Process Safety and Environmental Protection xxx (xxxx) 107089

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.elsevier.com

Attention-guided marine debris detection with an enhanced transformer framework using drone imagery

L. Minh Dang [a, b], ASM Sharifuzzaman Sagar [c], Ngoc Dung Bui [d], Luong Vuong Nguyen [e], Tri-Hai Nguyen [f, *]

[a] Institute of Research and Development, Duy Tan University, Da Nang 550000, Viet Nam
[b] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Viet Nam
[c] Department of Intelligent Mechatronics Engineering, Sejong University, Seoul 05006, Republic of Korea
[d] Faculty of Information Technology, University of Transport and Communications, Hanoi 100000, Viet Nam
[e] Department of Artificial Intelligence, FPT University, Danang, Viet Nam
[f] Faculty of Information Technology, School of Technology, Van Lang University, Ho Chi Minh City, Viet Nam

ARTICLE INFO

Keywords:
Drones
Transformer
Marine litter
Detection
Waste management

ABSTRACT

Marine debris poses a critical threat to coastal ecosystems, public health, and global tourism. Although various strategies have been proposed to tackle the issue, they are labor-intensive and time-consuming. To address these challenges, this paper introduces a novel drone-based marine debris detection framework based on a state-of-the-art Transformer model designed to capture global context. By training the model on high-resolution aerial imagery, the proposed framework accurately identifies eight different types of marine litter in complex marine environments. To further improve the framework's performance, a preprocessing module is implemented to enhance image quality, while extensive fine-tuning on a large-scale dataset ensures robust performance under diverse conditions. Finally, attention weights from the model's decoder layers are visualized to support the interpretability of detection results. The experimental results reveal that the proposed framework outperforms existing detection models in real-time inference speed while achieving a state-of-the-art mean average precision (mAP) of 81.5 %.

1. Introduction

Marine debris, often referred to as marine litter, represents a significant global threat to the health of marine ecosystems and the well-being of coastal communities. Marine debris has worsened at an alarming rate, primarily driven by rapid industrialization, unsustainable consumption, and inadequate waste disposal and management practices (Sugianto et al., 2023). They may range from plastic bottles and metals to fishing nets and microplastics. As debris accumulates in open waters and along shorelines, it disrupts water quality, jeopardizes ecological balance, and damages the beauty of coastlines. Moreover, marine debris is an alarming environmental concern that endangers wildlife through entanglement and ingestion (Woods et al., 2021). Over time, these pollutants can infiltrate entire food chains, potentially introducing harmful substances into human diets and raising public health concerns. The negative effects also significantly impact economic sectors such as tourism and fisheries. Therefore, there is an urgent need for effective recycling and waste management strategies due to the interlinked ecological and socio-economic consequences of marine debris (Gong et al., 2022; Bishop et al., 2020).

Recent technological advancements in unmanned aerial vehicles (UAVs) and computer vision (CV) have led to a powerful and efficient solution for monitoring marine debris (Taddia et al., 2021). By exploiting high-resolution imaging capabilities and integrating advanced machine learning (ML) algorithms, UAVs can detect, locate, and classify debris in real time with remarkable precision (Escobar-Sánchez et al., 2022; Nguyen et al., 2024). The drone-based marine debris framework not only accelerates the work of rapid response teams, such as swift planning and execution of targeted cleanup operations, but also mitigates threats to marine life through reduced risk of entanglement and habitat degradation. Moreover, the scalability and flexibility of drone-based approaches allow for frequent, cost-effective surveillance of extensive coastal areas in order to gather comprehensive data on floating and shoreline pollution Tightiz et al. (2024). The collected large-scale data enhances scientific understanding of marine debris distribution and origins, and enables researchers and experts to develop targeted mitigation measures. As a result, UAV-based marine litter detection sys-

* Corresponding author at: Van Lang University, Ho Chi Minh City, Vietnam.
E-mail address: hai.nguyentri@vlu.edu.vn (T.-H. Nguyen).

https://doi.org/10.1016/j.psep.2025.107089
Received 20 January 2025; Received in revised form 7 March 2025; Accepted 29 March 2025
0957-5820/

tems ultimately contribute to the long-term preservation of marine ecosystems and the socioeconomic well-being of coastal communities.

Recent advanced deep learning (DL) architectures have demonstrated state-of-the-art speed and accuracy on a wide range of CV tasks (Li et al., 2023; Nadeem et al., 2023). State-of-the-art object detection models, such as You Only Look Once (YOLO) (Shen et al., 2024), Faster R-CNN (Wang et al., 2023), or Mask R-CNN (Hong et al., 2020), can automatically distinguish between different types of litter, including plastics, metals, polyethylene terephthalate (PET) bottles, and discarded buoys. Moreover, recent studies have focused on lightweight DL models that are fine-tuned to further enhance the performance and efficiency of marine debris detection (Dosset et al., 2024; Dang et al., 2023). However, common DL-based detection models often rely mainly on hand-engineered components like anchors and proposals during training and still require an additional bounding box refinement process to remove identical predictions (Khan et al., 2022).

In contrast, Transformer models have emerged as a promising alternative to conventional object detection models like YOLO or Faster R-CNN (Sun et al., 2021). By performing a self-attention mechanism on an entire image, Transformer models capture global context more effectively and handle complex visual relationships in complex scenes. For example, the Detection Transformer (DETR) and its variants offer an end-to-end training scheme with simpler network architectures and no hand-crafted modules such as region proposals (Carion et al., 2020). D-FINE is a novel real-time transformer-based object detection model that redefines the bounding box regression module of DETR-based models using two key components: Global Optimal Localization Self-Distillation (GO-LSD) and Fine-grained Distribution Refinement (FDR) (Peng et al., 2024). In addition, D-FINE applies various optimizations in previous computationally intensive components. The experimental results on the COCO dataset reveal that it substantially boosts the performance of numerous DETR models by up to 5.3 % average precision (AP). D-FINE-L and D-FINE-X achieve 54.0 % and 55.8 % AP at 124 frames per second (FPS) and 78 FPS, respectively. The attention mechanism is particularly advantageous in marine debris detection, where marine litter can be scattered, partially submerged, or heavily occluded. In addition, the flexibility and scalability of Transformer models facilitate efficient adaptation to new datasets, which makes them particularly well-suited for autonomous drone-based monitoring tasks.

Motivated by the state-of-the-art performance of Transformer models and the urgent need for an accurate and robust marine debris detection framework, this study introduces a transformer-based framework for identifying eight common marine litters. First, a pre-processing module is implemented to improve the quality of drone-captured images. After that, the D-FINE model is trained and fine-tuned on a large-scale dataset containing about 364,361 images. Finally, the study extracts and visualizes D-FINE's attention weights from the decoder's layers to facilitate a more comprehensive interpretation of the model's predictions and offer valuable insights into the model outputs.

The remainder of this study is categorized as follows: Section 2 describes the collection of the large-scale marine debris dataset using UAVs. Section 3 provides a detailed description of the Transformer-based marine litter detection framework. Section 4 presents an evaluation of the proposed approach to the collected dataset. Section 5 discusses the results obtained from the proposed model. Finally, Section 6 summarizes the research with key remarks and suggestions for future research directions.

2. Marine debris dataset

Previous datasets on marine debris identification are insufficient in both scale and diversity to comprehensively address the complexities of marine debris. For instance, the marine debris dataset by Politikos et al. comprises 635 images (Politikos et al., 2021), the DSDebris dataset contains 15,000 images (Huang et al., 2023), and the JAMSTEC database includes 5352 images (Deep-sea debris database, 2023). In contrast, this research introduces a large-scale marine debris dataset of approximately 364,361 images with 2,268,950 annotated objects, which covers eight common types of marine litter. It is nearly 20 times larger than the DSDebris dataset and includes a broader range of litter categories. The dataset contains images from diverse environments, such as coastal and floating terrains (Politikos et al., 2021; Huang et al., 2023). Although the specific composition of marine debris varies depending on geographical location and local industrial activities, the chosen categories represent a significant portion of the debris commonly encountered worldwide. As a result, the dataset is highly relevant and applicable to a wide range of coastal monitoring and cleanup initiatives.

