weights for each scale; next, the normalized weights are applied to different feature maps so that the detector can dynamically select the most informative scale at each spatial location; finally, the fused feature is sent to the corresponding detection head for classification and localization. In this way, ASFF reduces feature conflict among different scales and improves

Figure 4

Architecture of ASFF for adaptive multi-scale feature fusion in the YOLOv13-LSA detection head.



the detection of garbage objects with large variations in size and shape.

2.5. Collaborative Mechanism of YOLOv13-LSA

In the YOLOv13-LSA pipeline, LSKA, ASFF, and SlideLoss play complementary roles in feature extraction, feature fusion, and loss optimization. Their integration is designed according to the degradation characteristics of water-surface and underwater garbage images, including blurred boundaries, weak textures, irregular shapes, and large scale variations, rather than simply stacking independent modules.

LSKA works at the feature-representation level. It uses large-receptive-field spatial attention to capture broader context and enhance weak target responses under light attenuation, turbidity, reflection, and low contrast.

ASFF works at the multi-scale fusion level. It adaptively assigns spatial weights to features from different scales, helping the detector select suitable spatial details or semantic information for targets with different sizes and reducing cross-scale feature conflict.

SlideLoss works at the optimization-supervision level. It strengthens difficult and boundary-ambiguous samples during training, improving regression stability and localization robustness under reflection, occlusion, and wave disturbance.

Overall, YOLOv13-LSA forms a feature enhancement, adaptive fusion, and boundary optimization pipeline, improving aquatic garbage detection through coordinated module design.

3. Multi-Phase Transfer Learning Framework

YOLOv13-LSA still faces challenges in water-surface and underwater garbage detection because aquatic imaging and data distributions are complex. Turbidity, light attenuation, ripples, and strong reflections weaken low-contrast targets and introduce false textures. Small debris, thin films, and elongated objects often have blurred boundaries, causing localization drift. In addition, variations in water bodies, devices, weather, and time create domain shifts, while long-tail category distributions weaken rare-class learning. To address these issues, a multi-stage transfer learning framework is developed.

In the first stage, the model is trained on source-domain aquatic garbage data to learn general representations. In the second stage, it is transferred to target-domain data with class-balanced sampling, hard-example resampling, and momentum-based optimization to improve feature alignment and convergence. LSKA enhances spatial context, ASFF improves adaptive multi-scale fusion, and SlideLoss strengthens hard-sample regression. In the third stage, SGD fine-tuning further improves small-target recognition, localization stability, and cross-environment robustness while maintaining real-time efficiency.

3.1. Transfer Learning Mechanism

The core idea of transfer learning [40] is to transfer knowledge from a source task to a target task, improving model performance when target data are limited or labeling costs are high. In underwater and water-surface garbage detection, data collection is constrained by location, lighting, water quality, and equipment conditions, while annotation is costly and class distributions are often imbalanced. Scattering, refraction, and surface highlights further change target appearance and texture, making training from scratch prone to overfitting and unstable convergence. Therefore, transfer learning can reuse generic visual features such as edges, textures, and shapes from pretrained models, accelerating convergence and improving adaptability under data scarcity, domain shift, and long-tail distributions.

From the perspective of domain adaptation theory, the source-domain underwater garbage dataset and the target-domain water-surface garbage dataset can be regarded as two related domains that share the same object detection task but follow different marginal data distributions. Let $D_s = \{X_s, Y_s\}$ and $D_t = \{X_t, Y_t\}$ denote the source and target domains, respectively. Although the label spaces are related to aquatic garbage detection, the image distribution $P_s(X)$ differs from $P_t(X)$ because underwater images are mainly affected by light attenuation, scattering, turbidity, and color shift, whereas water-surface images are more strongly influenced by specular reflection, wave disturbance, local overexposure, and large-scale variations of floating objects. Such distribution discrepancy may cause feature bias and decision-boundary mismatch when the detector is directly transferred across domains.

The proposed multi-stage transfer learning framework does not use explicit domain adaptation losses, such as adversarial alignment, MMD, or correlation alignment. Instead, it achieves implicit alignment through progressive parameter transfer and supervised target-domain optimization. Source-domain pretraining enables the model to learn transferable aquatic features, including contours, weak textures, blurred boundaries, irregular shapes, and water-related background patterns, providing better initialization than random training.

During target-domain fine-tuning, the feature extractor, multi-scale fusion layers, classification branch, and regression branch are recalibrated toward the target distribution. LSKA enhances shared spatial context, ASFF reduces scale-distribution mismatch, and SlideLoss improves boundary regression under localization drift. Thus, the strategy performs supervised progressive alignment in feature and parameter spaces, reducing the gap between underwater and water-surface garbage images.

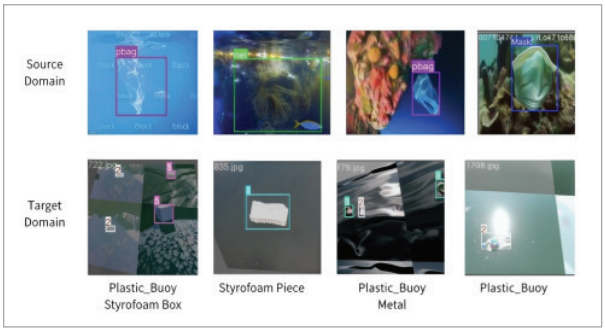
Transfer learning has improved object detection, especially in water-surface and underwater garbage detection, but several limitations remain. Many methods rely on public pretraining datasets such as COCO or ImageNet, whose object characteristics differ greatly from garbage in aquatic environments. Under scattering, light attenuation, reflection, low contrast, and irregular object shapes, these pretrained features may generalize poorly to real-world garbage detection.

In addition, some methods directly transfer full network weights without sufficient task-specific fine-tuning. Although this can provide initial gains, the domain shift between source and target data may cause unstable convergence and reduced accuracy in practical scenes. Therefore, targeted fine-tuning strategies are still needed to address distribution differences and improve adaptability in complex aquatic environments.

3.2. Source and Target Domain Datasets

This study uses a public underwater plastic pollution detection dataset as the source-domain data [25]. It contains 5130 YOLO-format annotated images affected by light attenuation, scattering, turbidity, and complex backgrounds, including 3628 training, 1001 validation, and 501 test images. The contrast-enhanced images cover 15 categories: Mask, can, cellphone, electronics, gbottle, glove, metal, misc, net, pbag, pbottle, plastic, rod, sunglasses, and tire, providing robust initialization for transfer learning.

Figure 5
Comparison of selected images from the source and target domain datasets.



A water-surface floating garbage dataset is used as the target-domain data [41]. It contains 2471 images at 2048 × 2048 resolution, including 1976 training, 248 validation, and 247 test images. The dataset includes six categories: Glass, Styrofoam_Piece, Metal, Rope, Plastic_Buoy, and Styrofoam_Box. Compared with the source domain, it has stronger ripples, reflections, overexposure, thin floating objects, and scale variation, making it suitable for evaluating cross-domain generalization and reflection robustness. Example comparisons are shown in Figure 5.

