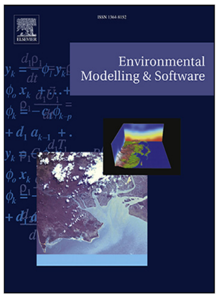


Automated marine litter investigation for underwater images using a zero-shot pipeline

Tri-Hai Nguyen, Minh Dang

PII: S1364-8152(24)00126-9
DOI: https://doi.org/10.1016/j.envsoft.2024.106065
Reference: ENSO 106065

To appear in: *Environmental Modelling and Software*

Received date : 22 January 2024
Revised date : 28 April 2024
Accepted date : 5 May 2024

Automated Marine Litter Investigation for Underwater Images using a Zero-shot Pipeline

Tri-Hai Nguyen[a], Minh Dang[b,c,*]

[a] Faculty of Computer Science, Ho Chi Minh City Open University, Ho Chi Minh City 700000, Vietnam
[b] Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
[c] Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam

Abstract

Accurate and automated identification of marine litter on the seafloor is crucial due to its detrimental effects on marine ecosystems. While advancements in underwater imaging have facilitated this task, the significant human involvement required in traditional approaches necessitates the development of more efficient and cost-effective solutions. This study presents an efficient zero-shot segmentation framework based on Segment-Anything (SAM) guided by Interpretable Contrastive Language–Image Pre-training (iCLIP) for identifying and segmenting eight common seafloor litter categories in realistic underwater environments without model training. The framework supports prompt input by design, which allows it to transfer its zero-shot capabilities to new types of marine litter. To further improve the framework's performance, two additional components were incorporated: an underwater image enhancement model that addresses the degraded image quality common in underwater environments, and a mask post-processing algorithm that reduces noise masks generated by the framework. The recorded mean intersection over union (mIOU) of 69.9% on the testing dataset suggested that zero-shot approaches have the potential to become a valuable technique for automatically detecting marine litter during surveys and enabling continuous and accurate litter monitoring.

Keywords: deep learning, zero-shot, marine litter, segmentation, waste management

*Corresponding author
Email address: `danglienminh@duytan.edu.vn` (L. Minh Dang)

1. Introduction

Marine debris, also known as marine litter or ocean trash, is a wide range of human-made waste that enters oceans, seas, and other water bodies. It includes items such as plastics, glass, metals, paper, textiles, and other materials, both macroscopic and microscopic. Marine debris is a pressing global environmental problem with far-reaching consequences for marine life, habitats, ecosystems, economies, and human health (Iñiguez et al., 2016). It harms marine creatures that can ingest or become entangled in debris, leading to injuries, suffocation, or death. The entanglement and ingestion of debris can disrupt the reproductive cycles and feeding habits of marine creatures, potentially impacting the abundance and quality of seafood available for human consumption (Jang et al., 2020). Additionally, it affects human health due to the consumption of seafood contaminated by litter. The sources and accumulation patterns of marine litter are highly diverse, influenced by factors such as geographical location, industrial activities, waste management practices, and human behavior. These multifaceted factors contribute to the complexity of the issue, necessitating comprehensive solutions (Jia et al., 2023; Galgani et al., 2019).

The growing technological capabilities of underwater observation technologies and computer vision advancements have led to the widespread adoption of photography-based monitoring for assessing the type, distribution, and abundance of marine litter (Radeta et al., 2022). This approach provides valuable insights into the severity of marine litter pollution, enabling the development of effective cleanup programs and fostering public awareness campaigns to reduce litter generation (Politikos et al., 2021; Kraft et al., 2021). The detection of seafloor litter in real-world underwater video footage presents a complex challenge due to the diverse and dynamic nature of marine environments (Jian et al., 2021). Video footage can exhibit varying lighting conditions, zoom levels, and camera angles, often causing marine litter to be barely visible (Raveendran et al., 2021). Moreover, the sheer variety of litter types, the diverse shapes within the same type of litter, the degradation of litter over time, its potential burial in the seabed, and the presence of complex background like rocks and seagrass can easily mislead detection algorithms (Schneider et al., 2018; Mæland and Staupe-Delgado, 2020). The development of algorithms capable of automatically detecting marine litter in underwater images would support analytical processes and play an imperative role in improving understanding of marine pollution and motivating targeted mitigation and management strategies (Sandra et al., 2023).

Deep learning (DL) has emerged as a powerful tool for image understanding across various domains, including computer vision (Marin et al., 2021), natural language processing, recommender system, and anomaly detection. Object segmentation, a subfield of DL, extends beyond object recognition by localizing, classifying, and segmenting objects within images (Minh et al., 2022). In recent efforts, researchers have implemented existing object segmentation models to identify the position of marine litter (Zhou et al., 2022; Teng et al., 2022; Chin et al., 2022), while others have focused on strengthening backbone networks to enhance marine litter feature extraction (Politikos et al., 2021; Corrigan et al., 2023). Additionally, several studies have suggested efficient and lightweight DL structures fine-tuned for marine litter detection (Deng et al., 2021; Ma et al., 2023). Despite these advancements, the supervised marine litter recognition approach still faces limitations. Models trained on a finite set of classes often exhibit restricted performance when encountering novel classes (Madricardo et al., 2020). This limitation stems from the reliance on labeled data, which can be scarce and expensive to acquire for the vast diversity of marine litter categories.

Zero-shot learning is a transformative machine learning (ML) paradigm that enables models to recognize classes or categories they have never encountered during the training phase (Sun et al., 2021). By leveraging a broader set of related information during training, ZSL enables models to generalize and make predictions for unseen or novel classes. The Segment Anything Model (SAM) (Kirillov et al., 2023), developed by Meta AI, represents a pioneering method in image segmentation, demonstrating remarkable generalization capabilities across various benchmark datasets without the need for additional training on unseen objects. ZSL holds particular promise for marine litter recognition due to the vast diversity of marine litter types and the challenges associated with collecting labeled data for each (Raveendran et al., 2021; Schneider et al., 2018). The development of an automatic marine litter recognition framework powered by ZSL has the potential to revolutionize litter assessment by providing a faster, more cost-effective alternative to standard manual data analysis approaches.

Therefore, the need for an efficient and accurate system for segmenting underwater objects is essential for the identification and cleanup of marine litter. This paper proposes a zero-shot pipeline for deep learning-based marine litter segmentation that overcomes the challenges of limited labeled data and complex seafloor environments. Our contributions include: (1) using underwater image enhancement (UIE) algorithms to improve dataset image quality; (2) developing a zero-shot segmentation approach based on SAM guided by Interpretable Contrastive Language–Image Pre-training (iCLIP) algorithms, which eliminates the need for manual data annotation; and (3) demonstrating that the proposed framework achieves comparable segmentation performance and inference speed to the supervised approach.