The dataset is provided by the National Information Society Agency of Korea (NIA)[1] and Irem Tech Co., Ltd.[2] for research purposes. It is collected, preprocessed, and labeled by Pukyong Maritime Technology Co., Ltd. and validated by Saltlux Innovation Co., Ltd.[3] Multiple survey missions are conducted on both coastal and sea surfaces using a DJI Mavic 2 Pro drone to assess marine debris comprehensively. The drone surveys more than 100 ha of coastal and 100 ha of sea surface with over 20 h of video footage. Data collection is strategically conducted on cloudless days between 11:00 AM and 1:00 PM, aligning with solar noon to maximize natural lighting and enhance contrast and clarity in the captured videos. To ensure the dataset's integrity and quality, video frames with high turbidity, significant color distortions, or light flares are manually excluded from the analysis. Each frame extracted from the videos has a resolution of 1200 × 800 pixels at 72 dpi. The final dataset comprises approximately 364,361 images, of which 291,488 (80 %) were used for training and validation, and 72,873 (20 %) for testing. Representative samples for each marine debris type are displayed in Fig. 1.

Fig. 2 provides a comprehensive visualization of the dataset distribution by showing the number of annotated marine debris objects across eight categories for both coastal and floating waste. The dataset is dominated by PET bottles (576,247 objects) and plastic debris (265,922 objects), reflecting their widespread prevalence in marine pollution. In contrast, glass debris (13,742 instances) and buoys (11,403 instances) are less frequently observed. The distribution is fairly balanced between coastal and floating waste. This variation is critical for ensuring model robustness, as the appearance and detectability of debris can vary significantly depending on factors, such as terrain, lighting, and water reflection.

3. Methodology

Fig. 3 describes the primary components of a comprehensive drone-based marine debris detection framework, abbreviated as MD-TR. In this context, MD indicates marine debris detection, while TR refers to the transformer-based real-time object detection model. The framework operates in two main stages. (I) Data Pre-Processing enhances and corrects the raw aerial images (e.g., adjusting color balance and removing geometric distortions). (II) Marine Debris Detection then applies a specialized transformer-based model ("D-FINE") to identify debris in real-time. After detection, an attention weight analysis step is introduced to interpret the detection results.

3.1. Image pre-processing

3.1.1. Image enhancement

Studies that utilize images captured by UAVs frequently encounter challenging environmental conditions, including poor lighting, color distortion, low contrast, and increased noise. These issues are primarily

[1] https://aihub.or.kr/
[2] http://iremtech.co.kr/
[3] https://www.saltluxinno.com/

Fig. 1. Sample image for eight marine debris types found in the dataset for both coastal and floating terrains.

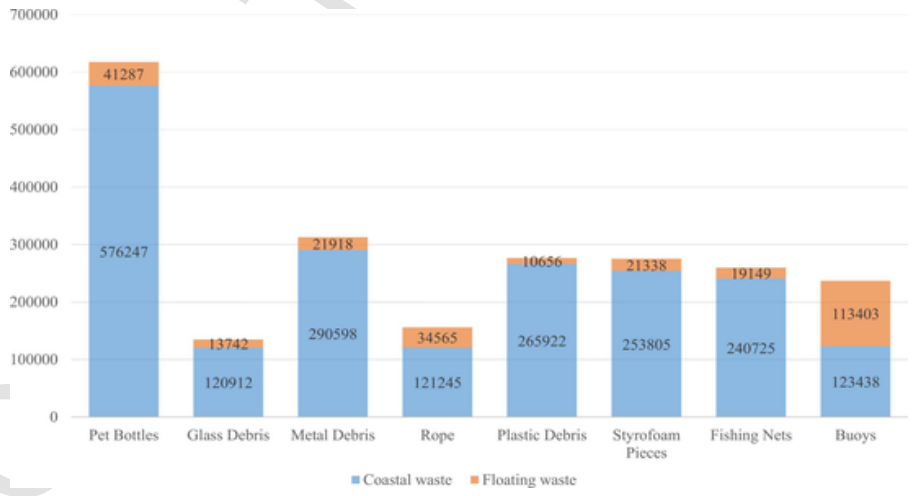

Fig. 2. A bar chart showing the distribution of annotated objects per each marine debris.

caused by reflection, scattering, and absorption at the water's surface (Zhang and Zhu, 2023). Moreover, factors such as glare, water turbidity, and harsh weather patterns can further reduce image clarity. As a result, the original images without preprocessing significantly degrade the performance of marine litter identification models during both the training and inference phases. Image enhancement is a widely used technique applied to address these challenges. By correcting color distortion, enhancing contrast, and improving overall clarity, image enhancement plays a crucial role in improving the performance of detection models (Qi et al., 2021). Image enhancement often involves a combination of noise reduction and color correction to counteract the visual degradation caused by the absorption and scattering of aquatic environments.

Traditional methods focus on filtering or statistical techniques to reduce noise and improve clarity, while color correction targets the hue shifts characteristic of aquatic settings. Researchers have increasingly employed ML, particularly DL, to tackle these challenges due to its adaptability and scalability as models learn from larger datasets. In this study, we implement a novel DL-based image enhancement algorithm (Chen et al., 2021). The method introduces parametric ReLU (PReLU)

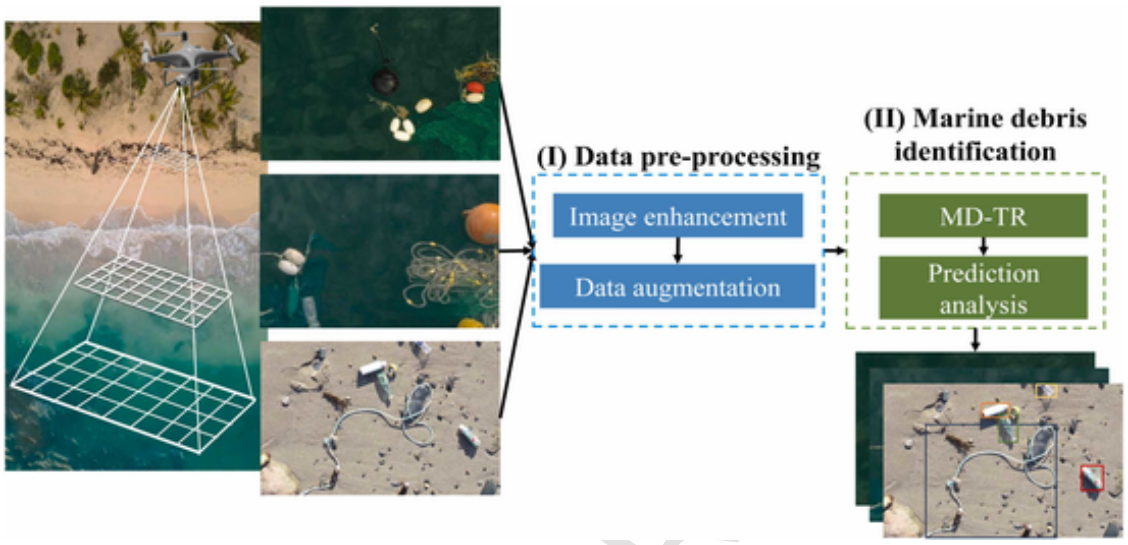

Fig. 3. Description of the primary processes of the proposed real-time transformer-based marine debris detection framework.

and dilated convolution to improve the neural network's fitting capacity. Through various experiments, the model is proven to enrich colors, clarify details, and increase both peak signal-to-noise ratio and structural similarity index measure metrics compared to existing approaches. Furthermore, the algorithm offers significantly faster computation speeds, which enables real-time processing requirements for the performance of subsequent high-level CV tasks.

