scientific reports

Check for updates

OPEN

An enhanced YOLOv8 model for accurate detection of solid floating waste

Juxing Di, Kaikai Xi & Yang Yang✉

To address the challenges in floating waste detection on water surfaces, such as small object scale, irregular shapes, and strong background interference, this study proposes an enhanced detection model based on the YOLOv8s frame work, named ES-YOLOv8. The new model optimizes the feature fusion strategy in the neck, constructing a refined "160-80-40-20" multiscale detection frame work. Integrated with the Efficient Multiscale Attention (EMA) module, it significantly improves the model's ability to extract features of small float ing objects. Additionally, an innovative Shape-IoU loss function is employed to optimize the bounding box regression accuracy of irregular targets through shape-sensitive constraints. This results in the development of an enhanced model that integrates feature enhancement, interference suppression, and localization optimization. Experimental results in a self-constructed floating waste dataset demonstrate that, compared to baseline YOLOv8s, the ES-YOLOv8 algorithm improves mAP@0.5 and mAP@0.5:0.95 by 5.4% and 6.1%, respectively. Compar ative experiments with state-of-the-art models further validate its superiority and effectiveness. Furthermore, experiments conducted on public datasets confirm the robustness and generalizability of ES-YOLOv8. This study aims to provide a high-precision, low-power-consumption technological solution for intelligent water governance, offering potential ecological and engineering applications.

Keywords YOLOv8s, Floating Waste Detection, Feature Extraction, Loss Functions, Attention Mechanism

With the acceleration of urbanization and the rapid development of industrialization, the problem of water Surface Garbage pollution is becoming increasingly severe. Floating waste severely damages water resources and the ecological environment and also has a significant impact on the river landscape. Therefore, the timely cleanup of floating waste is one of the key tasks in water environmental management[1]. Traditional methods of cleaning floating waste primarily rely on manual patrols and salvage. However, these approaches are often inefficient and costly, making it challenging to meet the growing demands for efficient detection and intelligent management. The advancement of computer vision technology presents an opportunity to transform the traditional manual management model. Currently, water surface cleaning robots and autonomous vessels are increasingly being deployed for floating waste removal[2]. These technologies enable the automated identification and processing of floating debris, significantly enhancing retrieval efficiency while reducing operational costs.

The primary task of floating garbage retrieval based on computer vision technology is to accurately detect and identify the floating garbage within complex riverine environments. Currently, floating object detection and recognition methods can be broadly classified into two main categories:traditional image processing techniques and deep learning-based object detection. Traditional image processing mainly relies on the foreground of the water surface environment, background features, and filtering theories for target recognition. Henriques et al.[3] used the Kernel Correlation Filter (KCF) algorithm to detect and track floating objects. By employing classifier training, target detection, and model update algorithms, improving the accuracy of tracking floating objects in complex river scenes. Xie et al.[4] employed the Kalman filter and Gaussian Mixture Model to enhance the effectiveness of motion tracking. Ding et al.[5] developed an Adaptive Pipeline Filter (APF) that leverages temporal correlation and motion information to refine the detection outcomes obtained from the Single Shot MultiBox Detector for infrared small target detection, achieving an impressive recall rate of 90%.

Traditional image processing offers the advantage of fast detection and recognition, however, it is highly susceptible to environmental interferences such as small target sizes, significant illumination variations, occlusions, and slow-moving objects. These challenges often result in false positives and missed detections, making it difficult to achieve the robustness required for reliable performance. Deep learning-based object detection algorithms leverage multi-layer convolutional neural networks to extract features of floating objects,

Hebei University of Architecture, Information Engineering College, Zhangjiakou 075000, China. ✉email: yy1977@hebiace.edu.cn

effectively addressing the challenges encountered in traditional image processing. These algorithms can be broadly categorized into two types: two-stage object detection methods based on region selection and single-stage object detection methods based on regression approaches. Two-stage object detection algorithms primarily include the Region-based Convolutional Neural Network (R-CNN)[6], Fast R-CNN[7], and related variants. Xie et al.[8] proposed an Oriented Region Proposal Network designed to generate oriented proposals that more accurately capture the geometric features of objects in images. Building upon the Oriented RPN, they developed a simple effective framework for oriented object detection, known as Oriented R-CNN. Cui et al.[9] proposed an improved Mask R-CNN algorithm, which integrates Mask R-CNN with a dust feature enhancement module based on the hue, lightness, and saturation color space. They demonstrated the feasibility and effectiveness of this framework, which significantly reduces the misjudgment and omission of dust areas, thereby improving the confidence level of detection. Minh et al.[10] proposed an algorithm based on the Mask R-CNN model for the detection and quantification of floating waste in images. This algorithm is designed to be effectively applied to the automated task of monitoring and quantifying floating waste along riverbanks.

Although the aforementioned research has significantly enhanced detection accuracy across various fields, achieving high precision, it suffers from relatively slow detection speed and high computational complexity. These limitations hinder its applicability for real-time detection of floating waste on water surfaces and impose stringent hardware requirements. For real-time applications such as water surface cleaning robots and autonomous vessels, these methods struggle to meet the demands of efficient and timely detection.

Single-stage object detection methods based on regression primarily include Single Shot MultiBox Detector (SSD)[11], you only look once (YOLO)[12], YOLOv5[13], YOLOv8[14], YOLOv10[15], YOLOv11[16], etc. These methods achieve significantly faster detection speeds compared to two-stage object detection algorithms. Jiang et al.[17] proposed the APM-YOLOv7 method for small target detection, which includes an adaptive algorithm for river contour extraction. This method enhances the model's ability to extract features from small targets, highlights the characteristics of small target debris, and reduces the probability of missed detections. However, its performance is limited by the scene constraints of the self-made dataset, and there is an issue of uneven detection performance for individual classes. Zhao et al.[18] introduced an enhanced YOLOX-S model capable of effectively recognizing various waste components within complex settings, achieving a notable mAP of 85.02%. Chen et al.[19] presented a streamlined YOLOv5 algorithm for water surface garbage detection, optimized for deployment on unmanned vessels. The model achieved its lightweight design primarily through the shufflenetv2 network architecture, resulting in a 93% reduction in parameters and a 9.5% FLOPs count relative to the original model. Shi et al.[20] developed a floating debris detection algorithm based on CDW-YOLOv8, enhanced by the Coordinate Attention mechanism and the Focaler Wise-IOU v3 loss function, which led to significant improvements across accuracy, recall rate, mAP@0.5, and mAP@0.5:0.95. Son et al.[21] evaluated the performance of cutting-edge AI models, including Mask R-CNN and YOLOv8, in enhancing plastic waste sorting, highlighting the criticality of choosing the right model to fit specific application needs.

For the detection of small targets, efficient image preprocessing algorithms can significantly enhance detection accuracy. Song et al.[22] introduced a Quantitative Augmentation strategy that effectively corrects the feature distribution of remote sensing data. Compared to traditional data augmentation techniques, this approach markedly improves the classification performance of Convolutional Neural Networks (CNN) and Vision Transformers (ViT). In addition, their research team has proposed numerous advanced methods in image preprocessing, including: RE-EfficientNet based on effective combination of data augmentation[23], the Variance Consistency Learning strategy[24], optimized Data Distribution Learning approach[25], Hybrid-Model Knowledge Distillation technique[26], Dual-Convolutional Neural Network Fusion method[27], the Quantitative Augmentation Learning strategy[28], and Quantitative Regularization combine with Vision Transformers[29]. These image preprocessing methods are essential for computer vision tasks involving image classification. By optimizing images for analysis and pattern recognition, they significantly enhance the performance and accuracy of classification models.

Although previous studies have significantly improved the accuracy of small target detection across various fields, the detection of floating debris on water surfaces still presents the following challenges: (1) In water surface environments, the high density and diversity of floating debris, along with background interference under varying lighting conditions, can affect feature extraction. As a result, the detection accuracy of the aforementioned methods still has significant room for improvement. (2) From the perspective of object detection, small floating waste contains limited features and undergoes significant appearance changes during its floating process, the samples are shown in Figure 1. These factors negatively impact the performance of deep learning algorithms, leading to suboptimal detection accuracy. (3) The limited availability of large-scale public datasets in this field leads to insufficient training data for deep learning models. This data scarcity hinders comprehensive algorithm validation and restricts the enhancement of their generalization capabilities. Therefore, in the field of floating waste detection, constructing comprehensive datasets and developing advanced methods and optimization strategies are essential. These efforts will improve the detection accuracy and generalization capability of deep learning models, thereby better supporting the effective management of aquatic floating waste.

To better address the aforementioned issues, this study selects YOLOv8s as the baseline model. YOLOv8[14] builds upon the strengths of the YOLO series and incorporates multiple significant improvements, enhancing its performance, flexibility, and robustness. In this work, we propose a multiscale feature fusion network, ES-YOLOv8, based on YOLOv8s. The proposed model effectively identifies multi-object solid waste in complex water surface environments, offering high accuracy, efficiency, versatility, and strong support for real-time applications and edge devices. The main contributions of this study are as follows.

