

YOLOv5-UOD: Underwater Object Detection Method Based on Improved YOLOv5

Xu Guan

*School of Computer Science and
Information Security, Guilin
University of Electronic Technology*
Guilin, China
631118211@qq.com

Xiaonan Luo

*School of Computer Science and
Information Security, Guilin
University of Electronic Technology*
Guilin, China
luoxn@guet.edu.cn

Dong Wang

*College of Mathematics and Informatics
South China Agricultural University*
Guangzhou, China
wdngng@163.com

Abstract—Underwater object detection presents a distinct and intricate challenge within the realm of computer vision. Although object detection methods have demonstrated exceptional performance on conventional datasets, the challenges associated with low visibility and color distortion in intricate underwater environments frequently lead to suboptimal image quality. Consequently, enhancing underwater object detection methods holds significant practical value. In this study, we take a pioneering approach by applying one of the most advanced object detection algorithms, YOLOv5, to underwater environments. We further tailor this approach with techniques designed specifically for underwater scenarios. Our improvements encompass the incorporation of the Swin Transformer as the foundational backbone network for YOLOv5, making the network particularly suitable for the challenges posed by underwater images featuring blurred objects. We enhance the process of feature extraction by harnessing multi-scale feature fusion, which strengthens the representation of object features in detection by integrating cross-scale connections and applying context information weighting operations. This is achieved by constructing a novel feature pyramid by amalgamating the fused feature maps. Additionally, we replace the traditional Non-Maximum Suppression (NMS) with Soft-NMS, enhancing the network's capability to recognize occluded and overlapping underwater objects. The experimental outcomes provide empirical evidence supporting the efficacy of our improved network model for underwater object detection, achieving an impressive average precision (mAP) of 87.3%. Our model outperforms traditional object detection models and proves its applicability in complex underwater environments.

Index Terms—deep learning; underwater object detection; YOLOv5; swin transformer; NMS

I. INTRODUCTION

The vast majority of the Earth's surface is covered by oceans, which contain valuable resources such as oil, natural gas, minerals, chemicals. Furthermore, in recent years, the depletion of terrestrial resources as a consequence of population expansion has prompted several developed countries to

shift their attention towards marine resources, including oil, natural gas, minerals, chemicals, and aquatic life[1,2], leading to increased marine exploration efforts. With the advancement of computer vision technology, utilizing it for ocean exploration has emerged as a novel approach. Within the realm of computer vision, optical-based object detection has gained significant importance in the marine domain. Researchers have increasingly delved into optical-based underwater object detection, achieving notable progress. At present, underwater object detection finds diverse applications in the field of marine ecosystems, including marine ecosystem research, estimating marine biological populations, marine species conservation, deep-sea fishing, detection of underwater unexploded ordnance[3], underwater archaeology and a range of other potential applications further underscore its efficacy as a pathway for the sustainable exploitation of marine resources.

Currently, mainstream object detection algorithms are typically categorized into two primary classes: two-stage object detection methodologies and one-stage object detection approaches. This classification includes two-stage object detection techniques introduced by Girschick et al, such as R-CNN[4], Fast R-CNN[5], and Faster R-CNN[6]. Building upon these algorithms, additional network stages have demonstrated exceptional performance, such as R-FCN[7], Mask R-CNN[8], and Cascade R-CNN[9]. While these algorithms deliver high detection accuracy, they are relatively slower in processing. In response to the need for faster object detection, Redmon et al. introduced one-stage algorithms like YOLO[10], YOLO9000[11], YOLOv2, and YOLOv3[12]. In 2016, Wei Liu and co-authors introduced the SSD (Single Shot MultiBox Detector)[13], which obviates the necessity for candidate box generation and conducts classification and regression operations directly on objects. This method notably amplifies real-time detection speed. Bochkovski and colleagues introduced YOLOv4[14], a model that integrates a bottom-up pathway within the Feature Pyramid Network (FPN) and leverages the Path Aggregation Network (PANet) to enhance the efficacy of semantic information propaga-

This research was supported by the Guangxi Science and Technology Project (GuikeAD23026034), the Bagui Scholar Program Fund of Guangxi Zhuang Autonomous Region of China (2019A40), the National Natural Science Foundation of China (61936002).

Corresponding author: Xiaonan Luo.

tion within the network. Glenn and collaborators introduced YOLOv5, an extension of YOLOv4, that integrates additional training techniques. A prominent improvement is the inclusion of the Focus structure within the backbone network, thereby reducing the model's parameter count and enhancing its overall efficiency. These innovative enhancements, superimposed on the foundation of YOLOv4, yield substantial improvements in both speed and accuracy.

II. RELATED WORK

Traditionally, object detection algorithms have been predominantly developed for land-based targets. Nevertheless, with the surge in underwater exploration endeavors, an escalating number of researchers are transitioning object detection and classification methodologies to underwater domains. In 2016, Ravanbakhsh and co-authors [15] introduced a learning-based methodology utilizing HOG in conjunction with SVM for the purpose of coral reef fish detection. This study emphasized the potential of deep learning in the context of underwater object detection. In a subsequent study conducted by Li and co-authors in 2015 [16], the utilization of Fast R-CNN was applied for the expedited detection and recognition of fish species. In a parallel effort, Faster R-CNN was employed specifically for fish detection [17]. To mitigate the challenge of a restricted pool of training images, Zeng and collaborators [1] incorporated the AON into the conventional Faster R-CNN detection algorithm. This strategic integration resulted in an expansion of the available training samples and a marked improvement in the model's resistance to occlusion-related challenges. In 2020, Long Chen and colleagues [18] proposed the Sample-Weighted Inter-Play Network (SWIPENET) to tackle image blurriness in underwater settings characterized by high levels of noise and distortion. Lin and co-authors [19] tackled the complexities associated with overlapping, occlusion, and blurriness in underwater object detection by conducting a comprehensive study involving simulated scenarios, leading to the development of a versatile model. In 2021, Wei and collaborators [20] presented a real-time and precise design tailored for underwater robots. They utilized a neural network that relies on Local Wavelet Acoustic Patterns and a MLP to address the challenges posed by the heterogeneity and complexities of passive target classification in underwater settings. However, the demanding underwater conditions present persistent challenges for practical underwater object detection algorithms. These challenges encompass issues such as suboptimal image quality, limited visibility, diminished contrast, texture deformations, and variations in color within underwater images. These factors collectively pose substantial hindrances to effective underwater object detection. While numerous object detection methods have made significant advancements, there is still a substantial gap in addressing these adversities effectively. In real-world applications, the deployment of models frequently occurs on mobile platforms like underwater robots, requiring models to possess qualities of robustness, durability, and portability. Within the realm of state-of-the-art YOLO series models, YOLOv5 stands out as

particularly well-suited for industrial applications owing to its compact dimensions and high level of encapsulation. Hence, this study introduces an enhanced version of YOLOv5 and employs it for the analysis of underwater images, thereby making the subsequent contributions:

(1) Incorporating the Swin Transformer as the fundamental backbone network for YOLOv5 augments the process of feature extraction in underwater images that involve blurry objects, resulting in the extraction of more relevant features and the emphasis of foreground targets.