3.1.2. Data augmentation

Various data augmentation strategies are performed to significantly increase both the volume and diversity of each marine debris class in the training dataset. First, a color jitter operation is applied to mitigate the effects of varying weather conditions by performing random adjustments to brightness, contrast, saturation, and hue. To create a broad spectrum of output variations, the selected contrast, brightness, saturation, and hue ranges are set to [0, 2], [0.5, 1.5], [0.9, 1.1], and [-0.5, 0.5], respectively. To further replicate changes in object orientation and camera viewpoints, images are randomly rotated at various angles, as well as flipped horizontally and vertically. Finally, Gaussian noise is injected to mimic equipment-induced distortions. The data augmentation module expands the dataset three-fold. Fig. 4 illustrates an example of these techniques applied to a single input image.

3.2. Transformer-based marine debris detection model

The proposed MD-TR model is based on D-FINE architecture (Peng et al., 2024), which solves the limitations of bounding box regression from the Real-time DETR (RT-DETR) model (Zhao et al., 2024). This is achieved through two key techniques: Global Optimal Localization Self-Distillation (GO-LSD) and Fine-grained Distribution Refinement (FDR), as illustrated in the flowchart in Fig. 5.

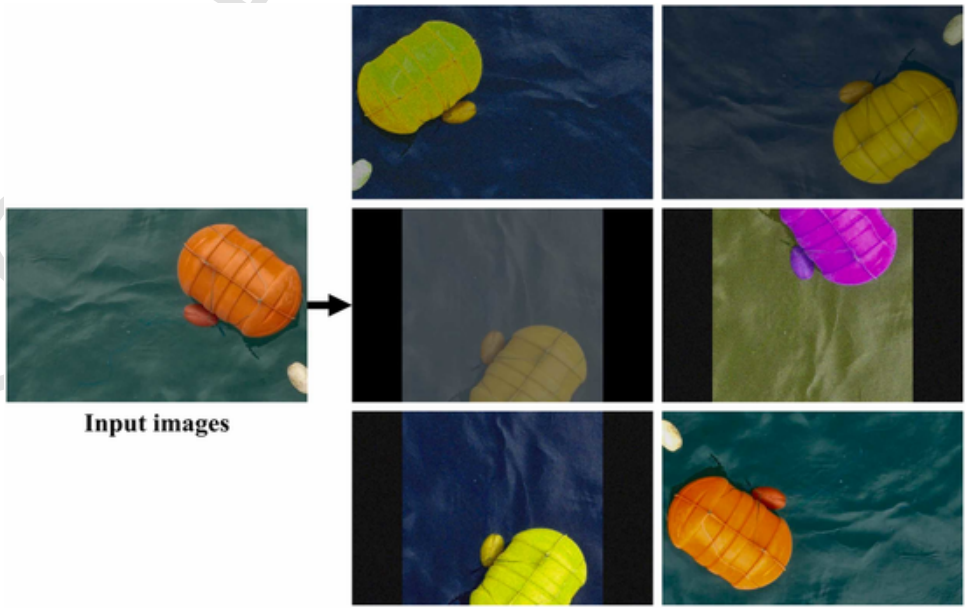

Fig. 4. Six augmented images generated by the data augmentation process.

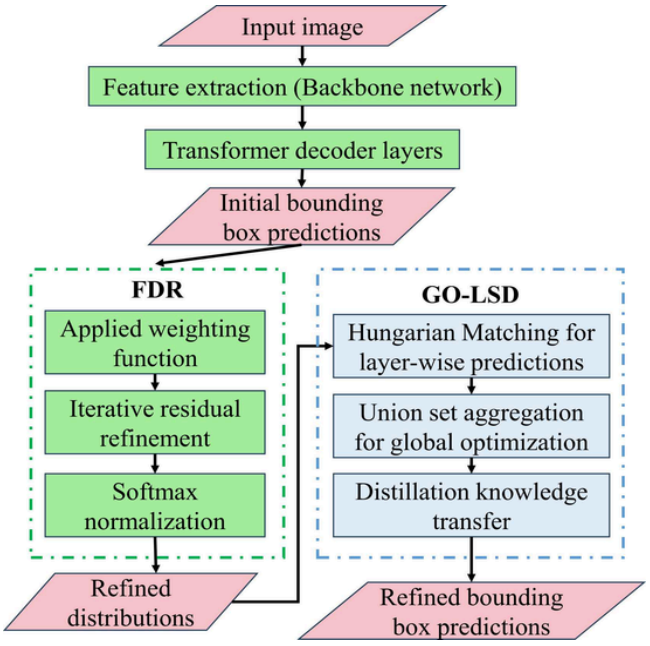

Fig. 5. A flowchart illustrating the refinement process for bounding box predictions using a transformer-based detection model based on FDR and GO-LSD modules.

The FDR module refines residuals iteratively by applying a weighting function and employing softmax normalization to improve the accuracy of edge distance distributions. Simultaneously, the GO-LSD module utilizes Hungarian Matching for precise layer-wise predictions, aggregates bounding box candidates to achieve global optimization, and transfers distilled knowledge to boost performance. The result is refined bounding box predictions with significantly improved localization accuracy. By integrating these techniques, D-FINE enhances the performance of the RT-DETR model while incurring only a minor increase in computational cost, both in terms of parameters and training time.

3.2.1. Real-time DETR (RT-DETR)

RT-DETR consists of three key components: a backbone, a hybrid encoder, and a Transformer decoder equipped with a box prediction head. Initially, the multi-scale outputs from the backbone's final three stages (S_3, S_4, S_5) are fed into the hybrid encoder, which fuses intra-scale interactions and cross-scale information to produce aligned image features. After that, an Intersection over Union (IoU)-aware query selection step is carried out to extract a fixed number of features from the encoder's output sequence, which serves as initial object queries for the decoder. Finally, the decoder, which includes the box prediction head, iteratively refines these queries to obtain box coordinates and confidence scores.

The hybrid encoder in RT-DETR includes two main modules: CNN-based Cross-scale Feature-fusion module (CCFF) and Attention-based Intrascale Feature Interaction (AIFI) module. AIFI operates at the S_5 scale, which applies D-dimensional self-attention for intra-scale interactions while minimizing computational redundancy. AIFI can effectively capture relationships among conceptual elements in the image because it processes higher-level features rich in semantic content. Therefore, AIFI enables downstream modules to perceive object boundaries and attributes more accurately. On the other hand, lower-level features are unsuitable for intra-scale interactions due to their limited semantic information and potential to introduce redundancy or noise when combined with higher-level features. The CCFM module was further optimized and integrated into the overall architecture based on a cross-scale fusion module, which merges the two scale features into a newly generated feature via multiple convolutional layers. The Fusion block consists of N RepBlocks, with two output paths, merged using element-wise addition (Zhao et al., 2024). The process can be mathematically formulated as follows:

$$Q = K = V = \text{Flatten}(S_5)$$
$$F_5 = \text{Reshape}(\text{Attn}(Q, K, V))$$
$$\text{Output} = \text{CCFM}(\{S_3, S_4, F_5\})$$

where *Attn* refers to multi-head self-attention, and *Reshape* restores the feature shape to to match those of S_5 (the inverse of Flatten).

