Marine Debris Detection in Satellite Surveillance Using Attention Mechanisms

Ao Shen, Yijie Zhu, Plamen Angelov, and Richard Jiang

***Abstract*—Marine debris poses a critical threat to environmental ecosystems, necessitating effective methods for its detection and localization. This study addresses the existing limitations in the literature by proposing an innovative approach that combines the instance segmentation capabilities of YOLOv7 with various attention mechanisms to enhance efficiency and broaden applicability. The primary contribution lies in the exploration and comparison of three attentional models: lightweight coordinate attention, combining spatial and channel focus (CBAM), and bottleneck transformer based on self-attention. Leveraging a meticulously labeled dataset of satellite images containing ocean debris, the study conducts a comprehensive assessment of box detection and mask evaluation. The results demonstrate that CBAM emerges as the standout performer, achieving the highest F1 score (77%) in box detection, surpassing coordinate attention (71%) and YOLOv7/bottleneck transformer (both around 66%). In mask evaluation, CBAM continues to lead with an F1 score of 73%, while coordinate attention and YOLOv7 exhibit comparable performances (around F1 scores of 68% and 69%), and bottleneck transformer lags behind at an F1 score of 56%. This compelling evidence underscores CBAM's superior suitability for detecting marine debris compared to existing methods. Notably, the study reveals an intriguing aspect of the bottleneck transformer, which, despite lower overall performance, successfully detected areas overlooked by manual annotation. Moreover, it demonstrated enhanced mask precision for larger debris pieces, hinting at potentially superior practical performance in certain scenarios. This nuanced finding underscores the importance of considering specific application requirements when selecting a detection model, as the bottleneck transformer may offer unique advantages in certain contexts.**

***Index Terms*—Deep learning, image segmentation, marine debris, marine pollution, remote sensing, satellite imagery, YOLO.**

I. Introduction

MARINE debris is any persistent solid material that has been abandoned directly or indirectly, intentionally or unintentionally, into the marine environment or the Great Lakes. Anything man-made and solid that is lost in these aquatic environments becomes marine debris [1]. According to [2], the main source of marine debris is plastic [2]. The slow degradation and light weight of plastic make it easier to stay and accumulate in nature than other garbage, and even form small "islands" in the ocean that can be easily seen and intercepted. But more often, it is a garbage patch—an area of the ocean made almost entirely of tiny plastic that's not always visible to the naked eye. The most obvious characteristic of this type of garbage belt is the change in the color of the water in the area, such as the famous Great Pacific Garbage Patch (GPGP). Within this 1.6 million km^2 area, there are 42 000 metric tons of giant plastics, 20 000 metric tons of large plastics, 10 000 metric tons of medium plastics, and 6400 metric tons of microplastics [3].

In addition to the pollution to the appearance of water bodies mentioned above, the most serious aspect of marine debris is the harm to the marine ecosphere. In the case of sea birds, the livers of black-footed albatrosses from Midway Atoll, near GPGP, were found to have high levels of perfluoroalkyl acid, much higher than those found in waterbirds of the continental United States and Arctic environments. The concentration of PFOS ranged from 22.91 to 70.48 ng/g wet weight, and PFUdA ranged from 8.04 to 18.70 ng/g wet weight, but the concentration of the latter was much higher than that of the former in contaminated albatross livers [4]. And that is because of the plastic waste in GPGP.

In addition, marine debris has an impact on our daily lives as it enters the food chain. Marine debris has been discovered in beer, salt, drinking water, vegetable-growing soil, and salt. Developmental, neurological, reproductive, and immunological issues can result from plastic materials' endocrine system disruption and carcinogenicity [5]. Toxic contaminants that frequently build up on plastic surfaces and are then ingested by people through seafood intake are another health risk.

In the past decade, the mainstream detection method for marine debris is still traditional manual methods. However, conducting on-site investigations requires extensive labor, equipment, and financial expenditures, thereby limiting the scope. With the improvement of machine learning, especially image segmentation technology [6], the plastic fragments in the visual data captured by the camera system are gradually accepted by researchers. These visual data are mainly collected from airplanes and drones. However, the data acquisition range and efficiency of aerial photography are limited, especially for the lack of detection capabilities for distant sea targets. Therefore, detection using satellite imagery is an effective option. So far, there have been numerous experiments in the field of satellite-based marine debris detection [27]. The technology of deep

Manuscript received 3 July 2023; revised 6 October 2023 and 19 November 2023; accepted 18 December 2023. Date of publication 3 January 2024; date of current version 12 February 2024. This work was supported in part by the European Space Agency under Grant 4000133854/21/NL/CBi, in part by the ESA's Φ–lab, and the Engineering, Physical Sciences Research Council (EPSRC) under Grant EP/P009727/2, and in part by the European Lighthouse on Secure and Safe AI by the European Union under Grant 101070617. *(Corresponding author: Richard Jiang.)*

The authors are with the LIRA Center, Lancaster University, LA1 4YW Lancaster, U.K. (e-mail: aoshen42@gmail.com; y.zhu43@lancaster.ac.uk; p.angelov@lancaster.ac.uk; r.jiang2@lancaster.ac.uk).

Digital Object Identifier 10.1109/JSTARS.2024.3349489

learning has been attempted to be applied to garbage detection [28] and classification [29].

Given certain limitations, such as the resolution of satellite imagery, most projects focus on the detection and classification of medium- to large-sized debris drifting in offshore areas or reaching the coast [30]. In this field, deep learning is used instead of traditional manual annotation for reasons including higher efficiency and lower costs, which is particularly important for small- to medium-sized environmental organizations. Therefore, one potential option would be adopting the use of YOLO architecture, which could provide greater speed, lower cost, and the ability for real-time processing capabilities, opening up new opportunities in remote sensing technologies.

Our research aims to address the challenge of detecting marine debris through the combination of state-of-the-art technology and innovative methods. We plan to leverage the power of the neural network model based on YOLOv7 instance segmentation to achieve accurate identification of marine debris in satellite imagery. By combining satellite technology and machine learning, the detection efficiency and cost of marine debris can be greatly improved. To further enhance the performance of our system, we intend to explore various attention mechanisms and evaluate their suitability for improving the results obtained from our baseline approach. By doing so, we hope to mitigate the shortcomings associated with other existing methods [31] and ultimately yield improved detection rates and spatial precision. Ultimately, our goal is to develop an effective solution for monitoring and managing the growing problem of marine pollution. The code is available at the link below.[1]

II. Related Work

A. Current Methods for Marine Debris Detection

Due to the increasing importance of environmental protection, there has been a lot of research on the source and location of marine debris, or marine plastic. As one might expect, a number of traditional, manual methods have been used to count this marine debris.

National oceanic and atmospheric administration (NOAA) has a marine debris monitoring and assessment program, or MDMAP. The program works by recruiting partners and volunteers around the world to conduct regular surveys of debris along coastlines. Monitoring kits will be made available for partners and volunteers to use uniform standards and methods to collect and record the amount and type of debris waste. In this way, the quality and comparability of data can be guaranteed. By analyzing the survey data, they were able to understand the scale and trends of the problem, as well as the most common types of debris waste in different regions. Since its inception in 2012, the work has conducted 4421 surveys at 335 monitoring sites in nine countries [8]. They found that, globally, the most common pieces of debris on coastlines were cigarette butts, food packaging, drink bottle caps, and plastic bags. The amount and type of debris on the shoreline are related to population density, wind direction, and tide. In addition, they analyzed the data and found that debris waste from Asia reaches Hawaii's coastline on ocean currents. In the Caribbean, debris waste on the coastline is mainly generated by residents and tourists, rather than by ocean currents [9].

