SCSE-YOLO: A High-precision Underwater Garbage Detection Model

Yanling Li, Tianyu Zhao, Jiaman Li, Qingqi Liang, Zhipeng Yang and Chongyang Chen

***Abstract*—To address the limitations of current underwater garbage detection algorithms, we propose SCSE-YOLO - an enhanced YOLOv8-based framework that improves detection performance through optimized feature fusion and detection mechanisms. The proposed methodology introduces two principal innovations: (1) Integration of Self-Calibrated Convolutions (SCConv) into the C2f module to enhance multi-scale feature fusion through spatial self-calibration, and (2) Implementation of the Self-supervised Equivariant Attention Mechanism (SEAM) in the detection head to mitigate feature degradation caused by alignment errors, local aliasing, and inter-class occlusions. Comprehensive evaluations on the Neural Ocean dataset demonstrate 4.1% precision and 3.2% mAP improvements over baseline YOLOv8. Cross-validation on the TrashCan dataset reveals consistent enhancements of 4.1% precision and 1.9% mAP. These results substantiate capability to maintain high detection accuracy of the model in complex underwater environments while preserving computational efficiency.**

***Index Terms*—Underwater garbage detection, Multi-object detection, Feature fusion network, Multi-scale feature detection**

I. Introduction

ACCORDING to the United Nations "2024 Global Waste Management Outlook" [1], the global generation of urban solid waste in 2023 surpassed a staggering 2.3 billion tons. Behind this number lies an overwhelming accumulation of discarded plastic products, paper, food scraps, and other waste materials, all contributing to a massive flood of garbage that places a heavy burden on the environment. Projections indicate that global municipal solid waste generation will escalate from 2.3 billion metric tons in 2023 to approximately 3.8 billion metric tons by mid-century, according to forecasts of the report. marking a 56% increase in less than a generation. Among these urban solid wastes, the mixed growth rate of plastic waste is projected to be 2-3%, with an estimated 510 million tons of plastic waste to be produced by 2050.

With the continued increase in plastic consumption, these non-biodegradable materials enter rivers and oceans through various pathways, including human activities, landfill seepage, illegal dumping by vessels, and incomplete wastewater treatment. These plastics gradually break down into micro-sized particles, known as "microplastics", which permeate every corner of the ocean, including five major concentrated plastic accumulation zones: one in the Indian Ocean and two each in the Atlantic and Pacific Oceans. These regions have become "hotspots" for microplastics, posing a long-term and profound threat to the health of underwater ecosystems.

In recent decades, waste generated by human activities has dramatically worsened the pollution of water bodies, severely contaminated precious water resources and significantly affected the survival and ecological balance of aquatic life. Despite the introduction of various environmental regulations worldwide to curb this trend, unfortunately, waste, particularly plastics, continues to be recklessly discarded into water bodies, causing irreversible damage to underwater ecosystems and posing unprecedented risks and challenges to the evolution of aquatic species.

Underwater debris poses significant threats to both wildlife and human activities. It endangers the survival of aquatic life in rivers, oceans, and coastal areas, causing injuries and fatalities, while also degrading natural habitats and contributing to their gradual destruction. Furthermore, underwater waste disrupts navigation, leading to safety hazards and substantial economic losses for the fishing and shipping industries, as well as diminishing the quality of life for coastal communities. Over time, it may also present serious risks to human health and safety [2]. In light of the widespread and profound social, economic, and ecological impacts of pollution [3], there has been a notable increase in research focused on developing systematic monitoring and automated collection frameworks for underwater debris in recent years [4].

Against this backdrop, debris detection methods have gradually diverged into two main approaches: one focused on detecting surface-floating debris and the other dedicated to detecting underwater debris [5]. While surface-floating debris detection presents relatively straightforward challenges, this research concentrates on the more complex domain of submerged debris identification. Within the realm of underwater object detection, several foundational algorithms have been developed: R-CNN [6] employs selective search for region proposal generation and convolutional neural networks for feature extraction, though its processing speed remains a limitation. Subsequent iterations, including Fast R-CNN [7] and Faster R-CNN [8], introduced refinements that enhanced both computational efficiency and detection accuracy, albeit at the

Manuscript received February 11, 2025; revised June 17, 2025.

This work was supported in part by the Science and Technology Project of Henan Province (252102211025), the Excellent Postgraduate Teaching Material Project of Henan Province (YJS2025JC30) and the Scientific Research Innovation Fund of Xinyang Normal University (2024KYJJ087).

Yanling Li is a professor of the School of Computer and Information Technology, Xinyang Normal University, Henan, China, 464000(e-mail: lyl75@163.com).

Tianyu Zhao is a postgraduate student of the School of Computer and Information Technology, Xinyang Normal University, Henan, China, 464000 (corresponding author to provide phone: +8618864507512; e-mail: zty0201520@163.com).

Jiaman Li is a postgraduate student of the School of Computer and Information Technology, Xinyang Normal University, Henan, China, 464000 (e-mail: jiaman0813@163.com).

Qingqi Liang is a postgraduate student of the School of Computer and Information Technology, Xinyang Normal University, Henan, China, 464000 (e-mail: qqliang2022@126.com).

Zhipeng Yang is a teacher of the School of Computer and Information Technology, Xinyang Normal University, Henan, China, 464000 (e-mail: yangzp@xynu.edu.cn).

Chongyang Chen is a teacher of the School of the Computer and Information Technology, Xinyang Normal University, Henan, China, 464000(e-mail: cychen@xynu.edu.cn).

cost of increased resource demands. In contrast, the YOLO [9] family of architectures represents a distinct paradigm as single-stage detectors, demonstrating notable proficiency in handling multi-scale target recognition tasks. In recent years, with the growing awareness of environmental protection in underwater ecosystems, many researchers have begun applying YOLO algorithms for underwater object detection. We chose to base our research on YOLOv8 [10] rather than the latest YOLO models, primarily because YOLOv8 offers faster inference speed and higher accuracy.

In this context, we present a series of methodological refinements to enhance underwater object detection. To tackle the inherent challenges of complex underwater environments, we integrate a Self-Calibrated Convolution (SCConv) module into the YOLOv8 architecture, thereby optimizing multi-scale feature representation and detection accuracy. Furthermore, we incorporate the Self-supervised Equivariant Attention Mechanism (SEAM) module to mitigate issues including inter-class occlusions, spatial misalignment, partial feature degradation, and information loss. Comprehensive evaluations conducted on established underwater debris detection benchmarks demonstrate that the proposed methodology surpasses the baseline YOLOv8 model, yielding measurable improvements in both detection precision and mean Average Precision (mAP) values. These experimental results validate the enhanced robustness and superior performance of model for challenging underwater debris recognition tasks.