The remainder of this paper is organized as follows. Section 2 introduces the large-scale marine litter dataset. Section 3 presents the zero-shot marine litter segmentation pipeline in detail. Section 4 evaluates the proposed approach on experimental data and discusses the obtained results of the zero-shot segmentation approach. Finally, Section 5 concludes the paper with some remarks.

2. Marine litter dataset

Previous marine litter studies have been limited by small datasets with few litter types. For example, the seafloor marine litter dataset (635 images) (Politikos et al., 2021), the JAMSTEC dataset (5352 images) (dat), and the DSDebris dataset (15K images) (Huang et al., 2023). As a result, this study uses a massive dataset of around 112K images that cover eight common marine litter types, surpassing previous datasets in both quantity and quality. Although the specific composition of marine litter can vary depending on geographical location and dominant industries, the chosen categories in the dataset represent a significant portion of debris found globally (Politikos et al., 2021; Huang et al., 2023), making it relevant for various coastal monitoring and cleanup scenarios.

The dataset was shared by the National Information Society Agency of Korea (NIA) for research purposes[1]. It was mainly collected by Pukyong Ocean Technology Co., Ltd. and labeled by the Pukyong University Industry-Academia Cooperation Division[2]. The marine litter dataset was collected using a commercial GoPro5 action camera (gop). Eight surveys were conducted to assess seafloor litter, covering more than 100 hectares of seafloor over 8 hours of underwater video footage. The collection was planned for cloudless days between 11:00 AM and 1:00 PM, during solar noon, to ensure the best

[1] `https://aihub.or.kr/`
[2] `https://www.pknu.ac.kr/eng`

contrast and clarity in the videos. Video frames with high turbidity, color shifts, or light flares were excluded from the analysis. Each image is 1920×1080 pixels at 96 dpi. Sample images for each litter type are shown in Figure 1. A total of 111,890 images were collected and annotated, of which 89,512 (80%) were used for training and validation, and 22,378 (20%) for testing.

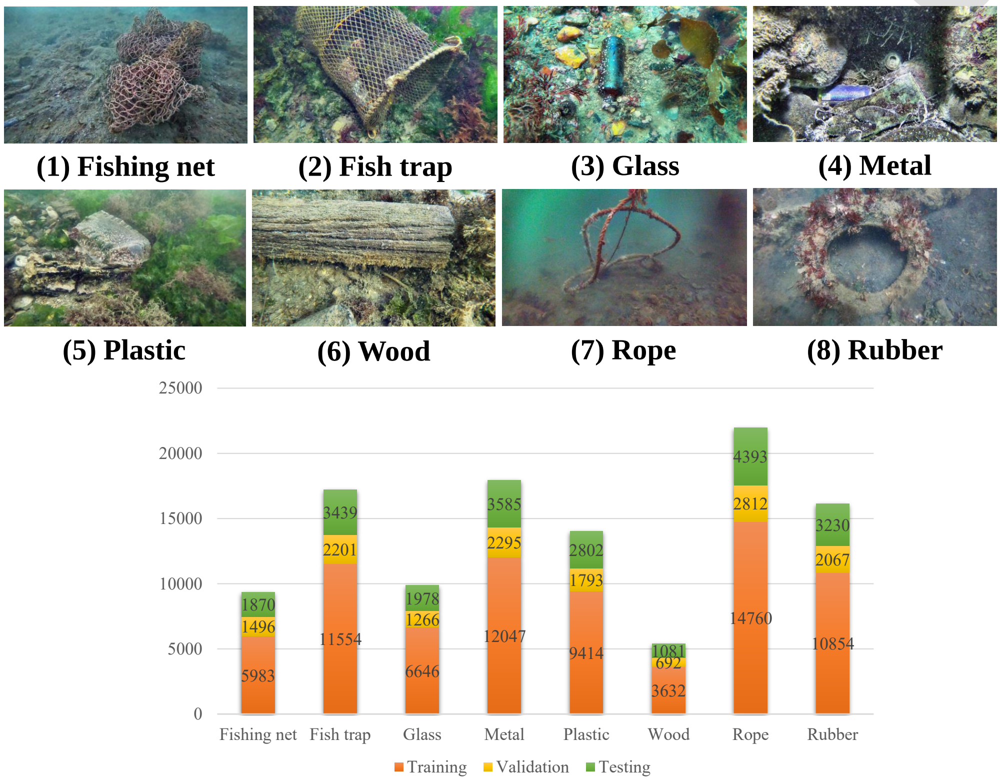


Figure 1: Visual representation of the eight most common types of marine litter in the dataset and a bar chart depicting the distribution of training, validation, and testing images across different marine litter types.

3. Methodology

3.1. Image pre-processing

Previous studies have shown that aquatic datasets often suffer from challenges such as poor lighting, color distortion, low contrast, and reduced visibility due to light scattering and absorption in water. These challenges can significantly degrade the performance of litter identification models during training (Radeta et al., 2022). Underwater image enhancement (UIE) is a common technique that can be implemented to mitigating these issues. UIE is essential for improving the performance of image recognition models, as it reduces the gap between the underwater and the terrestrial domains and enhances the discriminative features of the objects (Huang and Belongie, 2017).

UIE aims to improve the visual quality of images captured in underwater environments, where factors like light attenuation, scattering, and color distortion severely degrade image clarity (Raveendran et al., 2021). One common approach involves the restoration of images using various noise reduction methods, such as filtering or statistical approaches, to improve the image's clarity. Additionally, color correction techniques are employed to mitigate the color shifts caused by the absorption and scattering of light in water. Another notable approach to underwater image enhancement involves leveraging ML algorithms, particularly DL. Such approaches offer the advantage of adaptability and scalability, as the models can continuously improve with more training data, making them suitable for various underwater imaging applications (Gong et al., 2021).