Fig. 1. Example of an instance segmentation model applied to drone imagery. This shows that the AI was able to identify the litter scattered on the ocean.

The NOAA's study highlights the benefits of in situ monitoring in measuring and analyzing the quantity and quality of plastic waste present in oceans and seas. On-site measurements allow for precise evaluation of the geographical and chronological distribution of litter, thereby shedding light on the behavior patterns and effects of ocean currents. In addition, through on-site observations, it becomes feasible to investigate the harmful consequences of plastic waste on aquatic species and habitats, facilitating risk assessments. Nonetheless, as mentioned previously, conducting fieldwork entails substantial expenditure of labor force, equipment, and funds, thus limiting coverage. Furthermore, even when tools are accessible, variations in proficiency and motivation among participants may result in unequal protocols and data inconsistencies. Therefore, alternative solutions must continue to be explored and developed.

With recent advancements in technology, a number of machine-learning approaches that rely on convolutional neural networks (CNNs) have emerged for identifying debris in visual data captured using camera systems, including those mounted on drones, satellites, and vessels, see Fig. 1. These methods share the characteristic of processing images or streaming media content to locate debris within the recorded footage. Although primarily focused on aerial and spaceborne imagery, along with shipboard observations, these studies have shown promising results in addressing the issue of marine debris pollution via computer vision techniques.

Typically, drones or aerial photography are used to observe Marine debris deposition along the coast. Ellipsis Earth uses drones equipped with cameras to map the location of plastic pollution. By applying image recognition based on deep learning, it can identify the type and the size of debris. In some cases, even the brand or source of the waste can be identified. The drone can take video at high speeds in locations where humans cannot go, and then fly back to read the data in its storage and feed it into an AI to identify debris waste in it. Some more advanced algorithms have been applied to support real-time recognition of images transmitted synchronously by the UAV. According to the research by Song et al. [10], conducted in 2022, the

[1][Online]. Available: https://github.com/Panperception/Student_2023_OceanDebris.

drone mapping method can be more accurate and universal than traditional methods. The segmentation model based on U-Net achieved good segmentation results in UAV image analysis, and the F1 scores of polypropylene foam and plastic, the most serious pollutants, were 0.97 and 0.93, respectively. However, there are limits to what the Ellipsis technique and UAV can detect. The drone approach is limited to coastal detection, not deep into the ocean. And it was unable to identify microplastics—plastic particles smaller than 5 mm, of which at least 14 million tons are estimated to be on the ocean floor [11].

Another way to target large debris waste is to use on-board cameras. This approach is an alternative to the traditional water collection detection method such as the neuston trawls. It uses an algorithm based on deep learning, which can detect plastic fragments from high-resolution optical image data collected from GoPro, and add a GPS tag to the image to quantify the debris floating on the surface of the ocean [12]. Since this method has a closer view than a drone, it can go deeper to provide more precise and detailed estimates of debris density in the ocean.

Obviously, the above methods are mostly suitable for a small range of detection, and have the characteristics of long time consuming. To detect large amounts of microplastics and track the location of debris at a macro level, satellites are required. A new method developed by researchers at the University of Michigan (UM) uses satellite data to map the concentrations of microplastics in the world's oceans [13]. Eight microsatellites that are a part of NASA's Cyclone Global Navigation Satellite System (CYGNSS) mission provided data to the researchers. In order to gauge the roughness of the ocean surface, the CYGNSS satellite collects signals reflected from global positioning system (GPS) satellites from the water. Waves are suppressed by plastic or other garbage in the ocean, causing a smaller roughness than anticipated. In a 2021 study, Jamali et al. [14] used multispectral satellites and deep learning to develop a large-scale marine debris waste detection framework. By combining Sentinel-2 satellite imagery with cutting-edge machine learning algorithms on the Sentinel Hub cloud API, they created a cloud-based framework for large-scale marine pollution monitoring (API). Although both satellite-level debris detection methods have achieved good results, the former is not universal and lacks detail, and the latter is still in the coastal part, which is not useful for deeper into the ocean.

Among current approaches for detecting marine debris, methods employing UAVs, vessel-mounted cameras, and satellite imagery possess varying strengths and weaknesses. While UAVs and ship-based cameras provide highly accurate and detailed information, their limited detection ranges and costs restrict their effectiveness in monitoring large regions. Satellite imaging, conversely, offers greater coverage but suffers from lower resolution and dearth of available data. Existing methods utilized for pinpointing plastic particles in satellite images mainly concentrate on coastal zones, leaving unexploited the unique capabilities of satellite imagery. As such, evaluating the performance of YOLO on high-resolution, surface-oriented satellite photographs represents a crucial step toward establishing a comprehensive system capable of detecting marine debris over extensive territories while capitalizing on the unique advantages of satellite imagery.

B. YOLO

YOLO is a CNN-based architecture similar to a fully CNN. CNN is an artificial neural network similar to deep neural networks. The inputs to the nodes in the neural network extract local features from each local corresponding field of the previous layer. After this feature extraction is completed, the positional relationships between local features and other features are mapped or plotted [15].

When the YOLO algorithm applies CNN, each convolutional network starts predicting multiple bounding boxes and bounding box class probabilities simultaneously. This means that instead of processing the image multiple times to detect different classes, YOLO simply passes the image through the neural network once to obtain a prediction or output. This optimizes the detection performance of a single algorithm run, thus reducing the latency of the process. This allows YOLO to identify and localize objects in near real-time, such as detecting objects in streaming videos.

As the first YOLO model with a new model head, YOLOv7 has both object detection and instance segmentation. The main advantage of YOLO is its speed. It can process images at 155 frames per second, which is much faster than other state-of-the-art algorithms [7]. Therefore, although its base implementation is Mask R-CNN, it is more efficient than traditional Mask R-CNN-based models and more suitable for real-time tasks.

In the recent COVID-19 pandemic, YOLO's efficiency makes it stand out and achieve excellent results in real-time mask detection tasks. In 2021, Liu and Ren [16] conducted a study to replace the traditional Faster R-CNN with YOLOv3 and YOLO achieves a higher F1 score and is nearly 50% faster than Faster R-CNN.

YOLO also has a small number of applications in marine debris detection, mainly its object detection function in the application of UAV aerial photography. A study by Watanabe et al. [17] used YOLOv3 as a deep learning object detection algorithm to detect debris floating on the beaches and the sea surface next to the beach. Through the detection of hand-taken debris waste photos, YOLOv3 obtains 77.2% of the mAP, which is obviously higher precision and faster than the 41.2% of Faster R-CNN. Although the research is focused on in-atmosphere detection methods with drones, it confirmed the feasibility of the YOLO algorithm for marine debris detection, and more importantly, it confirmed the excellent accuracy and identification speed of YOLO.

For the instance segmentation of YOLO, there are still relatively few practical applications, more are improved versions of YOLOv3 era, such as Poly-YOLO or Insta-YOLO [18], [19]. These YOLO instance segmentation versions are based on backward YOLO versions and lack evaluation for real-world application scenarios. This highlights the importance of testing the application for YOLOv7 instance segmentation.