II. RELATED WORKS

The performance of underwater object detection is significantly affected by the optical characteristics of the aquatic environment. To address these challenges, recent research has increasingly incorporated deep learning methodologies and customized network architectures to enhance detection accuracy. Chen et al. [11] introduced the SWIPENet architecture, initially validated using the URPC2017 dataset. Experimental results demonstrated that the proposed framework achieved mAP values of 46.3% under these constrained conditions. Subsequent dataset expansion in URPC2018, which incorporated additional object categories, correlated with enhanced detection capabilities, yielding an improved mAP values of 64.5%. This progression in performance motivated the development of advanced methodologies, including an optimized Single Shot MultiBox Detector (SSD) variant by Jiang et al. [12] and a sophisticated high-capacity CNN framework by Han et al. [13], which respectively attained mAP values of 66.9% and 91.2% on standardized evaluation protocols.

Subsequent research endeavors built upon established detection frameworks, notably Faster R-CNN and YOLO architectures. Lin et al. [14] enhanced model performance through RoIMix data augmentation strategies, achieving a 74.92% mAP, while Liu et al. [15] incorporated water quality assessment modules to attain 63.83% mAP under turbid conditions. Xu et al. [16] subsequently proposed the Spatial Attention Feature Pyramid Network (SA-FPN) architecture, validated on the PASCAL VOC benchmark, which demonstrated environmental robustness through 76.27% mAP performance across varied imaging conditions. More recently, YOLOv5-based adaptations have yielded significant improvements: optimized backbone network by Wang et al. [17] achieved 69.3% mAP, and YOLOv5s-CA variant by Wen et al. [18] attained 80.9% mAP through contextual attention mechanisms. These progressive advancements collectively substantiate the efficacy of specialized deep learning architectures in addressing underwater detection challenges.

Teng et al. [19] adopted the YOLOv5 architecture as the foundational detection framework in their study, emphasizing improvements in its predictive accuracy. To this end, they applied an enhanced KMeans++ clustering algorithm to reinitialize the model's anchor boxes, thereby aligning them more effectively with the distribution of target objects. Moreover, the original box regression loss was substituted with the Complete Intersection over Union (CIoU) metric during the training phase. This replacement aimed to enhance bounding box localization by incorporating additional geometric factors such as aspect ratio and center point deviation, beyond conventional overlap measures. Experimental results on the Trash_ICRA19 dataset demonstrated a detection accuracy of 88% and a mean Average Precision (mAP) of 90.6%, reflecting a substantial performance improvement. In parallel, Zhu et al. [20] proposed a modified detection algorithm based on YOLOv8. Notably, the proposed method realized 63.6% in precision (P) and 47.1% in mAP on the TrashCan dataset.

The evolution of underwater object recognition systems reflects a paradigm transition from conventional deep learning frameworks to purpose-built architectures engineered to confront the distinctive challenges inherent in subaquatic environments. This trend signals the movement towards more refined and accurate detection capabilities in underwater environments. The accurate discrimination of diverse underwater object categories while maintaining low false positive rates and preventing missed detections continues to present formidable challenges in underwater perception systems. In this context, this study introduces an improved algorithm: the SCSE-YOLO algorithm, which builds upon the base YOLOv8 detection framework and is specifically designed for underwater debris recognition. To address the complexities of underwater environments, two key enhancements are proposed in this work: the enhancement of multi-scale calibration detection through the C2f module and the reinforcement of detection outputs. The core work of this paper is explained below:

(1) Introduction of SCConv module: The combination of the C2f and Self-Calibrated Convolutions (SCConv) modules, enabling further enhancement of performance of YOLOv8. The C2f module strengthens nonlinear representation and feature extraction capabilities of the model, while the SCConv module reduces computational cost and enhances performance by minimizing feature redundancy. Together, the C2f_Self-Calibrated Convolutions (C2f_SCConv) module significantly improves feature representation capability of YOLOv8 through multi-scale feature fusion and self-calibrated feature extraction. This combination allows YOLOv8 to accurately recognize targets while maintaining efficient computation.

(2) Introduction of the SEAM module: The SEAM attention mechanism can effectively learn the dependencies between features and enhance their expressiveness. By incorporating SEAM into YOLOv8, the model becomes better at handling features of occluded objects, reducing the impact of occlusions on detection accuracy. This enables YOLOv8

to more accurately recognize and localize targets in complex scenarios, such as densely populated categories or objects that overlap.

(3) Comprehensive Benchmarking and Evaluation: Extensive benchmarking and systematic evaluations have substantiated the superior performance of the proposed network. The experimental findings validate the efficacy of the presented approach, demonstrating a remarkable balance between computational efficiency and detection accuracy when compared with state-of-the-art models.

III. MATERIALS AND METHODS

We delve into the datasets that form the backbone of our study, as well as the strategic enhancements made to our algorithmic approach. These datasets are crucial as they encompass the essential samples required for training and validating our algorithms. Moving forward, we will elucidate the array of strategies we have implemented to refine various facets of our algorithms, with the aim of bolstering their precision and efficiency. The execution of these strategies is pivotal in ensuring that our algorithms maintain superior performance across a spectrum of scenarios.

Concluding this segment, we present a detailed exposition of the datasets employed in this paper, as depicted in Table I. This table outlines critical information about the datasets, including their names, sizes, the categories they cover, and the number of samples within each category. Such details are instrumental for readers to grasp our methodology and to assess the performance of our algorithms effectively. Through this meticulous presentation, we aspire to furnish readers with a lucid framework that elucidates the construction and validation of our research.

A. DATASET INTRODUCTION

The Neural Ocean dataset [21] represents a specialized repository of underwater imagery depicting anthropogenic marine debris. Curated by Nurzihan Reya and publicly available via the Roboflow Universe platform, this corpus comprises 5,127 annotated images categorized into 15 distinct classes, including personal protective equipment (masks), metallic containers, glass packaging, polymer-based sacks, discarded tires, and various other submerged waste items. The imagery was collected across heterogeneous aquatic environments, encompassing diverse salinity levels, turbidity conditions, and ecological zones. It is an invaluable resource for researchers developing automated marine debris detection systems.

The TrashCan [22] dataset is an instance segmentation dataset specifically designed for underwater trash detection, with the goal of advancing waste recognition technology in marine environments. The dataset consists of many carefully annotated images covering various types of underwater debris, including plastic bags, bottles, fishing nets, and other common waste items, as well as images of underwater vehicles and related objects. The TrashCan dataset contains 7212 images, each thoroughly annotated by professionals to ensure the accuracy and completeness of the annotations. These images not only capture the appearance and distribution of underwater waste but also document real-world scenarios of unmanned underwater vehicles (AUVs) performing trash detection tasks on the seafloor. This makes the TrashCan dataset suitable not only for waste detection tasks but also for research in fields such as AUV navigation and obstacle avoidance.

TABLE I
DESCRIPTION OF THE DATASET

Dateset	Environment	No.images	No.categories
Neural Ocean	Underwater(Mask, can, bottle, glove,etc.)	5127	15
TrashCan	Underwater (Bag, clothing, rope, can, etc.)	7212	22

B. YOLOV8 MODEL ARCHITECTURE

YOLOv8, an proven object detection framework, was officially released in January 2023 as a stable iteration of the renowned YOLO series, which is widely recognized for its high detection speed and accuracy. The architecture of YOLOv8 is composed of three principal modules: the Backbone, responsible for feature extraction; the Neck, which performs multi-scale feature fusion; and the Head, which executes object classification and bounding box regression. It adopts the FPN-PANet [23] structure pattern, as illustrated in Fig. 1.