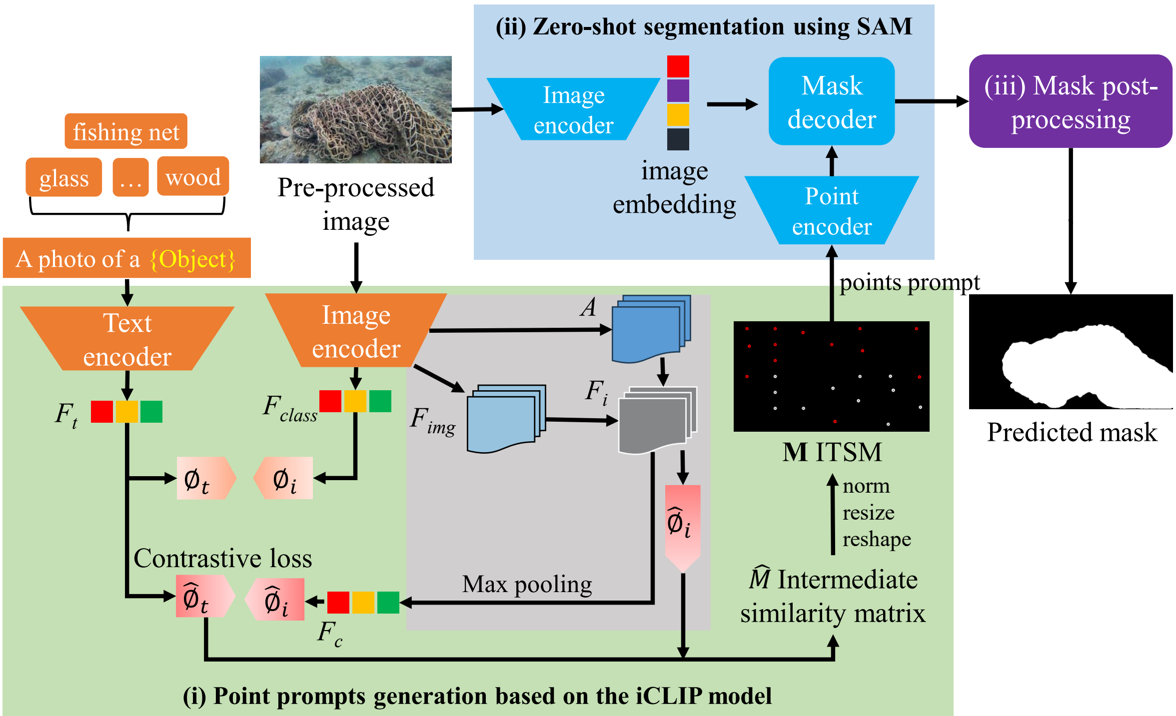
The main challenge of previous DL-based UIE is obtaining high-quality ground truth images (Raveendran et al., 2021). Most existing methods generate approximate reference images and train deterministic enhancement networks that cannot handle the ambiguity of reference mapping. To address the challenge of obtaining high-quality ground truth images for UIE, we implemented a probabilistic network for underwater image enhancement (P-UIE) trained on real-world datasets (Fu et al., 2022). The P-UIE model has two main branches, each implementing a U-Net model with modified SE-ResNet blocks (Gong et al., 2021) for enhanced image feature extraction. The first branch estimates the prior distribution of a single raw underwater image, while the second branch constructs the posterior distribution of UIE using the raw underwater image and corresponding reference image as input. The key component of P-UIE is PAdaIN that uses a conditional variational autoencoder (CVAE) (Sohn et al., 2015) and adaptive instance normalization (AdaIN) (Huang and Belongie, 2017) to create a model of the enhancement distribution. During training, random samples from the posterior distribution of the enhanced underwater image are injected into the AdaIN module to transform the enhanced representation. During testing, random samples from the prior distribution are used to make predictions.

One of P-UIE's strengths is its ability to handle the uncertainty of ground truth labels in UIE data.

Traditional UIE methods often struggle with this challenge due to noisy and inaccurate ground truth labels. P-UIE's robustness to uncertainty makes it a more reliable UIE method. Another strength of P-UIE is its ability to generate diverse enhanced images from a single input underwater image. This versatility makes P-UIE suitable for a variety of uses, including underwater photography, inspection, and surveillance. As a result, this study implemented a pretrained P-UIE model for enhancing the marine litter dataset before performing the marine litter identification.

3.2. Zero-shot seafloor litter segmentation pipeline

Figure 2 provides a schematic diagram of the proposed zero-shot seafloor litter segmentation pipeline, which consists of three main phases: iCLIP model for point prompt generation, SAM for zero-shot segmentation, and mask post-processing process for removing duplicated masks. iCLIP generates point prompts from an input image and text description of the object of interest, which guide the SAM segmentation model to focus on those regions. However, the obtained masks may contain duplicate masks and noise blobs from the background. To address this, we propose a mask post-processing module to eliminate duplication.



Note: Class features are denoted as F_{class}, image features as F_{img}, expanded mean attention map as A, text features as F_t, pooled features as F_c, and masked features as F_i. Additionally, there are dual projections, namely ϕ_i and $\hat{\phi}_i$, along with corresponding text projections, ϕ_t and $\hat{\phi}_t$, used in computing the contrastive losses.

Figure 2: Depiction of the text to points prompt from iCLIP to guide SAM for generating the mask of various types of seafloor litter.

3.2.1. Point prompts generation

Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) is a self-supervised learning approach that learns to encode images and text into a common representation where semantically similar images and text are mapped to nearby points. CLIP models have outperformed all other methods on a variety of downstream vision tasks, including zero-shot classification (Wei et al., 2023), image retrieval (Saito et al., 2023), and object segmentation (Kirillov et al., 2023). However, the visual interpretability of CLIP models has been a relatively underexplored area.

Li et al. (Li et al., 2022) propose a new interpretable CLIP (iCLIP) model that visualizes the feature maps of CLIP models. The iCLIP model introduces an Image-Text Similarity Map (ITSM) that computes the similarity between each image's feature map and the embedding for the corresponding text description. The ITSM can be used to identify the image regions most relevant to the text description. Additionally, the authors replace CLIP's original global pooling layer with a masked max pooling layer that pools over only the image regions relevant to the text description, as determined by the ITSM.

Given an image sample x and text supervision y, the self-supervised image encoder f_i and linear projection ϕ_i (a function that learns to project the output of the encoder to a lower dimension) produce L^p-normalized image token features $\boldsymbol{X} \in \mathbb{R}^{1+N_i,C}$, as shown in Equation 1.

$$\hat{\boldsymbol{X}} = f_i(x) \cdot \phi_i, \boldsymbol{X} = \frac{\hat{\boldsymbol{X}}}{\|\hat{\boldsymbol{X}}\|_p} \tag{1}$$

where the class token 1 and the image token N_i are represented as vectors in an embedding space of width C. The feature matrix $\hat{X}$ contains the features of the image before they are normalized. Similarly, the normalized text features $\boldsymbol{Y} \in \mathbb{R}^{N_t,C}$, are computed as shown in Equation 2. These features are used to train the ITSM model during training and become weights for the ITSM model during inference.

$$\hat{\boldsymbol{Y}} = f_t(y) \cdot \phi_t, \boldsymbol{Y} = \frac{\hat{\boldsymbol{Y}}}{\|\hat{\boldsymbol{Y}}\|_p} \tag{2}$$

After that, the intermediate similarity matrix $\hat{M} \in \mathbb{R}^{N_i,N_t}$ is computed by inner production between image features $\boldsymbol{X}_{1:,:}$ (excluding the class token $\boldsymbol{X}_{:1,:}$) and the transposed text features $\boldsymbol{Y}^\top$, as shown in Equation 3.

$$\hat{M} = \boldsymbol{X}_{1:,:} \times \boldsymbol{Y}^\top \tag{3}$$

The ITSM feature map $\mathbf{M} \in \mathbb{R}^{H,W,N_t}$ is then reconstructed reshaping and resizing it to the input image's size using bicubic interpolation, with width and height W and H, respectively. Additionally, min-max normalization $Norm$ is applied to the H and W dimensions to improve contrast for visualization. The obtained ITSM can be formulated as follows:

$$\mathbf{M} = \text{Norm}(\text{Resize}(\text{Reshape}(\hat{\boldsymbol{M}}))) \tag{4}$$

For interactive segmentation, iCLIP's output points with similarity scores higher than 0.8 are used to guide the SAM model (Li et al., 2022), and the same number of last-ranked points are assigned as background points. This helps to reduce the need for manual labeling and avoid the poor performance of SAM with text prompts only.