To improve experimental reproducibility, the dataset splits and augmentation settings are specified. The source-domain dataset contains 5130 images and was split into 3628 training, 1001 validation, and 501 test images, corresponding to approximately 70.72%, 19.51%, and 9.77%. The target-domain dataset contains 2471 images and was split into 1976 training, 248 validation, and 247 test images, corresponding to approximately 79.97%, 10.04%, and 10.00%. During training, fixed augmentation settings, including input resizing, HSV adjustment, translation, scaling, flipping, mosaic, mixup, and copy-paste, were applied. All experiments used the same splits and augmentation configuration for fair and reproducible evaluation.

Although the source and target datasets are both related to aquatic garbage detection, they differ significantly in imaging conditions, background distribution, and target morphology. To avoid relying only

on qualitative observations, a quantitative domain discrepancy analysis based on feature-level MMD and CORAL distances was further conducted in the experimental section. This analysis provides direct evidence for the existence of domain shift between the source and target datasets and supports the necessity of the proposed transfer learning strategy.

3.3. Multi-Stage Transfer Learning Framework

To address imaging variation, feature distribution gaps, and target-scale changes between source and target domains, this paper proposes a three-stage transfer learning framework. The model first learns general aquatic features from the source domain and then adapts to limited target-domain samples, improving generalization and adaptability in complex aquatic environments. The framework is shown in Figure 6.

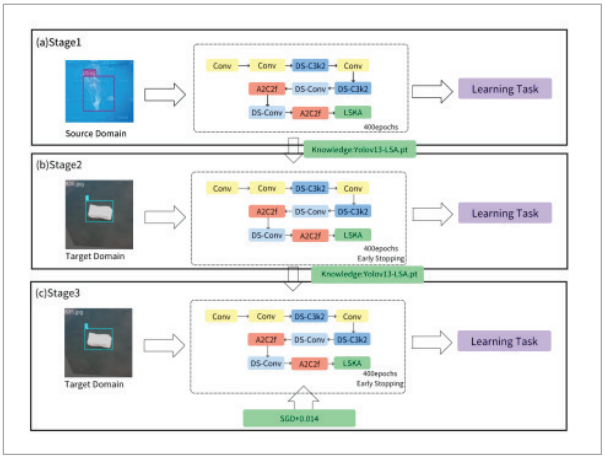
Figure 6(a) shows source-domain pretraining, where YOLOv13-LSA is trained for 400 epochs on 5130 underwater plastic pollution images to learn general aquatic features such as textures, contours, blurred boundaries, and low-contrast targets. LSKA, ASFF, and SlideLoss enhance spatial modeling, multi-scale fusion, and hard-sample regression, producing transferable source-domain weights.

Figure 6(b) shows target-domain transfer, where the pretrained weights are fine-tuned on 2471 water-surface floating garbage images with reflection, waves, overexposure, small debris, and scale variation. Early Stopping and momentum-based SGD are used to improve convergence and reduce overfitting.

Figure 6(c) shows final target-domain fine-tuning, which further adapts the model to complex water-surface backgrounds and weak target responses. This "generalization-alignment-specialization" process improves convergence, small-target detection, reflection robustness, and cross-domain stability.

Overall, the framework follows a "generalization-alignment-specialization" path: source-domain training learns general aquatic features, target-domain transfer aligns the distribution, and final fine-tuning strengthens task-specific detection, improving convergence, small-target detection, reflection robustness, and cross-domain stability.

Figure 6
Transfer learning model structure.



4. Experiment and Evaluation

4.1. Simulation Environment

4.1.1. Experimental Environment

For reproducibility, all experiments were conducted on the AutoDL platform under the same environment: Ubuntu 22.04/Linux, Python 3.10, PyTorch 2.1.0 with CUDA 12.1, one NVIDIA GeForce RTX 4090D GPU with 24 GB memory, Intel(R) Xeon(R) Platinum 8474C CPU with 15 vCPUs, 80 GB RAM, a 30 GB system disk, and a 50 GB data disk. This environment was used for ablation, comparison, transfer learning, and inference speed tests unless otherwise specified.

4.1.2. Parameter Settings

To improve convergence and detection accuracy, the model was trained on an RTX 4090D GPU for 400 epochs with batch size 32 and input size 640. SGD was used as the optimizer with lr = 0.0141, momentum = 0.937, weight_decay = 0.0005, warmup_epochs = 3, and lrf = 0.01. Cache and AMP were disabled, 8 workers were used for data loading, and checkpoints were saved every 10 epochs. Color, geometric, and MixUp augmentation were applied to improve generalization.

To improve reproducibility and generalization, fixed augmentation settings were used: imgsz = 640; hsv_h/s/v = 0.015/0.700/0.400; degrees/translate/scale/shear/perspective = 10.000/0.100/0.500/2.000/0.001; flipud/

fliplr = 0.000/0.500; mosaic/close_mosaic = 1.000/10; mixup/cutmix/copy_paste = 0.100/0.000/0.000; auto_augment/erasing = randaugment/0.400. Cache, AMP, multi_scale, and rect were disabled.

4.1.3. Evaluation Metrics

Precision and Recall evaluate detection correctness and target coverage. mAP@50 measures accuracy at IoU = 0.5, while mAP@50:95 averages results from IoU 0.5 to 0.95 for stricter localization assessment. Efficiency is evaluated by GFLOPs, training time, and FPS, which reflect computational cost, convergence efficiency, and real-time inference capability, respectively.

Precision (P) calculates the ratio of true positives (TP) to all positive detections (TP + FP) in the results, and is defined as:

$$P=\frac{TP}{TP+FP}, \tag{14}$$

where, TP denotes the number of true positives, and FP denotes the number of false positives. The higher the precision, the more reliable the model's detection results are. Recall (R) measures the proportion of true positives detected by the model among all actual positives (TP+FN), and is defined as follows:

$$R=\frac{TP}{TP+FN}, \tag{15}$$

where FN represents the number of false negatives. A higher recall rate indicates a more robust model because it is better able to cover target objects. In object detection, Mean Average Precision (mAP) provides a comprehensive evaluation. Specifically, mAP@50 uses an IoU threshold of 0.5 to measure the average accuracy across categories. It is calculated as follows and reflects how well the model detects and classifies objects across several categories:

$$mAP@50=\frac{1}{N}\sum_{i=1}^{N} AP_i,, \tag{16}$$

where N denotes the number of classes, and APi represents the average precision for class i. On top of that, mAP@50:95 represents the mean average precision determined across a spectrum of Intersection over Union thresholds spanning from 0.5 to 0.95, with the final figure being the average across ten distinct points in this range. This metric is used to comprehensively assess the model's detection capability. The calculation formula is as follows:

$$mAP@50:95=\frac{1}{10}\sum_{i=50}^{59} m\,AP_i. \tag{17}$$

Giga Floating Point Operations (GFLOPs) are a key metric used to measure the computational complexity of model inference, offering insight into network efficiency. For convolutional neural networks, the computational expense of a single layer can be represented as:

$$FLOPs=2\times C_{in}\times C_{out}\times K_H\times K_W\times H_{out}\times W_{out}, \tag{18}$$

where c_{in} and C_{out} represent the input and output channel counts, K_H and K_W are the kernel dimensions, and H_{out} and W_{out} indicate the dimensions of the output feature map:

The computing costs for each layer add up to the model's total FLOPs, and the computation for GFLOPs is written as follows:

$$GFLOPs=\frac{Total\ FLOPs}{10^9}. \tag{19}$$

A lower GFLOPs value indicates higher computational efficiency and faster inference speed of the model. Training time is used to measure the total duration required for the model to converge from the start of training, reflecting the training efficiency of the model. The calculation formula is expressed as:

$$T_{train}=\sum_{i=1}^{E} t_i/3600. \tag{20}$$

where E represents the total number of training epochs, t_i denotes the training time for the i-th epoch (in seconds), and the result is converted to hours by dividing by 3600. A shorter training time indicates higher model optimization efficiency, which allows the model to achieve faster convergence in fewer iterations, thereby improving overall experimental efficiency.