C. SCSE-YOLO MODEL ARCHITECTURE

The architecture of SCSE-YOLO is illustrated in Fig. 2. In this design, the C2f module within the feature fusion network of YOLOv8 is replaced by a newly proposed C2f_SCConv module. The integration of the C2f_SCConv block represents a key innovation in adapting the YOLOv8 framework for underwater object detection tasks, offering enhanced feature representation capabilities in challenging aquatic environments. By combining the strengths of the C2f module and the SCConv module, it not only enhances detection performance and feature representation capability of the model but also optimizes computational efficiency. This improvement makes YOLOv8 more adept at handling complex and dynamic detection tasks.

Next, we replace the second convolution operation of the three detection heads in the Head network with the SEAM module [24]. The main role of the SEAM module in YOLOv8 is to enhance detection accuracy of the model in occlusion scenarios, particularly when dealing with occlusion issues in multi-object detection tasks and complex backgrounds. Additionally, the SEAM module enhances feature expression ability of the model, enabling it to more accurately utilize information in subsequent detection and recognition tasks.

D. SCCONV MODULE

The C2f module was redesigned. Our approach leverages Self-Calibrated Convolution (SCConv) [25] to augment the receptive field of the C2f module and facilitate self-calibration through the integration of information across multiple spatial scales [26]. Fig. 3 illustrates the core concept of

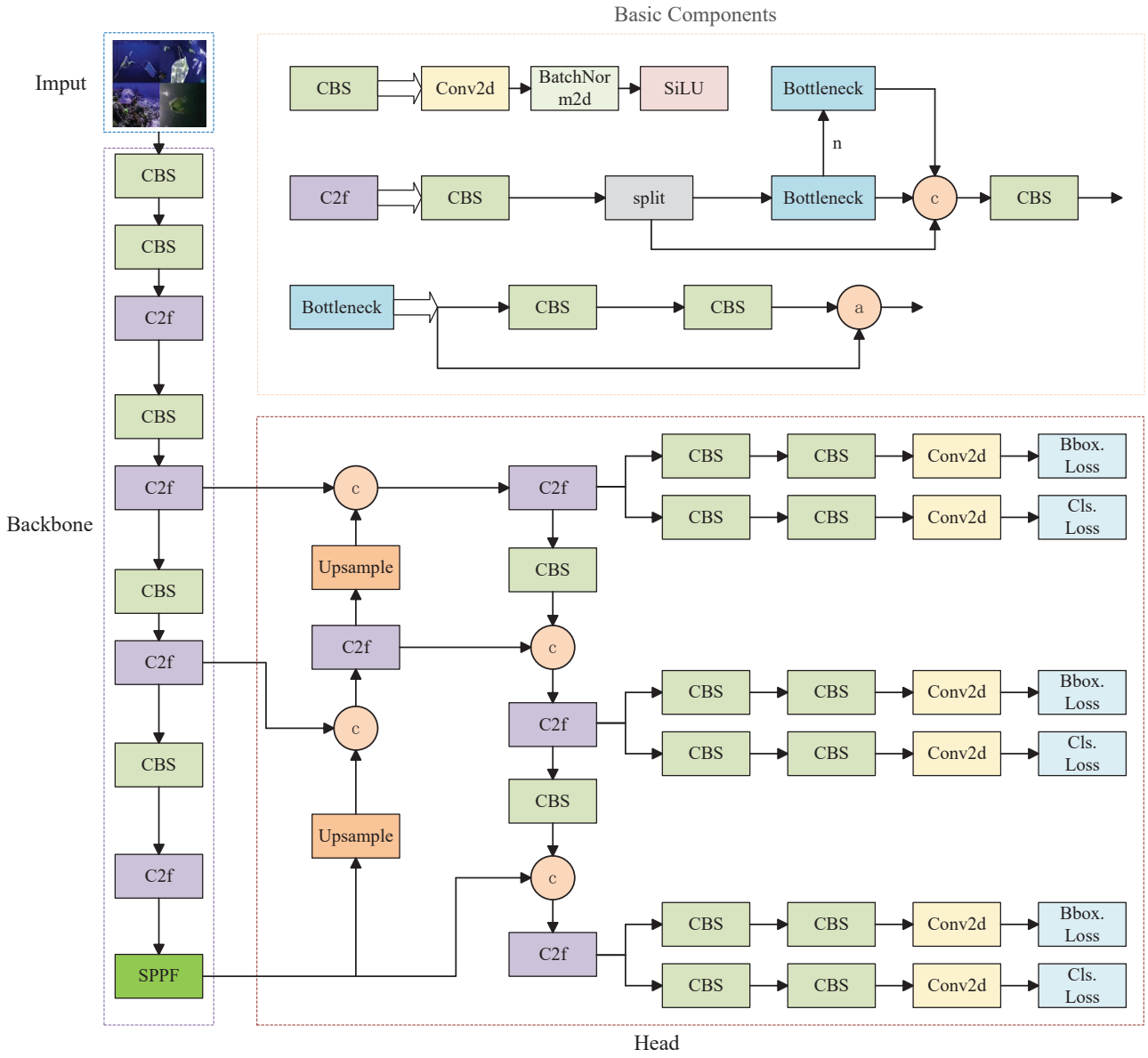

Fig. 1. Architecture for YOLOv8 module

SCConv, which focuses on refining the essential mechanism of convolution-based feature extraction in CNNs without requiring modifications to the existing network architecture.

The SCConv technique adopts grouped convolutions to facilitate multi-scale feature extraction by dividing the channel dimension into two parallel branches. The first branch follows the conventional convolution-based hierarchical feature extraction process, whereas the second incorporates downsampling operations to expand the receptive field. This dual-pathway approach is instrumental in broadening receptive field of the network. Consequently, each spatial location within the network is enabled to perform self-calibration by assimilating information from two disparate spatial scales, thereby enriching the feature representation.

As illustrated in Fig. 3, the architecture operates on input and output feature maps with a channel dimension of C. A kernel tensor K is defined with dimensions (C, C, k_h, k_w), where k_h and k_w denote the kernel height and width, respectively. Initially, the kernel is partitioned into four groups of filters, denoted as $\{K_i\}_{i=1}^{4}$, each with dimensions $(C/2, C/2, k_h, k_w)$. Subsequently, the input feature map X is equally divided into two subsets, $\{X_1, X_2\}$. The self-calibration procedure is then applied to X_1 using the filter set $\{K_1, K_2, K_3\}$, yielding the intermediate output Y_1. In the second branch, a standard convolution is applied to X_2 using K_1, formulated as $Y_2 = F_1(X_2) = X_2 \times K_1$, which is designed to preserve the original spatial context. Finally, Y_1 and Y_2 are concatenated along the channel dimension to produce the output Y. Specifically, the self-calibration process begins apply to X_1 with a window size of $r \times r$ and a step size of r, expressed as:

$$T_1 = AvgPool_r\left(X_1\right) \tag{1}$$

next, T_1 undergoes a feature transformation process utilizing filter K_2 :