3.2.2. Zero-shot segmentation

The Segment Anything Model (SAM) is a new prompt engineering-based semantic segmentation method introduced by Kirillov et al. (Kirillov et al., 2023). SAM is a promptable model, which means that it can segment objects in images using a simple prompt. SAM is trained on the SA-1B dataset, a large dataset of images and text descriptions, and can segment a wide variety of objects, even those not explicitly defined in the training data.

As illustrated in Figure 2(ii), The SAM model architecture consists of three key modules: an image encoder, a prompt encoder, and a mask decoder. The image encoder processes the input image and extracts essential visual features that are versatile enough to apply across various object classes in the context of zero-shot segmentation. It uses vision transformers (ViTs) (Dosovitskiy et al., 2020) to divide the image into patches and extract features from each patch, capturing both object-specific details and background information. The prompt encoder can handle two types of prompts: sparse (points, boxes, and texts) and dense (masks) prompts. Since the location of marine litter in the input image is unknown, we used the points prompt proposed by the iCLIP model to feed into the prompt encoder, which encodes the points prompt into a latent representation. Finally, the prompt encoder output is concatenated with the image encoder output and fed into the mask decoder, which predicts a segmentation mask for the input image.

SAM is trained using a supervised learning approach. The training data consists of images and text descriptions, where each text description indicates the objects that are present in the image. SAM is trained to find a minimum cross-entropy loss between the segmentation mask and the actual segmentation mask.

3.2.3. Mask post-processing

When points are used as input prompts, the resulting masks often contain many duplicate masks and noise blobs from the background. To tackle this issue, we implemented a mask post-processing algorithm (Algorithm 1) in Pseudocode (Nguyen et al., 2023). The algorithm works based on two parameters: (i) mask area and (ii) overlap ratio. The mask area threshold aims to eliminate excessively large or small masks that can be considered noise. On the other hand, the overlap ratio is used to merge masks that are highly similar or substantially overlap into a single mask.

Algorithm 1 Mask post-processing

$selected\_masks \leftarrow []$
for each $mask$ **in** sam_output_masks **do**
    $mask, mask\_area \leftarrow$ find_largest_contour($mask$)
    if $min\_area \leq mask\_area \leq max\_area$ **then**
        $selected\_masks \leftarrow selected\_masks \cup mask$
    end if
end for
$final\_results \leftarrow []$
while $selected\_masks \neq \emptyset$ **do**
    $pivot\_mask \leftarrow selected\_masks.pop()$ ▷ Assign the last mask from the list to pivot_mask
    for each $mask$ **in** $selected\_masks$ **do**
        $iou, overlap\_ratio \leftarrow calc\_mask\_overlap(pivot\_mask, mask)$
        if $(iou > iou\_threshold)$ **or** $(overlap\_ratio > overlap\_threshold)$ **then**
            $pivot\_mask \leftarrow pivot\_mask \cup mask$
        end if
    end for
    $final\_results \leftarrow final\_results \cup pivot\_mask$
end while

Algorithm 1 refines predicted object masks in several steps. Initially, it filters out very small or large ones based on their area (referred to as *mask_area*). It then iteratively merges overlapping masks (*selected_mask*). After that, it selects the last mask from the list and stores it in a variable called *pivot_mask* using the "pop;; operation (which removes the last element from a list). The algorithm keeps track of the *pivot_mask* and compares it to other masks. If the overlap between the pivot and another mask exceeds a threshold (either based on IoU or a custom overlap ratio), they are merged together. This process continues until all masks have been processed. The outcome is a refined list of masks with reduced noise and merged overlapping detections.

4. Experimental results

This section describes a series of experiments conducted on the seafloor litter dataset to comprehensively access the performance of the zero-shot segmentation pipeline under different testing conditions. Section 4.1 details the evaluation metrics used to evaluate the model's performance on various dimensions, whereas Section 4.2 reports the hardware and programming environment used to implement the model.

4.1. Evaluation metrics

Semantic segmentation models are evaluated using a confusion matrix, which has four components: true positive (TP), true negative (TN), false negative (FN), and false positive (FP). The terms TP, TN, FP, and FN refer to the number of pixels that are correctly or incorrectly classified, respectively. TP is the number of pixels that are correctly predicted to belong to the class of interest, TN indicates pixels correctly classified as background. FP is the number of pixels that are incorrectly predicted to belong to a certain class, and FN is the number of pixels that are incorrectly classified as background. TP, FN, and FP are used to calculate intersection over union (IoU), a popular metric for assessing model performance. IoU measures how much the predicted segmentation mask overlaps with the ground truth segmentation mask. Mean IoU (mIoU) is the average IoU over all classes.

$$\begin{aligned} IoU &= \frac{TP}{TP + FP + FN} \\ \text{mIoU} &= \frac{\text{IoU}}{\text{N}} \end{aligned} \tag{5}$$

where N is the total number of classes in the dataset, which is 8 in this study.

4.2. Implementation descriptions

The zero-shot marine litter segmentation framework was developed using PyTorch[3], a popular Python machine learning library, on a Linux system with two Nvidia Tesla V100 GPUs, each with 32 GB of memory. All DL models and hyperparameters, except for the zero-shot segmentation model,

[3]https://pytorch.org/

were implemented using open-source code from the original papers. To ensure reliable comparisons with the zero-shot approach, all supervised segmentation models used a pre-trained ViTs model on ImageNet as their backbone architecture.

4.3. Performance assessment of zero-shot marine litter segmentation framework

4.3.1. Pre-processing module analysis

Figure 3 shows eight input images from the dataset, which contain various challenges such as low light, blurriness, and poor illumination. The corresponding outputs from the pre-processing process show a significant improvement in image quality after passing through the P-UIE model. For example, marine litter in raw seafloor images with low contrast or poor illumination conditions can be challenging to see, but the P-UIE model significantly enhances image quality, making marine litter more visible. In addition, the pre-processing process does not add noise to input images or degrade the quality without any of the mentioned issues.