(1) **Dataset Expansion and Fine-Grained Annotation.** To address the limitations of existing public datasets for floating debris detection, such as small dataset sizes, limited diversity in annotated targets, and the risk of overfitting during training, this study expands upon the River Floating Debris Dataset and IWHR-AI-

(a) Dense floating waste (b) Marginal waste (c) Occluded floating waste

Fig. 1. Complex floating waste detection scene. (**a**) Dense floating waste obscured by abundant aquatic plants. (**b**) Marginal waste in a wide water area. (**c**) Tall vegetation along the water's edge obstructs sight.

Label-Floater-V1. A new dataset is constructed, incorporating images captured under diverse environmental conditions, including overcast, rainy, and high-glare scenarios, as well as challenges such as water disturbances, tree shadows along riverbanks, and varying floating debris sizes. The dataset is expanded to a total of 2,711 images, with a fine-grained multiscale, multi-object annotation scheme. The annotation categories are refined from a single-object class into 12 distinct target classes, resulting in a total of 9,088 labeled floating debris instances. Such a comprehensive dataset is relatively rare in publicly available water surface debris datasets.

(2) **multiscale Feature Enhancement for Small Object Detection.** Given the limitations of YOLOv8s in detecting small objects, this study proposes a multiscale feature enhancement method. Specifically, the Neck module is optimized with an improved feature fusion strategy, and an additional detection head with a resolution of 160×160 pixels is introduced to better capture small objects. By integrating an attention mechanism, the model's ability to extract features from small floating debris is significantly enhanced, reducing both missed detections and false positives.

(3) **Integrating the EMA Module into YOLOv8s.** To further improve detection robustness, an EMA module is incorporated into the Neck module. The EMA attention mechanism smooths and weights feature maps, allowing the model to focus more effectively on task-relevant information. This enhances the model's ability to refine low-quality anchor boxes and accelerates convergence. Consequently, the model demonstrates improved capability in extracting key features from complex backgrounds, reducing background interference, and enhancing the accuracy and robustness of detecting small floating debris.

(4) **Shape-IoU Loss function for Irregular Floating Debris.** Given the irregular shapes of floating debris, this study employs the Shape-IoU loss function to improve target recognition and sensitivity. By emphasizing the shape and size of floating debris during loss computation, the proposed method optimizes the bounding box regression process. This approach significantly enhances the precision of localization and classification for water surface debris, improving the robustness of the model against morphological variations and ensuring greater stability during the training process.

Related work

Deep learning-based object detection algorithms employ multi-layer convolutional neural networks to extract features of floating debris, followed by target classification and position regression. Object detection-based waste classification and recognition play a crucial role in advancing sustainable solid waste management. However, when target objects are small in scale and densely distributed, feature information may become incomplete or insufficient, posing a significant challenge in distinguishing different categories of solid waste in multi-target detection.

With a strong focus on small and dense object detection, numerous experts and scholars have made significant contributions to this field through comprehensive research. Liu et al.[30] have developed an Inverted Residual multiscale Dilation Network. This model leverages efficient feature transformations and multiscale dilated attention mechanisms to diminish the interference of background noise, broaden the receptive field, and enhance the detection of small targets, thereby overcoming the limitations of local contrast methods in complex scenes. Zhang et al.[31] have conducted research on small object detection by building upon the YOLOv8 framework. In their study, they introduce the GIoU loss function to alleviate class imbalance and to bolster the model's robustness, particularly in environments with skewed class distributions. Hui et al.[32] have introduced a novel small target detection algorithm for UAV remote sensing images. This model integrates the SwinTransformer with CNNs to create an innovative convolutional architecture that fortifies feature extraction, catering to the demand for swift and precise recognition of small objects. Addressing the challenge of boundary discontinuity in synthetic aperture radar ship detection, Peng et al.[33] have outlined the necessary conditions that encoding methods and loss functions of detection networks must meet to tackle this issue. Furthermore, they have devised a continuous encoding method known as the Coordinate Decomposition Method to achieve optimal detection outcomes.

In addition, attention mechanisms are widely used in small object detection within computer vision tasks. They enable models to autonomously learn and weight different input features by computing the correlation between them, making the model more sensitive to important regions in an image. Zhou et al.[34] have developed

an anchor-based object detection system specifically for identifying solid waste in aerial photographs. Their research introduces the Efficient Attention Fusion Pyramid Network, which is designed to extract contextual and multiscale geospatial information through a process of attention fusion. The proposed detector attains a mAP of 63.12%, showcasing its impressive performance in the detection of solid waste from aerial imagery, which is considered a high detection accuracy in the field of small target detection. Ma et al.[35] have introduced an enhanced model for solid waste detection, which integrates the Convolutional Block Attention Module (CBAM) and Contextual Transformer Networks into the YOLOv5 architecture. CBAM significantly enhances the model's ability to extract deep channel-related features and spatial attention cues, which are crucial for accurately identifying small or partially obscured waste in complex background image scenes. Cao et al.[36] have developed a streamlined algorithm for target detection in side-scan sonar imagery, termed multiscale Attention-based YOLO, which harnesses the power of multiscale feature fusion in conjunction with an attention mechanism. The EMA module[37] is employed to enhance the model's feature extraction proficiency while simultaneously reducing computational overhead. Zhou et al.[38] have devised a YOLO-based model for the detection of marine organisms that incorporates a dual-terminal attention mechanism, which adaptively compresses noisy feature map channels, resulting in a 10% improvement in mAP@0.5.

In YOLO-based object detection algorithms, the loss function measures the discrepancy between the model's predictions and the ground truth labels. By optimizing the loss function during training, the model gradually refines its parameters, enhancing prediction accuracy. With a keen focus on in-depth research into loss functions, numerous experts and scholars have made substantial contributions to the field. Yang et al.[39] proposed an improved YOLOv8 object detection algorithm integrating feature enhancement and attention mechanisms. By utilizing the slide loss function to refine classification loss, the algorithm better captures challenging example features, achieving a 3.4% increase in mAP@0.5. Wang et al.[40] proposed an improved model based on YOLOv8n, replacing the Complete Intersection over Union (CIoU) loss function with the Shape-IoU[41] bounding box loss function. This modification enhances the model's object localization capability and accelerates convergence. As a result, the new model achieved a mAP of 92.4%. Zheng et al.[42] proposed a full-stage network based on YOLO with an auxiliary focal loss and multi-attention modules for underwater garbage detection. The auxiliary focal loss function addresses the issue of imbalance between positive and negative samples, focusing on learning from difficult samples while improving overall detection accuracy. This approach is suitable for real-time object detection of underwater garbage in complex backgrounds. Yue et al.[43] proposed a small target detection algorithm for complex environments, integrating Shape-IoU with the YOLOv8n framework. Shape-IoU improves localization precision and shape matching, resulting in a 1.5% increase in small target detection accuracy.

Although the aforementioned studies have made progress in the detection of floating waste on water surfaces, several challenges remain. Background noise issues, such as water wave disturbances, reflections, and strong light glare, along with limited target feature information, diverse shapes, and uneven distribution, continue to hinder the detection accuracy in complex environments. Additionally, floating waste detection faces challenges such as high computational resource consumption, limited dataset diversity, and poor generalization capability. To further enhance the accuracy, robustness, and generalization capability of water floating waste detection, this study proposes an enhanced multiscale feature fusion network for solid waste detection, named ES-YOLOv8, based on YOLOv8s. Built upon a self-constructed dataset, the proposed approach significantly enhances detection accuracy while preserving real-time performance. In addition, it strengthens the robustness and generalization capacity of the model.

Principles and methods

Overview of the enhanced YOLOv8 network

YOLOv8, an open source release by Ultralytics in 2023, represents a major update after YOLOv5. As one of the most advanced object detection algorithms available, YOLOv8 features a lightweight design, high precision, and efficiency. Considering its outstanding practical performance, such as stability, robustness, scalability, and lightweight nature,this study conducts further research based on the YOLOv8 framework.

The YOLOv8 series comprises five versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l and YOLOv8x. These models differ in terms of depth, width, and maximum number of channels, resulting in variations in accuracy, parameter size, and computational complexity. Considering the deployment requirements for floating waste detection in edge devices such as embedded systems and autonomous vessels,where model lightweighting and real-time performance are critical,this study selects YOLOv8s as the baseline model for algorithmic improvements, balancing detection accuracy and computational complexity.

To enhance the precision of floating waste detection, this study proposes a multiscale feature fusion network, ES-YOLOv8, based on YOLOv8s. The proposed model effectively identifies multiple solid waste objects in complex water environments, offering advantages such as high accuracy, efficiency, versatility, and strong support for real-time applications and edge devices. The general framework of the algorithmic model is illustrated in Figure 2.