(2) We enhance the process of feature extraction for the specific target by utilizing multi-scale feature fusion, integrating cross-scale connections and context information weighting operations. This leads to the creation of a unique feature pyramid composed of the consolidated feature maps.

(3) We enhance the Non-Maximum Suppression (NMS) algorithm by incorporating soft NMS, which results in an improved recognition capability of the network when dealing with occluded and overlapping underwater targets.

(4) Our enhanced network has been subjected to rigorous experimentation using a dataset containing over 5000 annotated images. The results demonstrate an impressive detection accuracy of 87.3% (mean Average Precision or mAP), which not only surpasses the baseline performance but also outperforms other conventional object detection models.

III. METHODS

YOLO, a widely recognized single-stage object detection algorithm, effectively reformulates the object detection challenge as a regression problem. It directly forecasts bounding box coordinates and class probabilities for each object, thus eliminating the necessity for Region of Interest (RoI) extraction. This innovative approach significantly enhances the speed of object detection compared to traditional methods such as Faster R-CNN. In 2020, Ultralytics introduced YOLOv5, the fifth iteration of YOLO, distinguished by its superior balance between detection speed and accuracy. YOLOv5 provides adaptability through adjustable width and depth parameters in its backbone network, yielding four distinct model versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The YOLOv5 architecture, as depicted in Figure 1, encompasses four principal components: the input module, the feature extraction backbone network, the cross-scale feature fusion neck network, and the object detection prediction network.

The Input Module within YOLOv5 is tasked with data ingestion and initial image preprocessing. It dynamically resizes input images to predetermined dimensions and integrates several data augmentation techniques. These techniques encompass mosaic data augmentation, random cropping, random scaling, and random shuffling, collectively contributing to dataset enrichment and the reinforcement of the model's resilience. In the YOLO series, meticulous attention is given to the establishment of initial anchor frame dimensions, thoughtfully tailored to improve alignment with objects present in diverse datasets. Throughout the network training process, the model generates predicted frames based on these initial anchor

frames, subjecting them to a subsequent comparison against the ground truth. This underscores the critical significance of the initial anchor framework in the YOLO series algorithms.

The Backbone Network within YOLOv5 is dedicated to the extraction of image features, implementing several strategies to optimize this critical task. First and foremost, it leverages a focus structure designed to periodically capture pixels from high-resolution images and subsequently reassemble them into lower-resolution images. This technique entails the amalgamation of information from four adjacent positions within the image, thus condensing the information spanning the WH dimensions into the C channel space. The fundamental objective here is to enhance the receptive field at each individual point while minimizing the loss of original information. The overarching design principle behind this strategy primarily revolves around the reduction of computational overhead, ultimately resulting in expedited processing. Inspired by the fundamental design principles found in CSPNet, YOLOv5 introduces the CSP1_x and CSP2_x modules into its architecture. These modules perform a division of the feature maps originating from the base layer into two separate segments. Subsequently, they combine these segments through a cross-stage hierarchy. This strategic integration achieves a dual purpose: it not only reduces computational overhead but also ensures the preservation of accuracy in the process. The concluding segment of the backbone network encompasses the Spatial Pyramid Pooling (SPP) network. This network plays a pivotal role in further expanding the receptive field while facilitating the separation of contextual features.

The Neck Network within YOLOv5 is meticulously crafted to augment feature extraction capabilities. Its structural design takes inspiration from PANet, ingeniously amalgamating key attributes from the FPN and PAN modules. It commences with the utilization of the FPN structure to expedite the transmission of robust semantic features from the upper layers to the lower ones. Subsequently, it harnesses the feature pyramid structure, meticulously constructed by the PAN modules, to facilitate the seamless transfer of potent localization features from the lower layers to the upper ones. This well-thought-out strategy enables the amalgamation of features spanning diverse layers, thereby enriching the collective feature extraction prowess.

The Prediction Network in YOLOv5 preserves the structural aspects from YOLOv3's head, encompassing three branches dedicated to object detection. These branches are tasked with delivering predictions pertaining to object coordinates, class labels, and confidence scores. A substantial improvement in this segment involves the adoption of the Complete Intersection over Union (CIoU) loss as the bounding box regression loss.

IV. PROPOSED MODEL

This section delineates the methodology for underwater object detection, leveraging an improved iteration of YOLOv5. Illustrated in Figure 1, the workflow commences with dataset preparation, involving activities such as data cleansing and annotation. Subsequently, we harness the upgraded YOLOv5

network to enhance the precision of object detection. To elaborate, we introduce a novel backbone network based on the Swin Transformer for YOLOv5, as expounded in Section 4.1. We further implement a more efficient multi-scale feature fusion method, as delineated in Section 4.2. Additionally, we augment the Non-Maximum Suppression (NMS) algorithm with soft NMS[28], as comprehensively discussed in Section 4.3.

1) *Backbone Network Utilizing Swin Transformer:* The process of capturing underwater images is influenced by the restricted light transmission through water, which leads to diminished object detection clarity and poses challenges in discerning objects displayed on the monitor. Hence, it is imperative to prioritize the enhancement of identified object characteristics while concurrently minimizing the prominence of background features throughout the detection procedure. One efficacious method to achieve this is through the application of self-attention mechanisms. Within the domain of natural language processing, the Transformer model [23] has demonstrated its efficacy by employing multi-head self-attention in lieu of the conventional recursive layers commonly found in encoder-decoder architectures. The Vision Transformer [24] extended the usage of the Transformer architecture to the domain of visual data. However, its direct translation to computer vision faces two significant challenges:

(1) *Variation in Feature Scales:* In contrast to natural language processing, computer vision deals with a wide range of diverse and variable feature scales. While feature scales are usually standardized and consistent in NLP, they can exhibit significant variations in computer vision.