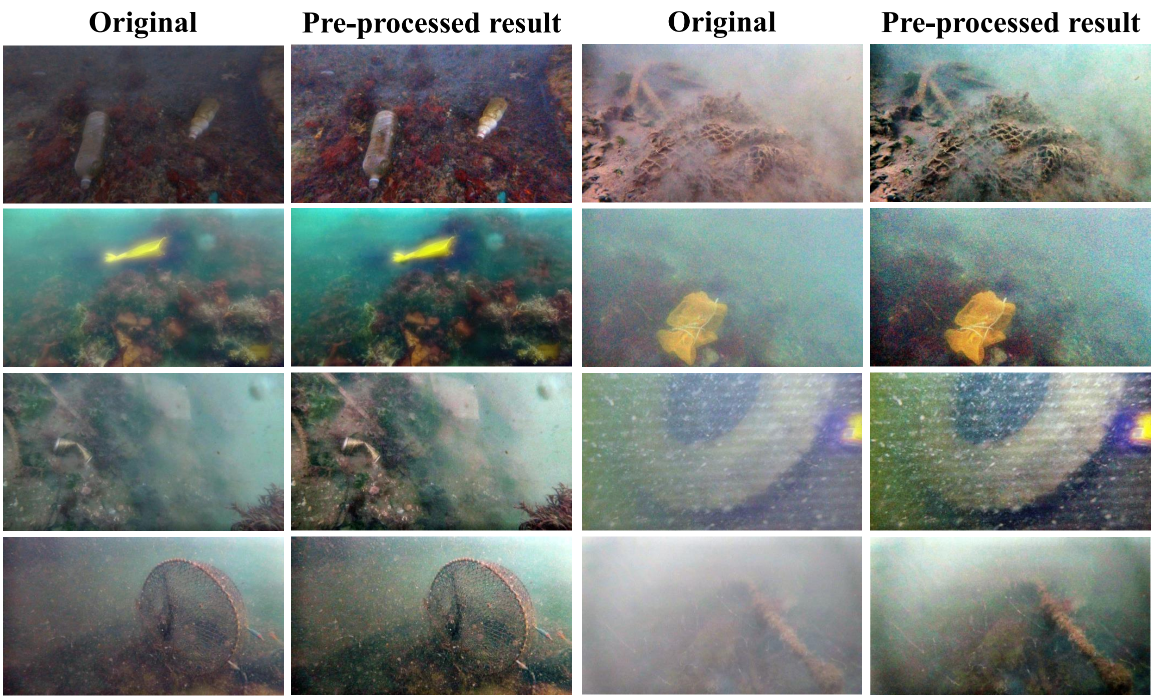


Figure 3: Comparison of the raw and pre-processed seafloor images.

As displayed in Table 1, P-UIE improved the marine litter segmentation performance of the zero-shot approach on three segmentation metrics, including mIoU, precision, and recall. Specifically, the mIoU score increased from 66.2% to 69.9%, the precision score increased from 65.9% to 69.6%, and the recall score increased from 66.7% to 69.8%. This suggests that the P-UIE model was able to effectively improve the quality of the input images, making it easier for the segmentation model to accurately identify and segment the marine litter. In addition, the P-UIE model was able to remove noise and artifacts from the input images, which can make it easier for the segmentation model to distinguish between the marine litter and the background. Finally, the P-UIE model also improved the contrast and color of the input images, which can also make it easier for the segmentation model to identify the marine litter.

Input data	mIoU (%)	Precision (%)	Recall (%)
Raw	66.2	65.9	66.7
Pre-processed	69.9	69.6	69.8

Table 1: The improvement of the pre-processing process on the models' performance.

4.3.2. Zero-shot segmentation performance analysis

Table 2 shows the performance of the proposed zero-shot marine litter segmentation approach for each of the eight marine litter types in the dataset. The table reports the IoU, precision, and recall scores.

	Fishing net	Fish trap	Glass	Metal	Plastic	Wood	Rope	Rubber	Average
IoU	61.5	63.7	75.1	77.2	74.8	70.2	66.4	70.7	70
Precision	60.9	63.3	76.5	75.8	72.2	69.1	68.2	71.4	69.7
Recall	58.9	63.1	77.3	75.7	74.4	71.5	68.7	69.3	69.9

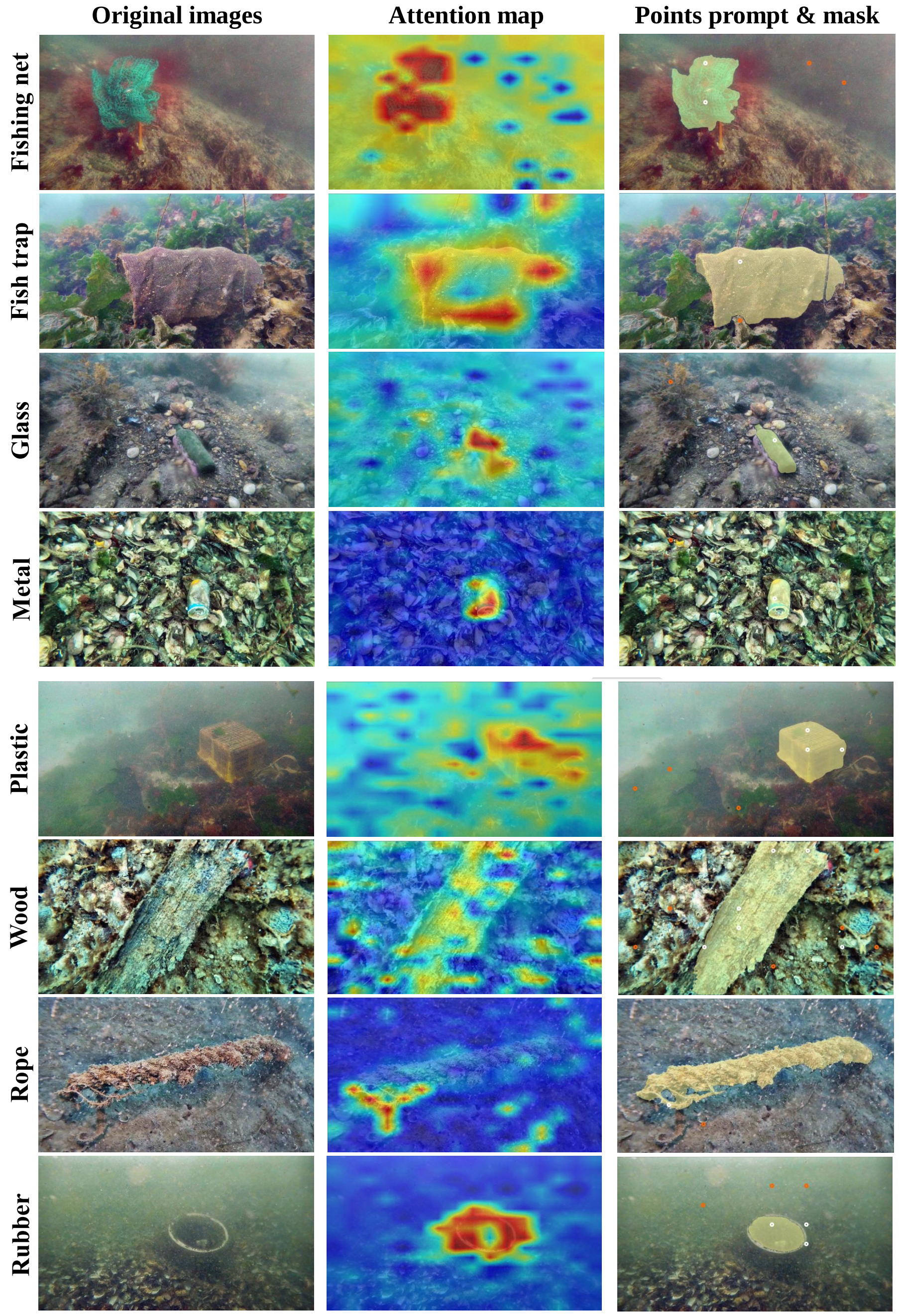
Table 2: Performance of the proposed approach for each marine litter type (IoU, precision, and recall).