In addition to GFLOPs, the actual inference speed was measured using Frames Per Second (FPS) to further evaluate the deployment efficiency of the

model. GFLOPs reflects the theoretical computational complexity, while FPS directly measures the number of images processed per second under a specific hardware and inference environment. Therefore, FPS can provide a more practical evaluation of real-time performance.

In this study, FPS was tested with a batch size of 1 and an input image size of 640 × 640. Before recording the inference time, 50 warm-up iterations were performed to reduce the influence of GPU initialization and cache loading. The FPS is calculated as follows:

$$FPS=\frac{N}{T}, \tag{21}$$

where N denotes the number of tested images or repeated inference runs, and T denotes the total inference time in seconds. The average inference latency is calculated as:

$$Latency=\frac{T}{N}\times 1000, \tag{22}$$

where Latency is measured in milliseconds per image. A higher FPS and lower latency indicate better real-time deployment capability.

To improve the consistency of experimental reporting, all detection metrics in the comparison and ablation experiments are reported using the same decimal format unless otherwise specified. Precision, Recall, mAP@50, and mAP@50:95 are reported as decimal values in the range of 0 to 1, while training duration is reported in hours and computational cost is reported in GFLOPs. For repeated experiments, the mean value, standard deviation, and paired statistical test are used to describe performance stability and statistical reliability.

4.2. Model Validation

4.2.1. Training Process Analysis

The improved model was evaluated on a water-surface and underwater garbage detection dataset. The evaluation focused on training efficiency, convergence speed, and multi-scale feature integration. Numerical analysis, loss-function comparison, mAP evaluation, and qualitative results show that the enhanced model improves detection of different waste categories in underwater plastic-pollution environments.

Figure 7
Classification loss convergence graph.

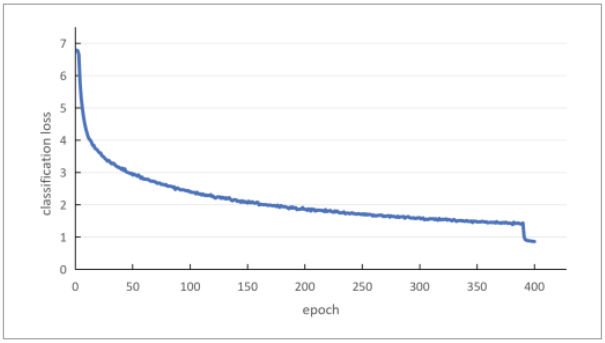
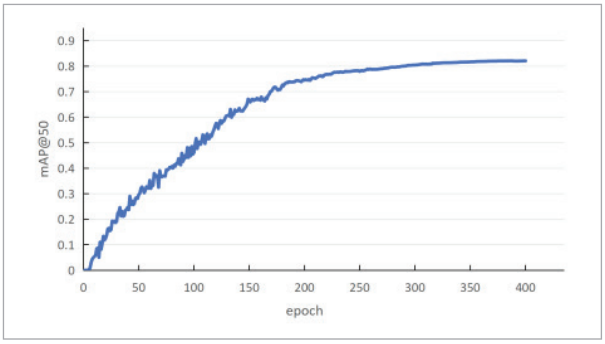


Figure 8
mAP@50 progression graph.



As shown in Figure 7, the classification loss decreases steadily during training, dropping from about 7.0 to around 3.0 within the first 50 epochs, then gradually converging to approximately 0.9-1.0 by 400 epochs with little fluctuation. Meanwhile, Figure 8 shows that mAP@50 increases rapidly in the early stage, reaches about 0.62 around 120 epochs, and finally stabilizes near 0.82. These trends indicate stable convergence and effective feature learning. YOLOv13-LSA was further evaluated using Precision, Recall, mAP@50, and mAP@50:95, showing improved detection performance, especially for small targets and complex aquatic backgrounds.

Table 1 shows the performance of YOLOv13-LSA in water-surface and underwater garbage detection. The model achieves a Precision of 0.830, a Recall of 0.754, and an mAP@50 of 0.821. Its mAP@50:95 reaches 0.467, indicating relatively stable localization under stricter IoU thresholds.

The low performance of Metal and Rod is mainly caused by visual ambiguity and data difficulty. Met-

Table 1
Evaluation metrics for each defect type of the final model.

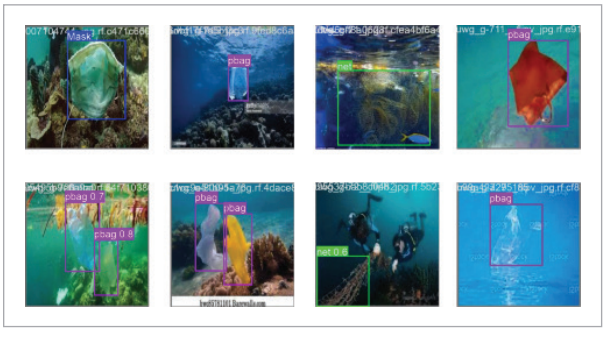
Class	Precision	Recall	mAP@50	mAP@50:95
Mask	0.98	0.552	0.804	0.561
Can	0.822	0.9	0.947	0.487
Cellphone	0.976	0.93	0.963	0.687
Electronics	0.685	0.75	0.738	0.366
Gbottle	0.719	0.683	0.716	0.464
Glove	0.884	0.831	0.945	0.665
Metal	0.589	0.455	0.488	0.228
Misc	0.714	0.831	0.774	0.389
Net	0.927	0.878	0.941	0.597
Pbag	0.852	0.976	0.975	0.695
Pbottle	0.833	0.793	0.837	0.452
Plastic	0.869	0.678	0.814	0.343
Rod	0.762	0.444	0.604	0.205
Sunglasses	1	0.98	0.995	0.532
Tire	0.837	0.624	0.772	0.334
Total	0.83	0.754	0.821	0.467

al objects often show strong reflection and background-like textures, while Rod objects are thin, elongated, and easily occluded, leading to weak features and localization uncertainty. In contrast, Sunglasses, Pbag, and Cellphone have clearer shapes or semantic boundaries, resulting in higher mAP@50 values. Thus, category-level differences reflect intrinsic detection difficulty rather than random model instability.

The Mask category has a relatively low Recall of 0.552, indicating missed detections. This is mainly due to visual ambiguity and limited sample diversity. In aquatic scenes, discarded masks are often small, thin, deformable, soaked, folded, partially submerged, or occluded by waves. Their weak edges and low-contrast textures may resemble foam, plastic bags, paper, or other light-colored debris, while reflection and background clutter further reduce discriminability. Future work will enrich mask samples and use category-balanced sampling, hard-example mining, targeted augmentation, and boundary-aware feature refinement to improve mask detection.