(2) *Computational Challenges:* Computer vision tasks frequently demand higher image resolutions compared to NLP. The computational complexity associated with the direct application of the Transformer in computer vision grows exponentially with the increase in image resolution. This can result in significant computational resource requirements. Additionally, the constrained computational resources typically accessible for underwater object detection render the direct use of the Transformer impractical in this scenario.

The Swin Transformer [25] represents a highly proficient method introducing self-attention to the realm of computer vision. Diverging from previous methodologies, it presents notable improvements, including: (1) It utilizes hierarchical structures, a characteristic often found in Convolutional Neural Networks (CNNs), to construct a layered Transformer; (2) It incorporates locality by executing self-attention calculations within non-overlapping windows; (3) It introduces a shift-based window partitioning technique that enhances connections between self-attention modules based on windows. These innovations establish a linear correlation between computational complexity and input image size. With the network's depth augmentation, image patches progressively combine to establish a versatile visual backbone network, featuring a hierarchical transformer.

The network structure in Figure 2 illustrates the replacement of the YOLOv5 backbone with the Swin Transformer. The

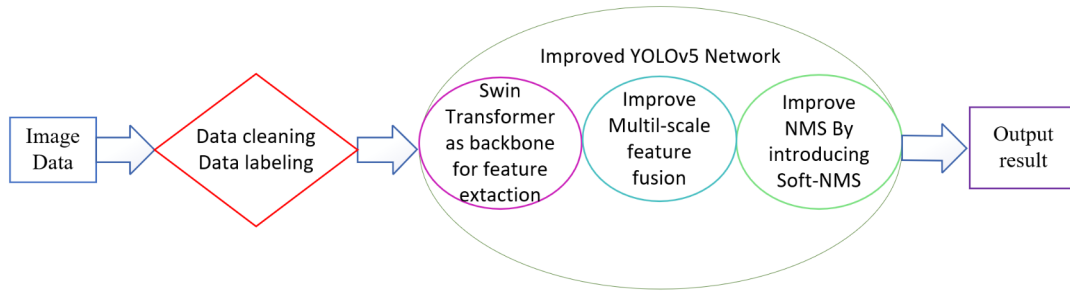


Fig. 1. The enhanced version of YOLOv5 is employed in the task of underwater object detection.

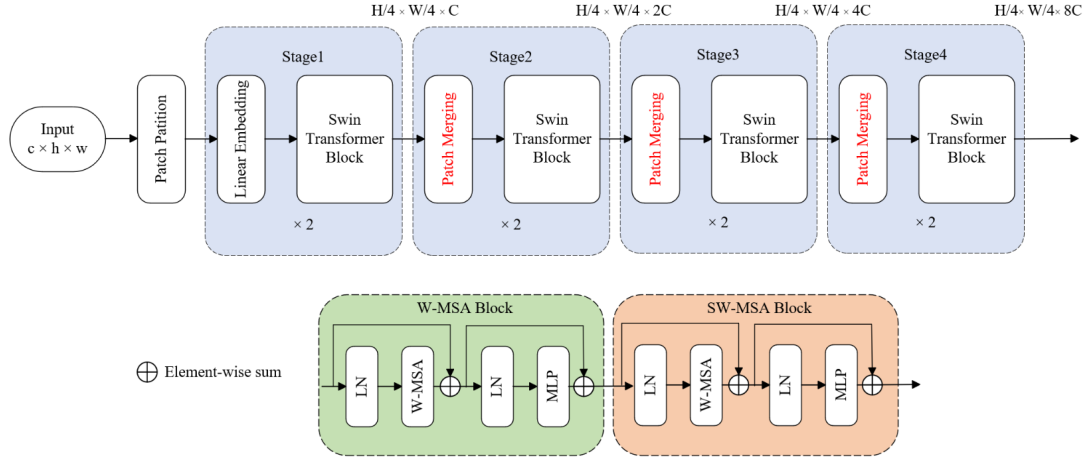


Fig. 2. The architectural framework of the Swin Transformer[25]

patch embedding comprises two crucial elements: The process of patch partitioning entails the division of the feature map into discrete, non-overlapping patches, while linear embedding is responsible for the conversion of these patches into a dimension of choice. The Swin Transformer block comprises both W-MSA and SW-MSA components. W-MSA plays a crucial role in alleviating computational complexity through feature map segmentation. In contrast, SW-MSA plays a pivotal role in enabling the exchange of information between distinct windows. The subsequent step of patch merging leads to a downscaling of the input feature map. The process initiates with the ingress of the initial feature map, characterized by dimensions $c \times h \times w$, into the patch embedding module. The feature map is subdivided into non-overlapping smaller blocks, resulting in a $96 \times (h \div 4) \times (w \div 4)$ feature map. Subsequently, this reduced feature map is processed through two consecutive Swin Transformer block, ultimately resulting in a feature map with dimensions of $96 \times (h \div 4) \times (w \div 4)$. Following this, feature maps P3, P4, and P5 are produced by utilizing three patch merging layers and Swin Transformer blocks. These resultant P3, P4, and P5 feature maps are subsequently employed as input feature maps for the neck module within the Feature Pyramid Network (FPN)[26]. The architectural depiction of YOLOv5 incorporating the Swin Transformer as the backbone network is presented in Figure 3.

2) *Enhancement of Cross-Scale Feature Fusion:* Considering the pyramid model's attributes, the refined procedure entails top-down processing (indicated by the green arrow), cross-scale connectivity (highlighted by the blue arrow), context information weighting (represented by the exchange between N3 and N4 within the red area, and the input flow weighting N5), and bottom-up processing (depicted by the orange arrow), as illustrated in Figure 4. Specific steps are as follows:

Upon input image size adjustment to 640×640 , successive convolution operations are applied to extract distinct-scale feature maps denoted as C1, C2, C3, C4, and C5. Subsequently, feature maps C3, C4, and C5 from the final three layers of the backbone network are integrated into a two-way feature weighted fusion network. The top-down process can be described as follows: The 1×1 feature map C5 is subjected to a convolution operation to adjust its channel dimension, resulting in the creation of feature map P5. Subsequently, P5 is subjected to dual sampling and combined with feature map C4. A local network is employed for cross-stage feature extraction, leading to the formation of feature map P4. After two additional sampling steps, P4 is fused with feature map C3, and once more, a local network across stages is used for feature extraction, yielding feature map P3.