The proposed zero-shot marine litter segmentation approach achieves good performance on all eight marine litter types, with average IoU scores of 70% and average precision and recall scores above 69%. This performance is particularly noteworthy considering that the dataset used in this study was collected on real-life seafloor conditions, which are often challenging for marine litter segmentation algorithms. The highest IoU scores are achieved for metal (77.2%), glass (75.1%), and plastic (74.8%), while the lowest IoU scores are achieved for fishing nets (61.5%) and fish trap (63.7%). This suggests that the proposed approach is a promising approach for zero-shot marine litter segmentation.

One possible explanation for the relatively low segmentation performance of fishing nets, fish traps, and ropes is that they can be difficult to distinguish from seaweed and other debris in the environment, especially in low-light or obscured conditions. Additionally, fishing nets and fish traps, which are often made of nylon, can have a similar appearance, further complicating accurate segmentation.

Figures 4 shows the model-predicted masks for the eight marine litter types. The first column shows the original image, the second column shows the interpretable CLIP attention masks, which highlight potential litter areas, and the third column shows the overlay of the predicted litter masks on the original image. Overall, the iCLIP attention masks are able to accurately highlight potential litter areas in the image, even in the presence of noise and occlusion. For example, in the case of fishing nets and traps, the attention masks accurately highlight the nets, even when partially obscured by seaweed. Similarly, the attention masks accurately highlight metal, wood, and rubber objects, even when they are similar in color to the surrounding environment.

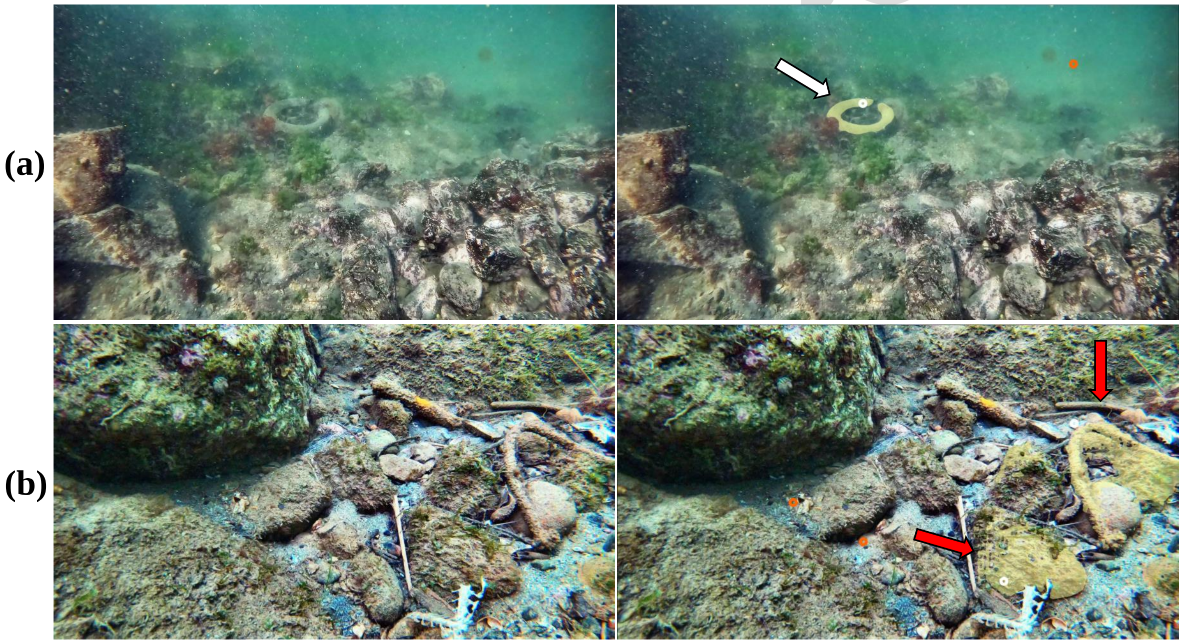
Based on the attention masks, the SAM model is guided to accurately segment the litter from the background, closely following the litter boundaries. The SAM model demonstrates potential for accurate object segmentation. While the SAM model generally performs well, it can occasionally exhibit shortcomings, such as incomplete object segmentation or slight over-segmentation.



Note: For each image in the third column, white dots indicating the potential litter areas extracted by the iCLIP model, with red dots indicating the background to guide the SAM model.

Figure 4: Visualization of the proposed zero-shot marine litter segmentation approach on four different types of marine litter.

Figure 5 shows the zero-shot segmentation results for two challenging cases where the marine litter resembles the surroundings. In Figure 5(a), the model correctly segmented the tire, even though it was small, far from the camera, and had the same color as the surrounding seafloor. However, in Figure 5(b), the model faced a more challenging scenario: the rope resembled the surrounding seafloor rocks, which is confusing. As a result, the model correctly segmented the rope, but it also falsely recognized some of the rocks near the rope as rope.



Note: The white arrow indicates correct segmentation, while the red arrow indicates wrong segmentation.

Figure 5: Two examples of marine litter that are challenging to identify due to the complex surrounding environment.

4.3.3. Mask post-processing analysis

Table 3 displays the mIoU on the testing set obtained from a grid search over various combinations of mask area (MA) within the range of [5%, 10%, 20%, 30%, 40%] up to 50% of the image area. Smaller MA values (e.g., 5% or 10%) allow the algorithm to focus on removing tiny masks, which can be noise or artifacts, whereas larger MA values (e.g., 30% or 40%) aim to eliminate excessively large masks that might cover significant portions of the image. Additionally, we explored overlap thresholds (OT) ranging from [0.6, 0.7, 0.8, 0.9, 1] for merging duplicated masks. A lower OT (e.g., 0.6 or 0.7) results in more conservative merging, preserving distinct marine litters. A higher OT (e.g., 0.8 or 0.9) merges masks more aggressively, potentially combining overlapping marine litters. Our objective was to identify the parameter combination that maximizes mIoU.