III. Methods

A. Data Collection and Labeling

According to previous work, we need images that can show small fragments clearly. While Sentinel-2 can only provide

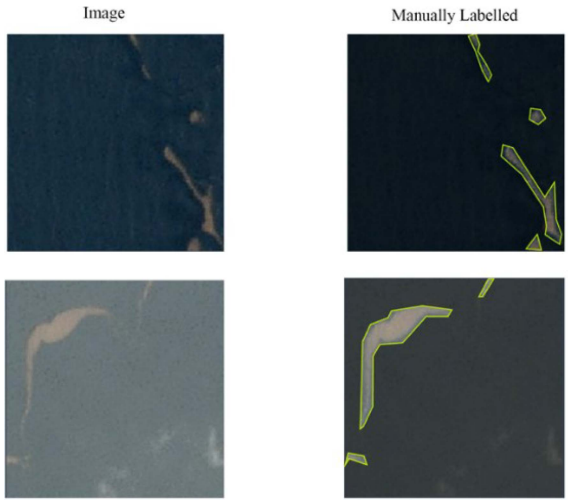

Fig. 2. Manually label process. On the left is the selected image, and on the right is the manually annotated image of the polygon suitable for instance segmentation.

resolution above 10 m per pixel, hindering the detection of smaller debris, we collect data based on NASA satellite images with a resolution of three meters [20]. The pixel size of these images is 256x256·, with four bands of red, green, blue, and near-infrared. Planet explorer was used to manually explore PlanetScope scenes and verify the presence of marine debris. Optical channels (RGB) are studied during this process. The images are mostly of the Gulf Islands in Honduras, where the trash is made up of plastic, wood, algae, and other man-made objects.

As shown in Fig. 2, although there are 707 images in this set, most of them have unclear garbage, lots of clouds, and coasts. Due to the focus of this project on ocean detection, images of the coastline and areas near the coast have been largely removed. Some debris that occupies only 2 to 9 pixels in the image has also been temporarily removed. A total of 321 images were manually selected and labeled in two rounds using Roboflow (see Fig. 2). Among them, 249 images were randomly selected to form the training set (79%) and 68 images were randomly selected to form the test set (21%) to form the dataset used in this work.

B. YOLOv7 Instance Segmentation

Instance segmentation works by using a CNN to detect and classify each object in an image. Methods for instance segmentation can be divided into two categories. One of them can be called a two-stage method, which uses a Region Proposal Network (RPN) to generate candidate boxes first, and then classify and segment each box. Single-stage methods, on the other hand, perform classification and segmentation directly on the whole image without generating candidate boxes.

YOLOv7 is a single-stage instance method, which uses a novel Overlapping double structure (Overlapping BiLayers), is able to handle overlapping problems, and improve accuracy and speed. The network of YOLOv7 mainly includes four parts: Input, Backbone, Neck, and Head. Firstly, the image was preprocessed by a series of operations such as data enhancement in the input part, and then sent to the backbone, and the backbone extracted the features of the processed image. Then, the extracted features were fused by the neck module to obtain features of large, medium and small sizes. Finally, the fused features are sent to the head, and the results are output after detection.

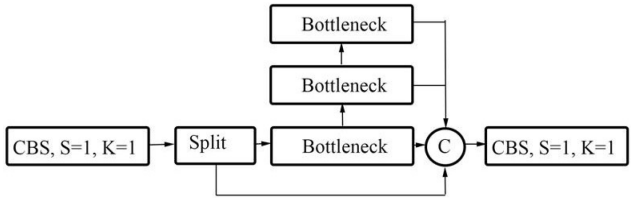

Fig. 3. Architecture diagram of C2f. Compared with the C3 module used in YOLOv7, it is more lightweight and has richer gradient information.

According to Wang et al. [21], the main backbone is mainly composed of convolution, E-ELAN module, MPConv module, and SPPCSPC module. ELAN is modified to E-ELAN in YOLOv7. Group convolutions are used by E-ELAN to increase the channels and cardinalities of the computation blocks. All computing blocks in the computation layer have the same channel multiplier and group parameters applied to them. Each computing block's feature maps are concatenated after being divided into groups of size g. To complete the combined cardinality, a shuffled group feature map will be added. The feature vectors are mapped to the number of classes here, which is two in this work (debris and no debris).

The Neck part of YOLOv7 consists of FPN feature pyramid and PANet. This part is the same as YOLOv5. The PANet part finally maps masks and boxes separately. Since this work is the detection of debris, only two classes exist, with debris and without debris, so only "0" and "1" will be output when mapping the feature layer of the box.

This work uses a YOLOv7 instance segmentation version based on OpenCV and PyTorch [22]. A data control group was created by training the dataset directly on that version.

C. YOLOv7 Instance Segmentation With C2f

Due to the release of YOLOv8, this work also attempts to incorporate a part of YOLOv8 C2f module. As shown in Fig. 3, the C2f module is a lightweight convolutional structure, which refers to the ideas of the C3 module and ELAN, and can obtain more abundant gradient flow information while ensuring lightweight. By replacing the specific convolution layer in YOLOv7, the C2f module is implanted without changing the overall structure of YOLOv7 to simulate a more lightweight scene.

As a variant of the basic YOLOv7, this version will be used as a base environment to test the effects of the attention mechanism. All models based on YOLOv7 with C2f added will be named "xx models with C2f" (where xx refers to an attention mechanism).

D. Attention on YOLOv7 Segmentation

Attention is a technique designed to mimic cognitive attention. It comes from the way humans perceive information about the environment. This effect enhances some parts of the input

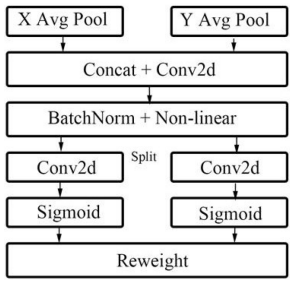

Fig. 4. Schematic expression of coordinate attention.

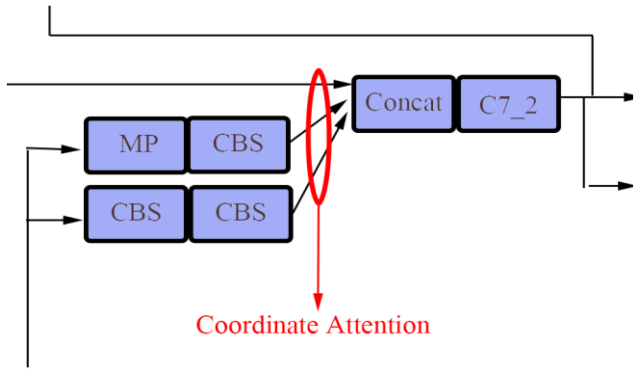

Fig. 5. Coordinate attention model is created by insert coordinate attention blocks into the head of instance segmentation.

data while reducing others—the motivation is that the network should pay more attention to smaller but important parts of the data.

E. Coordinate Attention Application

As a new mechanism based on channel attention, Hou et al. [23] proposed a new attention mechanism, called coordinate attention, by embedding location information. Contrary to channel attention, which splits up feature tensors into a single feature vector by two-dimensional (2-D) global pooling, coordinate attention splits up channel attention into two 1-D feature encoding procedures, each of which aggregates features along two different spatial directions [23]. This allows for the acquisition of remote dependencies along one spatial direction and the retention of accurate position data along the other. The direction-aware and location-sensitive attention maps that are produced from the resulting feature maps may then be complementarily applied to the input feature map to improve the representation of the object of interest. Fig. 4 shows the structure of a coordinate attention block.