The selection of feature maps involves choosing M4 for

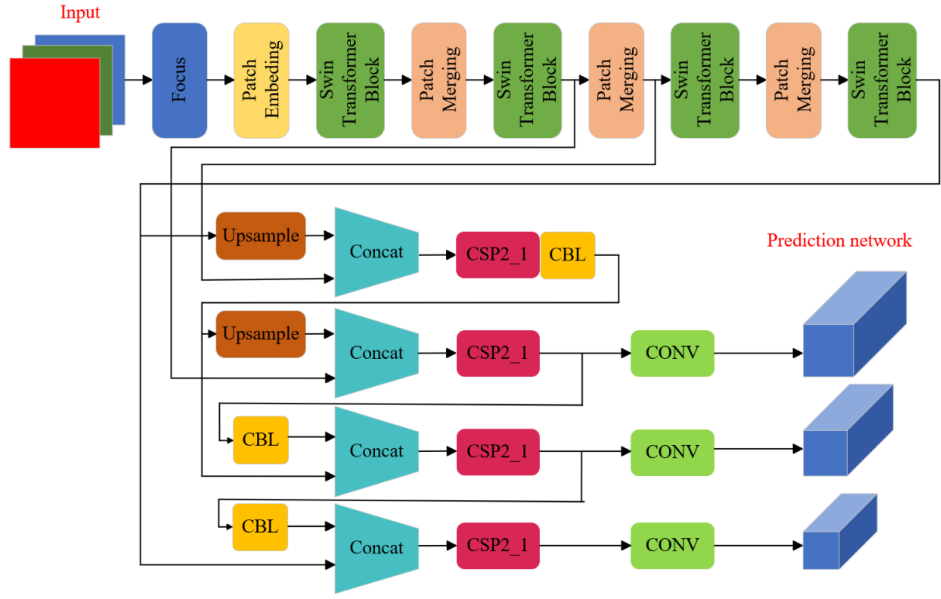


Fig. 3. The configuration of YOLOv5 employing the Swin Transformer as its foundational backbone network.

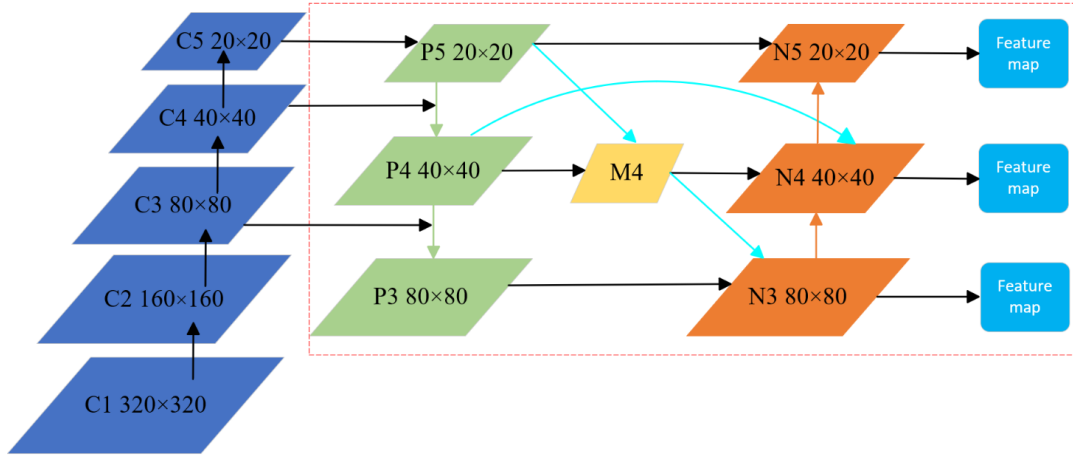


Fig. 4. Weighted bidirectional feature fusion[29]

feature map P4 and P5, achieved through a cross-scale 3 x 3 convolution operation. It's important to note the input flow relations as follows: N3 receives input from feature map P3 and M4, N4 interchanges input flow with feature maps P4, M4, and N3, while N5 receives input from feature map P5 and N4. To facilitate the interchange between N3 and N4, a rapid normalization fusion method [27] is employed. This enables the distribution of weighted information for N5's input stream, effectively learning the importance of different features. By combining both the top-down and bottom-up processes, the interaction between information outputs is integrated, resulting in the generation of the final updated feature map.

Rapid normalization of fusion method [27] training speed, high efficiency, its computation formula is shown below.

$$O = \sum_i \frac{\omega_i \cdot X_i}{\varepsilon + \sum_j \omega_j} \quad (1)$$

In the given context, O represents the output feature map, ω_i denotes the weight coefficients of the respective layer's feature map, and X_i represents the feature map that requires fusion. $\varepsilon \leq 0.001$.

To 3 characteristic figure of C4, for example, formula (2) and formula (3) describes the characteristics of adjacent two layers of figure in the fourth layer of fusion situation:

$$M_4 = \text{Conv} \left(\frac{\omega_1 \cdot P_4^{\text{in}} + \omega_2 \cdot \text{Resize}(P_5^{\text{in}})}{\omega_1 + \omega_2 + \varepsilon} \right) \quad (2)$$

$$N_4 = \text{Conv} \left(\frac{\omega'_1 \cdot P_4^{\text{in}} + \omega'_2 \cdot P_4^{\text{td}} + \omega'_3 \cdot \text{Resize}(P_3^{\text{out}})}{\omega_1 + \omega_2 + \omega_3 + \varepsilon} \right) \quad (3)$$

M4 represents the intermediary feature within the top-down fusion's fourth layer, whereas N4 signifies the intermediary feature within the fourth layer of the bottom-up fusion. The weights for different hierarchical feature maps are denoted as ω_i and ω'_i . P_i^{in} and P_i^{out} represent the input and output of the i -th layer feature map in the bottom-up fusion process, respectively. P_i^{td} represents the output of the intermediate node in the top-down fusion process for the i -th layer feature map. Resize refers to the operation of upsampling or downsampling the input feature map to maintain consistent dimensions. Conv typically represents a convolutional operation used for feature processing. ε is a parameter, set to 0.0001 in this case.

The bottom-up process is as follows: The N3 feature layer is downsampled by a 3×3 convolution with a downsampling factor of 2, then stacked with M4. The fused feature map is subjected to a 3×3 convolution operation to eliminate aliasing effects. Simultaneously, the N4 feature layer undergoes feature extraction using a cross-stage local network. This process results in a feature layer of size $40 \times 40 \times 512$. Similarly, the N4 feature layer is downsampled by a 3×3 convolution with a downsampling factor of 2, then fused with P5 to eliminate aliasing effects. The N5 feature layer undergoes feature extraction using a cross-stage local network. The resulting feature layer has dimensions of $20 \times 20 \times 1024$.