The object query in DETR is optimized within the decoder and mapped by the prediction head to learnable embeddings that produce both classification scores and bounding box predictions. However, these object queries lack explicit physical meaning, which makes them difficult to interpret and optimize. Recent research addresses this by expanding the notion of an object query into a combination of content and position queries (i.e., anchors) (Cui et al., 2023). To solve this problem, RT-DETR proposes IoU-aware query selection (Chen et al., 2023), which promotes high classification scores for features with high IoU and lower classification scores for those with low IoU. The model selects bounding box predictions with both high classification and IoU scores from the encoder's top K features. The detector's objective function is reformulated as follows:

$$\mathscr{L}(\hat{y}, y) = L_{\text{box}}(\hat{b}, b) + L_{\text{cls}}(c, c, \text{IoU}).$$

where $\hat{y} = \{\hat{c}, \hat{b}\}$ and $y = \{c, b\}$ represent the predicted and ground-truth classes and bounding boxes, respectively. $\hat{c}$ and c denote the categories, while $\hat{b}$ and b represent the bounding boxes. IoU scores are applied to the classification branch to enforce consistency between classification and localization for positive samples.

3.2.2. Fine-grained distribution refinement (FDR)

As shown in Fig. 6, FDR tackles the problem of iterative bounding-box refinement for object detection by producing and repeatedly refining probability distributions for each box edge (Peng et al., 2024). Unlike single-pass methods that regress bounding-box coordinates directly, FDR maintains a set of discrete bins representing potential offsets for each edge (top, bottom, left, right). At the first decoder layer, an initial bounding box $\mathbf{b}^0 = \{x, y, W, H\}$ is generated, and then converted into center coordinates $\mathbf{c}^0$ and edge distances $\mathbf{d}^0 = \{t, b, l, r\}$. The network also outputs preliminary probability distributions (via a D-FINE head), which will be iteratively updated by subsequent layers to progressively improve bounding-box accuracy.

Each bounding box is assigned four probability distributions for four edge directions. At layer l, the refined edge distances $\mathbf{d}^l$ are computed by adding residual offsets (scaled by box height/width) to $\mathbf{d}^0$. For the l-th layer, the refined edge distances $\mathbf{d}^l = \{t^l, b^l, l^l, r^l\}$ are computed as:

$$\mathbf{d}^l = \mathbf{d}^0 + \{H, H, W, W\} \cdot \sum_{n=0}^{N} W(n)\mathbf{Pr}^l(n), \quad l \in \{1, 2, \ldots, L\}$$

where $\text{Pr}^l(n) = \{\text{Pr}_t^l(n), \text{Pr}_b^l(n), \text{Pr}_l^l(n), \text{Pr}_r^l(n)\}$ indicates four distributions for four edge. These offsets come from a weighted sum of the distributions $\text{Pr}^l(n)$, where n runs over discrete bins, each mapped to a continuous offset value by a piecewise weighting function $W(n)$. The weighting function is designed with parameters a and c to control the shape of its curve, allowing fine-grained adjustments when the box is nearly accurate and larger shifts when the box is far from its target. To enable iterative refinement, FDR employs a residual approach on the logits.

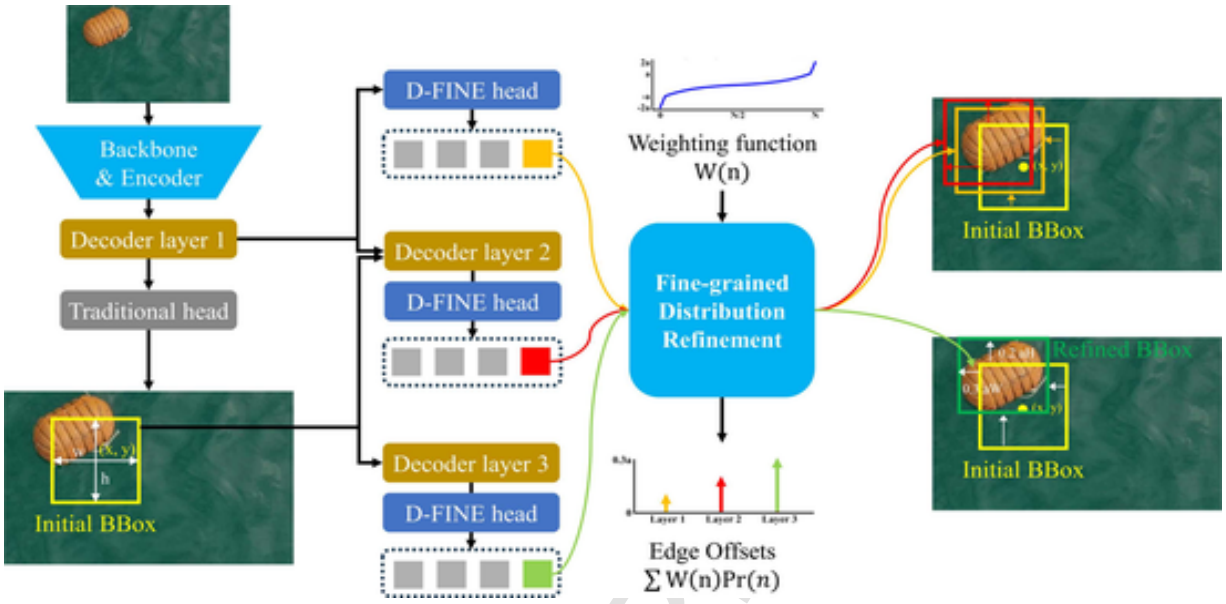

Fig. 6. Overview of D-FINE with FDR. **Note:** The decoder layers iteratively refine the probability distributions, which serve as fine-grained intermediate representations, in a residual manner. Non-uniform weighting functions are applied to enable more precise localization.

$$\Pr^{l}(n) = \text{Softmax}\left(\text{logits}^{l}(n)\right) = \text{Softmax}\left(\Delta\text{logits}^{l}(n) + \text{logits}^{l-1}(n)\right), \quad (1)$$

Specifically, each decoder layer predicts residual adjustments $\Delta\text{logits}^{l}(n)$, which are added to the logits from the previous layer $\text{logits}^{l-1}(n)$. A softmax normalization is then applied to these updated logits to generate refined probability distributions $\Pr^{l}(n)$. By operating in a residual fashion, the model incrementally refines predictions rather than recomputing distributions from scratch at each layer. This strategy leads to more stable and accurate bounding-box updates, as incremental corrections reduce the risk of abrupt changes in the optimization process.

The weighting function $W(n)$ is then defined as follows to facilitate precise and flexible adjustments:

$$W(n) = \begin{cases} 2 \cdot W(1) = -2a & n = 0 \\ c - c\left(\frac{a}{c} + 1\right)^{\frac{N-2n}{N-2}} & 1 \le n < \frac{N}{2} \\ -c + c\left(\frac{a}{c} + 1\right)^{\frac{-N+2n}{N-2}} & \frac{N}{2} \le n \le N-1 \\ 2 \cdot W(N-1) = 2a & n = N, \end{cases}$$

a and c are hyper-parameters that control the upper bounds and the curvature of the function. As illustrated in Fig. 6, the shape of $W(n)$ is designed to allow fine adjustments when the bounding box prediction is close to accurate, due to the minimal curvature in $W(n)$ near its center. On the other hand, for predictions that are significantly off, the steeper curvature near the edges and the sharp transitions at the boundaries of $W(n)$ provide the necessary flexibility for substantial corrections.

Training FDR module utilizes the Fine-Grained Localization (FGL) Loss, which builds on the principles of Distribution Focal Loss. The ground truth offsets $\mathbf{d}^{\text{GT}}$ are compared to the model's predicted distributions using bin-adjacent cross-entropy, where interpolation weights are determined based on proximity of the predicted bounding box edges to their true positions. To further prioritize spatial accuracy, an IoU-based weighting factor IoU_k amplifies the influence of predictions that are already spatially close to ground-truth edges. Altogether, FDR's multi-layer distribution updates, weighting function, and specialized loss function enable a more flexible and precise iterative bounding-box refinement process.