For an input X, as an example of coordinate information embedding, two spatial extents of pooling kernels (H,1) or (1, W) are carried to encode each channel through both horizontal and vertical positions. The output z of channel c at height h should be written as follows:

$$z_c^h\ (h) = \frac{1}{w} \sum_{0 \le i < W} x_c\,(h, i)\,. \tag{1}$$

Thus, the output of channel at width w can be similarly formulated as follows:

$$z_c^w\ (w) = \frac{1}{H} \sum_{0 \le j < H} x_c\,(j, w)\,. \tag{2}$$

In this way, coordinate attention aggregates features along both spatial directions, yielding direction-aware feature maps in pair. Therefore, in this work, the coordinate attention mechanism will be loaded as an additional module and replace a certain convolutional layer after the MP module of the YOLOv7 head (see Fig. 5). This type of improved structure is called the coordinate attention model in the following chapter in this report, which is the first improvement. The coordinate attention model is also established on a version of YOLOv7 instance segmentation which includes the addition of C2f.

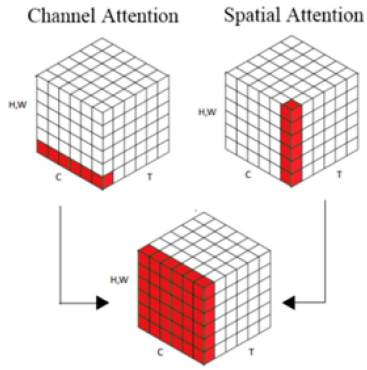

Fig. 6. What the combination of channel and spatial attention looks like. Where C represents the channel domain, H and W represent the spatial domain, and T represents the temporal domain.

F. Convolutional Block Attention Module Application

Convolutional block attention module (CBAM) stands for attention mechanism module of convolutional module, which is an attention mechanism module that combines spatial and channel (see Fig. 6).

Compared with the Senet's attention mechanism which only focuses on channels, it can achieve better results. Fig. 7 shows the general structure of CBAM. As observed, the convolutional layer's output result will first proceed to the channel attention module to obtain a weighted result, then pass through the spatial attention module to obtain the final result by weighting. [24].

The channel attention module compresses the feature map in the spatial dimension to obtain a 1-D vector before operation. Not only the average pooling but also the max pooling is taken into account when compressing in the spatial dimension. When conducting the gradient backpropagation computation, average pooling provides feedback to every pixel on the feature map, but maximum pooling only provides feedback to the gradient where the response is the biggest in the feature map. The mechanism can be expressed as follows:

$$M_c\ (F) = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(F\right)\right) + \mathrm{MLP}\left(\mathrm{MaxPool}\left(F\right)\right)\right) \tag{3}$$

where MLP is a multilayer perceptron, F is two different spatial context descriptors, and σ is a sigmoid function.

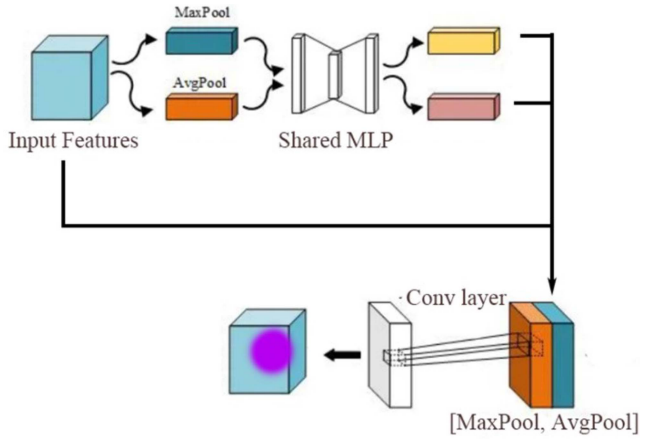

Fig. 7. Brief structure diagram of CBAM. The channel submodule uses the max-pooling output and the average pooling output of the shared network, and then obtains the weight of one channel submodule. The features weighted by the channel module will go to the spatial submodule, which takes two similar outputs, pooling them along the channel axis and forwarding them to the convolutional layer.

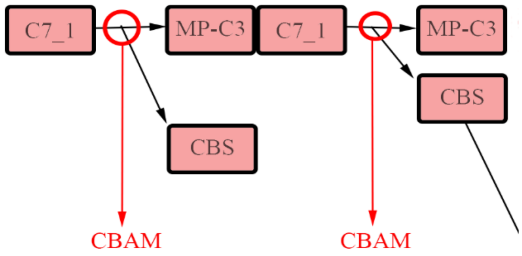

Fig. 8. CBAM model is created by replacing Conv layer in CBS modules with CBAM block. These two CBS modules are exits to branches in the backbone. A layer that is inserted before SPPCSPC module is tested, but has a poor result.

The feature map produced by the channel attention module serves as the input feature map for the spatial attention module. The spatial attention module compresses the channel and performs average pooling and max pooling in the channel dimension respectively. The max pooling operation extracts the maximum value from the channel by multiplying the height by the width. The average pooling operation simply extracts the average of the channels, again multiplying the height by the width. Then, all the feature maps (all with 1 channel) are combined to produce a 2-channel feature map. The mechanism can be summarized as follows:

$$M_s\ (F) = \sigma\left(f^{7\times7}\left([\text{AvgPool}\,(F)\,;\text{MaxPool}\,(F)]\right)\right) \quad (4)$$

where F is feature maps aggregated from channel information, σ is the sigmoid function, and f is a convolution operation that has a filter size of 7×7.

By combining two attention modules in order, they can be complemented. In this way, the whole attention block is able to focus on both location and content. Among them, channel attention needs to be arranged before spatial attention. Thus, a new layer combined with CBAM block and original convolutional layer will be added in this work, replacing two original convolutional layers in the YOLOv7 backbone (as shown in Fig. 8). A type of model will be formed in this way, which will be called CBAM model in the rest of this article. A version of the model based on C2f is also created for comparison.

G. Self-Attention Application

The self-attention mechanism or transformer is a variant of the attention mechanism that lessens reliance on outside input and excels at detecting internal correlations of features or data. The self-attention mechanism has been previously applied in text to solve the long-distance dependence problem by calculating the interaction between words. The core part of a self-attention mechanism in text processing is three vectors: query, key, and value. For a single word, query is dot multiplied by the key to get the score for the word, then a softmax calculation will provide the relevance of each word with respect to the current word. By adding the product of Value and softmax, value of self-attention at the current position can be obtained.

When used in computer vision, self-attention is often replaced with a convolutional layer [25]. It works similarly to text processing. For a pixel input x_{ij}, a local area in positions $ab \in N_k(i, j)$ is extracted, where k is spatial extent centered around x_{ij}. Then a singled-headed self-attention with pixel output y_{ij} can be roughly written as follows:

$$y_{ij} = \sum_{a,b\in N_k(i,j)} \text{softmax}_{ab}\left(q_{ij}k_{ab}\right) v_{ab} \quad (5)$$

where q_{ij}, k_{ab}, v_{ab} are queries, keys, and values, relatively. They provide linear transformations of pixels in position *ij* and nearby. softmax_{ab} illustrates a softmax based on computing all the logits in the neighborhood of *ij*.

Compared with the convolutional layer, which is a local operation and can only focus on a small region in the input feature map, the self-attention layer is a global operation and can focus on any position in the input feature map.