	$MA = 5$	$MA = 10$	$MA = 20$	$MA = 30$	$MA = 40$
$OT = 0.6$	55.5	52.3	54.4	43.9	39.7
$OT = 0.7$	65.1	54.8	51.7	48.1	42.4
$OT = 0.8$	**69.9**	59.0	58.8	49.6	47.3
$OT = 0.9$	67.3	58.2	58.3	48.8	45.3
$OT = 1$	65.2	55.1	57.5	45.7	44.2

Table 3: Identifying optimal parameters for post-processing process via grid search (MA and OT ranges).The highlighted value shows the best mIoU of the framework on the testing set.

The highlighted value represents the best mIoU achieved by the framework on the testing set. The combination of MA=5% and OT=0.8 yields the highest mIoU of 69.9, indicating effective post-processing for marine litter segmentation. This setting effectively removes small noise masks while maintaining reasonable merging thresholds. This optimal parameter combination represents the best choice for the marine litter dataset.

4.3.4. Comparison study for zero-shot segmentation

Table 4 shows the performance of the proposed zero-shot segmentation approach on four metrics: mIoU, precision, recall, and frames per second (FPS), compared to three supervised approaches, including Deeplabv3 (Chen et al., 2017), Mask-RCNN (He et al., 2017), and Mask2Former (Cheng et al., 2022), on the annotated testing dataset. Higher mIoU, precision, and recall indicate more accurate and complete detection, while higher FPS indicates faster processing speed.

Model	mIoU	Precision	Recall	FPS
DeepLabv3 (Chen et al., 2017)	74.2	75.6	73.7	18
Mask-RCNN (He et al., 2017)	**76.1**	**75.8**	**77.5**	12
Mask2Former (Cheng et al., 2022)	73.8	71.2	72.4	**22**
Ours (iCLIP+SAM)	69.9	69.6	69.8	16

Table 4: Performance of the zero-shot approach compared to supervised approaches on the annotated testing dataset.

Among evaluated models, Mask-RCNN achieves the highest mIoU (76.1%), precision (75.8%), and recall (77.5%). However, it has the slowest inference speed of 12 FPS, making it suitable for offline litter segmentation where time is not a crucial factor. While DeepLabv3, Mask2Former, and the proposed zero-shot approach offer faster inference speeds, their mIoU scores are lower (74.2%, 73.8%, and 69.9%, respectively).

The key distinction lies in data requirements. Supervised models like Mask-RCNN, Mask2Former, and DeepLabv3 demand a large-scale labeled dataset for training, which limits their applicability when such data is scarce or unavailable. In contrast, the proposed zero-shot learning approach based on SAM and iCLIP baselines offers a distinct advantage in scenarios where annotated training data is limited or unavailable. In summary, the choice between models depends on the trade-offs between performance and efficiency. Supervised models excel in performance when labeled data is readily available, while the proposed zero-shot approach provides an effective alternative in situations where labeled data is unavailable.

5. Discussion

Previous studies have shown that the marine environment can have a big impact on the performance of litter detection models. However, they did not offer a solution to this problem. We introduced a pre-processing module based on the P-UIE model to improve the performance of the marine litter segmentation framework. This module is crucial for handling the complexities inherent in marine datasets, leading to a remarkable 3.7% increase in the segmentation model's mIoU, compared to the baseline performance of 66.2% on the raw dataset. While seemingly modest, this improvement translates to a substantial reduction in missed or misclassified marine litter objects within a large-scale dataset. This translates to more precise marine pollution assessments, directly impacting cleanup efforts, ecosystem health monitoring, and research on pollution sources and impacts. While the pre-processing module needs more computing power and time, it can be easily turned on or off depending on the specific needs of the application. Overall, this pre-processing model makes it possible to segment marine litter more effectively in challenging seafloor environments.

In this study, we aimed to verify the effectiveness of the zero-shot approach for marine litter segmentation. We compared the zero-shot approach with three different DL-based segmentation models on a manually annotated marine litter test set. Our key finding was that the zero-shot approach achieved slightly lower performance (mIoU 69.9%) than other supervised models. The zero-shot approach achieved an inference speed of around 16 FPS, which is affected by the processing speed of the two different models, iCLIP and SAM, as well as the mask post-processing process.

Our results are similar to those of previous research on zero-shot approaches. The zero-shot marine litter segmentation pipeline based on iCLIP and SAM has several advantages over traditional supervised learning approaches, such as Mask-RCNN and DeepLabv3. First, it can be implemented without a large dataset of labeled images, which can be expensive and time-consuming to collect. Second, it is more robust to changes in the appearance of marine litter, such as variations in size, shape, and color. Third, it generalizes well to new environments, such as different water depths and different types of underwater terrain. The proposed zero-shot segmentation algorithm demonstrates promising results in automatically detecting and segmenting marine litter objects in underwater images. The proposed zero-shot marine litter detection framework represents one specific approach within the broader effort to combat marine litter. Our method offers valuable contributions in the field of large-scale monitoring of coastal areas, where the model is ready for use for various types of marine litter without the time-consuming process of labeling data and training the model.

6. Conclusions and future works

Marine litter on the seafloor poses a significant threat, but monitoring it traditionally requires extensive human labor. This research presents a simple yet remarkably efficient framework for automated seafloor litter monitoring. We leverage recent advancements in DL models, particularly in image registration, segmentation, and classification. These advancements have enabled pre-trained models to perform remarkably well in zero-shot learning scenarios. By harnessing the capabilities of these models, we have created a marine litter detection system that stands out for its ability to operate effectively without relying on labeled data or model training.

One of the notable innovations proposed in this framework is the utilization of a zero-shot segmentation pipeline that includes iCLIP. This technique generates potential points indicating the position of marine litter and the background. These points are then fed into the point prompt encoding of SAM to guide the automated segmentation process. Additionally, we implemented P-UIE, a DL model designed to enhance the visual quality of images captured in underwater environments. Given the challenging underwater conditions, which can sometimes deceive the framework into generating incorrect object masks, we also introduced a mask post-processing algorithm. This algorithm eliminates erroneous masks based on a carefully fine-tuned IoU threshold and overlap ratio.

The framework has been robustly tested and successfully detects eight types of marine litter, even in challenging seafloor environments where distinguishing litter from the background is difficult. It achieved an impressive mIoU of 69.9% and an inference speed of 16 frames per second (FPS). The P-UIE model further improves the mIoU of the pre-processed input from 66.2% to 69.9%. In addition, the extracted attention map serves as a visualization of the model's attention weights. These weights indicate the importance of each pixel in the input image for the model's prediction. This information can be used to understand the model's decision-making process and identify the key features it relies on for making accurate predictions.