3) *Improvement of NMS by introducing Soft-NMS.*: Non-Maximum Suppression (NMS) stands as a widely employed method in cutting-edge detectors to enhance results, effectively mitigating the incidence of false positives. Figure 5 outlines the flowchart of the NMS algorithm. The process initiates by organizing all bounding boxes in accordance with their individual confidence scores. The bounding box with the highest score, denoted as B_M , is subsequently incorporated into the final detection list D, while the remaining bounding boxes are assigned distinct labels represented as B_i . Subsequently, any predicted box B_i exhibiting an overlap area exceeding a predefined threshold N_t with B_M is excluded. This procedure is repeated for the remaining B_i bounding boxes until the original list is exhausted. However, in intricate underwater scenarios, NMS may lead to overlooked detections. To combat this issue, we introduce Soft-NMS[28] as an alternative to NMS for refining detection outcomes. The NMS algorithm can be formally represented utilizing a scoring function, as delineated in Equation (4):

$$S_i = \begin{cases} S_i, & \text{IoU}(B_M, B_i) < N_t \\ 0, & \text{IoU}(B_M, B_i) \geq N_t \end{cases} \quad (4)$$

Within the NMS procedure, S_i symbolizes the score assigned to the predicted box B_i , B_M signifies the highest-scoring predicted box, and N_t denotes the overlap threshold. NMS employs a stringent threshold to decide which boxes proximate to B_M should be preserved and which ones should

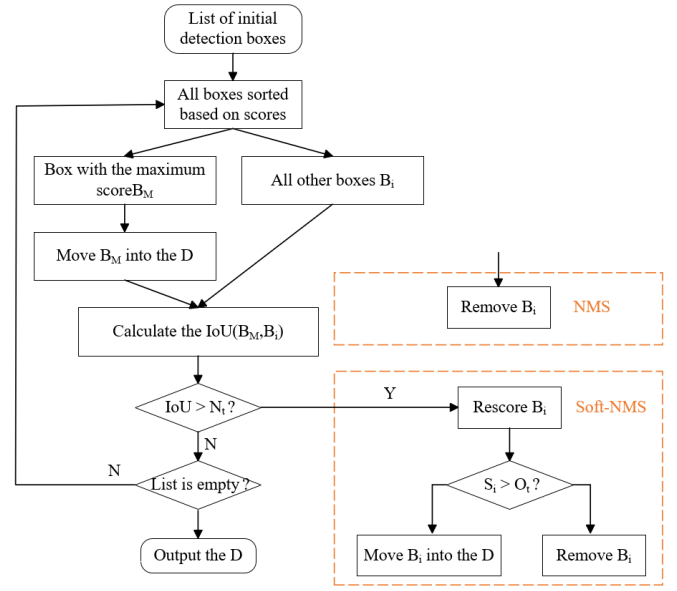


Fig. 5. The algorithmic sequence of NMS and Soft-NMS[28].

be suppressed. If an actual object exists but displays an overlap exceeding N_t with B_M , its detection is dismissed. The underlying principle of Soft-NMS entails applying a penalty function to decrease the scores of prediction boxes overlapping with B_M , as opposed to an outright score reduction to zero. The Soft-NMS scoring function, denoted as S_i , is defined in the subsequent Equation (5):

$$S_i = \begin{cases} S_i, & \text{IoU}(B_M, B_i) < N_t \\ S_i e^{-\frac{\text{IoU}(B_M, B_i)^2}{\sigma}}, & \text{IoU}(B_M, B_i) \geq N_t \end{cases} \quad (5)$$

The Gaussian penalty function, mathematically expressed as $e^{-\frac{\text{IoU}(B_M, B_i)^2}{\sigma}}$, constitutes a critical element of Soft-NMS. The hyperparameter σ is typically determined empirically. This function serves to significantly diminish the scores of prediction boxes with substantial overlap with B_M , while those boxes that are spatially distant from B_M remain largely unaffected. If the scores of these penalized boxes still surpass the confidence threshold O_t , they are retained rather than eliminated.

V. EXPERIMENTS

In this section, we validate the augmented efficacy of the enhanced YOLOv5 algorithm in underwater detection scenarios. The experimental findings corroborate that this algorithm enhances the precision of underwater target detection, rendering it exceptionally suitable for challenging underwater conditions.

A. Data Set

The experimental dataset employed in this research was sourced from the Object Recognition category of the Underwater Robot Professional Contest (URPC) in China. This dataset

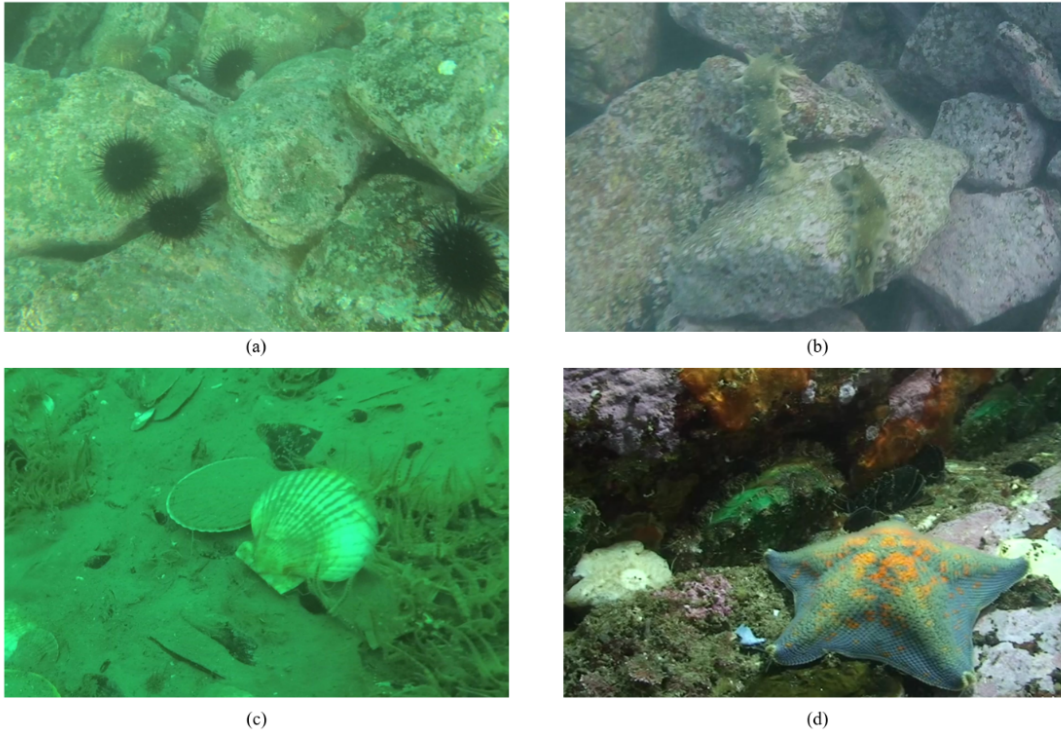


Fig. 6. The dataset comprises four distinct biological categories, Holothurian(a), Echinus(b), Scallop(c), and Starfish(d).