This work uses one of these transformers: bottleneck transformer, which, similar to YOLO itself, is efficient and low-overhead. This mechanism is derived from the BoTNet architecture. The BoTNet design is simple: replace the last three spatial (3×3) convolutions in the ResNet with a multihead self-attention (MHSA) layer [26]. The core part of this architecture, MHSA, performs all2all attention on a 2-D feature map (see Fig. 9). It performs with split relative position encoding R_h and R_w for height and width, respectively. The attention logits are qk^T and qr^T in the diagram, where q, k, r stands for query, key, and position encoding, respectively.

The self-attention used in the bottleneck network is a transformer block within the backbone. However, in order to facilitate the combination with other attention mechanisms, it is modified here to apply to the head part of the network. A new structure is created by inserting three segments of self-attention in the head, and a model is obtained through training. It will be called the self-attention model. A version based on the C2f module was also created since comparison. Fig. 10 shows where the self-attention blocks are inserted.

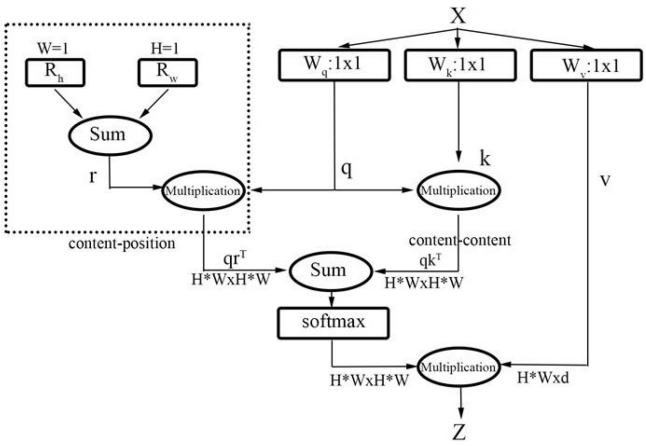

Fig. 9. Self-attention block used in BoTNet.

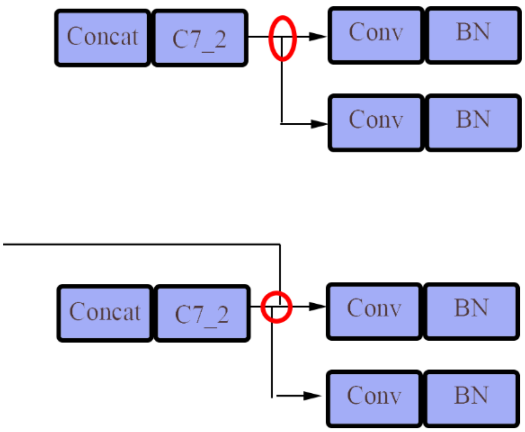

Fig. 10. Where the self-attention added.

H. CBAM Self-Attention

In the preliminary training, CBAM model shows good performance, and the self-attention model shows good plasticity. Therefore, an ensemble attempt was made to combine the two into a better structure. This design simultaneously considers the ability of the two attention mechanisms in the adoption of external information and internal information, making them complementary. The input feature map will first go to CBAM to pass the attention block based on external information, and then return to self-attention to pass the attention block based on internal information. In order to test likelihood and maintain consistency, a layer combining CBAM block and convolutional layer is inserted into the CBS module after SPPCSPC module, see Fig. 10.

This tentative architecture will be referred to as Dual-attention model in this paper. There is also a C2f module-based version of this structure for comparison.

IV. Results

A. Experimental Setup

Each improved model, including YOLOv7 and YOLOv8 itself, is implemented in the Google Colab Python environment. The default number of epochs for YOLOv7 is 299; however, through several training runs, it is found that the last 100 epochs hardly change. So all models are trained for 199 epochs with a batch size of 4 to obtain the highest efficiency within the video memory limit. In addition, hyp.scratch-high is used as the hyperparameter setting. Finally, considering that lightweight and low cost is one of the performance indicators, it is necessary to ensure that other training parameters are consistent to observe the training time. The CBAM model has twice the training time of the other models, about 4 h.

B. Evaluation

Since this is an instance segmentation task with only one class, false negative (FN) is to identify the ocean in the background as the background. However, for satellite images containing debris, the debris is only a small part of the image in most cases, so even if the debris detection error occurs, the accuracy rate is high (greater than 95%) due to the large area of the ocean background. Therefore, this part makes a more objective and accurate analysis of the performance of machine learning by using both precision and recall rate. In addition to precision and recall rates, which are commonly used to evaluate machine learning models, there is an intersection over union (IoU), which is commonly used for object recognition.

Confusion matrices commonly used in machine learning are introduced to compute the results. It is divided into two dimensions, "actual" and "predicted," and divided into four categories: true positive (TP), true negative (TN), false positive (FP), and FN. Precision is the fraction of all samples that are predicted to be correct that are actually correct. In this work, the precision is the ratio when the pixel data in the prediction map and the pixels on the ground truth are all equal to one. Colloquially speaking, the precision representation is actually the ability of the model to correctly identify marine debris as Marine debris.

For the BOX part of instance segmentation, the self-attention model has the best precision score of 0.832, which means that this model has the most accurate positioning ability for marine debris. For the MASK part, CBAM model has the best precision score of 0.787, which means that this model has the most accurate ability to describe the size and shape of marine debris.

Recall is the proportion of all positive samples that are correctly identified as positive. In the context of this work, recall represents the proportion of marine debris that is correctly identified. For the BOX aspect and the MASK aspect, the coordinate attention model with C2f achieves the best Recall scores of 0.787 and 0.773, respectively, which indicates that the model can correctly identify more marine debris.

Mean average precision (mAP) is a statistic to evaluate the performance of an object detection model, which represents the mean of the average precision across different classes. Average precision (AP) is the area under the precision curve at different recall rates. Since this work has only one class (marine debris), mAP is equivalent to AP. In terms of BOX, the best performance of mAP_0.5 is the dual-attention model with C2f, with 0.773. The best performance in MASK is the coordinate attention model with C2f, and its mAP_0.5 is 0.743.

TABLE I
STATISTICAL RESULT OF BOX

Model	Precision	Recall	mAP_0.5	F1-Score	IoU
YOLOv7	0.695	0.635	0.674	0.66	0.50
YOLOv8	0.756	0.656	0.746	0.70	0.54
Coordinate attention model	0.750	0.672	0.765	0.71	0.55
Coordinate attention model + C2f	0.622	**0.787**	0.762	0.69	0.53
CBAM model	0.821	0.721	0.756	**0.77**	**0.62**
CBAM model + C2f	0.665	0.717	0.718	0.69	0.53
Self-attention model	**0.832**	0.541	0.675	0.66	0.49
Self-attention model + C2f	0.693	0.557	0.616	0.62	0.45
Dual-attention model	0.795	0.639	0.751	0.71	0.55
Dual-attention model + C2f	0.706	0.721	**0.773**	0.71	0.56

The bold values denote the best one among the comparison.