As shown in Figure 9, YOLOv13-LSA achieves remarkable results in the underwater and water surface garbage detection task, with particularly out-

Figure 9
Object detection results of the YOLOv13-LSA model.



standing performance in detecting small targets. The figure displays the model's detection results for various types of water surface garbage in different aquatic scenes. Specifically, YOLOv13-LSA can consistently recognize targets even in more complex backgrounds. It performs well in terms of detection accuracy and recall across different categories, effectively handling variations in water surface floating garbage. Overall, the YOLOv13-LSA model demonstrates robust detection capabilities in the underwater and water surface garbage detection task, further proving its adaptability and transferability in real-world environments, especially in scenes with significant lighting and background variations, showing better potential for practical applications.

4.2.2. Ablation Study

To evaluate the engineered integration of the selected components, ablation experiments were conducted on the water-surface and underwater garbage dataset using YOLOv13 as the baseline. SlideLoss was first introduced to replace the original CIoU loss, followed by LSKA for fine-grained feature modeling and ASFF for adaptive feature fusion, forming YOLOv13-LSA. The results in Table 2 demonstrate the contribution of each component to detection performance.

Table 2 shows that baseline YOLOv13 achieves 0.8322 Precision, 0.7145 Recall, 0.7954 mAP@50, 0.4714 mAP@50:95, and 6.1682 GFLOPs. SlideLoss improves Recall to 0.7425 and mAP@50 to 0.8019 without extra computation. LSKA further raises Recall to 0.7638 and mAP@50:95 to 0.4939, while ASFF gives YOLOv13-LSA the highest mAP@50 of 0.8209 and restores Precision to 0.8300.

The ablation results confirm the collaborative design of YOLOv13-LSA. SlideLoss improves regression stability and target coverage, LSKA enhances weak-feature representation and strict localization, and ASFF strengthens multi-scale detectability. These modules address different aquatic degradation problems at different detection stages.

The decrease in mAP@50:95 after ASFF reflects a trade-off. This stricter metric is sensitive to small boundary deviations, and ASFF may smooth fine boundary details during scale alignment. However, it improves mAP@50 and target detectability, which are important for robust aquatic garbage monitoring.

The ablation results do not improve monotonically because each metric emphasizes different aspects. Precision reflects false positives, Recall measures target coverage, mAP@50 evaluates general detectability, and mAP@50:95 stresses strict localization. Therefore, the Precision drop after LSKA and the mAP@50:95 decrease after ASFF are trade-offs rather than contradictions. YOLOv13-LSA is selected because it achieves the highest mAP@50 with balanced Precision and Recall.

Each module contributes differently. SlideLoss improves hard-sample regression for blurred boundaries without adding inference cost. LSKA expands the receptive field and enhances weak responses of small, low-contrast, or occluded targets. ASFF adaptively weights multi-scale features to improve scale variation handling, but it also adds computation and may slightly affect high-IoU localization.

The GFLOPs increase mainly comes from ASFF. SlideLoss adds no inference cost, and LSKA only increases GFLOPs from 6.1682 to 6.2231, while ASFF raises the final cost to 8.4064. Compared with baseline YOLOv13, YOLOv13-LSA increases GFLOPs by 36.29%, but improves mAP@50 from 0.7954 to 0.8209 and Recall from 0.7145 to 0.7536, which is acceptable for GPU-assisted real-time monitoring.

Table 2
Ablation study results.

Method	SlideLoss	LSKA	ASFF	P	R	mAP@50	mAP@50:95	GFLOPs
YOLOv13	-	-	-	0.8322	0.7145	0.7954	0.4714	6.1682
	√	-	-	0.8308	0.7425	0.8019	0.4666	6.1682
	√	√	-	0.8052	0.7638	0.8042	0.4939	6.2231
	√	√	√	0.8300	0.7536	0.8209	0.4670	8.4064

Table 3
Comparison results with YOLO-based, Transformer-based, and two-stage detectors.

Method	P	R	mAP@50	mAP@50:95	GFLOPs
NBA-YOLO	0.7943	0.7748	0.8137	0.4804	8.5855
AAE-YOLO	0.8423	0.7477	0.8182	0.4754	8.3822
FSA-YOLO	0.8303	0.7415	0.8090	0.4598	8.3518
ACG-YOLO	0.8207	0.7714	0.8205	0.5000	5.8143
RT-DETR	0.8285	0.7493	0.7922	0.5243	103.50
Faster R-CNN	0.8336	0.7051	0.7320	0.4220	134.00
Ours	0.8300	0.7536	0.8209	0.4670	8.4064

4.2.3. Comparison Experiment

To demonstrate the effectiveness of the YOLOv13-LSA model, we conducted a comprehensive comparative experiment on the surface/waterborne garbage dataset. This experiment aimed to evaluate the performance of the improved model in engineering detection applications. The comparison models include: NWD_BRA_ASFF (NBA-YOLO), ASFF_ATF_EMA (AAE-YOLO), FCA_Slide_ASFF (FSA-YOLO), A2C2f_SHSA+C3K2_EFAtt+GFPN (ACG-YOLO) [36], RT-DETR [39], Faster R-CNN [27], and the YOLOv13-LSA model proposed in this study. ACG-YOLO is retained only as a general YOLOv13 architectural variant for controlled comparison, while RT-DETR and Faster R-CNN were added as representative Transformer-based end-to-end and classical two-stage detectors, respectively. The results of the comparison experiments are recorded in Table 3.

Among the compared models, YOLOv13-LSA achieves the highest mAP@50 value of 0.8209, followed closely by ACG-YOLO with 0.8205 and AAE-YOLO with 0.8182, while NBA-YOLO obtains the highest Recall and AAE-YOLO obtains the highest Precision. However, the mAP@50 advantage of YOLOv13-LSA over ACG-YOLO is very small, and ACG-YOLO performs better in Recall, mAP@50:95, and GFLOPs. Therefore, the comparison results indicate that YOLOv13-LSA provides competitive overall detectability, but different models show different trade-offs in Precision, Recall, strict localization accuracy, and computational cost.

Compared with non-YOLO detectors, YOLOv13-LSA improves mAP@50 over RT-DETR from 0.7922 to 0.8209 while reducing computation from 103.50 to 8.4064 GFLOPs. RT-DETR achieves higher mAP@50:95, but requires much greater computational cost. Compared with Faster R-CNN, YOLOv13-LSA achieves higher Recall, mAP@50, and mAP@50:95 with far lower GFLOPs, showing better suitability for real-time aquatic garbage monitoring.

In terms of computational cost, RT-DETR requires approximately 12.31 times the GFLOPs of YOLOv13-LSA, while Faster R-CNN requires approximately 15.94 times. This indicates that although Transformer-based and two-stage detectors may have advantages in certain localization metrics, their computational burden is much higher, making YOLOv13-LSA more suitable for real-time aquatic monitoring scenarios.

The performance differences are related to module functionality. The retained YOLO-based comparison models mainly represent feature-fusion or YOLOv13 architectural variants; ACG-YOLO is included as an architectural reference rather than as an aquatic-garbage-specific detector. In contrast, YOLOv13-LSA combines SlideLoss for boundary-aware optimization, LSKA for long-range spatial representation, and ASFF for adaptive multi-scale fusion, which explains its competitive mAP@50 and Recall in complex aquatic scenes.