includes underwater images featuring four different marine products: "holothurian," "echinus," "scallop," and "starfish" (as shown in Figure 6). To ensure data quality, we conducted data cleaning, removing the seagrass category and its associated images. This resulted in a total of 5543 valid images, which were subsequently annotated for object detection. Each sample image underwent processing and was subsequently organized and stored in accordance with the format utilized for the PASAL VOC2007 dataset. Figure 7 offers comprehensive insights into the dataset. In Figure 7a, you can find an overview of the object distribution across categories. Notably, the "echinus" category exhibits the highest frequency, followed by "scallops," "starfish," and "holothurian." It's important to observe that the majority of objects within the dataset are characterized by relatively diminutive sizes. For the sake of ensuring uniform data distribution, the dataset was subject to random partitioning into two subsets: a training set and a test set, adhering to a 7:3 ratio. The training set encompasses 3880 images, whereas the test set encompasses 1663 images.

B. Model Evaluation Metrics

In the context of object detection, we assess predictions by considering four possible outcomes characterized as TP, FP, TN, and FN. To classify a detected bounding box, it is designated as a TP if its IoU with a ground truth box exceeds a predefined threshold (typically set to 0.5 in our experiments). Conversely, it is labeled as a false positive (FP) if it does not meet this criterion. In the absence of any detected box aligning with a true box, it is categorized as a

false negative (FN). In this context, TP corresponds to the quantity of accurately recognized objects, FP denotes the tally of incorrectly identified objects, and FN signifies the count of overlooked objects. The assessment of model performance usually involves the computation of Pr and Re, as determined by equations (6) and (7) below, respectively.

$$Pr = \frac{TP}{FP + TP} \quad (6)$$

$$Re = \frac{TP}{FN + TP} \quad (7)$$

Precision (Pr) and recall (Re) are interconnected metrics. An increase in recall while maintaining high precision suggests improved model performance. In contrast, a model with subpar performance might trade off a substantial portion of precision for enhanced recall. To amalgamate these two metrics, we introduce the concept of average precision (AP) as a measure of detection accuracy, defined by equation (8).

$$AP = \int_0^1 Rr(Re) dRe \quad (8)$$

The AP value corresponds to the area under the precision-recall curve, with a higher AP indicating a higher level of network accuracy. In multi-class object detection tasks, the model's detection accuracy is evaluated by computing the average precision for all classes. This metric is denoted as mean average precision (mAP) and is defined in equation (9).

$$mAP = \frac{1}{C} \sum_{c=1}^C AP(c) \quad (9)$$

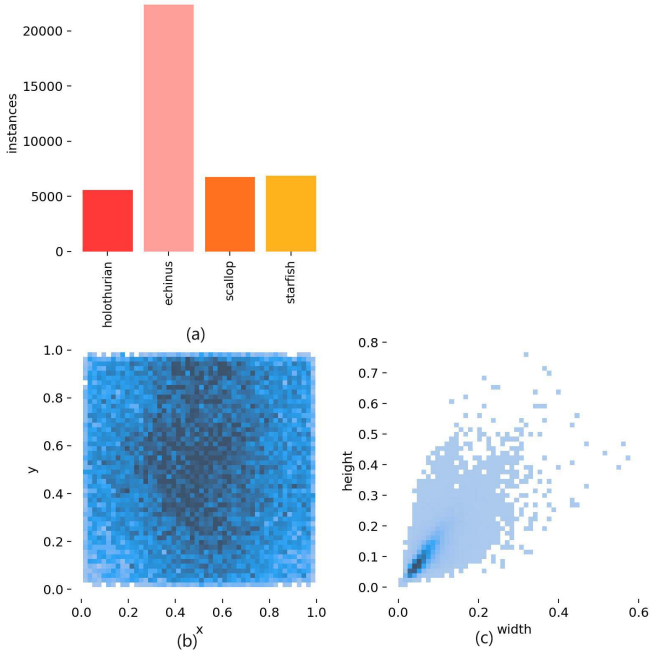


Fig. 7. Dataset statistics are depicted in the following manner: (a) a bar chart illustrating the object count distribution across different classes; (b) a normalized map illustrating the spatial distribution of objects; (c) a normalized map representing the size distribution of objects.

Where C represents the number of target classes.

C. Experimental Settings

The experiments were conducted on a system equipped with an AMD EPYC 7551P 32-Core Processor with 32GB of RAM, accompanied by an NVIDIA GeForce RTX A4000 graphics processing unit with 16GB of RAM. The software environment included CUDA 11.8, CUDNN 8.6, and Python 3.9. Model optimization utilized the stochastic gradient descent (SGD) method. Detailed network training hyperparameters are provided in TABLE I, with a train duration of 300, a batch size of 12, an initial learning rate of 0.011, a weight decay of 0.00051, and an SGD momentum of 0.9.

TABLE I
NETWORK TRAINING HYPERPARAMETERS.

Train Epochs	Batch Size	Learning Rate	Weight Decy	Momentum
300	12	0.011	0.00051	0.9

D. Experimental Results

YOLOv5 is categorized into four distinct models, specifically YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, with distinctions in terms of depth and width. TABLE II provides the parameter configurations, demonstrating a linear relationship between the number of model parameters and size with increasing depth and width multipliers. Within this set of models, YOLOv5s stands out as the most lightweight, featuring

the fewest parameters, rendering it particularly suitable for real-world applications. In contrast, YOLOv5x is the most resource-intensive, boasting the highest number of parameters and the largest dimensions, which results in increased computational demands during training. Our experiments encompassed a range of models, as indicated in the table, demonstrating that heightened depth and width lead to improved average precision (mAP). The transition from the smallest to the largest models yields a 0.8% mAP enhancement, albeit at the expense of significant computational resources. In light of these resource considerations, we opted for YOLOv5s as our baseline model for subsequent enhancements and experiments.