3.2.3. Global optimal localization self-distillation (GO-LSD)

GO-LSD is a novel technique that utilizes refined distribution predictions from the final decoder layer to enhance localization accuracy in shallower layers. The method employs a two-stage process:

- Local Matching: Hungarian Matching (Wang et al., 2024b) is applied to predictions at each decoder layer to establish localized correspondences between predicted and ground-truth bounding boxes.
- Global Aggregation: These layer-specific matches are integrated into a unified union set spanning all decoder layers, which consolidates the most accurate candidate predictions for each target.

The union set not only optimizes the localization performance globally but also ensures that unmatched predictions are refined during training. The classification branch enforces a strict one-to-one matching policy to preserve the distinct roles of classification and localization and prevent redundant bounding box proposals. The union set often contains low-confidence predictions with high localization accuracy. To handle this challenge, GO-LSD introduces the Decoupled Distillation Focal (DDF) Loss, a mechanism designed to balance contributions from matched and unmatched predictions during distillation. The DDF Loss involves weighting strategies customized based on the IoU and classification confidence of predictions.

$$\mathcal{L}_{\text{DDF}} = T^2 \sum_{l=1}^{L-1} \left(\sum_{k=1}^{K_m} \alpha_k \cdot \mathbf{KL}\left(\Pr^{l}(n)_k, \Pr^{L}(n)_k\right) + \sum_{k=1}^{K_u} \beta_k \cdot \mathbf{KL}\left(\Pr^{l}(n)_k, \Pr^{L}(n)_k\right) \right.$$

$$\alpha_k = \text{IoU}_k \cdot \frac{\sqrt{K_m}}{\sqrt{K_m} + \sqrt{K_u}}, \quad \beta_k = \text{Conf}_k \cdot \frac{\sqrt{K_u}}{\sqrt{K_m} + \sqrt{K_u}}$$

Mathematically, the loss is expressed using Kullback-Leibler (KL) divergence between probability distributions predicted by shallower decoder layers and the refined distribution from the final layer. T is a temperature parameter that softens logits for smoother distribution align-

ment. For matched predictions, the weight α_k is proportional to the IoU and a factor balancing matched and unmatched predictions (K_m and K_u). For unmatched predictions, the weight β_k depends on classification confidence and inverse balancing factor. This decoupling weighting strategy allows GO-LSD to emphasize high-IoU, low-confidence predictions, which are critical for refining localization while maintaining stability. By aligning shallower layers with the final layer's refined outputs, the method achieves robust, layer-wise localization improvements without compromising classification integrity.

4. Experimental results

This section provides various experiments to validate the performance of the proposed detection framework using the collected marine litter dataset. Section 4.1 provides a detailed explanation of the evaluation metrics used to measure the model's performance. After that, Section 4.2 describes the hardware setup and programming environment utilized for the model's implementation. Finally, Section 4.3 reports the experimental results of the proposed models on a series of experiments to evaluate the proposed model in various settings.

4.1. Evaluation metrics description

In this study, the performance of the marine litter detection framework is evaluated using three well-established metrics, including mean average precision (mAP), precision, and recall. These metrics provide a comprehensive understanding of the model's ability to detect and accurately localize marine litter under realistic conditions. Precision measures the portion of correct positive predictions out of all positively predicted instances, while recall tracks how many of the actual positive instances are successfully identified. Precision and recall metrics can be formulated as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \end{aligned} \tag{2}$$

where true positive (TP), false positive (FP), and false negative (FN) are important confusion matrix components.

In addition, we employ the widely recognized mAP@0.5 as the main object detection evaluation measure. In this context, a confidence threshold of 0.5 is applied to determine how accurately the model identifies different classes of marine litter. Initially, a precision-recall curve is generated for each class at the 0.5 threshold, and the area under the curve is computed to obtain the AP. The mAP is then calculated by averaging these AP values for all classes, as defined:

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP}_i \tag{3}$$

where n is the total number of classes and AP_i denotes the average precision of the $i-th$ marine debris class.

4.2. Description of the implementation environment of the proposed model

The marine debris detection system is based on PyTorch,[4] a widely used Python library. The operating system is Linux equipped with two Nvidia A100 GPUs (40 GB each GPU). All models and hyperparameters are configured according to open-source implementations provided in the original publications. Each model utilizes a pre-trained HGNetV2 backbone on ImageNet to facilitate fine-tuning Zhao et al. (2024).

[4] https://pytorch.org/

4.3. Comprehensive evaluation of the proposed marine debris detection framework

4.3.1. Analysis of the pre-processing process

Fig. 7 presents four input marine debris images, which exhibit various challenges such as blurriness, poor illumination, and low light. After being processed through the preprocessing module, the outputs show a marked improvement in image quality. For example, marine litter in drone-captured images with poor lighting or low contrast becomes significantly easier to identify after enhancement by the preprocessing stage. Moreover, the preprocessing module preserves the integrity of input images by avoiding the introduction of additional noise or degradation, ensuring that images without these issues remain unaffected.

As described in Table 1, the pre-processing module significantly improves the MD-TR marine litter detection model's detection performance on three metrics, including mAP, precision, and recall. Specifically, the mAP value improved from 76.2 % to 81.5 %, while precision rose from 75.3 % to 79.4 % and recall increased from 77.6 % to 82.1 %. The model performance after applying the pre-processing module demonstrates that it effectively improves input image quality, in terms of contrast and color, which allows the detection model to more accurately locate and identify marine litter. Finally, the introduction of noise and artifacts through data augmentation improves model robustness and adaptability.

4.3.2. Detection model performance evaluation

Fig. 8 provides a detailed performance evaluation of the MD-TR model. In Fig. 8(a), the validation mAP is tracked during the 12 training epochs. The mAP begins at about 0.62 and steadily rises to around 0.81 by the 12th epoch. The consistent improvement indicates that the learning process improves the model's ability to detect marine litter over time. The graph also suggests good generalization and convergence toward high performance in later epochs.

On the other hand, Fig. 8(b) illustrates the bounding box loss (loss_bbox) and IoU loss (loss_iou) during training. Both losses drop sharply during the early stages, which reflects rapid parameter optimization as the model adapts to the training data. After the initial sharp decrease, the losses stabilize and continue to decrease gradually. By the end of the training, both losses converge to low, stable values, which indicates that the model has effectively minimized errors associated with bounding box predictions and overlap estimations.

Table 2 presents the performance of the proposed MD-TR on eight marine debris types in the dataset. Overall, the model demonstrates strong detection capabilities, with an average mAP of 81.5 %, precision of 79.4 %, and recall of 82.1 %. Metal debris shows the highest mAP at 83.1 %, while Rope and Styrofoam pieces obtain the highest precision (81.2 %) and recall (83.4 %), respectively. Although some classes, such as Styrofoam pieces, show relatively lower precision (77.9 %), the obtained performance remains close to the overall average in all categories. These results indicate that the proposed framework can be effectively applied to marine litter detection in real-world scenarios.