Compared with RT-DETR and Faster R-CNN, YOLOv13-LSA offers a better accuracy-efficiency balance. RT-DETR has strong localization but much higher GFLOPs, while Faster R-CNN requires more

computation and shows lower Recall and mAP@50. Therefore, YOLOv13-LSA is more suitable for real-time aquatic garbage detection.

Overall, the extended comparison demonstrates that YOLOv13-LSA is competitive not only with YOLO-based variants but also with representative non-YOLO detectors. Although RT-DETR shows an advantage in mAP@50:95, YOLOv13-LSA achieves the highest mAP@50 with significantly lower GFLOPs. Therefore, the proposed method provides a favorable accuracy-efficiency trade-off for water surface and underwater garbage detection, especially in scenarios where real-time or near-real-time deployment is required.

4.2.4. Inference Speed Analysis

To evaluate deployment feasibility, YOLOv13-LSA was tested on different hardware platforms. Since GFLOPs only reflects theoretical complexity, FPS and average latency were also measured to assess actual inference efficiency, which is affected by hardware, memory bandwidth, inference framework, preprocessing, and post-processing.

The input size was fixed at 640 × 640 with batch size 1. Tests were conducted on an NVIDIA GeForce RTX 4090D GPU and the CPU of the same server. After 50 warm-up iterations, average FPS was calculated from repeated inference runs, as shown in Table 4.

Table 4 reports YOLOv13-LSA inference efficiency. On the RTX 4090D, it achieves 113.35 FPS with 8.82 ms latency, meeting GPU-assisted real-time detection requirements. On CPU, speed drops to 9.47 FPS with 105.56 ms latency, showing the higher burden without GPU acceleration. Together with GFLOPs analysis, these results indicate that YOLOv13-LSA provides a feasible accuracy-efficiency balance for GPU-assisted aquatic garbage monitoring, while CPU-only or low-power edge deployment still requires further model compression and acceleration.

Inference speed shows the deployment trade-off. Although YOLOv13-LSA adds computation over the baseline, it reaches 113.35 FPS with 8.82 ms latency on the RTX 4090D, meeting GPU-assisted real-time monitoring needs. However, CPU speed is only 9.47 FPS, which may be insufficient for high-frame-rate applications. Thus, low-end embedded or CPU-only deployment still requires pruning, quantization, or knowledge distillation.

Table 4
Inference speed of YOLOv13-LSA on different hardware platforms. YOLOv13-LSA was tested with input size 640 x 640, batch size 1, FP32 precision, 8.4064 GFLOPs, and an 8.08 MB model size.

Item	NVIDIA GeForce RTX 4090D	CPU
Warm-up iterations	100	20
Test iterations	500	100
Latency/ms ↓	8.82	105.56
FPS ↑	113.35	9.47

4.3. Transfer Learning Results Analysis

The transfer learning results further support the above implicit domain alignment mechanism. The improvement in mAP@50 and recall after loading source-domain weights indicates that the aquatic representations learned from the source domain can be effectively adapted to the target-domain water-surface garbage distribution. Meanwhile, the reduction in training duration suggests that source-domain pretraining provides a more favorable optimization starting point, thereby accelerating convergence while maintaining target-domain detection performance.

4.3.1. Quantitative Analysis of Domain Discrepancy

Before evaluating transfer learning, domain discrepancy between the underwater source dataset and water-surface target dataset was quantified. ResNet-50 was used as a fixed feature extractor, and MMD and CORAL distance were calculated to measure feature-level distribution gaps, where larger values indicate stronger domain shift.

Intra-domain comparisons, including Source train vs Source val and Target train vs Target val, measured split consistency within each domain. Cross-domain comparisons, including Source train vs Target train and Source val vs Target val, measured the gap between source and target domains. The results are shown in Table 5.

According to Table 5, intra-domain discrepancies are very small, with MMD values of 0.001260 for Source train vs Source val and 0.000370 for Target train vs Target val. In contrast, cross-domain MMD increases to 0.195779 and 0.208731, while CORAL

Table 5
Quantitative domain discrepancy analysis between source and target datasets.

Compared datasets	Feature extractor	Number of images A	Number of images B	MMD ↓	CORAL ↓
Source train vs Source val	ResNet-50	3628	1001	0.001260	5.7952e-11
Target train vs Target val	ResNet-50	1976	247	0.000370	8.8463e-11
Source train vs Target train	ResNet-50	3628	1976	0.195779	6.7299e-10
Source val vs Target val	ResNet-50	1001	247	0.208731	7.7287e-10

distance shows the same trend, with cross-domain values about one order of magnitude larger. These results confirm a clear distribution gap between source and target datasets.

This supports the need for transfer learning. Target-only training may suffer from limited data and overfitting, while direct source transfer may introduce feature bias. Therefore, the proposed framework first learns general aquatic representations and then adapts them to the target domain through supervised fine-tuning.

4.3.2. Transfer Framework Validation

Based on the domain discrepancy analysis, the proposed transfer learning framework was evaluated on the target-domain dataset. Using the same data and training settings, target-only training (Before) was compared with the proposed source-pretraining and target-fine-tuning strategy (After). Each strategy was repeated five times to reduce the effects of random initialization and training fluctuation. The results are shown in Table 6.

Table 6 shows that the transfer learning framework improves water-surface and underwater garbage detection. Across five repeated experiments, average Precision increases from 80.11% to 82.07%, Recall from 72.84% to 78.75%, and mAP@50 from 80.30% to 82.91%. These gains indicate that source-domain pretraining provides better initialization and feature alignment for the target domain, improving detection under low contrast, reflection, and texture interference.

Training efficiency also improves, with average training time decreasing from 1.573 h to 1.482 h, a reduction of about 5.79%. The consistent gains across repeated runs show that the proposed transfer strategy improves accuracy, stability, and reproducibility without increasing training overhead.

To further evaluate the statistical reliability of the transfer learning improvement, the mean and standard deviation of the five repeated trials were calculated. A paired t-test was also conducted between the Before and After strategies for each metric, as shown in Table 7.

Table 6
Comparison of results before and after applying the transfer learning framework.

	Before				After			
Trial	P	R	mAP@50	Duration (h)	P	R	mAP@50	Duration (h)
1	0.8011	0.7372	0.8015	1.573	0.8203	0.7882	0.8287	1.488
2	0.8012	0.7682	0.8112	1.575	0.8201	0.7888	0.8306	1.469
3	0.8003	0.7011	0.8074	1.578	0.8223	0.7845	0.8307	1.472
4	0.8015	0.7201	0.7987	1.571	0.8219	0.7890	0.8299	1.493
5	0.8014	0.7152	0.7962	1.567	0.8189	0.7871	0.8257	1.486
Mean	0.8011	0.7284	0.8030	1.573	0.8207	0.7875	0.8291	1.482

Table 7
Statistical analysis of repeated transfer learning experiments.

Metric	Before (mean ± std)	After (mean ± std)	p-value	Significance
Precision	0.8011 ± 0.0005	0.8207 ± 0.0014	1.32e-5	$p < 0.01$
Recall	0.7284 ± 0.0257	0.7875 ± 0.0018	0.0057	$p < 0.01$
mAP@50	0.8030 ± 0.0062	0.8291 ± 0.0021	2.58e-4	$p < 0.01$
Duration (h)	1.5728 ± 0.0041	1.4816 ± 0.0105	1.20e-4	$p < 0.01$

Table 8
Comparison of performance with transfer learning.