TABLE II
RESULTS FOR VARIOUS YOLOV5 MODELS.

Model	Depth	Width	Parameters	Storage size	mAP
YOLOv5s	0.33	0.50	7.2×10^6	14.1	84.9
YOLOv5m	0.67	0.75	21.2×10^6	40.5	85.2
YOLOv5l	1.0	1.0	46.5×10^6	89.4	85.6
YOLOv5x	1.33	1.33	8607×10^6	167.0	85.7

Figure 8 presents the precision-recall curve for the enhanced YOLOv5s, showcasing its robust detection capabilities across various target categories. Notably, the model achieves a substantial AP of 90.5% for sea urchins, contributing to a mAP of 87.3%. Ablation experiments were carried out to evaluate

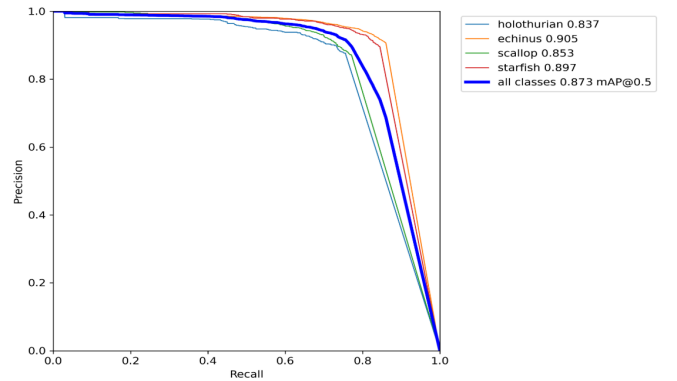


Fig. 8. The precision-recall performance graph for the enhanced YOLOv5s.

the influence of different enhancements on the performance of our model. In order to tailor the model to our underwater dataset, we initially established anchor frames based on domain-specific knowledge, as the conventional YOLOv5 anchor frames derived from the COCO dataset were found to be unsuitable. Furthermore, we introduced a customized backbone network founded on Swin Transformers to enhance feature extraction and made adjustments to the PANET multi-scale fusion network. Additionally, improvements were applied to the confidence loss function. The experimental results

TABLE III
EXPERIMENTAL OUTCOMES FROM VARIOUS ALGORITHMS.

Method	Backbone Network	mAP(%)	Holothurian AP(%)	Echinus AP(%)	Scallop AP(%)	Starfish AP(%)
SSD[13]	VGG-16	60.9	59.5	73.8	41.1	69.3
Faster-RCNN[6]	VGG-16	73.2	73.0	74.0	72.5	73.5
R-FCN[7]	VGG-16	77.5	75.5	80.0	76.1	78.5
YOLOv4[14]	Darknet53	82.2	71.8	89.6	82.3	85.2
YOLOv5s	CSPDarknet53	84.9	74.9	91.0	85.1	88.4
YOLOv5x	CSPDarknet53	85.7	76.5	91.4	86.1	88.6
ours	SwinTransformer	87.3	83.7	90.5	85.3	89.7

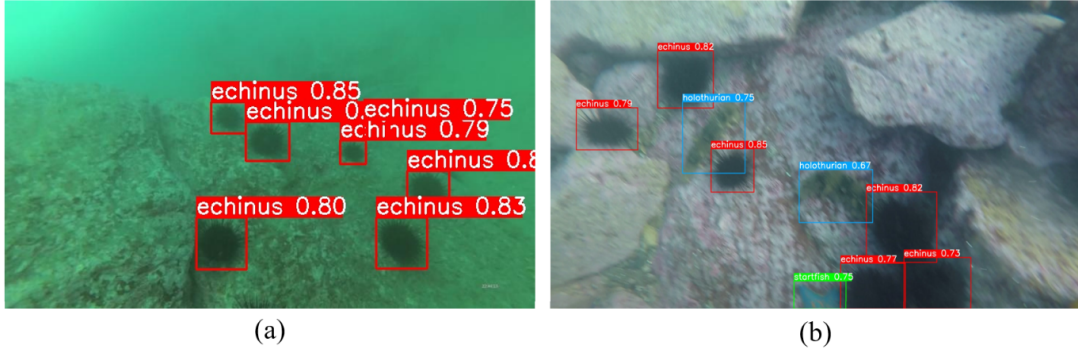


Fig. 9. Results for object detection under various distances are presented in Figure 9, with distinct colored squares denoting different target categories. In single-class object detection (Figure 9a), the red squares represent "echinus," the green squares represent "starfish," and the blue squares represent "holothurian." In multi-class object detection (Figure 9b), the model successfully identifies and distinguishes between these different target categories, demonstrating its capability to handle diverse underwater scenarios.

are presented in TABLE IV. Notably, among the various enhancements, the adoption of the Swin Transformer as the backbone network for enhanced feature extraction proved to be the most significant, yielding a noteworthy 1.5% improvement in the model's mAP. The preliminary definition of anchor frames before the experiments resulted in a 0.3% enhancement in the network's mAP. Furthermore, optimizations in the multi-scale feature fusion methodology and the confidence loss function contributed to additional increases in mAP, with gains of 0.2% and 0.4%, respectively. To validate the

in TABLE III, highlight that the enhanced YOLOv5s achieved the highest Mean Average Precision (MAP). Significantly, the mAP of the enhanced YOLOv5s (87.3%) outperformed SSD by 26.1%, surpassed YOLOv4 by 5.0%, and even exceeded the standard YOLOv5s by 2.3%. Moreover, it demonstrated a 1.5% enhancement in performance compared to the largest model, YOLOv5x, which achieved a MAP of 85.7%. These outcomes undeniably emphasize the outstanding superiority of the proposed approach in the domain of underwater object detection.

VI. CONCLUSIONS

In the contemporary context, object detection algorithms designed for terrestrial environments often encounter difficulties when applied to intricate underwater scenarios. This paper presents an enhanced underwater object detection algorithm rooted in the YOLOv5 framework. The refined algorithm encompasses three pivotal stages. Initially, it integrates the Swin Transformer as the fundamental backbone network within the YOLOv5 architecture, thereby elevating the representation of target features. Subsequently, it refines the multi-scale feature fusion method to enhance the amalgamation of images captured at different resolutions. Additionally, through the augmentation of the non-maximum suppression (NMS) algorithm with Soft-NMS, the network exhibits enhanced recognition

TABLE IV
ABLATION EXPERIMENTS.