4.3.3. Comparison with other state-of-the-art detection models

The results presented in Table 3 demonstrate the superior performance of the proposed MD-TR framework for marine debris detection compared to existing state-of-the-art models. With an mAP of 81.5 %, MD-TR outperforms all other models, including RT-DETR (80.2 %) and DINO (80.7 %), which are known for their robust detection capabilities. The improvement is particularly notable in terms of recall (82.1 %), which indicates that MD-TR is highly effective at identifying a larger proportion of marine debris objects, even in challenging scenarios such as cluttered backgrounds or poor lighting conditions. In addition, MD-TR achieves a precision of 79.4 %, indicating a strong balance between accurately detecting debris and minimizing false positives. The model's

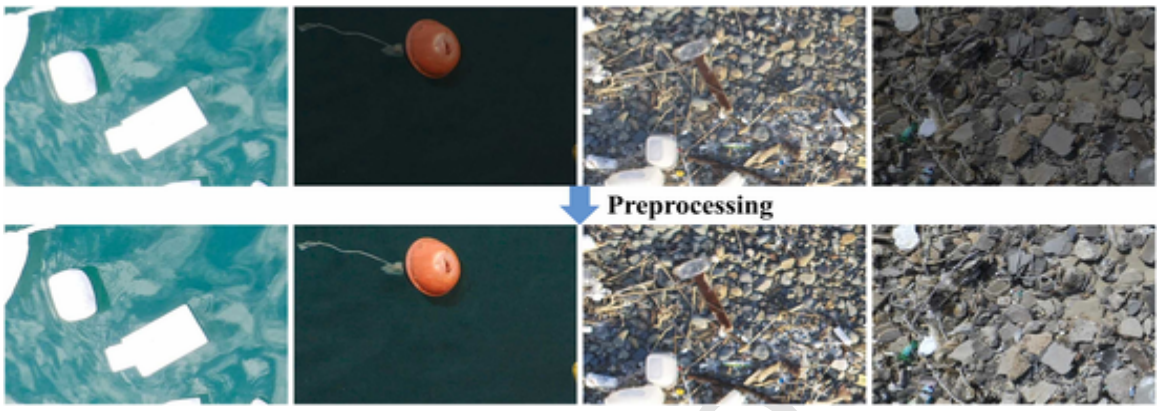

Fig. 7. Comparison of the original and pre-processed marine litter images.

Table 1
Improvements in the MD-TR model's performance following the pre-processing step.

Input data	mAP (%)	Precision (%)	Recall (%)
Raw	76.2	75.3	77.6
Pre-processed	81.5	79.4	82.1

high FPS at 117 further highlights its practical applicability for deployment in dynamic marine environments where rapid detection is critical.

The comparative analysis also highlights the incremental advancements achieved by recent transformer-based architectures like Deformable-DETR, RT-DETR, and DINO, which consistently outperform traditional models such as SSD and Faster R-CNN. For example, Deformable-DETR achieves an mAP of 78.9 %, because it can effectively handle complex spatial relationships through deformable attention mechanisms. Similarly, RT-DETR and DINO further refine different modules and achieve mAP scores of 80.2 % and 80.7 %, respectively. However, MD-TR distinguishes itself not only through its higher accuracy but also through its integration of FDR and GO-LSD. These techniques enhance localization accuracy and robustness. Overall, the results confirm that MD-TR represents a significant step forward in marine debris detection, which is essential for large-scale monitoring and cleanup initiatives.

4.3.4. Qualitative evaluations

This section provides a quantitative analysis of the predictions of the proposed MD-TR framework by visualizing detection performance and attention weight analysis for various real-life samples. Fig. 9 illustrates the detection performance of the MD-TR model on four real-life scenarios, including both floating and coastal waste. Although each example contains variations in scene complexity, lighting conditions, and background textures, the model successfully identifies all marine litter items, such as buoys, nets, ropes, plastic bottles, metal debris, and glass. Notably, the bounding boxes are tightly fitted around target objects, with confidence scores indicating high model certainty.

In the floating waste examples, the model demonstrates a strong ability to handle reflective surfaces and partial occlusions as it accurately detects and distinguishes closely clustered plastic objects. Similarly, in the coastal scenes, the model demonstrates stable detection of various materials scattered on different terrains. These results highlight the versatility and robustness of the MD-TR model in detecting diverse debris types in dynamic marine environments. By utilizing a Transformer structure and an effective pre-processing pipeline, the model shows that it is well-suited for real-world applications where water reflection, variable weather conditions, and cluttered shorelines can complicate detection tasks. The consistent performance observed in both nearshore and offshore contexts highlights the effectiveness of the model's design and training strategy.

The attention mechanism in the MD-TR model, as visualized in Fig. 10 and Fig. 11, demonstrates remarkable effectiveness in identifying and prioritizing marine debris in diverse real-life scenarios. The

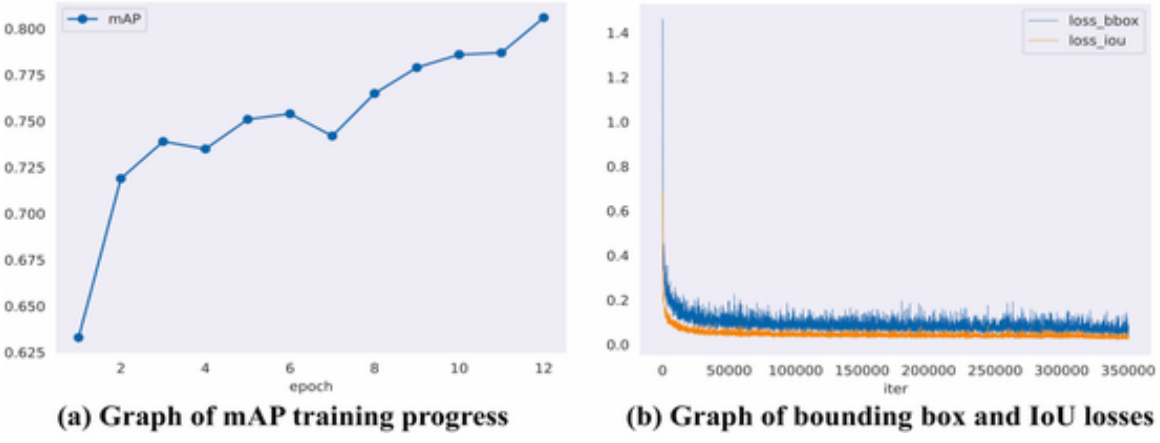

Fig. 8. Detailed performance evaluation of the MD-TR model using different evaluation metrics.

Table 2
Detailed MD-TR performances on each marine debris class.

Litter type	mAP	Precision	Recall
Buoys	82.4	78.9	81.4
Fishing nets	80.2	79.8	82.2
Glass debris	80.9	80.5	82.9
Metal debris	83.1	78.6	80.7
PET bottles	81.6	79	81.2
Plastic debris	81.7	79.3	81.9
Rope	79.8	81.2	83.1
Styrofoam pieces	82.3	77.9	83.4
Average	81.5	79.4	82.1

Table 3
Performance of the proposed framework compared to other state-of-the-art detection models using the testing set.

Model	mAP	Precision	Recall	FPS
SSD (Liu et al., 2016)	72.6	72.1	73.8	120
Faster-RCNN (Ren et al., 2016)	76.3	76.9	75.2	50
YOLOv10 (Wang et al., 2024a)	79.1	77.5	78.4	140
DETR (Carion et al., 2020)	78.5	77.8	79.6	45
Deformable-DETR Zhu et al. (2020)	78.9	79.2	80.5	55
RT-DETR (Zhao et al., 2024)	80.2	78.8	80.7	90
DINO Zhang et al. (2022)	80.7	79	80.6	51
MD-TR (Our)	81.5	79.4	82.1	117

heatmaps on the right side of each image pair highlight areas where the model focuses its attention, with warmer colors (red and yellow) indicating higher attention weights and cooler colors (blue) showing lower emphasis. Across rocky coastlines, grassy shorelines, and open-water regions, the model demonstrates robust detection capabilities by focusing on marine debris (e.g., bottles, cans, buoy, and plastic fragments) while effectively filtering out environmental distractors like rocks, vegetation, reflection, and wave patterns. This selective focus suggests that the attention mechanism successfully captures global context, enabling the model to discern subtle patterns and object boundaries in complex, cluttered environments, which is critical for accurate debris detection. Moreover, the attention weights reveal the model's ability to adapt to varying environmental conditions, such as different terrains, lighting, and water reflections, as seen in the distinct heatmap patterns across the samples. For instance, the model's attention is tightly concentrated around debris objects in the water, even amidst distracting wave patterns, indicating robust feature extraction and localization. The consistency and precision of these attention visualizations highlight the proposed model's capacity to enhance interpretability and performance.