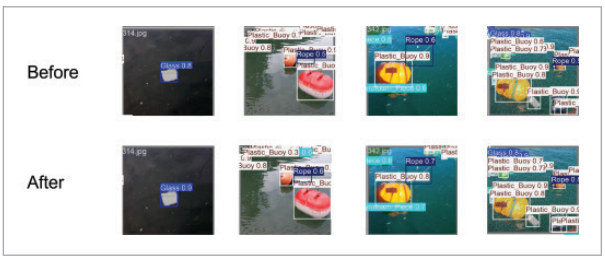
	P	R	mAP@50	Duration (h)
YOLOv13-LSA	0.8011	0.7372	0.8015	1.573
YOLOv13-LSA + LSA.pt	0.8203	0.7882	0.8287	1.488
YOLOv13-LSA + LSA.pt + Fine-tuning	0.8714	0.7731	0.8346	1.454

The statistical results confirm significant improvements after transfer learning. mAP@50 increases from 0.8030 ± 0.0062 to 0.8291 ± 0.0021 with p = 2.58e-4, showing that the gain is stable across repeated trials. Recall also improves significantly (p = 0.0057), while training duration decreases significantly (p = 1.20e-4), supporting the reliability and reproducibility of the framework.

As shown in Figure 10, fine-tuning improves detection of small and difficult categories such as Plastic Buoy, Rope, and Glass, increasing their detection rate from 0.8 to 0.9. The fine-tuned model also maintains stable accuracy across target scales, demonstrating better robustness in complex aquatic scenes.

Transfer learning has demonstrated significant performance improvements in water surface/underwater garbage detection. It not only enhanced category recognition and target localization accuracy but also expanded the effective detection range, particularly improving the recognition capability for small targets. The experimental results further validate its effectiveness in aquatic environments, providing solid technical support for intelligent garbage detection systems.

Figure 10
Comparison of detection results before and after fine-tuning.



4.3.3. Ablation Study of Transfer Learning

To evaluate the proposed transfer learning framework, three training strategies were compared using YOLOv13-LSA: (1) training from scratch on the target domain, (2) loading source-domain pretrained weights without further fine-tuning, and (3) loading source-domain pretrained weights followed by supervised target-domain fine-tuning. This comparison analyzes the effects of weight initialization and fine-tuning on aquatic garbage detection. Precision, Recall, mAP, and training duration are reported in Table 8.

The results in Table 8 indicate that using source-domain pretrained weights without fine-tuning improves Precision, Recall, and mAP@50 over the baseline by 1.92%, 5.10%, and 2.72%, respectively. This suggests that pretrained aquatic features provide useful initialization and accelerate adaptation to water-surface and underwater garbage detection.

After target-domain fine-tuning, YOLOv13-LSA + LSA.pt + Fine-tuning achieves an mAP@50 of 83.46%, improving by 0.59% over the baseline and

Table 9
Performance comparison of different models.

Model+Weights	P	R	mAP@50	Duration (h)
NBA-YOLO +s.pt	0.8508	0.7170	0.7994	1.760
AAE-YOLO +n.pt	0.8676	0.7305	0.8284	1.486
FSA-YOLO +x.pt	0.8670	0.7166	0.7906	1.496
ACG-YOLO +x.pt	0.8299	0.7251	0.7724	4.15
YOLOv13-LSA + LSA.pt	0.8203	0.7882	0.8287	1.488

showing better adaptation to the target distribution. Its training time is reduced from 1.573 h to 1.454 h, saving 0.119 h, or about 7.57%. These results indicate that transfer learning improves detection accuracy and training efficiency while supporting faster convergence.

4.3.4. Comparison Experiment of Transfer Learning

As shown in Table 9, this study examines the impact of transfer learning and model optimization using a controlled approach. Four comparison models--NBA-YOLO, AAE-YOLO, FSA-YOLO, and ACG-YOLO--were compared with YOLOv13-LSA. A detailed evaluation of precision, recall, mAP@50, and training time was conducted to assess the proposed method under a more task-relevant comparison setting.

Table 9 shows that YOLOv13-LSA achieves 82.03% Precision and 78.82% Recall in water-surface and underwater garbage detection. Although its Precision is lower than AAE-YOLO and FSA-YOLO, it maintains reliable accuracy while providing relatively strong target coverage. This indicates good performance for complex backgrounds and small-target detection.

YOLOv13-LSA achieves the highest mAP@50 among all comparison models, reaching 82.87%, which indicates strong overall detection accuracy. Its training time is 1.488 h, maintaining good efficiency while avoiding excessive training cost. Although NBA-YOLO and AAE-YOLO achieve slightly higher Precision, they do not exceed YOLOv13-LSA in Recall or mAP@50. Therefore, YOLOv13-LSA provides a better balance between accuracy and training efficiency for water-surface and underwater garbage detection.

4.4. Discussion

4.4.1. Advantages of the Proposed Method

The proposed YOLOv13-LSA integrates LSKA, ASFF, SlideLoss, and a multi-stage transfer learning strategy to improve detection accuracy, inference efficiency, and training stability in complex aquatic environments.

1. Detection accuracy: LSKA enhances spatial context modeling and improves sensitivity to small or weak targets. ASFF adaptively fuses multi-scale features to handle targets with different sizes and shapes. SlideLoss strengthens boundary alignment for elongated, blurred, or irregular objects. Together, these modules improve detection under low contrast, reflection, and complex backgrounds.
2. Detection speed: YOLOv13-LSA maintains a lightweight structure and avoids excessive computational overhead. The efficient designs of LSKA and ASFF help preserve fast inference, making the model suitable for GPU-assisted real-time aquatic garbage monitoring.
3. Training efficiency: The multi-stage transfer learning strategy first learns general aquatic representations from the source domain, then adapts the model to the target domain through fine-tuning. This process accelerates convergence, reduces training instability, and supports faster deployment.

4.4.2. Limitations and Future Work

Although YOLOv13-LSA improves aquatic garbage detection, limitations remain in robustness, annotation efficiency, and deployment adaptability.

1. Complex backgrounds: Performance may still degrade under strong reflection, backlighting, turbidity, shallow-water refraction, surface fluctuation,

and dense clutter. These factors blur boundaries and cause missed or false detections, so robustness in low-contrast and reflective scenes requires further improvement.

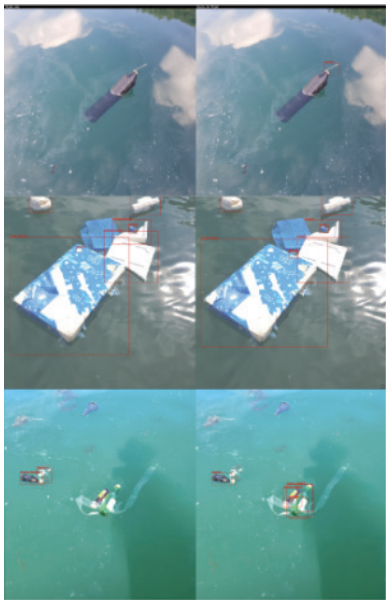
2 Annotation dependence: Accurate labels are still needed for small, blurred, or weakly salient objects. Although multi-stage transfer learning reduces this dependence by transferring general aquatic representations, representative target-domain annotations remain necessary. A comparison is shown in Figure 11.