Pre-Set Anchor	Swin Transformer	Improved Multi-Scale Feature Fusion	Improved NMS	mAP(%)
				84.9
✓				85.2(+0.3)
✓	✓			86.7(+1.5)
✓	✓	✓		86.9(+0.2)
✓	✓	✓	✓	87.3(+0.4)

efficacy of the enhanced techniques implemented in YOLOv5, comparative experiments were conducted, involving alternative models such as YOLOv4 and SSD. The results, as presented

capabilities, particularly for obscured and overlapping underwater objects. We conduct experiments to compare detection results among various YOLOv5 models, perform ablation experiments to evaluate the effectiveness of the proposed improvement strategies, and compare these results with those obtained from other models. The experimental results clearly demonstrate that the aforementioned enhancements substantially improve the detection performance of the YOLOv5 in challenging underwater environments. The enhanced model outperforms conventional object detection models and shows increased robustness in intricate underwater scenarios.

Nevertheless, it is imperative to acknowledge that our experiments were conducted exclusively on a singular dataset, and the dataset itself may exhibit limitations in terms of both the quantity and diversity of underwater data. These limitations stem from the inherent challenges associated with the collection of comprehensive underwater datasets. Additionally, the less-than-optimal quality of underwater images, attributed to light diffusion in aquatic environments, underscores the importance of implementing underwater image enhancement to improve the image quality and diversity of such datasets. Furthermore, it is worth noting that the dimensions of our model are relatively substantial. In our future work, we plan to procure datasets that encompass a more extensive spectrum of underwater object detection scenarios. Additionally, we aim to integrate advanced image enhancement techniques into our models to ameliorate data quality. Furthermore, we will prioritize the development of lightweight networks that can expedite inference speed while upholding detection accuracy.

REFERENCES

- [1] L. Zeng, B. Sun, D. Zhu, "Underwater target detection based on faster r-cnn and adversarial occlusion network," *Engineering Applications of Artificial Intelligence*, vol. 100, pp. 104190, 2021.
- [2] L. Chen, M. Zheng, S. Duan, "Underwater target recognition based on improved yolov4 neural network," *Electronics*, vol. 10, no. 14, pp. 1634, 2021.
- [3] M. Czub, L. Kotwicki, T. Lang, "Deep sea habitats in the chemical warfare dumping areas of the baltic sea," *Science of the Total Environment*, vol. 616, pp. 1485-1497, 2018.
- [4] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, pp. 580-587, 2014.
- [5] R. Girshick, "Fast r-cnn," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1440-1448, 2015.
- [6] S. Ren, K. He, R. Girshick and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.
- [7] J. Dai, Y. Li, K. He, "R-fcn: object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.
- [8] K. He, G. Gkioxari, P. Dollár, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969, 2017.
- [9] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 6154-6162, 2018.
- [10] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: unified, real-time object detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788, 2016.
- [11] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 6517-6525, 2017.
- [12] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," *arXiv preprint arXiv*, vol. 1804, pp. 02767, 2018.
- [13] W. Liu, D. Anguelov, D. Erhan, "Ssd: single shot multibox detector," *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October pp.11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, pp. 21-37, 2016.
- [14] A. Bochkovskiy, C. -Y. Wang, H. -Y. M. Liao, "Yolov4: optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004*, pp. 10934, 2020.
- [15] S. Villon, M. Chaumont, G. Subsol, "Coral reef fish detection and recognition in underwater videos by supervised machine learning: comparison between deep learning and hog + svm methods," *International Conference on Advanced Concepts for Intelligent Vision Systems*. Cham: Springer International Publishing, pp. 160-171, 2016.
- [16] Xiu Li, Min Shang, H. Qin and Liansheng Chen, "Fast accurate fish detection and recognition of underwater images with fast r-cnn," *OCEANS 2015 - MTS/IEEE Washington*, Washington, DC, pp. 1-5, 2015.
- [17] X. Li, M. Shang, J. Hao and Z. Yang, "Accelerating fish detection and recognition by sharing cnns with objectness learning," *OCEANS 2016 - Shanghai, Shanghai, China*, pp. 1-5, 2016.
- [18] L. Chen, F. Zhou, S. Wang, "SWIPENET: Object detection in noisy underwater images," *arXiv preprint arXiv*, vol. 2010, pp. 10006, 2020.
- [19] W. -H. Lin, J. -X. Zhong, S. Liu, T. Li and G. Li, "Roimix: proposal-fusion among multiple images for underwater object detection," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 2588-2592, 2020.
- [20] W. Qiao, M. Khishe, S. Ravakhah, "Underwater targets classification using local wavelet acoustic pattern and multi-layer perceptron neural network optimized by modified whale optimization algorithm," *Ocean Engineering*, vol. 219, pp. 108415, 2021.
- [21] C. -Y. Wang, H. -Y. Mark Liao, Y. -H. Wu, P. -Y. Chen, J. -W. Hsieh and I. -H. Yeh, "Cspnet: a new backbone that can enhance learning capability of cnn," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, pp. 1571-1580, 2020.
- [22] K. Wang, J. H. Liew, Y. Zou, D. Zhou and J. Feng, "Panet: few-shot image semantic segmentation with prototype alignment," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp. 9196-9205, 2019.
- [23] A. Vaswani, N. Shazeer, N. Parmar, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, "An image is worth 16x16 words: transformers for image recognition at scale," *arXiv 2020. arXiv preprint arXiv*, vol. 2010, pp. 11929, 2010.
- [25] Z. Liu, "Swin transformer: hierarchical vision transformer using shifted windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, pp. 9992-10002, 2021.
- [26] T. -Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature pyramid networks for object detection," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 936-944, 2017.
- [27] M. Tan, R. Pang and Q. V. Le, "Efficientdet: scalable and efficient object detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 10778-10787, 2020.
- [28] N. Bodla, B. Singh, R. Chellappa and L. S. Davis, "Soft-nms — improving object detection with one line of code," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 5562-5570, 2017.
- [29] L. Pei, G. Cheng, X. Sun, Q. Li, M. Zhang and S. Miao, "Multi-scale bidirectional feature fusion for one-stage oriented object detection in aerial images," 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, pp. 2592-2595, 2021.