Specifically, ES-YOLOv8 introduces several network optimizations and design improvements over the original YOLOv8s model:

(1) **Reconstruction of the multiscale Network Structure:** To address the challenges associated with floating waste detection, including small target size and sparse distribution, this study proposes a multiscale feature enhancement method. By optimizing the feature fusion strategy in the neck section and introducing an additional small object detection head at a resolution of 160×160, the model significantly enhances its ability to perceive small targets. This improvement extends the original FPN-PAN structure by incorporating the reuse of shallow features from the backbone output, corresponding to a feature map size of 160×160. By fully leveraging the high-resolution detail information from shallow features and integrating it with the semantic information from

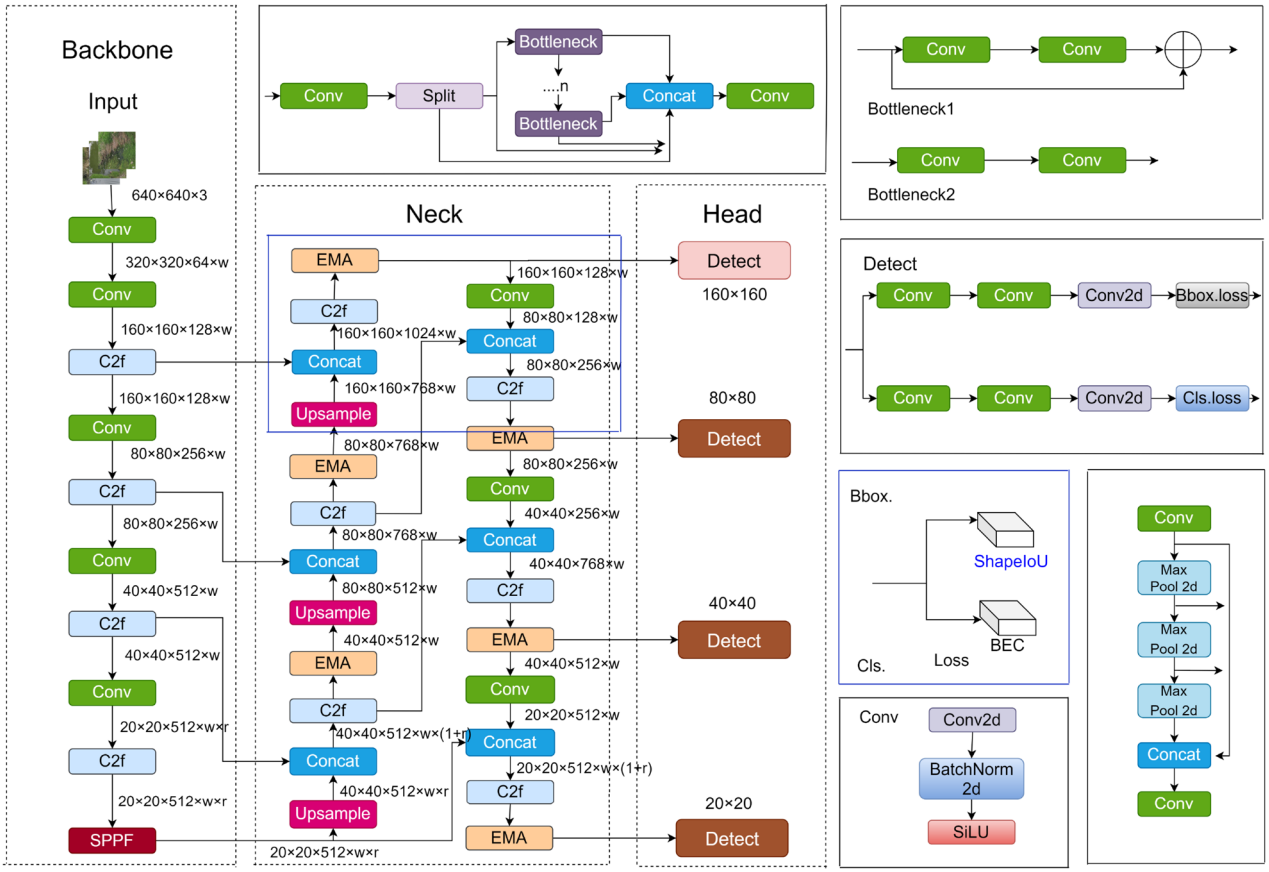


Fig. 2. ES-YOLOv8 network structure.

deep features, the model constructs a more refined "160-80-40-20" multiscale detection framework, enabling more comprehensive coverage of floating waste objects across various scales.

(2) **Neck Structure Optimization Based on EMA:** To address challenges such as illumination variations and complex background textures in water environments, this study integrates EMA modules into the neck section of YOLOv8s. This module explicitly models multiscale contextual differences through dilated convolutions and attention mechanisms, enhancing the model's focus on small targets. By parallelizing multiscale feature extraction and cross-dimensional attention interactions, the EMA module dynamically increases the weight of small target related features while suppressing background noise interference. This adaptive enhancement improves the saliency representation of floating waste and significantly enhances the detection accuracy and robustness of multiscale objects in complex scenes.

(3) **Shape-IoU Loss Function Design for Complex-Shaped Objects:** Due to the irregular shapes of floating waste (e.g., clothing, fragmented foam), the Conventional Complete IoU (CIoU) loss function used in YOLOv8 struggles to achieve precise bounding box regression. To address this issue, this study introduces the Shape-IoU loss function into the detection head of YOLOv8s. By incorporating a shape-awareness factor and contour alignment constraint, this function optimizes the bounding box regression process, allowing the model to perceive the local geometric characteristics of object edges. This mitigates boundary ambiguity issues caused by wave occlusion or reflections, thereby significantly improving detection accuracy for irregular floating waste. Furthermore, as Shape-IoU is a loss function improvement, it does not require modifications to the network's forward propagation structure and remains fully compatible with the original detection head, making it suitable for deployment on edge computing devices.

In summary, through the proposed enhancements to the YOLOv8s model, the ES-YOLOv8 framework significantly improves small target perception, strengthens focus on small objects, and enhances feature extraction for floating waste detection. These improvements collectively contribute to greater detection accuracy and model robustness, forming a novel improvement structure that integrates feature enhancement, interference suppression, and localization optimization.

Network structure improvement based on multiscale feature enhancement

The floating waste detection dataset constructed in this study exhibits a significant dominance of small targets. Experimental findings indicate certain limitations of the original YOLOv8s model when applied to this dataset:

(1) Insufficient representation capability of shallow features: Deep convolutional downsampling causes severe attenuation of critical features, such as edges and textures of small objects, during forward propagation through the Backbone.

(2) Suboptimal adaptation of multiscale detection heads: The size of the anchor box of the original detection heads (80×80/40×40/20×20) range from 8×8 to 32×32 pixels, which is significantly mismatched with the distribution of small targets (2×2 to 16×16 pixels). This mismatch leads to inefficient candidate box generation.

To address the challenges of variable target scales and loss of small-object feature in floating waste detection, this study proposes a multiscale feature enhancement approach, focusing on structural optimization of the Neck component in YOLOv8s and the expansion of high-resolution detection heads. As illustrated in Figure 3, the improved ES-YOLOv8 model incorporates the following key technical innovations:

(1) Cross-Level Feature Fusion Reconstruction in the Neck: Based on the dual-pyramid FPN-PAN structure of the original YOLOv8s, this study strengthens the feature map reuse mechanism for shallow features. High-resolution features are extracted from output feature map size of 160×160 in the Backbone, injected into the FPN network through lightweight cross-layer connections, which include 1×1 convolution for channel compression and upsampling. This design better preserves fine-grained edge and texture details of small targets while mitigating the loss of shallow semantic information caused by deep downsampling.

(2) 160×160 Small-Object Detection Head Design: In addition to the existing 80×80, 40×40, and 20×20 detection heads, a new 160×160 resolution detection branch is introduced to enhance multiscale feature coupling. The 160×160 feature map, which integrates shallow details with deep semantics from the Neck output, is fed into the new detection head. Multi-granularity feature extraction is performed using parallel 3×3 convolutions and dilated convolutions, thereby improving the model's ability to represent small objects.

In summary, the proposed multiscale collaborative detection framework, incorporating four detection heads at resolutions of 160×160, 80×80, 40×40 and 20×20, significantly enhances the model's sensitivity to small-object features through shallow feature injection. This approach enables more efficient utilization of multiscale features, leading to a notable improvement in detection accuracy. However, it also results in an increased model parameter count.

The EMA module

In floating waste detection, YOLOv8s' standard feature fusion struggles with multi-scale targets due to limited hierarchical representation. While attention mechanisms improve feature discrimination, conventional channel reduction risks losing spatial details. Our study integrates the EMA module into YOLOv8s' neck, preserving channel completeness through partial reconstruction and spatial redistribution. Unlike dimensionality-reducing attention methods, EMA achieves global channel recalibration while maintaining pixel-wise correlations via cross-dimensional interactions. This resolves multi-scale detection challenges with significant performance gains and preserved efficiency.

Figure 4 depicts the EMA module's architecture, where the 1×1 convolution shared components are labeled as 1×1 branches. For multi-scale spatial information integration, a 3×3 convolutional kernel parallel to the 1×1 branch is added, known as the 3×3 branch, enhancing adaptability to spatial changes.

(1) Feature grouping. For any given input feature map $X \in \mathbb{R}^{C\times H\times W}$, EMA divides X into G sub-features along the channel dimension to learn distinct semantic representations. This grouping can be denoted as $X = [X_0, X_i, \ldots, X_{G-1}]$. The group G is reshaped and integrated into the batch dimension, thereby redefining the shape of the input tensor to $C//G \times H \times W$.