Figure 11 shows that the source-to-target transfer model gives more stable responses than target-only training under weak boundaries, reflection, and background interference. However, transfer learning cannot replace representative target-domain labels; extremely scarce annotations may still cause overfitting and unstable generalization.

3 Hardware limitations: YOLOv13-LSA is efficient, but deployment on mobile, embedded, or in-vehicle devices may still be constrained by memory and computation. Future work will explore pruning, quantization, knowledge distillation, and hardware-aware optimization.

Figure 11

Qualitative comparison between target-only training and source-to-target transfer learning under limited target-domain annotations.



Overall, future work will improve robustness, reduce annotation cost, and enhance deployment through few-shot, semi-supervised, pseudo-labeling, active learning, domain adaptation, and adaptive image enhancement methods.

Future work will further investigate category imbalance and hard-category optimization, especially for visually ambiguous classes such as Metal and Rod. Class-balanced sampling, hard-example mining, boundary-aware annotation refinement, and category-specific data augmentation will be explored to improve the robustness of these difficult categories.

Future work will also include a more systematic comparison with self-supervised and large-scale pretraining-based detection paradigms. In particular, MAE-, DINO-, and DINOv2-pretrained backbones combined with controlled detector heads will be evaluated under the same aquatic garbage datasets and training schedules, so as to further analyze the role of representation learning in cross-domain small-object detection.

4.4.3. Practical Deployment Feasibility

In practical water-surface and underwater garbage monitoring, deployment feasibility depends on accuracy, real-time speed, hardware resources, and annotated target-domain data. YOLOv13-LSA achieves 113.35 FPS with 8.82 ms latency on the RTX 4090D, indicating feasibility for GPU-assisted real-time warning, floating-garbage tracking, and online inspection.

However, resource-constrained deployment remains challenging. Although the model has 8.4064 GFLOPs and a size of 8.08 MB, CPU inference reaches only 9.47 FPS, which may limit use on low-power embedded or CPU-only platforms. Pruning, quantization, TensorRT/ONNX optimization, lightweight backbones, and knowledge distillation are needed to improve efficiency.

Annotated data scarcity is another obstacle. Reflection, turbidity, occlusion, motion blur, small targets, and ambiguous boundaries make aquatic images difficult to label. Multi-stage transfer learning reduces this dependency through source-domain pretraining and target-domain fine-tuning, but representative target annotations are still required. Future deployment should combine semi-supervised learning, pseudo-labeling, active learning, and continual

domain adaptation to reduce labeling cost and handle changing water conditions.

4.4.4. Scene Expansion

YOLOv13-LSA may be extended to scenarios with weak boundaries, scale variation, and complex backgrounds, such as industrial defect detection, crop disease monitoring, and traffic sign detection.

1. Industrial defects: SlideLoss can improve boundary matching for cracks, bubbles, and scratches, while ASFF and LSKA enhance multi-scale and contextual feature representation.
2. Crop diseases: For spots, wilting, and insect damage with varied size and contrast, YOLOv13-LSA may improve localization through boundary-aware regression, multi-scale fusion, and spatial context modeling.
3. Traffic signs: For small, distant, blurred, or illumination-affected signs, ASFF, SlideLoss, and LSKA can support multi-scale detection, boundary refinement, and contextual enhancement.

Overall, YOLOv13-LSA shows potential for broader object detection tasks after scenario-specific optimization.

5. Conclusion

This study proposes YOLOv13-LSA with a multi-stage transfer learning framework to improve localization, reduce annotation dependence, and enhance adaptability in water-surface and underwater garbage detection. On the test dataset, YOLOv13-LSA achieves 0.8300 Precision, 0.7536 Recall, and 0.8209 mAP@50, with the main improvements observed in Recall and mAP@50, while slight trade-offs remain in Precision and mAP@50:95. Combined with transfer learning, the model improves robustness and generalization in complex aquatic scenes, providing an effective framework for intelligent garbage monitoring.

Future work will improve robustness, data efficiency, and deployment. Image enhancement, reflection suppression, and adaptive preprocessing will be explored for reflection, turbidity, low illumination, waves, and occlusion. Category-balanced sampling, hard-example mining, boundary-aware augmentation, and fine-grained features will be used to improve difficult categories such as Mask, Metal, and Rod. Semi-supervised learning, pseudo-labeling, active learning, few-shot detection, and continual domain adaptation will reduce annotation dependence. Lightweight ASFF, pruning, quantization, knowledge distillation, and TensorRT/ONNX acceleration will be explored to support future low-power edge deployment.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 62406297)

References

1. Chen, J., Kao, S. H., He, H., Zhuo, W., Wen, S., Lee, C. H., Chan, S. H. G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 12021-12031. https://doi.org/10.1109/CVPR52729.2023.01157
2. Chen, L., Zhu, J. Water Surface Garbage Detection Based on Lightweight YOLOv5. Scientific Reports, 2024, 14(1), 6133. https://doi.org/10.1038/s41598-024-55051-3
3. Chu, Y., Wang, J., Ma, L., Wu, C. LMSFA-YOLO: A Lightweight Target Detection Network in Remote Sensing Images Based on Multiscale Feature Fusion. Journal of King Saud University Computer and Information Sciences, 2025, 37(4), 63. https://doi.org/10.1007/s44443-025-00069-4
4. Đuraš, A., Wolf, B. J., Ilioudi, A., Palunko, I., De Schutter, B. A Dataset for Detection and Segmentation of Underwater Marine Debris in Shallow Waters. Scientific Data, 2024, 11(1), 921. https://doi.org/10.1038/s41597-024-03759-2
5. Feng, Y., Huang, J., Du, S., Ying, S., Yong, J. H., Li, Y. Hyper-YOLO: When Visual Object Detection Meets Hypergraph Computation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 47(4), 2388-2401. https://doi.org/10.1109/TPAMI.2024.3524377
6. Folkman, L., Pitt, K. A., Stantic, B. A Data-Centric Framework for Combating Domain Shift in Under-

water Object Detection with Image Enhancement. Applied Intelligence, 2025, 55(4), 272. https://link.springer.com/article/10.1007/s10489-024-06224-0 https://doi.org/10.1007/s10489-024-06224-0