$$X_1^{'} = Up(F_2(T_1)) = Up(T_1 \times K_2) \tag{2}$$

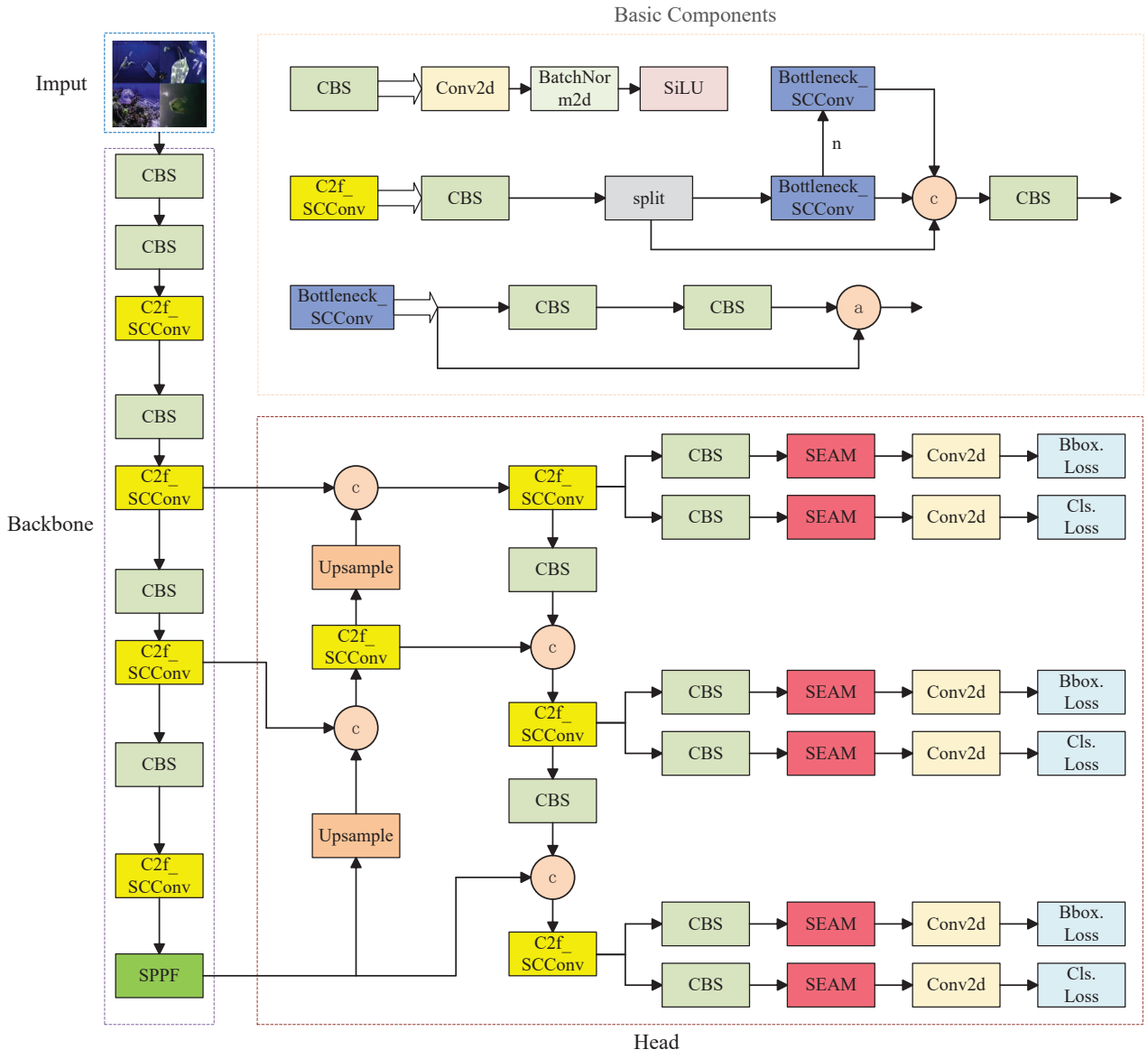

Fig. 2. Architecture for SCSE-YOLO module

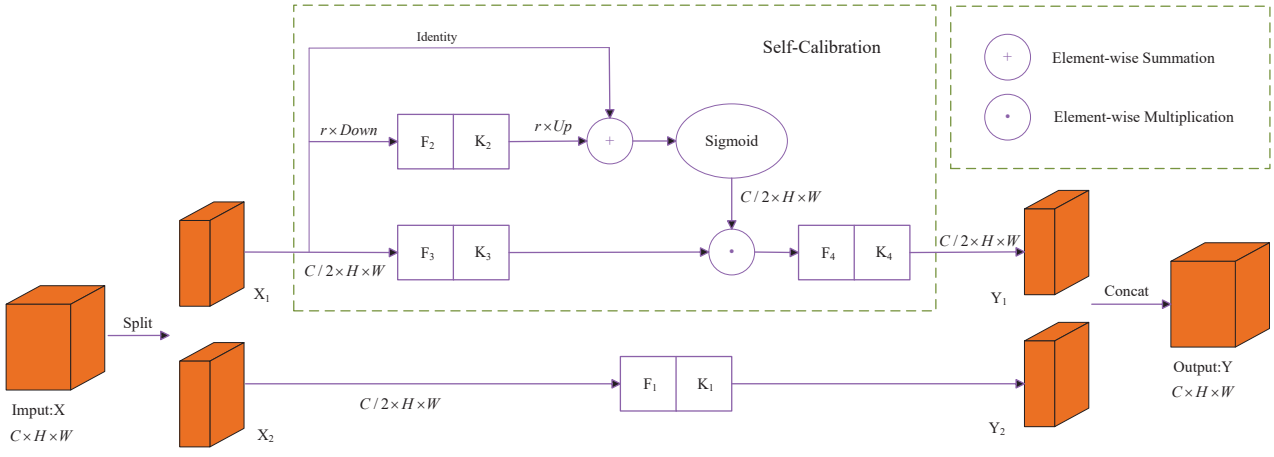

Fig. 3. Architecture for SCConv module

here, Up(·) represents a bilinear interpolation operation that reconstructs intermediate features from a lower-resolution scale to the initial feature resolution. Consequently, the self-calibration mechanism can be defined as:

$$Y_1^{'} = F_3(X_1) \cdot \sigma(X_1 + X_1^{'}) \tag{3}$$

where $F_3(X_1) = X_1 \times X_3$, with σ denoting the activation function. The term X_1' acts as a residual component utilized for generating the self-calibration weights. The ultimate output resulting from the self-calibration process can be expressed as:

$$Y_1 = F_4(Y_1^{'}) = Y_1^{'} \times K_4 \tag{4}$$

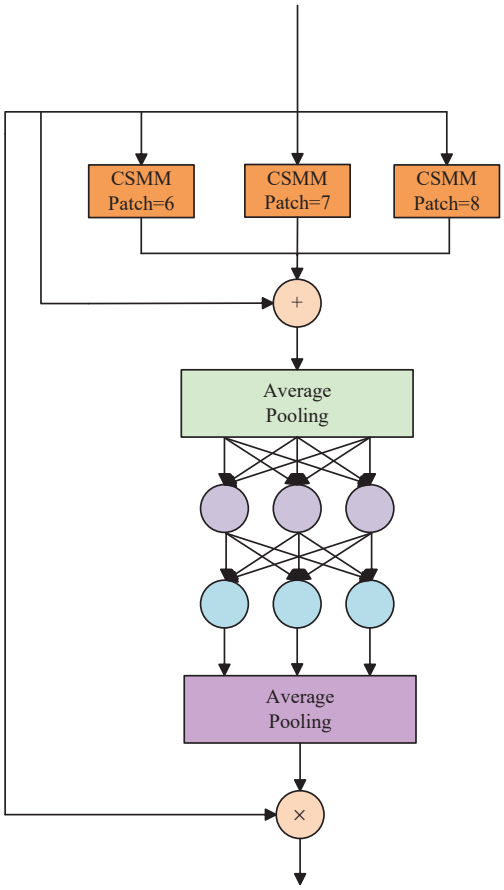

Fig. 4. Architecture for SEAM module

E. SEAM MODULE

The SEAM module [27] aims to achieve multi-scale object detection, highlight object regions in the image, and simultaneously suppress background regions. The module architecture is shown in Fig. 4. The initial components of SEAM require depth-separable convolutions with residual connections. Depth-separable convolutions operate on channels, separating them individually. While this technique effectively reduces the parameter count and enhances the model's understanding of channel importance, it does not fully account for the inter-channel relationships, which can limit its ability to capture complex feature interactions. To address this limitation, we introduce pointwise (1x1) convolutions, which are applied to combine the outputs of convolutions at different depths. These 1x1 convolutions enable the model to integrate feature maps from different layers, allowing for more efficient dimensionality reduction and ensuring that crucial information from various stages of processing is preserved.

To further strengthen the inter-channel relationships, we employ a dual-layer densely network that operates on the concatenated outputs of the pointwise convolutions. This network enables the fusion of data throughout all channels, thus boosting the interconnection among them. Through analyzing the relationships and interrelations among various feature channels, the model becomes more adept at capturing intricate patterns that may be missed when channel interactions are ignored. Additionally, the model is designed to compensate for the loss of information in occlusion scenarios. By learning the relationships between occluded and non-occluded targets during training, the network becomes more resilient to occlusions, which are common in object detection tasks. The model can infer and predict the likely locations and characteristics of occluded objects based on the features of non-occluded objects for the same location, improving detection accuracy in challenging environments.

Conclusively, the output of the Spatial-Enhanced Attention Module (SEAM) is multiplied by the original feature maps, effectively re-weighting the features based on the enhanced channel relationships. This step enables the model to better handle misalignment errors, local aliasing, and feature loss that may arise due to inter-class occlusions. By emphasizing relevant features and downplaying irrelevant ones, the model can more effectively detect objects, even when they are partially obscured or misaligned, leading to improved overall detection performance [28].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL PLATFORM AND MODEL PARAMETERS

Presented in Table II are the hardware platform parameters and configuration utilized for the experimental training phase.

TABLE II
TRAINING PLATFORM PARAMETER CONFIGURATION

Parameters	Configuration
Operational platform	PyCharm
Compilers	Python 3.8
Network construction method	PyTorch 2.1.0+cu121
CPU	Intel Core i5-12400F
GPU	NVIDIA GeForce RTX4060

Presented in Table III are some essential setup of parameter.

B. EVALUATION METRICS

A range of standard evaluation indices were employed to objectively measure the efficacy of the proposed approach. Specifically, these included Precision (P), Recall (R), Average Precision (AP), mean Average Precision (mAP), billion floating-point operations (GFLOPs), and the number of parameters (Params). Collectively, these indices offer a

TABLE III
SOME ESSENTIAL PARAMETERS SETUP

Parameters	Setup
Epochs	300
Batch size	16
Workers	8
Input image size	640x640
Optimizer	SGD
Data enhancement strategy	Mosaic

thorough assessment of effectiveness of the model. Precision measures the accuracy of the model's predictions, Recall evaluates the completeness of its detections, and mAP offers an overall performance metric that balances both precision and recall across all classes. The formulas for P, R, AP, and mAP are presented below:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{5}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{6}$$

$$AP = \int_0^1 P(R)dR \tag{7}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{8}$$

C. EXPERIMENT RESULTS

The experimental results derived from the Neural Ocean dataset are systematically detailed in Table IV. Through a meticulous comparative analysis of classic YOLO architectures and our proposed SCSE-YOLO model—an innovative extension of the YOLOv8 framework—we demonstrate the transformative impact of targeted architectural enhancements on marine debris detection. Specifically, our SCSE-YOLO model integrates context-aware spatial-channel refinement mechanisms, designed to enhance feature discriminability and scale invariance in underwater environments. Comparative analysis reveals a 4.1% accuracy improvement over the baseline YOLOv8, coupled with a 3.2% increase in mAP. These advancements are particularly significant in the context of marine debris detection, where complex backgrounds, variable lighting, and object occlusion pose substantial challenges. The mAP enhancement underscores not only improved detection rates but also more precise object localization, critical for operational applications requiring accurate debris positioning.

To rigorously validate adaptability and practical relevance of the model, we conducted cross-dataset evaluations using the TrashCan dataset (Table V), which introduces additional environmental complexities and debris diversity. Notably, SCSE-YOLO maintained a consistent 4.1% accuracy advantage while achieving a 1.9% mAP improvement. The attenuated mAP gain compared to Neural Ocean may reflect higher object density of TrashCan, smaller debris sizes, and increased visual clutter, necessitating finer-grained feature discrimination. This nuanced performance variation highlights robustness of the model across heterogeneous scenarios while underscoring remaining challenges in small object detection under extreme environmental conditions.

Collectively, these findings establish SCSE-YOLO as a vital framework for marine debris detection, demonstrating significant progress in both detection accuracy and precision across distinct datasets. The consistent performance improvements suggest strong generalizability, positioning the model as a valuable tool for advancing marine pollution monitoring and mitigation efforts. The enhanced detection confidence and reduced false negative rates are particularly critical for operational deployments where reliable debris identification directly impacts cleanup efficiency and ecological protection.

D. ABLATION EXPERIMENT

This manuscript presents a systematic ablation studies designed to assess the incremental contributions of each enhancement introduced to the YOLOv8s model. Effectiveness of each component was subjected to stringent testing to affirm the validity of the collective improvements. The outcomes of these ablation studies, which are pivotal for understanding the impact of each enhancement, are meticulously documented. Specifically, the results pertaining to the Neural Ocean dataset are encapsulated in Table VI, offering a detailed data for each enhancement introduced to the YOLOv8s model. Similarly, the findings for the TrashCan dataset are elaborated in Table VII, providing a comparative analysis of the detection capabilities across different model configurations.