5. Discussion

The proposed MD-TR model demonstrates significant advancements in marine debris detection by combining a robust Transformer architecture with a carefully designed pre-processing pipeline. The pre-processing module is critical in addressing the complex marine environment. The module allows the proposed framework to obtain a remarkable 5.3 % increase in the mAP compared to its mAP performance on the original dataset (76.2 %). As a result, the pre-processing module substantially reduces missed or misclassified marine debris from the collected dataset. Although the preprocessing module requires additional computational resources, it offers flexibility by being easily activated or deactivated depending on the specific application requirements. Overall, the proposed pre-processing significantly enhances the detection of marine litter in challenging marine environments.

In addition, the proposed MD-TR model achieves a high detection mAP of 81.5 %, which highlights its effectiveness in detecting various marine debris in diverse and challenging marine environments. When compared to state-of-the-art models such as SSD, YOLOv10, DINO, and RT-DETR, the MD-TR model shows high mAP and a competitive inference speed of 117 FPS on an NVIDIA A100 GPU. The high performance and real-time efficiency make it particularly well-suited for large-scale

Fig. 9. Visualization of the proposed MD-TR model on four different real-life samples. Two samples for coastal waste and two samples for floating waste.

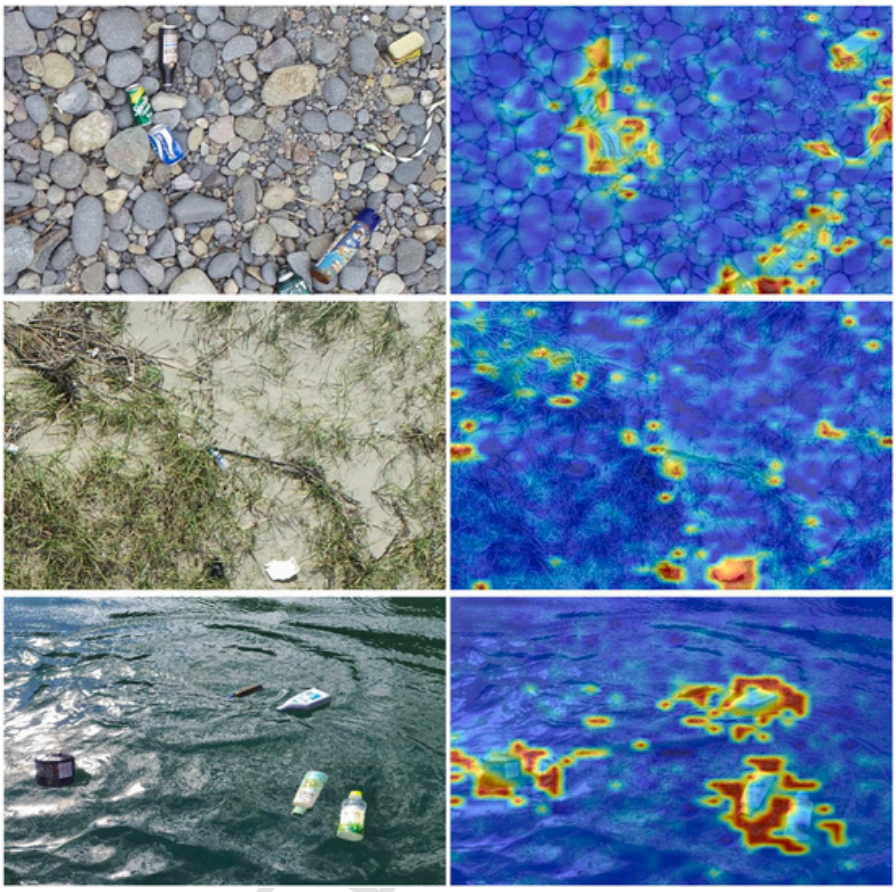

Fig. 10. The MD-TR model's attention weights analysis results for six real-life samples.

marine debris monitoring programs. Moreover, experimental results demonstrate the model's robustness to variations in the appearance of marine debris, such as differences in shape, color, and size, as well as its ability to generalize effectively on different marine terrains, including coastal waste and floating waste.

Over the past decade, while numerous studies on DL-based marine debris detection have demonstrated their superior performance over traditional ML algorithms, the interpretability of these models has often been overlooked. Interpretability is a crucial factor in marine debris monitoring, as it enhances trust among experts and stakeholders in the automated decision-making process. This study emphasizes the importance of interpretability in marine debris detection frameworks by relying on the attention mechanism of Transformer architectures. The MD-TR model effectively extracts and visualizes multi-scale attention weights from the decoder, a distinctive feature that enhances transparency and provides insights into the model's predictions.

6. Conclusions and future works

In this study, we proposed the MD-TR model, a Transformer-based framework for detecting marine debris with high accuracy and efficiency. By utilizing an effective pre-processing pipeline and Transformer-based D-FINE approach with two key components FDR and GO-LSD, the MD-TR framework outperformed other state-of-the-art detection models with an average mAP of 81.5 %. Its competitive inference speed of 117 FPS makes it suitable for real-time deployment, where there is an urgent need for efficient and scalable solutions in marine debris monitoring. In addition, the model demonstrated robustness to variations in marine litter appearance, including size, shape, and color, and generalization across diverse environmental conditions such as coastal and floating waste scenarios. These advantages position MD-TR as a reliable and practical tool for large-scale marine pollution assessment and ecosystem health monitoring. A key innovation of this framework is its emphasis on model interpretability, which enables the analysis and visualization of learned attention weights from the Transformer architecture. This unique feature enhances transparency and helps users to comprehend and improves trust in the model's automated decision-making process.

Despite its strong performance, there are still opportunities for improvement. For example, future work could focus on improving the detection of small or occluded objects through advanced multi-scale feature extraction or the integration of additional context-aware modules. Moreover, while the model achieves impressive results with the current dataset, expanding the dataset to cover more diverse conditions and debris types could improve its generalizability to other environments. Future work could also explore the integration of lightweight MD-TR with autonomous drones equipped for continuous and large-scale monitoring of marine ecosystems.

CRediT authorship contribution statement

L. Minh Dang: Writing – review & editing, Writing – original draft, Data curation. **Ngoc Dung Bui:** Investigation, Formal analysis. **ASM Sharifuzzaman Sagar:** Visualization, Data curation. **Tri-Hai Nguyen:** Supervision, Methodology, Writing – review & editing. **Luong Vuong Nguyen:** Validation, Software.