(2) Parallel Subnetworks. The large local receptive fields of neurons enable them to gather multiscale spatial information. EMA utilizes three parallel pathways to extract attention weight descriptors from the grouped feature maps. Two of these pathways are within the 1x1 branch, while the third is in the 3x3 branch.

(3) Cross-spatial learning. EMA offers a cross-spatial information aggregation method across different spatial dimensions to achieve richer feature aggregation. Global spatial information is encoded into the output of the 1x1 branch through a 2D global average pooling operation. Meanwhile, the output of the 3x3 branch is directly transformed to match the corresponding dimensional shape required before the joint activation mechanism of

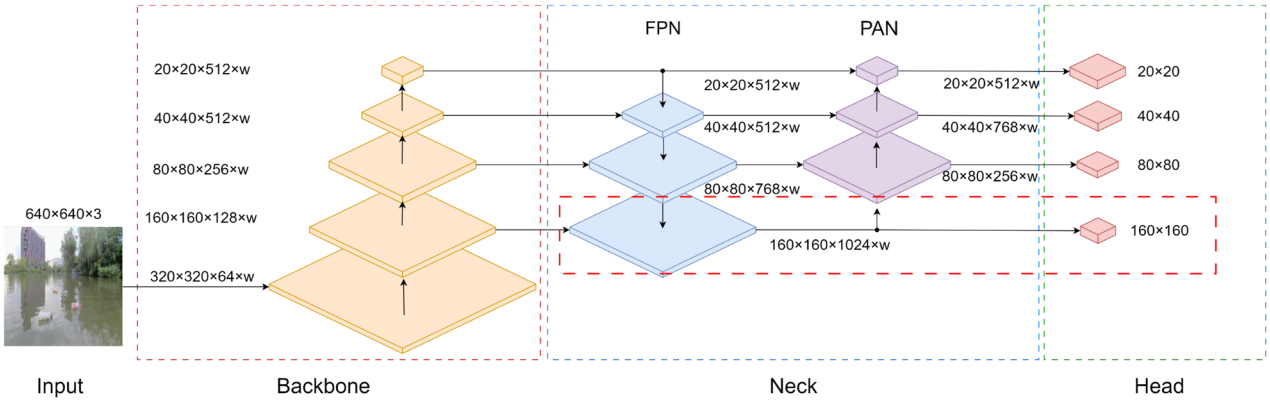


Fig. 3. Multiscale detection framework.

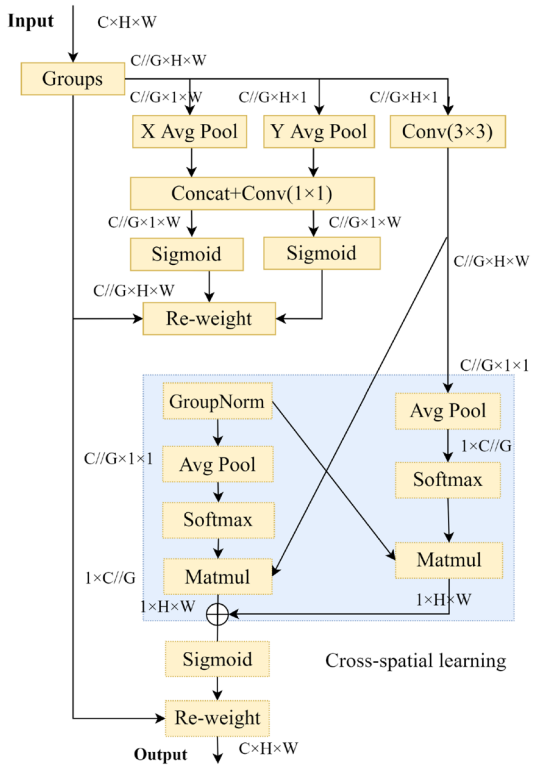


Fig. 4. Structure diagram of the EMA module.

channel features, i.e., $\mathbb{R}_1^{1\times C//G} \times \mathbb{R}_3^{C//G\times H\times W}$. The formula for the 2D global pooling operation is illustrated in Equation (1).

$$Z_c = \frac{1}{H*W}\sum_j^H \sum_j^W x_c(i,j) \tag{1}$$

The output feature map within each group is calculated as the aggregation of the two generated spatial attention weight values, followed by a Sigmoid function. This process captures pairwise relationships at pixel-level and emphasizes the global context for all pixels. The final output of EMA matches the size of *X*, making it efficient for integration into modern architectures.

Shape-IoU loss function

The purpose of bounding box regression is to fine-tune the detection model's output candidate boxes to maximize their overlap with the true boundaries of the target objects. Therefore, the Intersection over Union (IoU) serves as a metric to measure the degree of overlap between the predicted and true boxes, as shown in Formula (2):

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{2}$$

The YOLOv8s model uses the CIoU metric for bounding box regression, which considers overlap, center distance, and aspect ratio but can be aspect-ratio sensitive and may overemphasize center point localization. To address these issues, we switch to the Shape-IoU loss function, which more comprehensively accounts for the geometric relationship, shape, and scale of the bounding boxes, leading to more accurate regression. The derivation of Shape-IoU is as follows.

$$ww = \frac{2\times (w^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{3}$$

$$hh = \frac{2\times (h^{gt})^{scale}}{(w^{gt})^{scale} + (h^{gt})^{scale}} \tag{4}$$

$$distance^{shape} = hh\times (x_c - x_c^{gt})^2/c^2 + ww\times (y_c - y_c^{gt})^2/c^2 \tag{5}$$

$$\Omega^{shape} = \sum_{t=w,h}(1-e^{-w_t})^{\theta}, \theta = 4 \tag{6}$$

$$\begin{cases} w_w = hh \times \frac{|w-w^{gt}|}{max(w,w^{gt})} \\ h_h = ww \times \frac{|h-h^{gt}|}{max(h,h^{gt})} \end{cases} \tag{7}$$

The scaling factor scale denotes the scaling factor related to the size of the targets in the dataset, ranging from 0 to 1.5. Here, ω, h, ω^{gt}, and h^{gt} signify the width and height of the predicted and ground-truth bounding boxes, respectively. $(\omega^{gt})^{scale}$ and $(h^{gt})^{scale}$ represent ω^{gt} and h^{gt} adjusted by the scaling factor. x_c, y_c, x_c^{gt}, and y_c^{gt} denote the center coordinates of the predicted bounding box and the ground-truth bounding box. c stands for the length of the diagonal of the minimum bounding rectangle enclosing both the predicted bounding box and the ground truth bounding box. $\omega\omega$ and hh indicate the weighting coefficients in the horizontal and vertical directions, with their values related to the shape of the ground truth bounding box. The $\mathrm{distance}^{\mathrm{shape}}$ represents the distance loss function while Ω^{shape} denotes the shape loss function. The corresponding bounding box regression loss, denoted Shape-IoU is formulated as follows:

$$L_{shape} = 1 - IoU + distance^{shape} + 0.5 + \Omega^{shape} \tag{8}$$

As shown in Figure 5, Shape-IoU focuses on adjusting the predicted bounding box to better match the size and shape of the GT box. By quantifying shape adaptability and scale adaptability, it provides a more comprehensive method for bounding box regression.

Experiment

Datasets

The self-collected dataset used in this paper consists of 2711 images, with a total of 9088 annotated floating objects. The dataset is diverse and rich, with part of it sourced from the public river solid waste dataset, another part from the IWHR-AI-Label-Floater-V1 surface floater dataset[44], and the remainder self-shot by us, with all images being independently annotated. It covers a variety of scenes, including complex weather conditions such as overcast, rainy, and sunny days, as well as wide waters, eutrophic waters, significant shore obstructions, and situations where it is difficult to identify objects under the shadow of trees. To enhance the accuracy and practicality of the dataset, we have conducted a more refined classification, which includes 12 categories of solid waste: cardboard, plastic bags, plastic bottles, milk-boxes, mess tin, cigarette-boxes, cardboard-boxes, cups, cans, cover, clothing, and foam. The dataset is divided into a training subset, validation subset, and a test subset, The respective ratios are 8.5 : 1 : 0.5. In the data splitting process, we strictly adopted stratified sampling to ensure that the class distribution in the training, validation, and test sets remains completely consistent with the original dataset. The specific quantities of various objects are detailed in Figure 6.

The algorithm in this paper also detects the subset of the public dataset FloW-IMG[45]. The FloW-IMG dataset is the first to be used for detecting floating waste in inland waters. It includes a vision-based subset named FloW-IMG and a multi-modal dataset called FloW-RI. The FloW-IMG subset comprises 2000 images with 5,271 annotated instances, which we have divided into training, testing, and validation sets with a ratio of 6 : 2 : 2. We have introduced the two datasets more clearer in Table 1. The generalization, robustness, and applicability of the algorithm proposed in this paper have been validated on this dataset.