The findings from these studies reveal that the incorporation of each improved method results in a discernible enhancement in model accuracy, varying in degree. This observation signifies a stepwise improvement in the detection capabilities of the YOLOv8 model as it undergoes refinement. Of particular note is the synergistic effect observed when the self-calibrated convolution and the SEAM attention mechanism were integrated into the YOLOv8 framework, leading to a marked enhancement in detection performance. These results not only underscore the incremental benefits of the proposed improvements but also highlight the potential of our enhanced YOLOv8 model to achieve superior detection accuracy in the context of underwater debris and marine litter. The ability to reduce false negatives and increase confidence in detection is crucial for environmental monitoring and cleanup operations. The manuscript effectively demonstrates that the enhancement to the YOLOv8s model is not only theoretically sound but also effective in practice, and the detailed ablation study also confirms the high efficiency of SCSE-YOLO model.

E. COMPARISON OF MODEL CHECKING EFFECTS

To elucidate the enhanced performance of our SCSE-YOLO model, Fig. 5 provides a comparative analysis of the detection capabilities across the Neural Ocean dataset for YOLOv5s, YOLOv8s, and our SCSE-YOLO model. The visualization clearly demonstrates that while all three models are capable of detecting marine debris, the detection accuracy for YOLOv5s and YOLOv8s are consistently lower compared to our SCSE-YOLO model. This disparity underscores the superior detection confidence of our model, thereby validating its more efficient detection performance. The

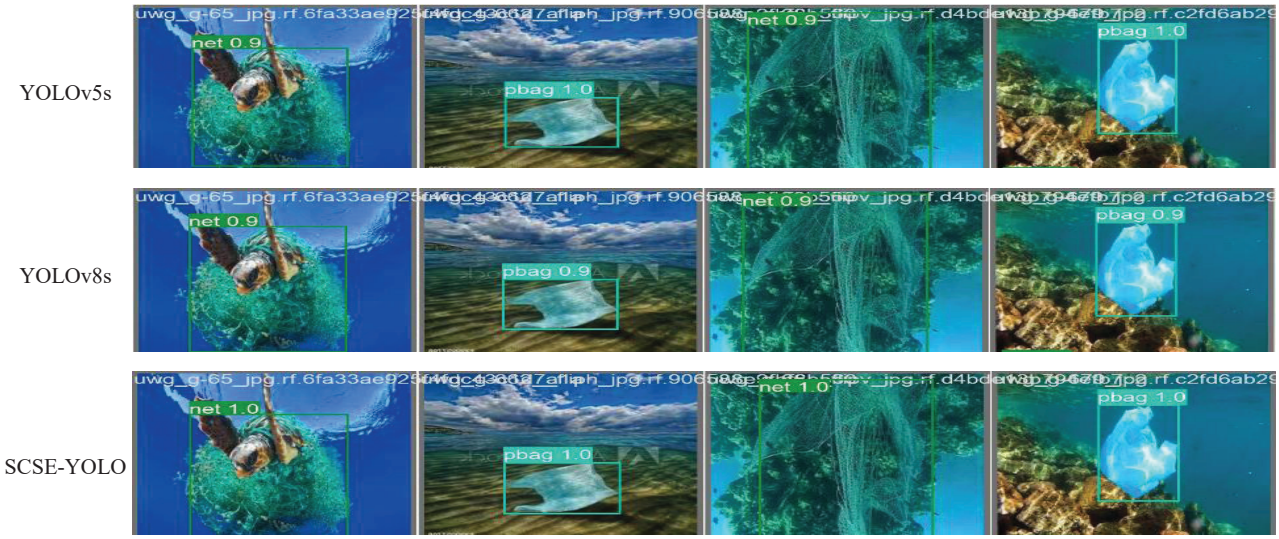

Fig. 5. Comparison of detection results on Neural Ocean dataset

Fig. 6. Comparison of detection results on TrashCan dataset

TABLE IV
COMPARISON OF EXPERIMENTAL RESULTS ON THE NEURAL OCEAN DATASET

Modules	P/%	R/%	mAP/%	Params/M	GFLOPS
YOLOv5s	84.2	77	82.3	9.12	24.1
YOLOv6s	78.9	67.3	73.5	16.3	44.2
YOLOv8s	84.5	77.3	82.6	11.1	27.8
YOLOv9s	83.7	77.5	82.4	7.29	27.4
YOLOv10s	84.3	72.6	80.4	8.04	24.5
YOLOv11s	84.6	76.3	82.4	9.43	21.6
SCSE-YOLO	88.6	82.7	85.8	12.4	28.7

higher confidence scores indicate greater reliability in real-world scenarios where decision-making based on detection outputs is critical. Fig. 6 extends this comparative analysis to the TrashCan dataset, revealing the detection outcomes for plastic waste. The results indicate that, although all models detect plastic waste to some extent, YOLOv5s and YOLOv8s suffer from significant missed detections and lower lower precisions.