7. Girshick, R., Donahue, J., Darrell, T., Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81
8. Girshick, R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2015, 1440-1448. https://doi.org/10.1109/ICCV.2015.169
9. Guo, L., Zhang, Y., Tian, Q., Ran, Y. Improved YOLOv5s and Transfer Learning for Floater Detection. Science Progress, 2025, 108(2), 00368504251342075. https://doi.org/10.1177/00368504251342075
10. Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., Zhang, S. H., Martin, R. R., Cheng, M. M., Hu, S. M. Attention Mechanisms in Computer Vision: A Survey. Computational Visual Media, 2022, 8(3), 331-368. https://doi.org/10.1007/s41095-022-0271-y
11. Hou, Q., Zhou, D., Feng, J. Coordinate Attention for Efficient Mobile Network Design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350
12. Jia, T., Kapelan, Z., De Vries, R., Vriend, R., Peereboom, E. C., Okkerman, I., Taormina, R. Deep Learning for Detecting Macroplastic Litter in Water Bodies: A Review. Water Research, 2023, 231, 119632. https://doi.org/10.1016/j.watres.2023.119632
13. Jiang, Z., Wu, B., Ma, L., Zhang, H., Lian, J. APM-YOLOv7 for Small-Target Water-Floating Garbage Detection Based on Multi-Scale Feature Adaptive Weighted Fusion. Sensors, 2023, 24(1), 50. https://doi.org/10.3390/s24010050
14. Zhang, J., Jiang, F., Chen, Y., Wu, W., Wu, Q. A Water Surface Garbage Recognition Method Based on Transfer Learning and Image Enhancement. Results in Engineering, 2023, 19, 101340. https://doi.org/10.1016/j.rineng.2023.101340
15. Khriss, A., Elmiad, A. K., Badaoui, M., Barkaoui, A. E., Zarhloule, Y. Exploring Deep Learning for Underwater Plastic Debris Detection and Monitoring. Journal of Ecological Engineering, 2024, 25(7), 58-69. https://doi.org/10.12911/22998993/187970
16. Lau, K. W., Po, L. M., Rehman, Y. A. U. Large Separable Kernel Attention: Rethinking the Large Kernel Attention Design in CNN. Expert Systems with Applications, 2024, 236, 121352. https://doi.org/10.1016/j.eswa.2023.121352
17. Lei, M., Li, S., Wu, Y., Hu, H., Zhou, Y., Zheng, X., Ding, G., Du, S., Wu, Z., Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. arXiv Preprint arXiv:2506.17733, 2025. https://doi.org/10.48550/arXiv.2506.17733
18. Li, N., Huang, H., Wang, X., Yuan, B., Liu, Y., Xu, S. Detection of Floating Garbage on Water Surface Based on PC-Net. Sustainability, 2022, 14(18), 11729. https://doi.org/10.3390/su141811729
19. Li, X., Tian, M., Kong, S., Wu, L., Yu, J. A Modified YOLOv3 Detection Method for Vision-Based Water Surface Garbage Capture Robot. International Journal of Advanced Robotic Systems, 2020, 17(3), 1729881420932715. https://doi.org/10.1177/1729881420932715
20. Lin, Y., Zhang, Y., Yang, D., Zhou, Y., Wang, L. DLFE-YOLO: An Enhanced Framework for Underwater Object Detection Based on YOLOv8. Intelligent Marine Technology and Systems, 2025, 3(1), 37. https://doi.org/10.1007/s44295-025-00088-x
21. Liu, S., Huang, D., Wang, Y. Learning Spatial Fusion for Single-Shot Object Detection. arXiv Preprint arXiv:1911.09516, 2019. https://doi.org/10.48550/arXiv.1911.09516
22. Liu, Y., Shao, Z., Teng, Y., Hoffmann, N. NAM: Normalization-Based Attention Module. arXiv Preprint arXiv:2111.12419, 2021. https://doi.org/10.48550/arXiv.2111.12419
23. Mittal, P. A Comprehensive Survey of Deep Learning-Based Lightweight Object Detection Models for Edge Devices. Artificial Intelligence Review, 2024, 57(9), 242. https://doi.org/10.1007/s10462-024-10877-1
24. Nabahirwa, E., Song, W., Zhang, M., Fang, Y., Ni, Z. A Structured Review of Underwater Object Detection Challenges and Solutions: From Traditional to Large Vision Language Models. arXiv Preprint arXiv:2509.08490, 2025. https://doi.org/10.48550/arXiv.2509.08490
25. Neural Ocean. NeuralOceanDataset, Apr. 2023. Accessed: 27 February 2026. https://universe.roboflow.com/neural-ocean/neuraloceandataset
26. Ren, S., Hao, H., Zhang, Y., Yu, Z. Underwater Complex Environment Domain Adaptation for Few-Shot Object Detection Based on Transfer Learning. Neurocomputing, 2025, 132341. https://doi.org/10.1016/j.neucom.2025.132341

27. Ren, S., He, K., Girshick, R., Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6), 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

28. Situ, Z., Teng, S., Feng, W., Zhong, Q., Chen, G., Su, J., Zhou, Q. A Transfer Learning-Based YOLO Network for Sewer Defect Detection in Comparison to Classic Object Detection Methods. Developments in the Built Environment, 2023, 15, 100191. https://doi.org/10.1016/j.dibe.2023.100191

29. Yi, Z., Lian, J., Liu, Q., Zhu, H., Liang, D., Liu, J. Boosting R-CNN: Reweighting R-CNN Samples by RPN's Error for Underwater Object Detection. Neurocomputing, 2023, 530, 150-164. https://doi.org/10.1016/j.neucom.2023.02.026

30. Tan, M., Le, Q. EfficientNetV2: Smaller Models and Faster Training. International Conference on Machine Learning. PMLR, 2021, 10096-10106. https://doi.org/10.48550/arXiv.2104.00298

31. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 7464-7475. https://doi.org/10.1109/CVPR52729.2023.00721

32. Wei, L., Tong, Y. Enhanced-YOLOv8: A New Small Target Detection Model. Digital Signal Processing, 2024, 153, 104611. https://doi.org/10.1016/j.dsp.2024.104611

33. Wu, C. M., Sun, Y. Q., Wang, T. J., Liu, Y. Underwater Trash Detection Algorithm Based on Improved YOLOv5s. Journal of Real-Time Image Processing, 2022, 19(5), 911-920. https://doi.org/10.1007/s11554-022-01232-0

34. Ye, Z., Peng, X., Li, D., Zhao, X., Shi, F. Adaptive Multi-Scale Attention Networks Fusion in YOLOv12-LSE: A Lightweight Framework for Efficient and Robust Underwater Object Detection. Applied Optics, 2025, 64(23), 6733-6745. https://doi.org/10.1364/AO.568505

35. Zhang, J., White, T. S., Zhang, H., Hu, W., Wunsch, D. C., Liu, J. MS-YOLO: Infrared Object Detection for Edge Deployment via MobileNetV4 and SlideLoss. 2025 International Joint Conference on Neural Networks (IJCNN). IEEE, 2025, 1-8. https://doi.org/10.1109/IJCNN64981.2025.11228885

36. Zhang, L., Zhao, S., An, D., Wang, P., Guo, B., Sun, H. Marine Ship Detection from GF-2 High-Resolution Remote Sensing Images with Improved YOLOv13 Model. Frontiers in Environmental Science, 2025, 13, 1707611. https://doi.org/10.3389/fenvs.2025.1707611

37. Zhang, Y., Ren, W., Zhang, Z., Jia, Z., Wang, L., Tan, T. Focal and Efficient IoU Loss for Accurate Bounding Box Regression. Neurocomputing, 2022, 506, 146-157. https://doi.org/10.1016/j.neucom.2022.07.042

38. Zhao, L., Ren, X., Fu, L., Yun, Q., Yang, J. UWS-YOLO: Advancing Underwater Sonar Object Detection via Transfer Learning and Orthogonal-Snake Convolution Mechanisms. Journal of Marine Science and Engineering, 2025, 13(10), 1847. https://doi.org/10.3390/jmse13101847

39. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q. DETRs Beat YOLOs on Real-Time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, 16965-16974. https://doi.org/10.1109/CVPR52733.2024.01605

40. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H. A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE, 2020, 109(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

41. 2401_88440984. Water Surface Floating Garbage YOLO Object Detection Dataset, 2023. Accessed: 27 February 2026. https://www.kaggle.com/search?q=Water+surface+floating+garbage+yolo+object+detection+dataset

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).