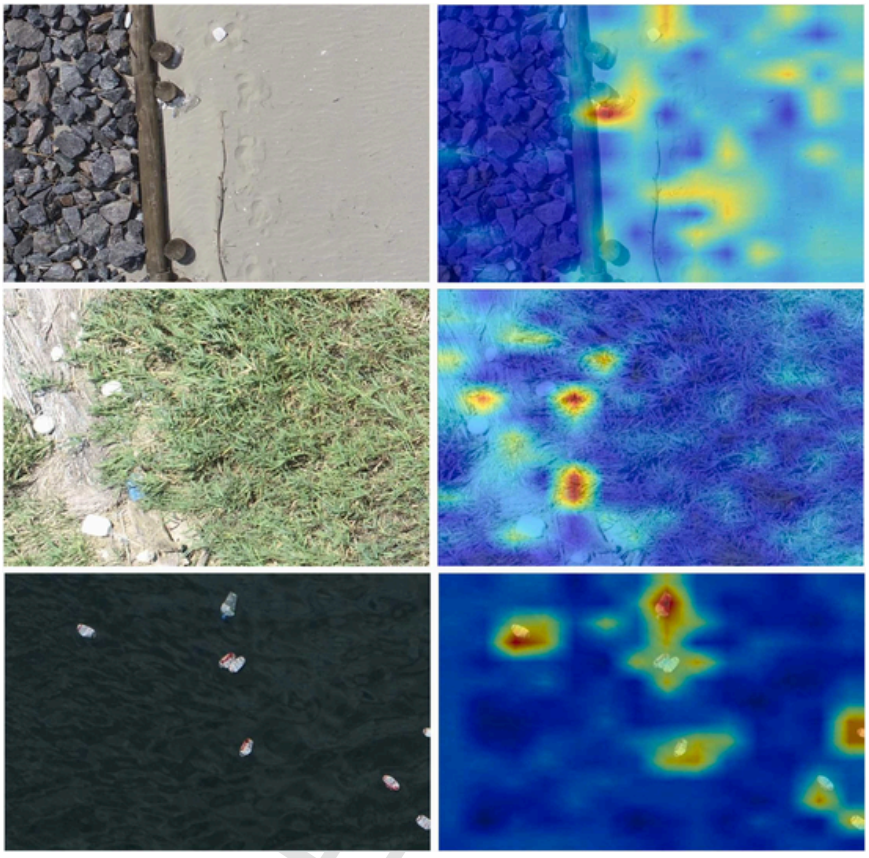

Fig. 11. The MD-TR model's attention weights analysis results for six real-life samples (continue).

Declaration of Competing Interest

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

Data availability

The data is publicly available on http://www.aihub.or.kr/.

References

Bishop, G., Styles, D., Lens, P.N., 2020. Recycling of european plastic is a pathway for plastic debris in the ocean. Environ. Int. 142, 105893.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, European conference on computer vision. Springer, pp. 213–229.

Chen, F., Zhang, H., Hu, K., Huang, Y.K., Zhu, C., Savvides, M., 2023. Enhanced training of query-based object detection via selective query recollection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23756–23765.

Chen, X., Zhang, P., Quan, L., Yi, C., Lu, C., 2021. Underwater image enhancement based on deep learning and image formation model.arXiv preprint arXiv:2101.00991.

Cui, Y., Yang, L., Yu, H., 2023. Learning dynamic query combinations for transformer-based object detection and segmentation, in: International Conference on Machine Learning, PMLR.6591-6602.

Dang, L.M., Wang, H., Li, Y., Nguyen, L.Q., Nguyen, T.N., Song, H.K., Moon, H., 2023. Lightweight pixel-level semantic segmentation and analysis for sewer defects using deep learning. Constr. Build. Mater. 371, 130792.

Deep-sea debris database, 2023. Accessed: September 21.⟨https://www.godac.jamstec.go.jp/dsdebris/e/index.html⟩.

Dosset, A., Dang, L.M., Alharbi, F., Habib, S., Alam, N., Park, H.Y., Moon, H., 2024. Cassava disease detection using a lightweight modified soft attention network.Pest Management Science.

Escobar-Sánchez, G., Markfort, G., Berghald, M., Ritzenhofen, L., Schernewski, G., 2022. Aerial and underwater drones for marine litter monitoring in shallow coastal waters: factors influencing item detection and cost-efficiency. Environ. Monit. Assess. 194, 863.

Gong, Y., Wang, Y., Frei, R., Wang, B., Zhao, C., 2022. Blockchain application in circular marine plastic debris management. Ind. Mark. Manag. 102, 164–176.

Hong, J., Fulton, M., Sattar, J., 2020. Trashcan: A semantically-segmented dataset towards visual detection of marine debris.arXiv preprint arXiv:2007.08097.

Huang, B., Chen, G., Zhang, H., Hou, G., Radenkovic, M., 2023. Instant deep sea debris detection for maneuverable underwater machines to build sustainable ocean using deep neural network. Sci. Total Environ. 878, 162826.

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M., 2022. Transformers in vision: a survey. ACM Comput. Surv. (CSUR) 54, 1–41.

Li, Y., Wang, H., Dang, L.M., Song, H.K., Moon, H., 2023. Attention-guided multiscale neural network for defect detection in sewer pipelines. Comput. -Aided Civ. Infrastruct. Eng. 38, 2163–2179.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37.

Nadeem, M., Dilshad, N., Alghamdi, N.S., Dang, L.M., Song, H.K., Nam, J., Moon, H., 2023. Visual intelligence in smart cities: a lightweight deep learning model for fire detection in an iot environment. Smart Cities 6, 2245–2259.

Nguyen, L.Q., Choi, J., Dang, L.M., Moon, H., 2024. Background debiased class incremental learning for video action recognition. Image Vis. Comput. 151, 105295.

Peng, Y., Li, H., Wu, P., Zhang, Y., Sun, X., Wu, F., 2024. D-fine: Redefine regression task in detrs as fine-grained distribution refinement.arXiv preprint arXiv:2410.13842.

Politikos, D.V., Fakiris, E., Davvetas, A., Klampanos, I.A., Papatheodorou, G., 2021. Automatic detection of seafloor marine litter using towed camera images and deep learning. Mar. Pollut. Bull. 164, 111974.

Qi, Y., Yang, Z., Sun, W., Lou, M., Lian, J., Zhao, W., Deng, X., Ma, Y., 2021. A comprehensive overview of image enhancement techniques. Arch. Comput. Methods Eng. 1–25.

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1137–1149.

Shen, A., Zhu, Y., Angelov, P., Jiang, R., 2024. Marine debris detection in satellite surveillance using attention mechanisms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.

Sugianto, E., Chen, J.H., Purba, N., 2023. Cleaning technology for marine debris: a review of current status and evaluation. Int. J. Environ. Sci. Technol. 20, 4549–4568.

Sun, Z., Cao, S., Yang, Y., Kitani, K.M., 2021. Rethinking transformer-based set prediction for object detection, Proceedings of the IEEE/CVF international conference on

computer vision. pp. 3611–3620.
Taddia, Y., Corbau, C., Buoninsegni, J., Simeoni, U., Pellegrinelli, A., 2021. Uav approach for detecting plastic marine debris on the beach: a case study in the po river delta (italy). Drones 5, 140.
Tightiz, L., Dang, L.M., Padmanaban, S., Hur, K., 2024. Metaverse-driven smart grid architecture. Energy Rep. 12, 2014–2025.
Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G., 2024a. Yolov10: Real-time end-to-end object detection.arXiv preprint arXiv:2405.14458.
Wang, J., Dong, J., Tang, M., Yao, J., Li, X., Kong, D., Zhao, K., 2023. Identification and detection of microplastic particles in marine environment by using improved faster r-cnn model. J. Environ. Manag. 345, 118802.
Wang, S., Xia, C., Lv, F., Shi, Y., 2024b. Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision.arXiv preprint arXiv:2409.08475.
Woods, J.S., Verones, F., Jolliet, O., Vázquez-Rowe, I., Boulay, A.M., 2021. A framework for the assessment of marine litter impacts in life cycle impact assessment. Ecol. Indic. 129, 107918.
Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y., 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection.arXiv preprint arXiv:2203.03605.
Zhang, Z., Zhu, L., 2023. A review on unmanned aerial vehicle remote sensing: Platforms, sensors, data processing methods, and applications. Drones 7, 398.
Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J., 2024. Detrs beat yolos on real-time object detection, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16965–16974.
Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: Deformable transformers for end-to-end object detection.arXiv preprint arXiv:2010.04159.