TABLE II
STATISTICAL RESULT OF MASK

Model	Precision	Recall	mAP_0.5	F1-Score	IoU
YOLOv7	**0.787**	0.609	0.696	0.69	0.52
YOLOv8	0.686	0.639	0.675	0.66	0.49
Coordinate attention model	0.737	0.639	0.671	0.68	0.52
Coordinate attention model + C2f	0.636	**0.773**	**0.743**	0.70	0.53
CBAM model	**0.787**	0.689	0.716	**0.73**	**0.58**
CBAM model + C2f	0.679	0.721	0.737	0.70	0.54
Self-attention model	0.710	0.459	0.525	0.56	0.39
Self-attention model + C2f	0.663	0.475	0.553	0.55	0.38
Dual-attention model	0.692	0.541	0.593	0.61	0.43
Dual-attention model + C2f	0.773	0.508	0.578	0.61	0.44

The bold values denote the best one among the comparison.

The *F*-score is a measure of the performance of a classification model, where in this project it is represented by two classes: "with debris" and "without debris." It is the harmonic mean of precision and recall. A higher *F*-score indicates that the model has a stronger ability to identify positive examples and a higher comprehensive ability. But the *F*-score is also affected by a parameter β, which indicates the importance of precision and recall rates. When β is greater than 1, more attention is paid to recall. When β is less than 1, more attention is paid to the accuracy. In our work, we assigned equal weight to precision and recall, namely $\beta = 1$, while from a practical standpoint, precision and recall are equally important. In terms of F1 score, the CBAM model has the best score, 0.77 and 0.73, respectively, from the perspective of BOX or MASK. This represents its best performance as an instance segmentation model.

A common evaluation measure for object detection, neural network detectors, and semantic segmentation techniques is called intersection of union (IoU). Its concept can be used to measure the amount of overlap between the predicted region and the ground truth. When applied to instances with irregular borders, the IoU value is calculated in terms of pixels rather than the area of the box.

In theory, an IoU of 1 is the perfect model, but in practice an IoU of 0.5 is considered a good model. On the IoU evaluation, the CBAM model still achieves the best performance, achieving 0.62 for BOX and 0.58 for MASK.

C. Statistical Comparison

Since the resulting data is too large, two tables are plotted separately in this part, where Table I is the performance in terms of BOX and Table II is the performance in terms of MASK. Each model was trained more than twice and the best value was averaged.

As can be seen from Table I, from the perspective of F1 score, some models with attention mechanisms have better comprehensive performance than the original YOLOv7 or the updated YOLOv8. The C2f module in the architecture of YOLOv7 will increase the recall rate and reduce the precision, which means that more debris waste can be accurately found in this task. However, while this can increase the average precision in some cases, it can lead to a decrease in overall performance in many cases.

As can be seen from Table II, the data of the Mask part is generally lower than that of the Box part, and YOLOv7 itself even has better performance than the later YOLOv8, which may be due to the fact that YOLO is more biased towards object detection. The addition of an attention mechanism does not bring significant improvement to the model. This may be due to the fact that implementing various attention mechanisms on the code refers more to object detection than semantic segmentation. Even so, CBAM model was able to achieve the best results.

Combining the two tables, it is clear that CBAM model has the best performance. Although the structure with the coordinate attention layer is weaker than the CBAM layer in performance, it has lower training cost, which is represented by half of the training time and lower occupancy than CBAM. Self-attention, which also has a lower training time, suffers the most despite its numerically worse performance, as will be shown in later sections. Theoretically, the C2F module can provide richer and lighter gradient information, while CBAM requires traditional coherent convolution layers, but judging from the results, the impact of the addition of the C2F module on the model is complex. Due to the poor numerical performance, it is difficult to judge whether the self-attention model or the C2f module have a good correlation.

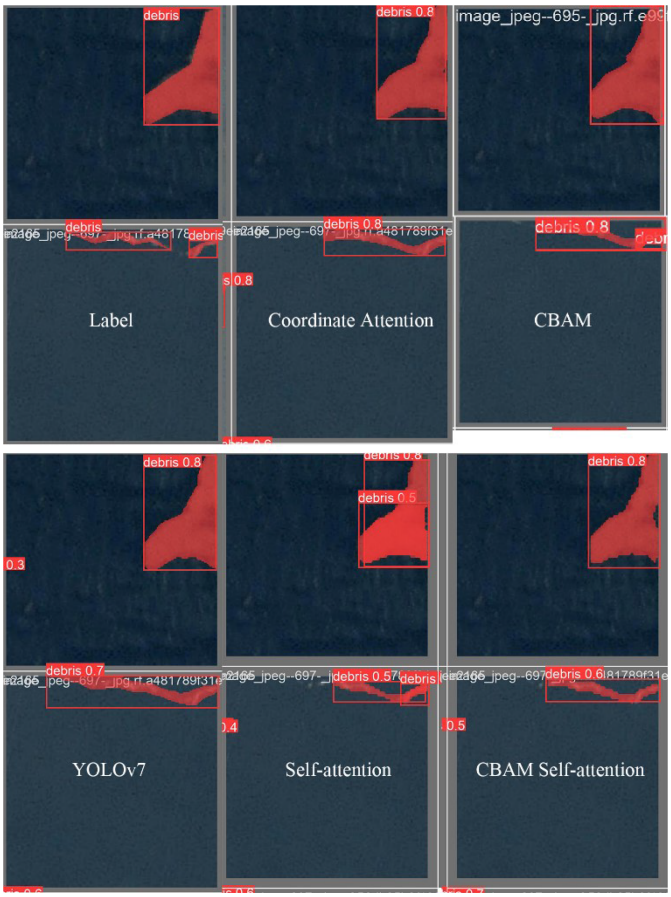

Fig. 11. General comparison of improved models.

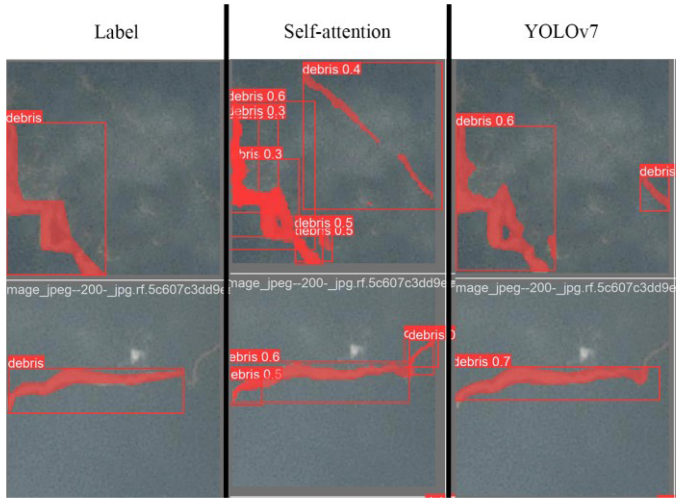

Fig. 12. Comparison shows self-attention model has a complete detection of Marine debris. The left column represents annotations, the middle column represents self-attention, and the right column represents YOLOv7.

Fig. 13. Detailed comparison between self-attention model and CBAM model in some scenes. Left is self-attention model and right is CBAM model.

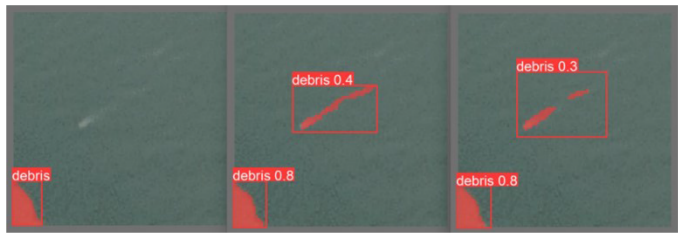

Fig. 14. Comparison between label, self-attention model and dual-attention model. The left one is label, the middle one is a self-attention model, and the right one is a dual-attention model.

D. Images Comparison

Fig. 11 is a comparison of the output of multiple models, which shows that these models generally have good predictive ability and have little deviation from each other. The images of the CBAM model use a different form of output due to the interruption by Google Colab. It can be seen that almost all models predict the exact location of the marine debris, and some of them have successfully detected small parts missed in the label. This led to all test images being reaudited to re-evaluate the model accuracy from an image perspective. The self-attention model obviously produces a lot of repeated judgments, however this allows it to get the most complete segmentation results in some scenes, as shown in Fig. 12.

Fig. 13 shows the mask difference between the self-attention model and the CBAM model for another debris fragment that is long and close to the border. The CBAM model losses a part of debris next to the border of image while the self-attention model correctly finds it.

The dual-attention model combining CBAM and self-attention inherits some of the advantages of self-attention, but is more conservative. Self-attention model detects a piece of unlabeled debris but it also identifies large swaths of the ocean as part of the debris. The error of this part is reduced through the dual-attention model, but there is still a part of the tail that is difficult to identify whether it is marine debris (see Fig. 14).

In any case, by comparing the output images, the performance and practical significance of the model can be more accurately evaluated. The self-attention model has a lower score when calculating the result due to the phenomenon of duplicate detection and detecting the background as debris. However, according to the image, a large proportion of these cases of classifying the background as debris are identifying the missing parts of the label, so the actual performance of the self-attention model should be better than the performance of data evaluation.

E. Result Analysis

Through the comparison of eight models including C2f, it can be found that C2f seems to have little effect on the comprehensive performance of the model. Its performance is mainly

reflected in the single-item value. Therefore, the model with the C2f module and the attention mechanism is counted as the same model as the model only with the attention mechanism. For example, the dual-attention model and the dual-attention model with C2f both count as dual-attention models. Therefore, although the gap between them seems not to be very large from the numerical point of view (especially in terms of Mask), it can be seen that the CBAM model has the best comprehensive performance, followed by the coordinate attention model, combined with the detailed comparison on the image. This is followed by the dual-attention model, then the YOLOv7 instance segmentation itself, and the worst is the self-attention model.

It can be seen that the performance of the dual-attention model with CBAM decreases to a level similar to that of the coordinate attention model due to the addition of the self-attention block. In terms of comprehensive performance and single attribute (precision and recall), the dual-attention model is more balanced than the coordinate attention model. However, the dual-attention model is more biased than the coordinate attention model when considering Mask and Box simultaneously. The training cost (training time) of the dual-attention model is almost the same as that of the coordinate attention model, which is about 25% longer than that of the self-attention model. Therefore, in cases where it is difficult to afford the higher training cost of the CBAM model, the coordinate attention model can be prioritized.

Unlike YOLOv8, the model of YOLOv7 instance segmentation can obviously be used for debris waste detection tasks in satellite images with only the sea surface, however, its comprehensive performance is indeed lower than several improved models proposed in this work. The self-attention model is affected by a series of restrictions, which has great limitations for its evaluation, so its data are not reliable and have less practical significance, and more work is still needed to test it.

V. Discussion

This research seeks to integrate YOLOv7 instance segmentation with diverse attention mechanisms toward enhancing the detection of marine debris in satellite imagery. The goal is to establish a reliable, efficient, and extensible real-time monitoring system for identifying marine debris waste. Overall, the nine models tested in this work, including YOLOv7, demonstrated their satisfactory ability to identify marine debris. While specific metrics like precision might seem subpar, factors like recall, F1-score, and visual analysis indicate promising results. Among all the modified versions evaluated, only the self-attention variant performed worse than YOLOv7, while others surpassed or matched its accuracy. The CBAM model yielded top outcomes. Coordinate attention and dual attention also exhibited competitive yet affordable capabilities. Nevertheless, keep in mind that these observations do not necessarily imply the inferiority of self-attention over YOLOv7; additional experiments with richer datasets would be necessary to draw definitive conclusions.

The dataset was insufficient and biased toward specific situations. Satellite images with a resolution of 3 m/pxl exclusively featuring sea surfaces remain hard to find online. A large portion of these datasets consists of images of land areas such as beaches. Websites like USGS Earth Explorer and Sentinel Open Access Hub display no data or only blurred images when zooming in on the ocean section. Consequently, the dataset remains quite scarce, and homogeneous, and fails to cover various weather settings, further reducing the chances of the attention mechanism performing optimally. Small sample sizes impede the model's capacity to discern among diverse lightning and cloud densities.

Furthermore, implementing and tweaking attention mechanisms is not always suitable for detecting maritime waste via satellite imagery. Just as YOLOv8, which performed better on COCO dataset, did not perform as well as YOLOv7 in this work. Although these attention mechanisms have been experimented with and their high performance has been verified, their application scenarios are still limited.

Two primary paths could branch off from this work: one focused on improving applications in the marine domain, and another centered on advancing deep learning techniques.

In terms of marine usage, there are limitations due to insufficient dataset diversity, which affects attention mechanism performance. To develop a robust and versatile marine debris recognition system, access to high-quality images from various environments, light conditions, and geographical regions would be essential. In addition, using spectrograms instead of RGB images, which capture more detailed information, could enhance the effectiveness of the solution. Moreover, integrating oceanographic knowledge, particularly regarding sea currents, would help forecast the movement of marine waste, making the platform even more valuable. Finally, this work uses YOLO as a base to expand the possibilities of real-time detection. This real-time does not exactly mean streaming media, but rather, depending on how the satellite collects the data, it can mean quick preliminary processing and detection of the images. This is also a feature worth exploring.

As for deep learning improvements, the viability of self-attention mechanisms in debris identification was proven during this investigation. Exploring novel ways of applying self-attention without relying on manually labeled data could streamline the process significantly. Unsupervised learning combined with self-attention holds enormous promise for cutting down human effort while boosting efficacy. Conducting extensive experimental evaluations of self-attention variants will pave the way forward toward developing advanced AI tools tailored specifically for addressing ecological issues, notably those involving the oceans. For future research, the problem of marine debris can also introduce tensor-based learning or semi-supervised learning methods to further expand the functions of related systems to adapt to more diverse functions that may be needed in the future.

VI. Conclusion

In conclusion, our research underscores the transformative impact of attention mechanisms on advancing the precision and efficacy of marine debris detection within satellite imagery. Through comprehensive experimentation, we establish CBAM as the optimal attentional model for this purpose, while acknowledging the potential practical advantages offered by the bottleneck transformer in specific scenarios. This study not

only highlights the immediate practical implications for marine debris detection but also charts a course for future investigations. Noteworthy areas for further exploration include the imperative for more diverse datasets, the potential advantages of incorporating spectrograms, and the integration of oceanographic knowledge. Emphasizing our commitment to advancing the field, our work makes a substantial contribution to the development of resilient and adaptable marine debris recognition systems. By doing so, we envision facilitating a significant stride in the global endeavor to mitigate and combat debris pollution in our oceans.

Acknowledgment

Views and opinions expressed in this paper are however those of the authors only and do not necessarily reflect those of the funders. None of the funders can be held responsible for them.