Experimental environment

The experiments in this paper were conducted on a PC with a Windows 10 operating system, an Intel(R) Core(TM) i9-9900K CPU, and an NVIDIA GeForce RTX 3090 GPU. The training was accelerated using CUDA 11.7, and the deep learning framework PyTorch 1.13.1 was used for training. The input image size was 640×640,

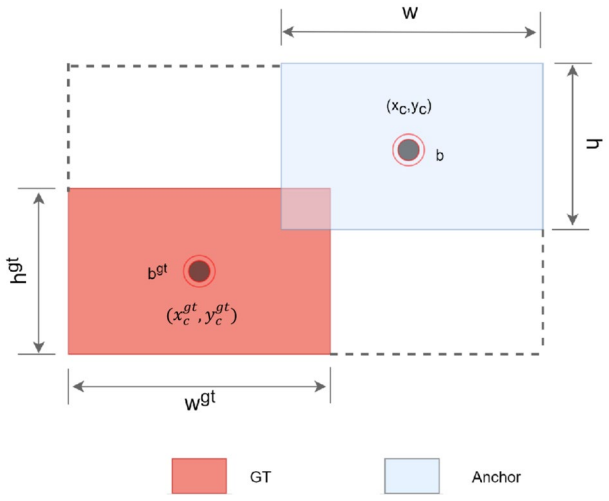


Fig. 5. Shape-IoU schematic diagram.

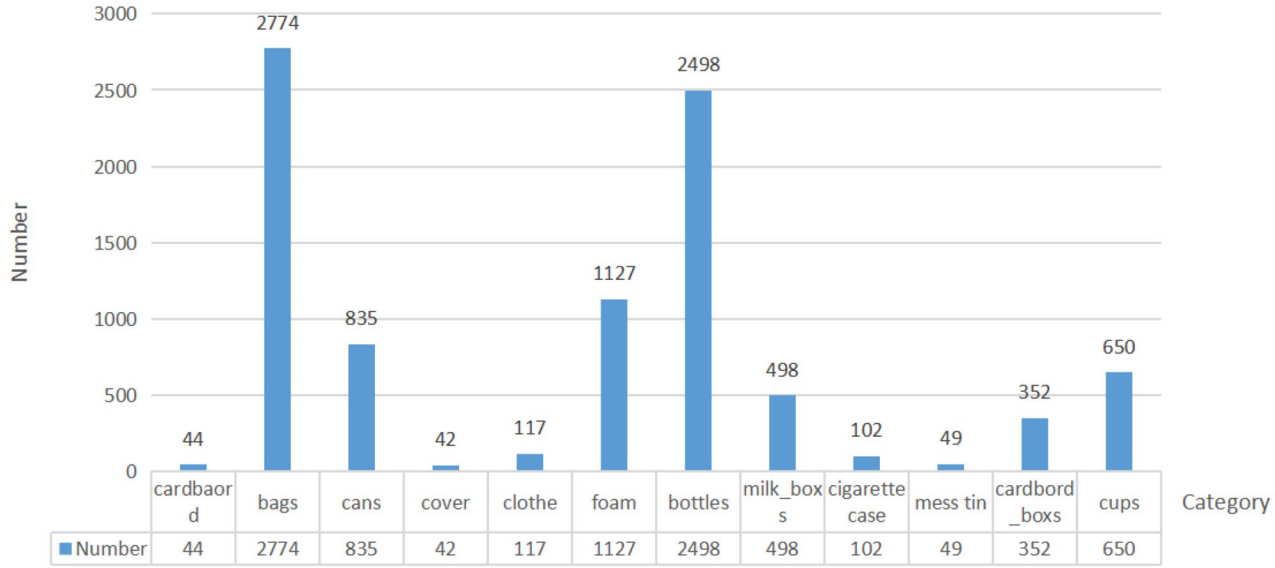


Fig. 6. Instances for each category.

Dataset	Total images	Train/Validation/Test Ratio	Category	Instances
Self-Made dataset	2711	8.5 : 1 : 0.5	12	9088
FloW-IMG	2000	6 : 2 : 2	1	5271

Table 1. Datasets introduction.

Component	Name/Value
Operating system	Windows 10
CPU	Intel(R) Core(TM) i9-9900K CPU
GPU	NVIDIA GeForce RTX3090
CUDA	11.7
PyTorch	1.13.1
Input image size	640*640
Training batch size	16
Optimizer momentum	0.937
Epoch	500
Learning rate	0.01
Learning rate decay strategy	Cosine Annealing and Linear Decay

Table 2. Experimental Configuration.

the training batch size was 16, the momentum for the SGD optimizer was 0.937, and the number of iterations was 500, initial learning rate was 0.01, learning rate decay strategy was Cosine Annealing and Linear Decay, as shown in Table 2.

Evaluation metrics

In this paper, we select precision(P), recall(R), average precision(AP), mAP, F1 score, and FPS as the evaluation metrics for this experiment.

Precision refers to the ratio of correct detections to all positive sample detections, and its formula is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall(R) represents the proportion of all true positive samples that are correctly detected by the model. Its formula is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Average Precision (AP) refers to the average precision achieved by the detection algorithm at different recall levels. Its formula is as follows:

$$AP = \int_0^1 precision(r)dr \tag{11}$$

mAP is a commonly used evaluation metric that measures the average precision of a model across various categories. Its formula is as follows:

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{12}$$

For comparative performance evaluation of models, F1 score and average precision are used as the primary metrics because they consider both precision and recall. The F1 score is the harmonic mean of precision and recall, and is defined as follows:

$$F1score = \frac{2 \times precision \times recall}{precision + recall} \tag{13}$$

FPS is commonly utilized as a metric to gauge the efficiency of object detection algorithms, serving as an indicator of the algorithm's detection speed when executed on a specific hardware setup. Typically, a higher FPS value corresponds to a quicker detection rate for the algorithm in question.

Ablation experiment

To evaluate the performance of the ES-YOLOv8 model under different optimization strategies, a series of ablation experiments were conducted under unified environmental conditions. These experiments aim to explore the specific impact of various improvement measures on the model's performance. The following is a summary of the results from the ablation experiments, as detailed in Table 3. In the table, A, B, and C represent the reconstruction of the multiscale network structure, replacement the loss function with Shape-IoU, and neck structure optimization based on EMA, respectively.

As shown in Table 3, compared to the original YOLOv8s model, the incorporation of small target detection led to a 4.7% increase in precision, a 3.3% increase in recall, a 3% improvement in mAP@0.5, a 5.6% improvement in mAP@0.5:0.95, and a 3.9% increase in the F1 score, demonstrating a significant overall performance enhancement. Furthermore, after replacing the Shape-IoU loss function and integrating the lightweight EMA attention mechanism, we re-evaluated the model's performance with the small target detection head. The results showed an additional 0.7% increase in precision, a 1.9% improvement in recall, a 2.4% boost in mAP@0.5, a 0.5% increase in mAP@0.5:0.95, and a 2.1% rise in the F1 score.

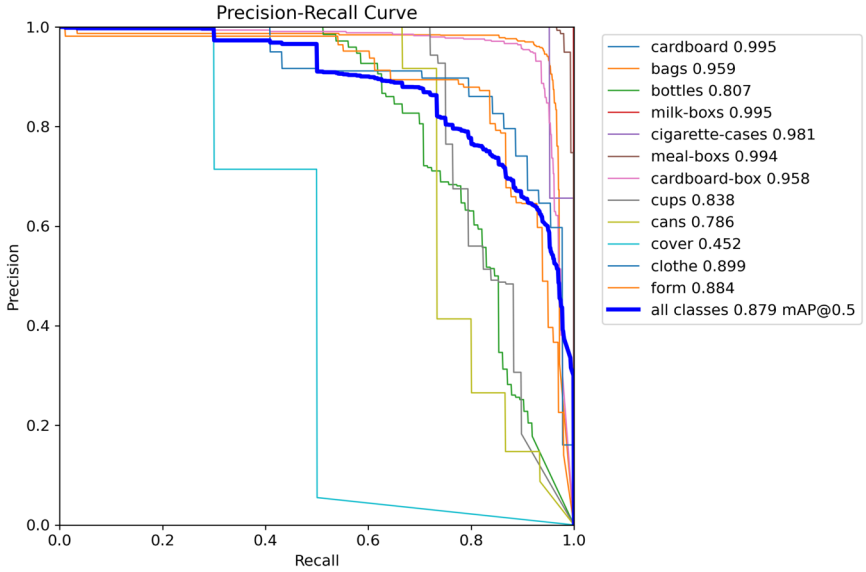
These findings confirm that the proposed algorithm achieves comprehensive improvements over the original YOLOv8 model. Therefore, it can be concluded that our ES-YOLOv8 outperforms the original model in solid floating waste detection.

Figure 7 (b) illustrates the P-R curves for individual floating waste categories with the ES-YOLO model. Compared to Figure 7 (a), the precision-recall curves and the mAP@0.5 curve in Figure 7 (b) enclose a noticeably larger area. Additionally, the balance point (P=R) is significantly higher. These observations further validate the novel network's accuracy and effectiveness.