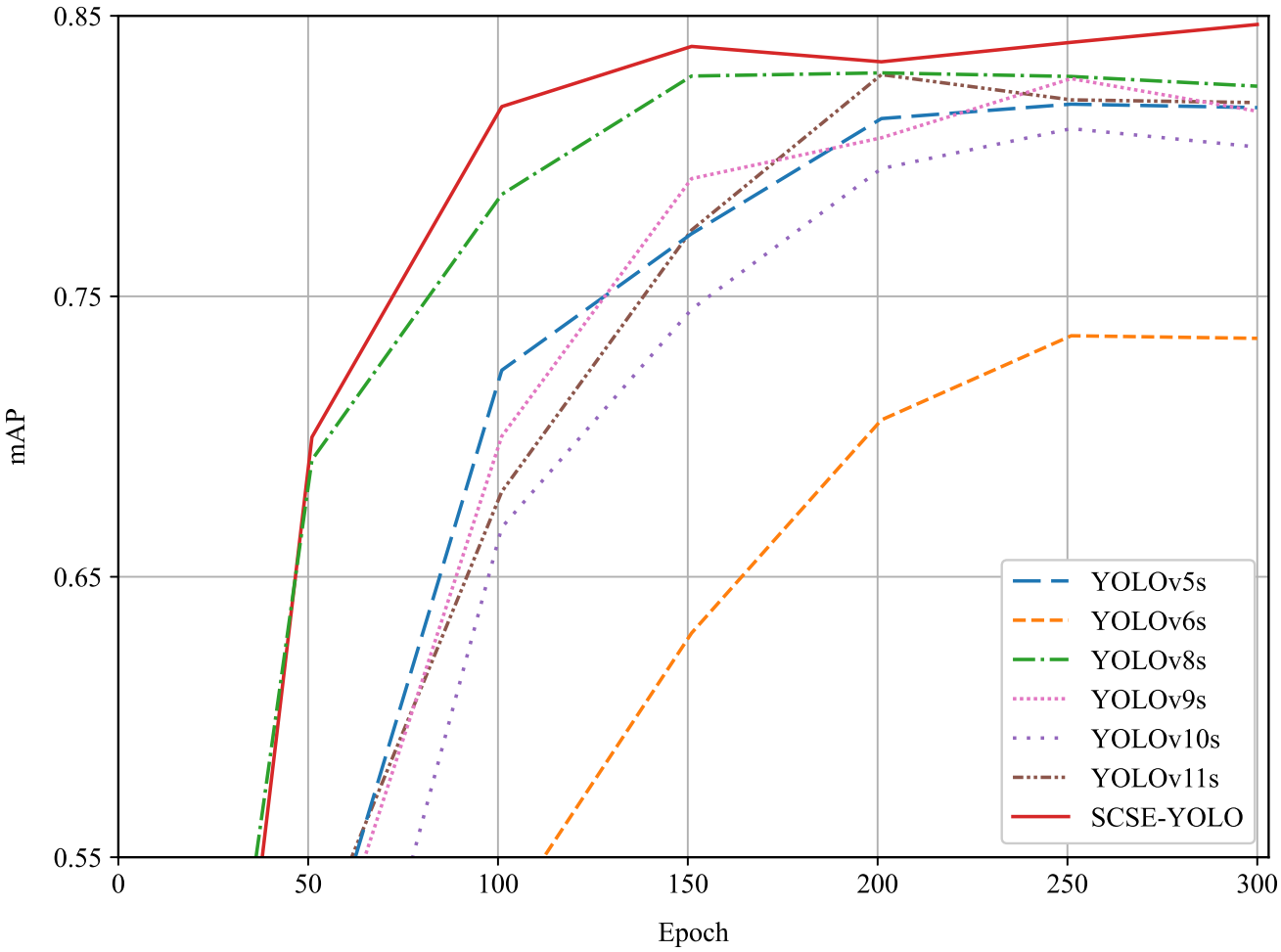

Fig. 7. Comparison of mAP curve results of Neural Ocean dataset

TABLE V
COMPARISON OF EXPERIMENTAL RESULTS ON THE TRASHCAN DATASET

Modules	P/%	R/%	mAP/%	Params/M	GFLOPS
YOLOv5s	77.2	62.4	66.7	9.12	24.1
YOLOv6s	70.8	62.5	64.7	16.3	44.2
YOLOv8s	77.2	63.8	69.7	11.1	27.8
YOLOv8-C2f-Faster-EMAv3[20]	63.6	44.8	47.1	-	-
YOLOv9s	73.7	64.8	67	7.29	27.4
YOLOv10s	75.9	61.9	66.2	8.04	24.5
YOLOv11s	77.6	63.1	67.8	9.43	21.6
SCSE-YOLO	81.3	64.9	71.6	12.4	28.7

TABLE VI
ABLATION EXPERIMENTS ON THE NEURAL OCEAN DATASET

SCConv	SEAM	P/%	R/%	mAP/%	Params/M	GFLOPS
		84.5	77.3	82.6	11.1	27.8
✓		85.1	78.7	83.2	12.9	30.7
	✓	85.4	78.5	83.5	10.6	26
✓	✓	88.6	82.7	85.8	12.4	28.7

In contrast, the SCSE-YOLO model can detect the target more comprehensively with higher accuracy and confidence. This comprehensiveness and accuracy are crucial for the reliable identification of marine debris, particularly in complex environments where debris may be partially obscured or exhibit varying morphological characteristics. The compara-

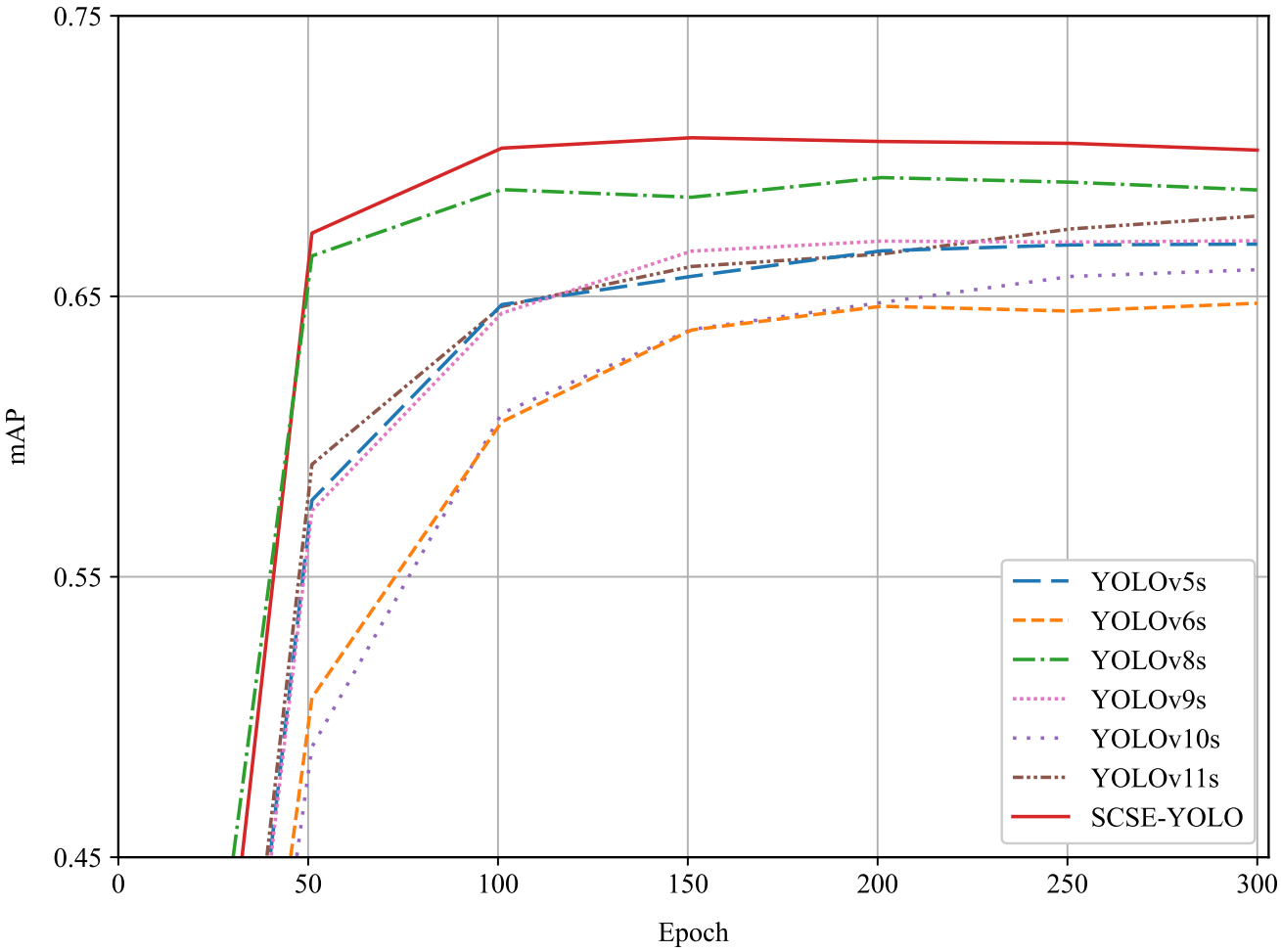

Fig. 8. Comparison of mAP curve results of TrashCan dataset

TABLE VII
ABLATION EXPERIMENTS ON THE TRASHCAN DATASET

SCConv	SEAM	P/%	R/%	mAP/%	Params/M	GFLOPS
		77.2	63.8	69.7	11.1	27.8
✓		78.8	62.9	69.9	12.9	30.7
	✓	77.6	63.1	70.9	10.6	26
✓	✓	81.3	64.9	71.6	12.4	28.7

tive results from both datasets highlight dual strengths of the SCSE-YOLO model: its high detection accuracy and its ability to substantially reduce the rate of missed detections.

F. COMPARISON OF MAP CURVE RESULTS

To substantiate the efficacy of the SCSE-YOLO algorithm employed in this research, a controlled experiment was conducted. The above two underwater debris datasets are used to train and evaluate some classical models of YOLO and SCSE-YOLO models, ensuring that the training conditions remained constant across all models. This approach allows for a direct comparison of their performance metrics. Fig. 7 illustrates the mAP curves for each model on the Neural Ocean dataset, while Fig. 8 presents the corresponding mAP curves for the TrashCan dataset. These visual representations of performance over iterations provide a clear indication of the detection capabilities of each model.

The experimental results for SCSE-YOLO show a progressive improvement in detection accuracy with each iteration. This trend is particularly pronounced when compared to the performance of several classical YOLO models, with SCSE-YOLO consistently outperforming its counterparts. These findings highlight the robustness of the SCSE-YOLO algorithm, demonstrating its ability to achieve higher detection accuracy as training progresses. This is especially significant as it not only validates the superiority of SCSE-YOLO in detecting underwater garbage but also emphasizes its potential for real-world applications where high accuracy and reliability are critical. The observed performance improvements further suggest that the integration of self-calibrated convolution and the SEAM module significantly enhances the detection capabilities about model, which further confirms the applicability of our model for underwater environments.

V. Discussion

This paper proposes a detection model, SCSE-YOLO, designed for efficient performance and robust detection results, even in complex scenes. The SCSE-YOLO model enhances multi-scale feature fusion and detection capabilities, enabling effective multi-object detection. Additionally, the model highlights target areas within the image while suppressing background areas, thus mitigating the effects of feature loss caused by alignment errors, local aliasing, and inter-class occlusion. This optimization ensures the model maintains strong feature detection performance, even in challenging underwater environments.

Although the SCSE-YOLO model has shown remarkable effectiveness in detecting underwater garbage across various datasets, there is still significant potential for further development and application. Subsequent investigations will be dedicated to enhancing the efficiency and real-time responsiveness of the model. Specifically, efforts will be directed at improving model speed while maintaining high accuracy to meet the real-time demands of underwater garbage detection.

References

[1] I. S. W. Association *et al.*, "Global waste management outlook 2024: beyond an age of waste, turning rubbish into a resource," 2024.

[2] M. McCoy, "Marine debris: The us federal role in a local and global problem," *Nat. Resources & Env't*, vol. 35, p. 9, 2020.

[3] L. Miao and Y. Tian, "Detection of small underwater organisms based on improved yolov8." *IAENG International Journal of Computer Science*, vol. 51, no. 8, pp. 1020–1026, 2024.

[4] B. Huang, G. Chen, H. Zhang, G. Hou, and M. Radenkovic, "Instant deep sea debris detection for maneuverable underwater machines to build sustainable ocean using deep neural network," *Science of The Total Environment*, vol. 878, p. 162826, 2023.

[5] H. Zhuang and W. Liu, "Underwater biological target detection algorithm and research based on yolov7 algorithm." *IAENG International Journal of Computer Science*, vol. 51, no. 6, pp. 594–601, 2024.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[7] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[9] M. Hussain, "Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection," *Machines*, vol. 11, no. 7, p. 677, 2023.

[10] M. Safaldin, N. Zaghden, and M. Mejdoub, "An improved yolov8 to detect moving objects," *IEEE Access*, 2024.

[11] L. Chen, Z. Liu, L. Tong, Z. Jiang, S. Wang, J. Dong, and H. Zhou, "Underwater object detection using invert multi-class adaboost with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[12] Z. Jiang and R. Wang, "Underwater object detection based on improved single shot multibox detector," in *Proceedings of the 2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, 2020, pp. 1–7.

[13] F. Han, J. Yao, H. Zhu, and C. Wang, "Marine organism detection and classification from underwater vision based on the deep cnn method," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 3937580, 2020.

[14] W.-H. Lin, J.-X. Zhong, S. Liu, T. Li, and G. Li, "Roimix: proposal-fusion among multiple images for underwater object detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2588–2592.

[15] H. Liu, P. Song, and R. Ding, "Wqt and dg-yolo: Towards domain generalization in underwater object detection," *arXiv preprint arXiv:2004.06333*, 2020.

[16] F. Xu, H. Wang, J. Peng, and X. Fu, "Scale-aware feature pyramid architecture for marine object detection," *Neural Computing and Applications*, vol. 33, pp. 3637–3653, 2021.

[17] H. Wang, S. Sun, X. Wu, L. Li, H. Zhang, M. Li, and P. Ren, "A yolov5 baseline for underwater object detection," in *OCEANS 2021: San Diego–Porto*. IEEE, 2021, pp. 1–4.

[18] G. Wen, S. Li, F. Liu, X. Luo, M.-J. Er, M. Mahmud, and T. Wu, "Yolov5s-ca: A modified yolov5s network with coordinate attention for underwater target detection," *Sensors*, vol. 23, no. 7, p. 3367, 2023.

[19] X. Teng, Y. Fei, K. He, and L. Lu, "The object detection of underwater garbage with an improved yolov5 algorithm," in *Proceedings of the 2022 International Conference on Pattern Recognition and Intelligent Systems*, 2022, pp. 55–60.

[20] J. Zhu, T. Hu, L. Zheng, N. Zhou, H. Ge, and Z. Hong, "Yolov8-c2f-faster-ema: An improved underwater trash detection model based on yolov8," *Sensors*, vol. 24, no. 8, p. 2483, 2024.

[21] M. Shah, D. Garg, R. Ghariya, V. Solanki, R. Rajput, and M. Chauhan, "Enhancing marine conservation: Yolov8-based underwater waste detection system," in *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2023, pp. 1396–1404.

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

[23] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8759–8768.

[24] Z. Yu, H. Huang, W. Chen, Y. Su, Y. Liu, and X. Wang, "Yolo-facev2: A scale and occlusion aware face detector," *Pattern Recognition*, vol. 155, p. 110714, 2024.

[25] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving convolutional networks with self-calibrated convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 096–10 105.

[26] W. Zhu and Z. Yang, "Csb-yolo: a rapid and efficient real-time algorithm for classroom student behavior detection," *Journal of Real-Time Image Processing*, vol. 21, no. 4, p. 140, 2024.

[27] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 275–12 284.

[28] Z. Zhang, L. Tao, L. Yao, J. Li, C. Li, and H. Wang, "Ldsi-yolov8: Real-time detection method for multiple targets in coal mine excavation scenes," *IEEE Access*, 2024.