References

[1] National Oceanic and Atmospheric Administration, "What is Marine Debris?." Accessed: Feb. 1, 2024. [Online]. Available: https://marinedebris.noaa.gov/discover-marine-debris/what-marine-debris

[2] T. Maes, et al, United Nations Environment Programme, "From pollution to solution: A global assessment of marine litter and plastic pollution," Nairobi, 2021. [Online]. Available: https://research.usc.edu.au/esploro/outputs/report/From-Pollution-to-Solution-A-Global/99584903702621

[3] L. Lebreton et al., "Evidence that the Great Pacific Garbage Patch is rapidly accumulating plastic," *Sci. Rep.*, vol. 8, 2018, Art. no. 4666, doi: 10.1038/s41598-018-22939-w.

[4] S. Chu, J. Wang, G. Leong, L. Ann Woodward, R. J. Letcher, and Q. X. Li, "Perfluoroalkyl sulfonates and carboxylic acids in liver, muscle and adipose tissues of black-footed albatross (*Phoebastria nigripes*) from Midway Island, North Pacific Ocean," *Chemosphere*, vol. 138, pp. 60–66, 2015, doi: 10.1016/j.chemosphere.2015.05.043.

[5] M. F. Fava, "Ocean plastic pollution an overview: Data and statistics," Ocean Literacy Portal, Jun. 9, 2022. [Online]. Available: https://oceanliteracy.unesco.org/plastic-pollution-ocean/

[6] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," *Pattern Recognit.*, vol. 34, no. 12, pp. 2259–2281, 2001, doi: 10.1016/S0031-3203(00)00149-7.

[7] R. Kundu, "YOLO: Algorithm for object detection explained [+Examples]," V7, Feb. 3, 2023. [Online]. Available: https://www.v7labs.com/blog/yolo-object-detection

[8] Marine Debris Program at National Oceanic and Atmospheric Administration, "New tools for collecting and exploring marine debris data." Accessed: Feb. 1, 2024. [Online]. Available: https://blog.marinedebris.noaa.gov/mdmap

[9] Marine Debris Program at National Oceanic and Atmospheric Administration, "Monitoring toolbox." Accessed: Feb. 1, 2024. [Online]. Available: https://marinedebris.noaa.gov/monitoring-toolbox

[10] K. Song, J. Y. Jung, S. H. Lee, S. Park, and Y. Yang, "Assessment of marine debris on hard-to-reach places using unmanned aerial vehicles and segmentation models based on a deep learning approach," *Sustainability*, vol. 14, no. 14, Jul. 2022, Art. no. 8311, doi: 10.3390/su14148311.

[11] "Drones are helping to clean up the world's plastic pollution," *CNN*, Jun. 23, 2021. [Online]. Available: https://edition.cnn.com/2021/06/23/Europe/ellipsis-drone-plastic-pollution-c2e-spc-intl/index.html

[12] R. De Vries, "Using AI to monitor plastic density in the ocean," The Ocean Cleanup, Feb. 8, 2022. [Online]. Available: https://theoceancleanup.com/updates/using-artificial-intelligence-to-monitor-plastic-density-in-the-ocean/

[13] E. Cassidy, "Tracking ocean plastic from space," Earthdata, Feb. 24, 2022. [Online]. Available: https://www.earthdata.nasa.gov/learn/articles/ocean-plastic

[14] A. Jamali and M. Mahdianpari, "A cloud-based framework for large-scale monitoring of ocean plastics using multi-spectral satellite imagery and generative adversarial network," *Water*, vol. 13, no. 18, Sep. 2021, Art. no. 2553, doi: 10.3390/w13182553.

[15] Tutor @ Eduonix, "Real-world implementations of YOLO algorithm - Eduonix blog," Eduonix Blog, Feb. 4, 2022, [Online]. Available: https://blog.eduonix.com/software-development/real-world-implementations-of-yolo-algorithm/

[16] R. Liu and Z. Ren, "Application of Yolo on mask detection task," in *Proc. IEEE 13th Int. Conf. Comput. Res. Develop.*, 2021, pp. 130–136, doi: 10.1109/ICCRD51685.2021.9386366.

[17] J. Watanabe, Y. Shao, and N. Miura, "Underwater and airborne monitoring of marine ecosystems and debris," *J. Appl. Remote Sens.*, vol. 13, no. 4, Oct. 2019, Art. no. 044509, doi: 10.1117/1.JRS.13.044509.

[18] E. A. Mohamed, A. M. Shaker, A. El-Sallab, and M. Hadhoud, "INSTA-YOLO: Real-time instance segmentation," Feb. 2021, *arxiv.2102.06777*.

[19] P. Hurtik et al., "Poly-YOLO: Higher speed, more precise detection and instance segmentation for YOLOv3," *Neural Comput. Appl.*, vol. 34, pp. 8275–8290, 2022, doi: 10.1007/s00521-021-05978-9.

[20] A. Shah, L. Thomas, and M. Maskey, "Marine debris dataset for object detection in planetscope imagery," *Version 1.0, Radiant MLHub*, 2021, doi: 10.34911/rdnt.9r6ekg.

[21] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," Jul. 2022, *arxiv.2207.02696*.

[22] R. Munawar, "GitHub - RizwanMunawar/yolov7-segmentation: YOLOv7 instance segmentation using OpenCV and PyTorch," GitHub. Accessed: Feb. 1, 2024. [Online]. Available: https://github.com/RizwanMunawar/yolov7-segmentation

[23] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," Mar. 2021, *arxiv.2103.02907*.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," Jul. 2018, *arxiv.1807.06521*.

[25] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," Jun. 2019, *arxiv.1906.05909*.

[26] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," Jan. 2021, *arxiv.2101.11605*.

[27] K. Topouzelis, D. Papageorgiou, A. Karagaitanakis, A. Papakonstantinou, and M. Arias Ballesteros, "Remote sensing of sea surface artificial floating plastic targets with Sentinel-2 and unmanned aerial systems (Plastic Litter Project 2019)," *Remote Sens.*, vol. 12, no. 12, Jun. 2020, Art. no. 2013, doi: 10.3390/rs12122013.

[28] L. Biermann et al., "Finding plastic patches in coastal waters using optical satellite data," *Sci. Rep.*, vol. 10, 2020, Art. no. 5364, [Online]. Available: https://doi.org/10.1038/s41598-020-62298-z

[29] M. M. Duarte and L. Azevedo, "Automatic detection and identification of floating marine debris using multispectral satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2002315.

[30] O. Karakuş, "Can we 'sense' the call of the ocean? Current advances in remote sensing computational imaging for marine debris monitoring," Oct. 12, 2022. [Online]. Available: https://arxiv.org/abs/2210.06090

[31] P. Bhadoria, S. Agrawal, and R. Pandey, "Image segmentation techniques for remote sensing satellite images," *IOP Conf. Ser., Materials Sci. Eng.*, vol. 993, no. 1, Dec. 2020, Art. no. 012050, doi: 10.1088/1757-899x/993/1/012050.

[32] K. Makantasis, A. Georgogiannis, A. Voulodimos, I. Georgoulas, A. Doulamis, and N. Doulamis, "Rank-r fnn: A tensor-based learning model for high-order data classification," *IEEE Access*, vol. 9, pp. 58609–58620, 2021.

[33] K. Makantasis, A. D. Doulamis, N. D. Doulamis, and A. Nikitakis, "Tensor-based classification models for hyperspectral data analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6884–6898, Dec. 2018.

[34] C. Baur, S. Albarqouni, and N. Navab, "Semi-supervised deep learning for fully convolutional networks," in *Proc. 20th Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2017, pp. 311–319.

[35] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," 2020, *arXiv:2006.05278*.