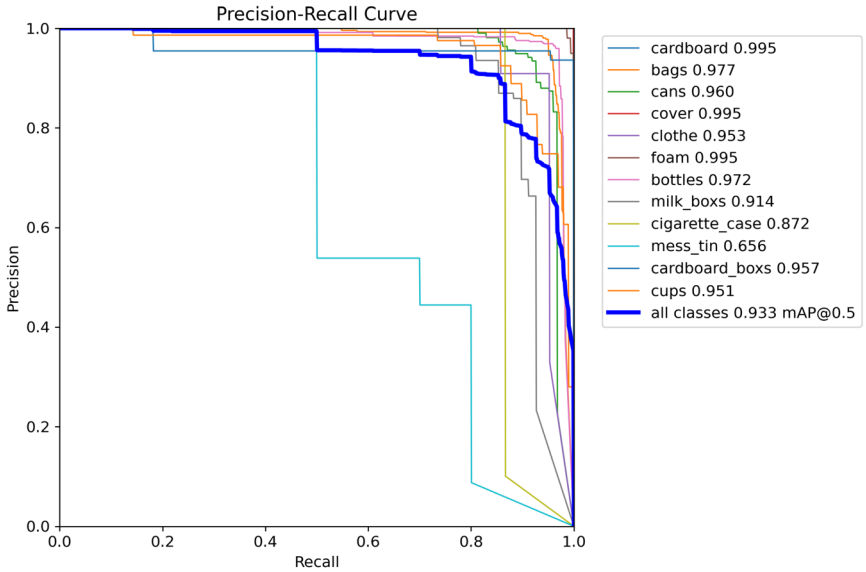
As illustrated in Figure 8, the YOLOv8s baseline model converges rapidly and initially achieves high precision, however, its final mAP@0.5 is significantly lower than that of other enhanced strategies. After incorporating the small object detection head, the model exhibits a slower convergence rate but achieves a notable improvement in mAP@0.5. In our proposed ES-YOLOv8 model, the convergence speed is accelerated, model complexity is reduced, and mAP@0.5 is further enhanced. These results demonstrate that the improved model is better suited for solid waste detection compared to the original YOLOv8.

Model	A	B	C	P(%)	R(%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1score (%)
YOLOv8s				86.6	82.4	87.9	68.9	84.5
	✓			91.3	85.7	90.9	74.5	88.4
		✓		85.4	84.4	88.5	68.4	85
Ours			✓	85.2	85	89.6	70.6	85
	✓	✓		91	**88.4**	93	74.9	**89.7**
	✓	✓	✓	**92**	87.6	**93.3**	**75**	**89.7**

Table 3. Ablation experiment. The boldface indicates the best values.



(a) P-R curve with YOLOv8s



(b) P-R curve with our method

Fig. 7. Comparison of P-R curves across various classes.

As demonstrated in Figure 9, the visualization results of the three images above indicate that YOLOv8s has a higher false detection rate and more misclassifications compared to ES-YOLOv8. In these three images, YOLOv8s exhibits one instance of misclassification, two instances of incorrect detection, and one instance of false detection. These errors are highlighted with black boxes in the figure.

Comparative experiment with state-of-the-arts models

To objectively assess the comprehensive performance of the proposed model in terms of average precision, recall, mAP@0.5, mAP@0.5:0.95, and F1 score, we compare its performance against seven state-of-the-art deep learning-based object detection algorithms, including one-stage methods SSD (VGG-16), YOLOv5s, YOLOv8s, CDW-YOLOv8[20], YOLOv9s[46], YOLOv10s, YOLOv11s, and the two-stage method Faster R-CNN (ResNet-50)[47]. The results are shown in Table 4. CDW-YOLOv8 is an advanced model built upon improvements to YOLOv8n, specifically designed for solid waste detection.

As indicated in the Table 4, it can be observed that YOLOv5s demonstrates commendable performance in solid waste detection, with its recall and F1 score slightly surpassing those of other algorithms. However, its mAP@0.5:0.95 is marginally lower. SSD exhibits higher precision compared to other algorithms but has the

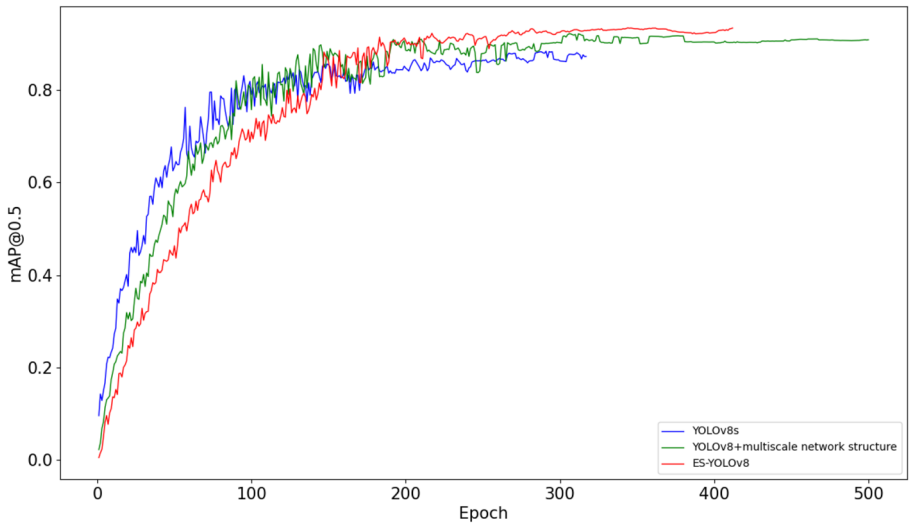
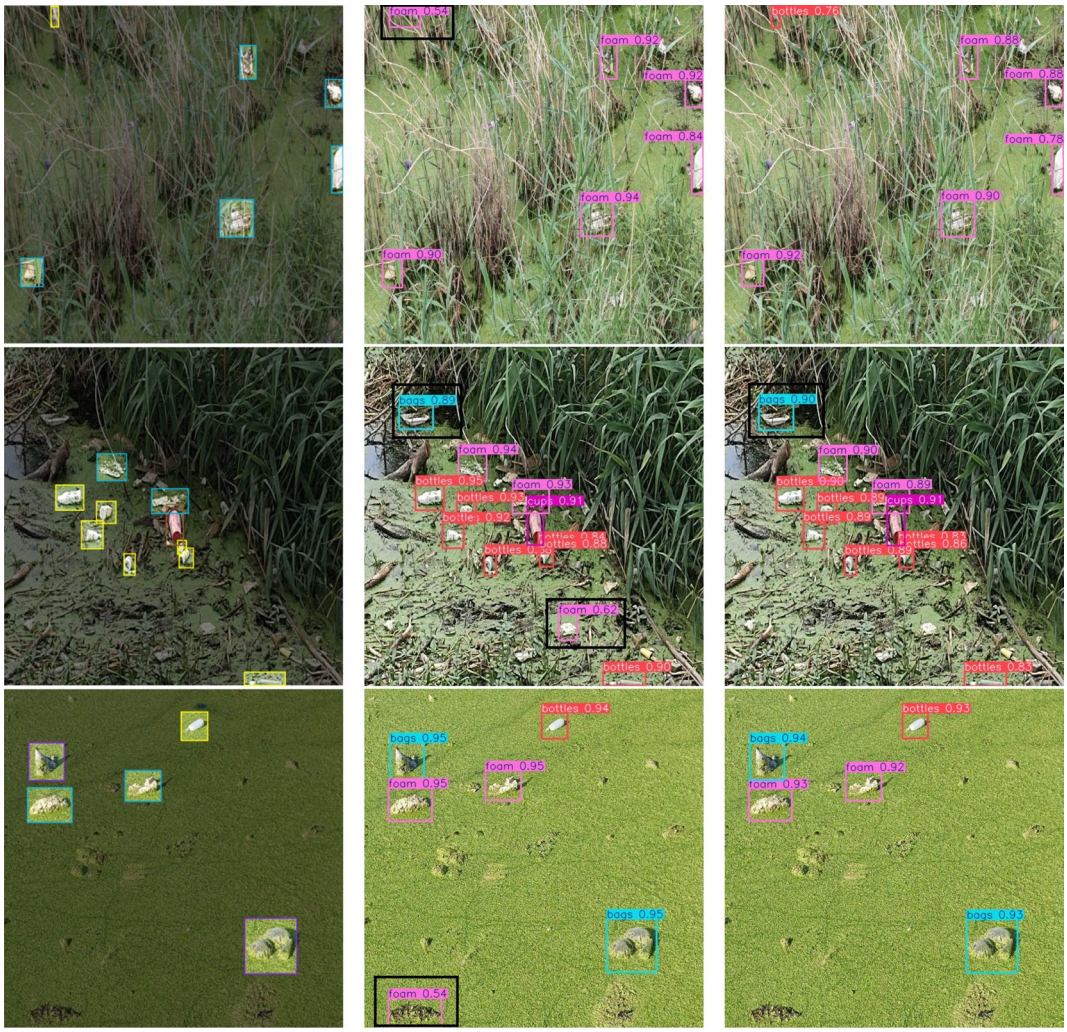


Fig. 8. Comparison of mAP@0.5 before and after improvement.



(a) Original annotated image (b) YOLOv8s (c) ES-YOLOv8

Fig. 9. Visualization of comparative results.

Algorithm	P(%)	R(%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	F1score (%)	FPS
SSD	97	39.9	52.5	–	51.4	187.65
YOLOv5s	91.2	89.3	91.8	71.7	90.2	255
YOLOv8s	**86.6**	**82.4**	**87.9**	**68.9**	**84.5**	**477.8**
CDW-YOLOv8	79.7	84.8	85.3	54.3	82.2	413.9
YOLOv9s	82.9	76.5	85.9	67.1	79.6	271.3
YOLOv10s	90.1	79	88.4	70.3	84.2	413.7
Faster R-CNN	55.1	58.8	53.6	–	55.2	31.4
YOLOv11s	91.5	83.3	89.9	73.5	86	487.7
Ours	**92**	**87.6**	**93.3**	**75**	**89.7**	**314.9**

Table 4. Comparative results of state-of-the-arts methods on the self-made Dataset. Boldface indicates the baseline metric values and the improved metric values.

Algorithm	P(%)	R(%)	mAP@0.5 (%)	F1score (%)	Size (MB)	FPS
RetinaNet	91.7	18	26.6	30	138.9	49.7
Faster R-CNN	32.8	62.1	45.1	43	108.2	31.4
SSD	94.5	36.3	82.2	52	112.8	187.65
YOLOv5s	85.3	82.6	86.9	83.9	16.5	255
YOLOv8s	**85.1**	**78.9**	**85.7**	**82**	**21.4**	**477.8**
YOLOv10s	84.5	78.4	84.8	81	15.7	413.7
YOLOv11s	87	79.8	86.4	83	18.2	487.7
Ours	**86.4**	**81.5**	**87.3**	**84**	**21.1**	**314.9**

Table 5. Comparative results of state-of-the-arts methods on FloW-IMG dataset. Boldface indicates the baseline metric values and the improved metric values.

lowest recall, leading to a relatively lower average precision. Faster R-CNN, on the other hand, performs poorly overall in solid waste detection.

Additionally, we compare our model with CDW-YOLOv8, a newly improved model specifically designed for solid waste detection. While its detection accuracy on the self-made dataset is relatively low, it demonstrates a 2.5% improvement over YOLOv8n as reported in their research[20]. YOLOv9s does not exhibit outstanding performance in solid waste detection. YOLOv10s and YOLOv11s outperform other versions on the custom dataset. However, compared to ES-YOLOv8, our proposed model demonstrates superior detection accuracy.

In summary, ES-YOLOv8 not only achieves the highest performance in terms of mAP@0.5 and mAP@0.5:0.95, but also outperforms most other algorithms across all evaluation metrics. The data presented in Table 4 clearly indicate that the proposed algorithm provides a distinct advantage in detecting various categories of solid floating waste. These findings demonstrate its superior performance and highlight its potential for application in diverse waterborne solid waste detection scenarios.

Robustness and generalization verification of ES-YOLOv8

To verify the robustness and generalization ability of the proposed model, we conducted experiments on the public dataset FloW-IMG to demonstrate the excellent performance of the algorithm proposed in this paper. Table 5 shows the experimental results on the public dataset. Figure 10 is the corresponding histogram for the table.

The findings from Table 5 indicate that on the FloW-IMG dataset, YOLOv5s continues to demonstrate strong performance in detecting solid waste on the water surface. It not only achieves the highest recall rate but also maintains consistently high overall scores, suggesting that the model possesses strong generalization capabilities and robustness.

The SSD algorithm attains an impressive precision of 94.5%, significantly higher than that of other algorithms. However, its recall rate is considerably lower, leading to an overall lower average precision and F1 score. The YOLOv10s algorithm also performs well, nevertheless, the proposed algorithm in this study still exhibits distinct advantages in detecting solid waste on FloW-IMG. It achieves relatively high overall evaluation metrics, with both average precision and F1 score remaining the highest among all compared methods.

Figure 10 presents a histogram illustrating the performance of different algorithms on the FloW-IMG dataset.

Discussion

The proposed ES-YOLOv8 model addresses key challenges in floating waste detection on water surfaces, such as small object scale, irregular shapes, and strong background interference. Through targeted improvements, the model significantly enhances detection performance.

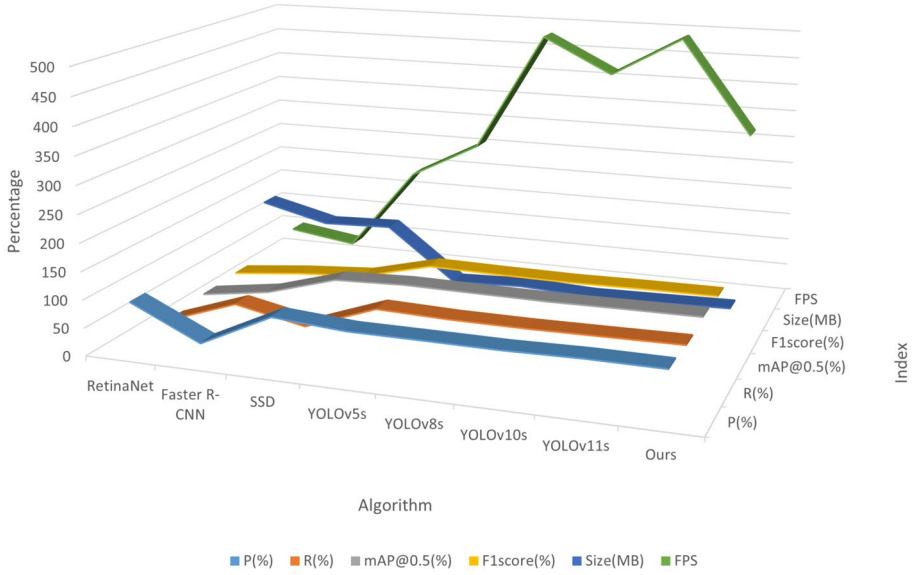


Fig. 10. Performance of different detection algorithms on FloW-IMG.

Firstly, in feature extraction, the optimized "160-80-40-20" multiscale detection framework, combined with the EMA attention mechanism, effectively mitigates the false detection and missed detection of small floating objects. The EMA module, incorporating dilated convolution and attention mechanisms, strengthens the model's focus on small objects. When integrated with refined multiscale feature fusion, it significantly enhances the semantic representation of small targets. Secondly, the introduction of the Shape-IoU loss function optimizes the bounding box regression process through shape-sensitive constraints. This effectively alleviates the sensitivity of traditional IoU-based methods to geometric deviations of irregular objects, thus improving localization accuracy. Additionally, the self-constructed dataset in this study partially addresses the limitations of existing publicly available floating waste datasets, which are relatively scarce and contain homogeneous annotations.

Experimental results demonstrate that on the self-constructed dataset, ES-YOLOv8 achieves mAP@0.5 and mAP@0.5:0.95 scores of 93.3% and 75%, respectively, representing improvements of 5.4% and 6.1% over the baseline model. Furthermore, compared to other state-of-the-art models, including SSD, YOLOv5s, YOLOv8s, CDW-YOLOv8, YOLOv9s, YOLOv10s, Faster R-CNN and YOLOv11s, ES-YOLOv8 exhibits mAP@0.5 improvements of 40.8%, 1.5%, 5.4%, 8.0%, 7.4%, 4.9%, 39.7%, and 3.4%, respectively. In addition, the model's generalization capability was validated on public FloW-IMG datasets, achieving an mAP@0.5 of 87.3%, a 1.6% improvement over the baseline, further demonstrating its robustness and applicability.

However, this study still has certain limitations:

1. Trade-off between computational efficiency and accuracy: While the incorporation of multiscale feature fusion and attention mechanisms enhances detection precision, it also increases model complexity. Future research should explore lightweight designs to facilitate deployment on edge devices.

2. Adaptability to extreme scenarios: The model's false detection rate remains an area for improvement under challenging conditions, such as strong light reflections or dense occlusions. Further research on dynamic interference suppression methods is needed.

3. Data diversity limitations: The current self-constructed dataset primarily targets static water surface scenarios. Future work should expand the dataset to include dynamic water flow environments to enhance the model's practicality.

Conclusions

This study proposes the ES-YOLOv8 algorithm, an improved model based on YOLOv8s, designed to meet the practical demands of floating waste detection on water surfaces. The proposed model integrates multiscale feature enhancement, attention-based interference suppression, and shape-sensitive localization optimization. Experimental results demonstrate that ES-YOLOv8 outperforms mainstream object detection models on both self-constructed and publicly available datasets, validating its robustness and generalization capability. This study provides a high-precision, low-power-consumption solution for intelligent water governance, contributing to water ecological protection and the engineering application of intelligent monitoring systems. Future research will focus on model lightweighting and adaptation to complex environments to facilitate real-world deployment.

Data availability

The datasets generated and analysed during the current study are available in the availability: https://github.com/666xkk/ES-YOLOv8/tree/master. https://github.com/666xkk/Floating-waste/tree/master.

Received: 16 January 2025; Accepted: 2 July 2025

Published online: 11 July 2025

References

1. Xu, Wei et al. Globally elevated greenhouse gas emissions from polluted urban rivers. *Nat. Sustain.*https://doi.org/10.1038/s41893-024-01358-y (2024).
2. N Shivaanivarsha, AG Vijayendiran, and M Ajay Prasath. " "WAVECLEAN"–An Innovation in Autonomous Vessel Driving Using Object Tracking and Collection of Floating Debris". In: *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. IEEE. 2024, pp. 1–6.
3. Henriques, Joao F.. et al. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.***37**(3), 583–596 (2015).
4. Xie, Yifan, Huang, Yuan & Song, Taek Lyul. Iterative joint integrated probabilistic data association filter for multiple-detection multiple-target tracking. *Digit. Signal Process.***72**, 232–243 (2018).
5. Lianghui Ding et al. "Detection and tracking of infrared small target by jointly using SSD and pipeline filter". In: *Digital Signal Processing* 110 (2020), p. 102949.
6. Ross Girshick et al. "Region-based convolutional networks for accurate object detection and segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 38.1 (2015), pp. 142–158.
7. Ross Girshick. "Fast r-cnn". In: arXiv preprint arXiv:1504.08083 (2015).
8. Xie, Xingxing et al. Oriented r-cnn and beyond. *Int. J. Comput. Vis.*https://doi.org/10.1007/s11263-024-01989-w (2024).
9. Cui, Yiming et al. An exploratory framework to identify dust on photovoltaic panels in offshore floating solar power stations. *Energy***307**, 132559 (2024).
10. Trinh Duc Minh, Nguyen Thi Ngoc Hoa, and Tat-Hien Le. "A Model for Floating Garbage Detection and Quantification Using Fixed Camera". In: *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*. 2022, pp. 389–393. https://doi.org/10.1109/NICS56915.2022.10013461.
11. Wei Liu et al. "Ssd: Single shot multibox detector". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.
12. J Redmon. "You only look once: Unified, real-time object detection". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
13. 2022.Available: https://github.com/ultralytics/yolov5/tree/v7.0.GitHub. "YOLOv5 release v7.0". In: (2020).
14. 2023. Available: https://github.com/ultralytics/ultralytics/tree/main. GitHub. "YOLOv8". In: (2023).
15. Ao Wang et al. "Yolov10: Real-time end-to-end object detection". In: arXiv preprint arXiv:2405.14458 (2024).
16. Rahima Khanam and Muhammad Hussain. "YOLOv11: An overview of the key architectural enhancements". In: arXiv preprint arXiv:2410.17725 (2024).
17. Zhanjun Jiang et al. "APM-YOLOv7 for Small-Target Water-Floating Garbage Detection Based on Multi-Scale Feature Adaptive Weighted Fusion". In: *Sensors* 24.1 (2024). ISSN: 1424-8220. https://doi.org/10.3390/s24010050. https://www.mdpi.com/1424-8220/24/1/50.
18. Zhao, Rui et al. Multi-target detection of waste composition in complex environments based on an improved YOLOX-S model. *Waste Manag.***190**, 398–408 (2024).
19. Chen, Luya & Zhu, Jianping. Water surface garbage detection based on lightweight YOLOv5. *Sci. Rep.***14**(1), 6133 (2024).
20. Shi, Chenan et al. Enhanced floating debris detection algorithm based on CDW-YOLOv8. *Phys. Scr.***99**(7), 076019. (2024).
21. Son, Junhyeok & Ahn, Yuchan. AI-based plastic waste sorting method utilizing object detection models for enhanced classification. *Waste Manag.*https://doi.org/10.1016/j.wasman.2024.12.014 (2025).
22. Song, Huaxiang et al. Pure data correction enhancing remote sensing image classification with a lightweight ensemble model. *Sci. Rep.*https://doi.org/10.1038/s41598-025-89735-1 (2025).
23. Song, Huaxiang & Zhou, Yong. Simple is best: A single-CNN method for classifying remote sensing images. *Networks Heterog. Media*https://doi.org/10.3934/nhm.2023070 (2023).
24. Song, Huaxiang et al. Variance consistency learning: Enhancing cross-modal knowledge distillation for remote sensing image classification. *Ann. Emerg. Technol. Comput.*https://doi.org/10.33166/AETiC.2024.04.003 (2024).
25. Song, Huaxiang et al. Optimized data distribution learning for enhancing vision transformer-based object detection in remote sensing images. *Photogramm. Rec.***40**(189), e70004 (2025).
26. Song, Huaxiang et al. Efficient knowledge distillation for hybrid models: A vision transformer-convolutional neural network to convolutional neural network approach for classifying remote sensing images. *IET Cyber-Systems and Robotics***6**(3), e12120 (2024).
27. Song, Huaxiang. Mbc-net: Long-range enhanced feature fusion for classifying remote sensing images. *Int. J. Intell. Comput. Cybern.***17**(1), 181–209 (2024).
28. Song, Huaxiang et al. QAGA-net: Enhanced vision transformer-based object detection for remote sensing images. *Int. J. Intell. Comput. Cybern.*https://doi.org/10.1108/IJICC-08-2024-0383 (2024).
29. Song, Huaxiang et al. Quantitative regularization in robust vision transformer for remote sensing image classification. *Photogramm. Rec.*https://doi.org/10.1111/phor.12489 (2024).
30. Liu, Biaohua et al. Irmsd-yolo: Multiscale dilated network with inverted residuals for infrared small target detection. *IEEE Sens. J.*https://doi.org/10.1109/JSEN.2025.3546966 (2025).
31. Zhang, Huiying et al. Fusion of multi-scale attention for aerial images small-target detection model based on PARE-YOLO. *Sci. Rep.*https://doi.org/10.1038/s41598-025-88857-w (2025).
32. Hui, Yanming, Wang, Jue & Li, Bo. STF-YOLO: A small target detection algorithm for UAV remote sensing images based on improved SwinTransformer and class weighted classification decoupling head. *Measurement***224**, 113936 (2024).
33. Li, Peng et al. Oriented bounding box representation based on continuous encoding in oriented SAR ship detection. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*https://doi.org/10.1109/JSTARS.2025.3541217 (2025).
34. Zhou, Liming et al. SWDet: Anchor-based object detector for solid waste detection in aerial images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **16**, 306–320 (2022).
35. Ma, Wanqi et al. DSyolo-trash: An attention mechanism-integrated and object tracking algorithm for solid waste detection. *Waste Manag.***178**, 46–56 (2024).
36. Cao, Yu. et al. MAL-YOLO: A lightweight algorithm for target detection in side-scan sonar images based on multi-scale feature fusion and attention mechanism. *Int. J. Digit. Earth***17**(1), 2398050 (2024).
37. Daliang Ouyang et al. "Efficient multi-scale attention module with cross-spatial learning". en-US. In: ().
38. Zhiyu Zhou et al. "YOLO-based marine organism detection using two-terminal attention mechanism and difficult-sample resampling". In: *Applied Soft Computing* 153.000 (2024), p. 15.
39. Yang, Jian et al. YOLO-RDFEA: Object detection in RD imagery with improved YOLOv8 based on feature enhancement and attention mechanisms. *IEEE Access*https://doi.org/10.1109/ACCESS.2024.3485499 (2024).
40. Wang, Yu. & Xiang, Xiaodong. GMS-Yolo: An enhanced algorithm for water meter reading recognition in complex environments. *J. Real-Time Image Process.***21**(5), 173 (2024).
41. Hao Zhang and Shuaijie Zhang. "Shape-IoU: More Accurate Metric considering Bounding Box Shape and Scale". en-US. In: (Dec. 2023).

42. Zheng, Hui et al. Full stage networks with auxiliary focal loss and multi-attention module for submarine garbage object detection. *Sci. Rep.*https://doi.org/10.1038/s41598-023-42896-3 (2023).
43. Yue, Y. R., Cui, S. H. & Shan, W. Apple detection in complex environment based on improved YOLOv8n. *Eng. Res. Express*https://doi.org/10.1088/2631-8695/ad9e6a (2024).
44. MX Yang et al. *Surface Floater Dataset (IWHR _AI _Lable _Floater _V1)*. [DB/OL]. China Institute of Water Resources and Hydropower Research. 2023.
45. Yuwei Cheng et al. "FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 10953–10962.
46. Chien Yao Wang, I Hau Yeh, and Hong Yuan Mark Liao. "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information". In: (2024).
47. Shaoqing Ren et al. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *IEEE transactions on pattern analysis and machine intelligence* 39.6 (2016), pp. 1137–1149.

Author contributions
J.D. designed the experiment, J.D. and K.X. conducted the experiment, J.D. and Y.Y. analyzed the experimental results, J.D. and K.X. were responsible for the production of Figures 1 to 10 and Tables 1 to 5, K.X. and Y.Y. wrote the main text of the paper, and all authors reviewed the paper.

Funding
The work is supported by the Zhangjiakou 2023 Municipal Science and Technology Program Funded by Fiscal Support under Grant 2311010A.

Declarations

Competing interests
The authors declare no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Yes.

Additional information
Correspondence and requests for materials should be addressed to Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025