One notable limitation of the framework is its computational complexity, which hinders real-time litter segmentation. Therefore, optimizing the zero-shot marine litter segmentation framework for both robustness and time efficiency is a crucial area for future research. Additionally, compared to fully supervised segmentation models like Mask-RCNN and DeeplabV3, the proposed model exhibited lower accuracy for recognizing underwater marine litter with fine-grained details and subtle variations. One possible direction is to combine the zero-shot framework with limited amounts of marine litter detection-specific labeled data (few-shot learning) or incorporating active learning strategies to improve accuracy and reduce reliance on large pre-trained models. In addition to technological advancements, promoting complementary strategies is essential for tackling marine litter effectively. These strategies can include improved waste management infrastructure, educational initiatives promoting responsible waste disposal, and policy changes encouraging sustainable practices.

CRediT authorship contribution statement

Tri-Hai Nguyen: Writing – original draft, Investigation, Formal analysis, Visualization. **L. Minh Dang**: Data curation, Validation, Writing – review & editing.

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 https://www.aihub.or.kr/

References

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

, . Hero5 black. `https://gopro.com/en/us/update/hero5-black`. Accessed: September 27, 2023.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 .

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290–1299.

Chin, C.S., Neo, A.B.H., See, S., 2022. Visual marine debris detection using yolov5s for autonomous underwater vehicle, in: 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS), IEEE. pp. 20–24.

Corrigan, B.C., Tay, Z.Y., Konovessis, D., 2023. Real-time instance segmentation for detection of underwater litter as a plastic source. Journal of Marine Science and Engineering 11, 1532.

Deng, H., Ergu, D., Liu, F., Ma, B., Cai, Y., 2021. An embeddable algorithm for automatic garbage detection based on complex marine environment. Sensors 21, 6391.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

Fu, Z., Wang, W., Huang, Y., Ding, X., Ma, K.K., 2022. Uncertainty inspired underwater image enhancement, in: European Conference on Computer Vision, Springer. pp. 465–482.

Galgani, L., Beiras, R., Galgani, F., Panti, C., Borja, A., 2019. Impacts of marine litter.

Gong, L., Du, X., Zhu, K., Lin, C., Lin, K., Wang, T., Lou, Q., Yuan, Z., Huang, G., Liu, C., 2021. Pixel level segmentation of early-stage in-bag rice root for its architecture analysis. Computers and Electronics in Agriculture 186, 106197.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

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. Science of the Total Environment 878, 162826.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE international conference on computer vision, pp. 1501–1510.

Iñiguez, M.E., Conesa, J.A., Fullana, A., 2016. Marine debris occurrence and treatment: A review. Renewable and Sustainable Energy Reviews 64, 394–402.

Jang, Y.C., Lee, G., Kwon, Y., Lim, J.h., Jeong, J.h., 2020. Recycling and management practices of plastic packaging waste towards a circular economy in south korea. Resources, Conservation and Recycling 158, 104798.

Jia, T., Kapelan, Z., de Vries, R., Vriend, P., Peereboom, E.C., Okkerman, I., Taormina, R., 2023. Deep learning for detecting macroplastic litter in water bodies: A review. Water Research , 119632.

Jian, M., Liu, X., Luo, H., Lu, X., Yu, H., Dong, J., 2021. Underwater image processing and analysis: A review. Signal Processing: Image Communication 91, 116088.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

Kraft, M., Piechocki, M., Ptak, B., Walas, K., 2021. Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle. Remote Sensing 13, 965.

Li, Y., Wang, H., Duan, Y., Xu, H., Li, X., 2022. Exploring visual interpretability for contrastive language-image pre-training. arXiv preprint arXiv:2209.07046 .

Ma, D., Wei, J., Li, Y., Zhao, F., Chen, X., Hu, Y., Yu, S., He, T., Jin, R., Li, Z., et al., 2023. Mldet: Towards efficient and accurate deep learning method for marine litter detection. Ocean & Coastal Management 243, 106765.

Madricardo, F., Ghezzo, M., Nesto, N., Mc Kiver, W.J., Faussone, G.C., Fiorin, R., Riccato, F., Mackelworth, P.C., Basta, J., De Pascalis, F., et al., 2020. How to deal with seafloor marine litter: an overview of the state-of-the-art and future perspectives. Frontiers in Marine Science 7, 505134.

Mæland, C.E., Staupe-Delgado, R., 2020. Can the global problem of marine litter be considered a crisis? Risk, Hazards & Crisis in Public Policy 11, 87–104.

Marin, I., Mladenović, S., Gotovac, S., Zaharija, G., 2021. Deep-feature-based approach to marine debris classification. Applied Sciences 11, 5644.

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review , 1–66.

Nguyen, L.Q., Shin, J., Ryu, S., Dang, L.M., Park, H.Y., Lee, O.N., Moon, H., 2023. Innovative cucumber phenotyping: A smartphone-based and data-labeling-free model. Electronics 12, 4775.

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. Marine Pollution Bulletin 164, 111974.

Radeta, M., Zuniga, A., Motlagh, N.H., Liyanage, M., Freitas, R., Youssef, M., Tarkoma, S., Flores, H., Nurmi, P., 2022. Deep learning and the oceans. Computer 55, 39–50.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR. pp. 8748–8763.

Raveendran, S., Patil, M.D., Birajdar, G.K., 2021. Underwater image enhancement: a comprehensive review, recent trends, challenges and applications. Artificial Intelligence Review 54, 5413–5467.

Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T., 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19305–19314.

Sandra, M., Devriese, L.I., Booth, A.M., De Witte, B., Everaert, G., Gago, J., Galgani, F., Langedock, K., Lusher, A., Maes, T., et al., 2023. A systematic review of state-of-the-art technologies for monitoring plastic seafloor litter. Journal of Ocean Engineering and Science .

Schneider, F., Parsons, S., Clift, S., Stolte, A., McManus, M.C., 2018. Collected marine litter—a growing waste challenge. Marine pollution bulletin 128, 162–174.

Sohn, K., Lee, H., Yan, X., 2015. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems 28.

Sun, X., Gu, J., Sun, H., 2021. Research progress of zero-shot learning. Applied Intelligence 51, 3600–3614.

Teng, C., Kylili, K., Hadjistassou, C., 2022. Deploying deep learning to estimate the abundance of marine debris from video footage. Marine Pollution Bulletin 183, 114049.

Wei, Y., Cao, Y., Zhang, Z., Peng, H., Yao, Z., Xie, Z., Hu, H., Guo, B., 2023. iclip: Bridging image classification and contrastive language-image pre-training for visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2776–2786.

Zhou, W., Zheng, F., Yin, G., Pang, Y., Yi, J., 2022. Yolotrashcan: A deep learning marine debris detection network. IEEE Transactions on Instrumentation and Measurement 72, 1–12.

Highlights

- We introduce a zero-shot marine litter segmentation framework
- An underwater image enhancement algorithm was applied to improve the dataset quality
- The framework achieves a test mIOU of 69.9% for eight common marine litter
- We perform detailed analysis of the model's robustness against complex background noise
- We demonstrate the potential of zero-shot approach for automated marine litter monitoring

Declaration of interests

☒ 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.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: