Enhancing Underwater Object Recognition Through the Synergy of Transformer and Feature Enhancement Techniques

Hoanh Nguyen*, Tuan Anh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

***Abstract*—Underwater object recognition presents a unique set of challenges due to the complex and dynamic characteristics of marine environments. This paper introduces a novel, multi-layered architecture that leverages the capabilities of Swin Transformer modules to process segmented image patches derived from aquatic scenes. A key component of our approach is the integration of the Feature Alignment Module (FAM), which is designed to address the complexities of underwater object recognition by enabling the model to selectively emphasize essential features. It combines multi-level features from various network stages, thereby enhancing the depth and scope of feature representation. Furthermore, this paper incorporates multiple detection heads, each embedded with the innovative ACmix module. This module offers an integrated fusion of convolution and self-attention mechanisms, refining detection precision. With the combined strengths of the Swin Transformer, FAM, and ACmix module, the proposed method achieves significant improvements in underwater object detection. To demonstrate the robustness and effectiveness of the proposed method, we conducted experiments on the UTDAC2020 dataset, highlighting its potential and contributions to the field.**

Keywords*—Underwater object recognition; swin transformer; self-attention; feature alignment

I. Introduction

Underwater object recognition is a specialized domain within computer vision and robotics that focuses on identifying and locating objects within aquatic environments. The complexities associated with this field are manifold, given the unique challenges posed by underwater conditions. These include limited visibility due to turbidity, light refraction and attenuation, and the dynamic nature of the aquatic medium with constantly moving particles and organisms. Detecting objects in such environments is crucial for a variety of applications, ranging from marine biology research, underwater archaeology, to defense and surveillance. Advanced techniques and algorithms in this area not only aim to improve the accuracy of detection but also enhance the real-time processing capabilities, making it possible for autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs) to perform intricate tasks with minimal human intervention. Traditional underwater object detection is usually based on handcrafted features of images for detecting objects [1], [2], [3]. In these conventional methods, detection was often based on basic image processing techniques. Specifically, thresholding, contour detection, and basic filter operations were commonly employed to differentiate objects from the surrounding environment. While these methods had their merits, especially in low-visibility conditions, they often struggled with false positives and lacked the precision needed for intricate tasks. Furthermore, these approaches were highly dependent on manual calibration and expert interpretation, making them labor-intensive.

In recent years, deep learning has revolutionized the field of object detection, ushering in a new era of accuracy and efficiency. These methods leverage complex neural network architectures, particularly Convolutional Neural Networks (CNNs), to automatically learn hierarchies of features from raw pixel data, eliminating the need for handcrafted feature extraction. Advanced architectures such as Faster R-CNN [4], YOLO [5], [6], SSD [7], R-FCN [8], Mask R-CNN [9] have emerged as frontrunners, offering real-time detection capabilities with impressive precision. These models have been applied in various vision applications such as depth estimation [10], intrusion detection [11], [12], vehicle license detection [13], and face mask detection [14]. With the success of deep learning-based object detection models, researchers have begun to apply deep learning to underwater object detection [15-24]. Although these methods have achieved certain successes, they encounter a number of issues. Firstly, all objects, regardless of their ambiguity, are subjected to the same supervisory signal. As a result, the classification scores obtained using simple cross-entropy loss don't accurately represent the ambiguity of the objects. This leads to misleadingly overconfident predictions. Secondly, these methods struggle with objects that are vague due to blurred boundaries or colors similar to their background. This similarity makes it challenging for the methods to distinguish such objects from their surroundings effectively.

Recognizing these limitations, our study aims to overcome these specific challenges. We propose an innovative approach for underwater object detection leveraging a multi-layered framework powered by the Swin Transformer. In the model, images pass through a patch partitioning process, segmenting them into smaller patches. These patches are then processed through several Swin Transformer layers. After each transformation, techniques like concatenation, upsampling, and convolution are utilized to enrich the feature maps. These enhanced feature maps pass through the FAM to amplify feature representation, ensuring precise object detection in complex underwater scenarios. Following the FAM, the framework integrates multiple detection heads, ensuring reliable localization of objects in underwater imagery. By

addressing the core issues of overconfidence in predictions and the struggle with vague object boundaries, our method seeks to provide a more robust and accurate solution for object detection.

The remainder of this paper is organized as follows: Section II provides an in-depth review of related work. Section III details our proposed methodology. In Section IV, we present a thorough analysis of our experiments. Finally, Section V concludes the paper with a summary of our findings and implications for future research.

II. Related Work

Underwater object detection, a crucial technology enabling AUVs to execute various tasks beneath the surface, has garnered significant interest globally among researchers. In [15], the authors introduced a method that utilizes a region proposal network from Faster R-CNN to enhance underwater object detection and recognition speed. This approach achieves quicker detection by employing convolutional networks to produce superior object candidates and integrating these networks with the primary detection systems. Chen et al. [16] proposed the Sample-WeIghted hyPEr Network (SWIPENET) and the Curriculum Multi-Class Adaboost (CMA) training paradigm to address challenges in underwater object detection, specifically blurry images with noise and small object detection. SWIPENET uses Hyper Feature Maps for enhanced resolution and detection of small objects, while its sample-weighted detection loss function emphasizes learning from high weight samples and disregarding low weight ones. Wei et al. [17] addressed challenges in underwater image target detection, particularly blur caused by water particles, by integrating squeeze and excitation modules into the YOLOv3 model after its deep convolution layers, enhancing semantic information. Zeng et al. [18] introduced the Faster R-CNN-AON network by integrating an adversarial occlusion network (AON) with the standard Faster R-CNN detection algorithm. The AON competes with the Faster R-CNN, teaching it to obscure targets, which in turn enhances the robustness of underwater seafood detection and prevents overfitting of the detection network.

In another approach, Lingyu et al. [19] adapted the YOLOv4 neural network for underwater target recognition by substituting its upsampling module with a deconvolution module and integrating depthwise separable convolution. Cao et al. [20] addressed underwater dynamic target tracking by developing a deep learning-based detection algorithm that uses the YOLO v3 network to identify targets in multibeam forward-looking sonar images and determine their positions. Huang et al. [21] introduced three specialized data augmentation techniques to address the scarcity of labeled samples in underwater environments: the inverse process of underwater image restoration for creating varied marine turbulence scenarios, perspective transformation to simulate different camera viewpoints, and illumination synthesis for replicating uneven lighting conditions underwater. In study [22], an innovative underwater salient object detection method that integrates both 2D and 3D visual features was introduced. This approach combines color and intensity (2D features) with 3D depth features, enhanced by a region-specific method that separately extracts these features in artificial and natural light regions, leading to more comprehensive and accurate detection results in three-dimensional underwater environments. Lin et al. [23] focused on augmentation policies designed to simulate overlapping, occluded, and blurred objects, constructing a model that achieves enhanced generalization. They introduce RoIMix, an innovative augmentation method that blends proposals from multiple images, unlike previous methods that operate on single images, thereby creating more complex and varied training data to improve model performance. Recently, Song et al. [24] introduced a two-stage underwater detector called boosting R-CNN, which features a novel region proposal network named RetinaRPN for high-quality proposals and models object prior probability through objectness and IoU prediction.

III. Proposed Model

A. Overview Pipeline

Fig. 1 illustrates the overall structure of our method for underwater object detection. The proposed method employs a multi-layered architecture that exploits the power of Swin Transformer modules. The input underwater image is first processed through a patch partition module, which segments the image into manageable patches. These patches are then sequentially passed through four layers of Swin Transformer modules. Specifically, Layers 1 and 2 involve two repetitions of the Swin Transformer module, Layer 3 contains six repetitions, while Layer 4 processes the patches twice through the Swin Transformer module. After the transformation process in each layer, specific operations including concatenation, upsampling, and convolution are performed to refine the feature maps. These refined feature maps are then passed through the Feature Alignment Module (FAM) to further enhance the feature representation, ensuring accurate object detection in the complex underwater environment. Following the FAM, the architecture incorporates multiple detection heads, which are responsible for the final object detection, ensuring robust identification of objects present in the underwater image. The details of each module are explained in the following subsections.

B. Swin-Transformer Backbone

- Transformers, initially introduced by Vaswani et al. in 2017 [25], are a type of neural network architecture primarily designed for handling sequence-to-sequence tasks in the field of natural language processing (NLP). They make use of attention mechanisms, notably self-attention, to weigh the significance of different parts of the input data. While Transformers have achieved remarkable success in NLP, their direct application to the vision domain presents challenges. One major reason is that unlike textual data which is inherently sequential, images are spatially structured with local patterns and hierarchies. Processing an image as a flat sequence of pixels loses this spatial coherence. Additionally, due to the high-dimensionality of images, Transformers can be computationally expensive and memory-intensive. To address these challenges, various adaptations, such as Vision Transformers (ViTs) [26] which divide images into fixed-size patches and then

linearly embed them, have been proposed to better suit the unique characteristics of visual data. However, they sometimes lack fine-grained local feature extraction. Recently, Swin Transformer [27] has emerged as a novel and powerful architecture that brings together the strengths of both classic CNNs and Transformers, and in some contexts, it has outperformed both. While CNNs have traditionally been strong at capturing local features through their hierarchical design of convolutional layers, they often struggle with long-range dependencies and global context. Swin Transformer tackles the issues of both CNNs and Transformers by hierarchically partitioning the image into non-overlapping windows and applying shifted windows across layers. This approach allows it to capture both local features within each window and global context across the entire image. The combination of local window-based processing with the global contextual understanding provided by the Transformer structure makes Swin Transformer particularly effective at feature extraction, offering advantages over traditional CNNs and basic Transformers. Underwater images often exhibit a range of complexities, including varying light conditions, attenuation, backscatter, and color distortions. Traditional architectures, like CNNs, can sometimes struggle with these irregularities, especially when it comes to recognizing objects that may be obscured or distorted due to water turbidity. Swin Transformer, with its hierarchical partitioning and shifted windows, can capture both local details and global contexts effectively. The local window-based processing ensures fine-grained feature extraction, which is crucial for identifying subtle characteristics of underwater objects. Meanwhile, the global contextual understanding inherent in the Transformer structure helps in identifying objects even when they are partially obscured or when the surrounding environment is cluttered. Additionally, the self-attention mechanism of the Swin Transformer can focus on long-range dependencies, which is beneficial for analyzing the spatial relationships between various underwater elements. Based on the features analyzed above, we chose Swin Transformer as the backbone network to perform feature extraction.

- The architecture of Swin Transformer is depicted in Fig. 2. It takes an image of dimensions $w \times h \times c$ (where c denotes channels, h is height, and w is width) and partitions it into non-overlapping patches. These patches pass through a linear embedding transformation and patch merging operations, converting them into token representations suitable for processing by transformer blocks. As the image progresses through the layers of the architecture (Layer 1 to Layer 4), the spatial resolution decreases, while the embedding dimensions increase. Specifically, the resolutions get adjusted by a downsampling factor ($w/4 \times h/4$) to ($w/32 \times h/32$) from Layer 1 to Layer 4. Each stage contains a specific number of Swin Transformer blocks, denoted by multipliers (i.e., ×2, ×2, ×6, ×2).

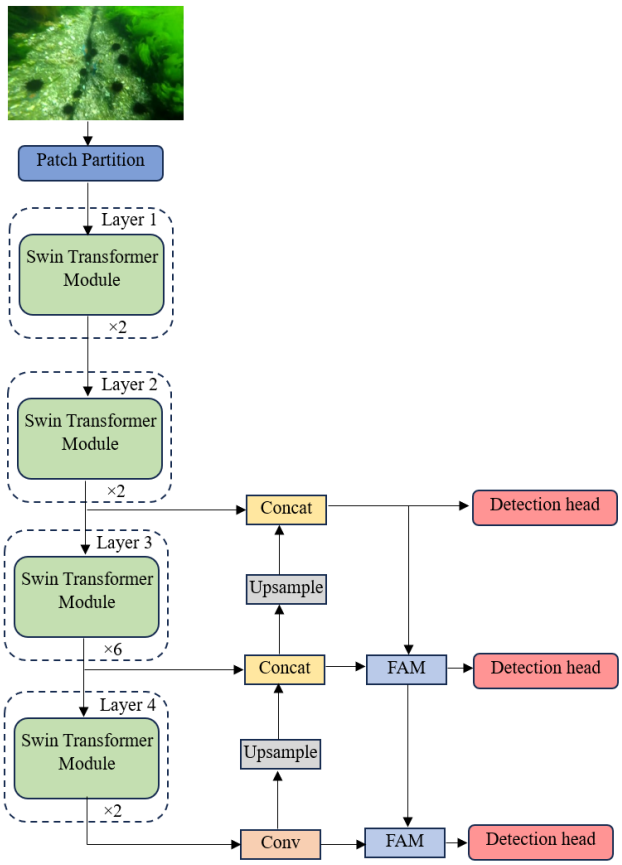

Fig. 1. Overview of the proposed model.

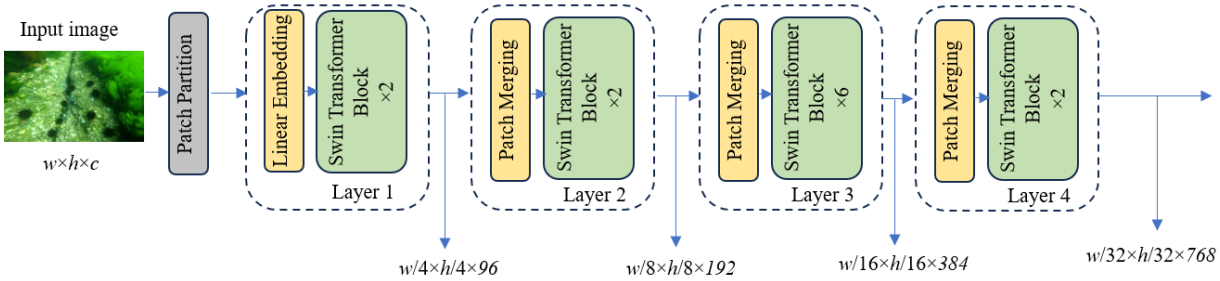

Fig. 2. Swin transformer architecture.

1) Patch partition block: Given an input image $I \in R^{w \times h \times c}$, the patch partition block divides this image into non-overlapping patches of size $p \times p \times c$. The total number of patches, N, produced by this partitioning is given by:

$$N = \left(\frac{w}{p}\right) \times \left(\frac{h}{p}\right) \quad (1)$$

Each patch is then flattened to produce a vector of dimension $p^2 \times c$. Thus, after the patch partition block, the image representation transforms from I to a matrix P of dimensions $N \times (p^2 \times c)$. In essence:

$$I \in R^{W \times H \times C} \rightarrow P \in R^{N \times (p^2 \times c)} \quad (2)$$

where, $P[i]$ represents the flattened vector for the i^{th} patch.

2) Linear embedding block: The linear embedding block applies a linear transformation to each of these flattened patches to project them into a specified embedding dimension D. This transformation can be represented by a matrix E of dimensions $(p^2 \times c) \times D$. Thus, the output L of the linear embedding block for each patch can be computed as:

$$L[i] = P[i] \times E \quad (3)$$

where, $L[i]$ represents the embedded vector for the i^{th} patch.

Given this, the entire input-output relationship for the linear embedding block can be represented as:

$$P \in R^{N \times (p^2 \times c)} \rightarrow L \in R^{N \times D} \quad (4)$$

where, P is the matrix of flattened patches, L is the embedding matrix.

3) Swin transformer block: The structure of two successive Swin Transformer blocks is depicted in Fig. 3. Each block consists of a sequence of operations: Layer Normalization (LN), Window-based Multi-head Self Attention (W-MSA) or Shifted Window-based Multi-head Self Attention (SW-MSA), and a Multi-Layer Perceptron (MLP) head. The LN standardizes the activations by calculating the mean and variance of the input image patch, thus stabilizing the training process. The W-MSA operates on a set of query (Q), key (K), and value (V) vectors. It matches the query to a set of key-value pairs to produce an output. This matching is achieved by computing the dot product between the query vector and every key vector. Subsequently, a softmax function is employed to scale these dot products, transforming them into weights denoted as k. The process is calculated as follows:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (5)$$

where, Q represents the Query matrix, K represents the Key matrix, V stands for the Value matrix, and d_k is the dimension of the keys. The divisions by $\sqrt{d_k}$ functions as a scaling factor, ensuring stability in the gradients during the training phase.

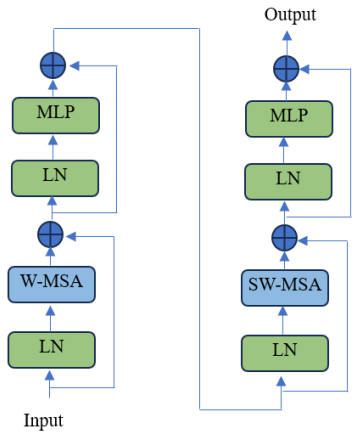

Fig. 3. The structure of two successive swin transformer blocks.

The locality of W-MSA might raise concerns about its ability to capture global context. To mitigate this, Swin Transformer employs multiple blocks of W-MSA and integrates a "shifting" strategy in subsequent blocks (SW-MSA), ensuring that tokens in one window in a certain block can interact with tokens in neighboring windows in the next block.

4) Patch merging block: This block is used to reduce the spatial dimensions of the input while augmenting the feature dimensions. Conceptually, this block aggregates neighboring patches from the previous layer and fuses them to form a larger patch. For instance, four adjacent patches of size p × p are merged to create a single patch of size 2p × 2p. This merging process typically employs a simple linear transformation. As a result, the spatial resolution of the feature map is halved in both height and width dimensions, but the depth or the number of channels is doubled. The purpose of this operation is twofold: firstly, it progressively reduces the computational requirements for subsequent layers, and secondly, it increases the receptive field, enabling the model to capture more global and abstract features as information flows deeper into the transformer.

C. Feature Attention Mechanism

In hierarchical models such as Swin Transformers and CNNs, lower-level features often capture fine-grained details, textures, and simple patterns. Meanwhile, higher-level features encompass more abstract, complex, and semantically rich information about objects, enabling the model to recognize more intricate and high-level attributes. By combining features from different layers, the model is equipped with a comprehensive and multi-scale representation of the input image. In addition, the underwater visuals are typically characterized by low-light conditions, varied light absorption and scattering, and blurry images due to particulate matter suspended in the water, which can result in a significant degradation of image quality and object distinguishability. This paper proposes a feature attention mechanism (FAM) to precisely address these challenges by enabling the model to selectively focus on important features and effectively integrate multi-level features from different stages of the network,

enhancing the representative power of deep features, especially in the challenging context of underwater object detection. The architecture of the FAM is illustrated in Fig. 4, which consists of two branches: the first branch directly processes the lower-level feature (F_1) through a batch normalization layer, ensuring the features are normalized and thereby enhancing the model's stability and convergence during training. Simultaneously, the second branch takes the higher-level feature (F_2) through a sophisticated pathway comprising a coordinate attention (CA) block, followed by a convolution layer, a max-pooling layer, and a batch normalization layer. The CA block [28] is notable for its capacity to encode both channel relationships and long-range dependencies with precise positional information, implemented through two crucial steps: coordinate information embedding and coordinate attention generation. Once the individual pathways of the two branches have processed the features, their outputs are aggregated by summation and then fed to a ReLU activation layer, which ensures the generation of a robust and hierarchically rich feature representation, designed to significantly enhance the underwater object detection capabilities of the system. The output of the FAM mechanism can be calculated as follows:

$$F_1' = BN(F_1) \quad (6)$$

$$F_2' = BN(MAXPOOL(CONV(CA(F_2)))) \quad (7)$$

$$F_{output} = ReLU(F_1' + F_2') \quad (8)$$

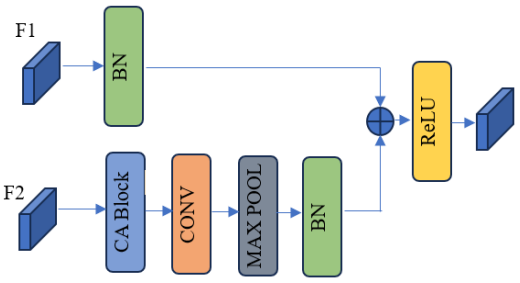

Fig. 4. Feature attention mechanism.

D. Detection Head with Combination of Convolution and Self- Attention

Pan et al. [29] highlighted the connection between convolution and self-attention mechanisms by highlighting computational similarities in both methods. Consequently, they designed a hybrid model, ACmix, which adeptly integrates the advantages of both self-attention and convolution, while maintaining minimal computational overhead relative to pure convolution or self-attention models. Given the potentially robust link between convolution and self-attention, the ACmix module is utilized in this paper to integrate the convolution and self-attention mechanisms. Fig. 5 illustrates the structure of the ACmix module. The module channels the data through a series of three 1×1 convolutional layers. These layers serve to capture local features and correlations in the data. Concurrently, the input is also processed through a self-attention mechanism equipped with a position encoder. This mechanism ensures the model can recognize and weigh global dependencies in the data effectively. Subsequent to their independent processing, the outputs from the convolutional and self-attention pathways pass through a 'Shift Operation' and are concatenated. This combined representation exploits the strengths of both paradigms, ensuring a comprehensive understanding of the data's local and global patterns. Finally, the concatenated output is summed to produce the final output. The output of each ACmix module is input into a YOLO detector head for location and classification.

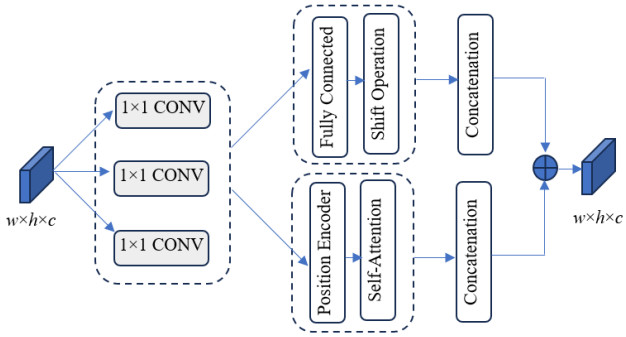

Fig. 5. The structure of ACmix.

IV. Experiments and Results

A. Dataset

The experiments utilize the UTDAC2020 dataset, which originates from the Underwater Target Detection Algorithm Competition in 2020 [18]. This comprehensive dataset comprises 5,168 training images and 1,293 validation images, focusing primarily on four specific marine species: echinus, holothurian, starfish, and scallop. The unique attribute of this dataset lies in its variety of resolutions, with images spanning four distinct sizes: 3840×2160, 1920×1080, 720×405, and 586×480. This dataset serves as an important resource for understanding and advancing underwater image analysis and target detection. For evaluation and comparison purposes, the standard COCO-style evaluation metric is employed.

B. Experimental Settings

The proposed model was implemented using the PyTorch deep learning framework and programmed in Python. All experiments were carried out on machines equipped with an NVIDIA RTX 4080 GPU. The backbone of our architecture is the base version of the Swin Transformer, which was pre-trained on the ImageNet-1K dataset and has an embedding dimension of C = 128. We chose this version because of its balance between computational efficiency, model size, and accuracy. The model was fine-tuned for 15 epochs, using a batch size of 2. For optimization, the AdamW optimizer [30] was employed, starting with a learning rate of 0.0002. This rate was adaptively adjusted based on the training progress, and a weight decay of 0.05 was implemented. Our data augmentation strategies included a variety of techniques such as random resizing, combined random resizing and cropping, as well as horizontal and vertical random flipping. For a comprehensive overview of the hyperparameters employed in the comparative models, (see Table I).

TABLE I. HYPERPARAMETERS OF ALL MODELS

Model	Initial learning rate	Regularizer	Optimizer	Batch size	Number of Epochs
Our model	0.0002	Weight decay of 0.05	AdamW	2	15
Deformable DETR [28]	0.00001	Weight decay of 0.0001	AdamW	2	40
RetinaNet [29]	0.01	L2	SGD with Momentum	10	20
Faster R-CNN with FPN [30]	0.02	Weight decay of 0.0001	SGD with Momentum	2	12
DetectoRS [31]	0.02	Weight decay of 0.0001	SGD with Momentum	2	20
FCOS [32]	0.01	Weight decay of 0.0001	SGD with Momentum	16	20
CenterNet [33]	0.0002	Weight decay of 0.0001	Adam	6	25

C. Comparison with Other Methods

The comparison results on the UTDAC2020 dataset are shown in Table II. Our underwater object detection model based on the Swin Transformer architecture obtains a significant improvement in performance when compared to other state-of-the-art models on the UTDAC2020 dataset. In terms of Average Precision (AP), the proposed model achieved a score of 51.6, which is notably higher than other models utilizing the ResNet50 backbone, such as Deformable DETR [31], RetinaNet [32], and Faster R-CNN with FPN [33], DetectoRS [34], FCOS [35], and especially CenterNet [36] which used ResNet18. Additionally, in the specific AP metrics (AP_{50}, AP_{75}), our Swin Transformer-based model also outperforms the competition, indicating a robustness in detecting objects at different Intersection over Union (IoU) thresholds. Remarkably, there is a significant jump in AP_{75} to 57.5, suggesting that the model is efficient at achieving a tighter fit around the detected objects. When analyzing the performance based on object size (AP_S, AP_M, AP_L), the proposed model consistently delivers superior results. The model's weakest performance is in AP_S at 23.2, which, while comparable to some models like Deformable DETR and DetectoRS, demonstrates that there might be challenges in detecting smaller underwater objects. Nonetheless, the model's AP_M and AP_L scores of 44.6 and 57.9, respectively, emphasize its efficiency in medium to large object detection. In summary, leveraging the Swin Transformer architecture and feature attention mechanisms has enhanced the efficacy of the proposed model in the challenging domain of underwater object detection.

Fig. 6 shows qualitative results of our model on the UTDAC2020 dataset. We can see a notable performance of the proposed model across diverse underwater scenarios. The detection results are evident across a range of images, from those where marine life is interspersed among a sea of green to those with rocky terrains. Even in images with dense clusters of organisms or potential overlapping instances, the model is efficient in differentiating between individual entities, avoiding much of the occlusion-related errors that often plague underwater detection tasks. Furthermore, the model's performance is evident in various lighting conditions and water turbidities, emphasizing its robustness.

D. Importance of Feature Attention Mechanism

We also conducted experiments to evaluate the impact of the FAM. Fig. 7 shows comparing the performance of the model with and without the FAM. When FAM is implemented, there is a noticeable improvement in all metrics. Specifically, the AP increases from 50.1% to 51.6%, indicating a more accurate model overall. This improvement is more pronounced in AP_{50} (from 82.2% to 85.1%), which measures precision at 50% IoU threshold, suggesting that FAM particularly enhances the model's ability to detect objects with a moderate overlap with the ground truth. The increase in AP_{75}, from 54.2% to 57.5%, also highlights better performance at a stricter IoU threshold, implying enhanced precision for more accurately localized predictions. The improvements in AP_S, AP_M, and AP_L are also noteworthy. The model with FAM achieves better results across these size-based categories, with the most significant jump observed in the medium-sized object category, from 40.3% to 44.6%. This suggests that FAM effectively enhances the model's capability to recognize and detect objects of various sizes, especially in challenging underwater environments where visibility and image quality are often compromised. Overall, the integration of FAM into the model clearly leads to better performance in object detection, making it a valuable addition to the model's architecture, particularly for tasks in complex and visually challenging environments like underwater object detection.

TABLE II. COMPARISON RESULTS ON THE UTDAC2020 DATASET

Model	Backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Deformable DETR [31]	ResNet50	46.6	84.1	47.0	24.1	42.4	51.9
RetinaNet [32]	ResNet50	43.9	80.4	42.9	18.1	38.2	50.1
Faster R-CNN with FPN [33]	ResNet50	44.5	80.9	44.1	20.0	39.0	50.8
DetectoRS [34]	ResNet50	47.6	82.8	49.9	23.1	41.8	54.2
FCOS [35]	ResNet50	43.9	81.1	43.0	19.9	38.2	50.4
CenterNet [36]	ResNet18	31.3	61.1	27.6	11.9	32.5	33.4
Our model	Swin Transformer	51.6	85.1	57.5	23.2	44.6	57.9

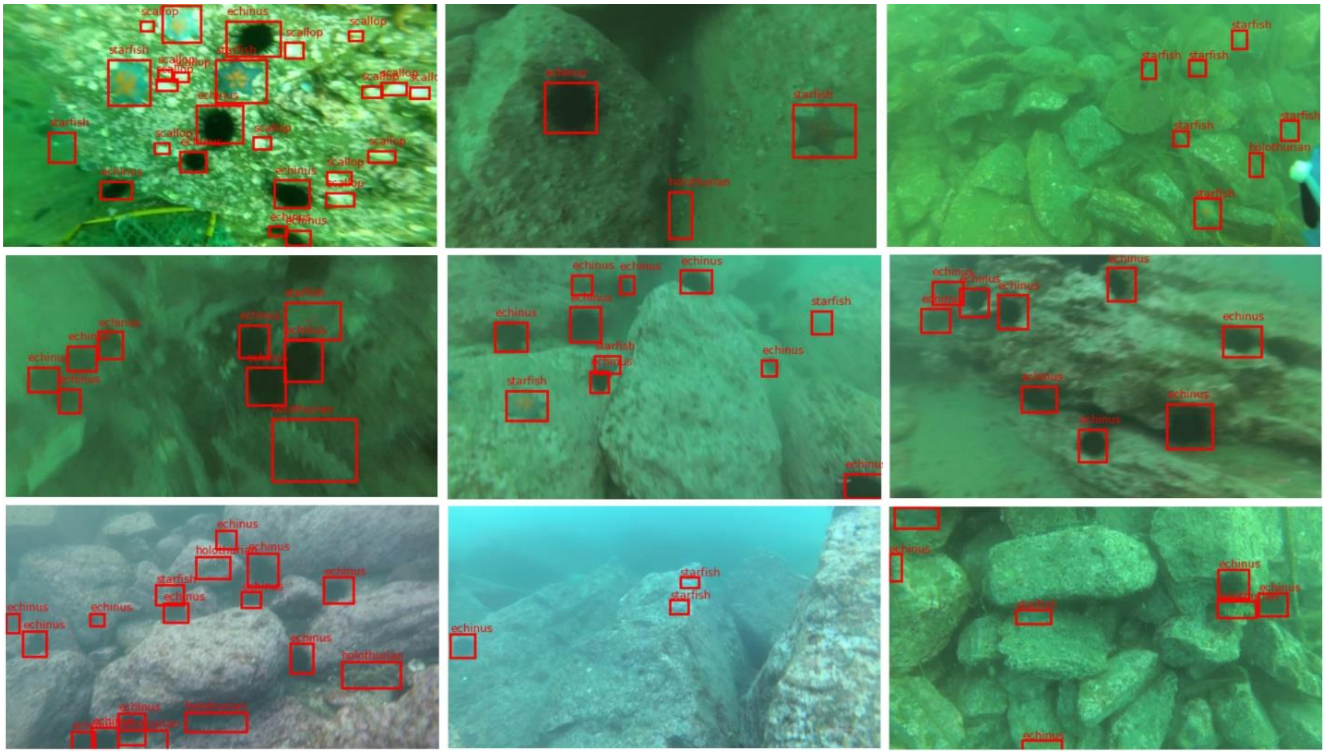

Fig. 6. Qualitative results on the UTDAC2020 dataset.

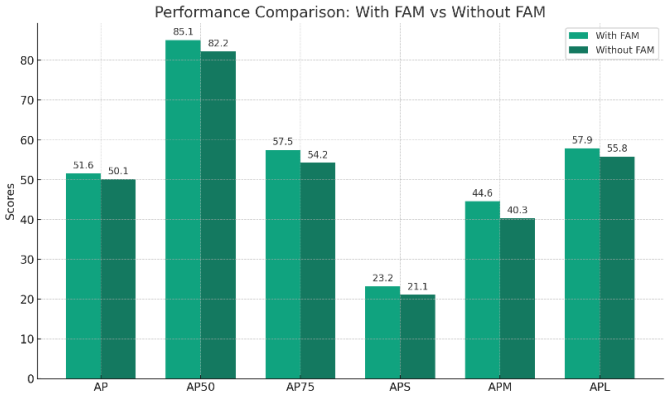

Fig. 7. Comparing the performance of the model with and without the FAM.

V. CONCLUSIONS

In this research, we addressed the challenges associated with underwater object detection by introducing a novel multi-layered architectural approach that effectively exploits the capabilities of Swin Transformer. Our method provides a structured approach to process segmented image patches derived from underwater scenes, ensuring accurate and efficient object detection. A significant contribution of our research is the Feature Alignment Module (FAM), specifically designed to address the complexities of marine environments. By focusing on essential features and integrating multi-level features across various network stages, the FAM substantially elevates the depth and precision of feature representation. Moreover, the incorporation of several detection heads, coupled with the ACmix module, represents a transformative approach to enhancing detection accuracy. The results on the UTDAC2020 dataset emphasize not only the efficacy of our proposed method but also its potential as a benchmark solution in the field of underwater object detection. In future, we envision further refining our model by integrating more advanced attention mechanisms and exploring its applicability in other complex environmental scenarios.

REFERENCES

[1] Shi, X. U. X., and J. L. Zhang. "Feature extraction of underwater targets using generalized S-transform." *J. Comput. Appl.* 32 (2012): 280-282.

[2] Liu, L. X., S. H. Jiao, and T. Chen. "Detection and recognition of underwater target based on feature matching." *Modern Electronics Technique* 34, no. 4 (2014): 73-76.

[3] Liu, L. I. K., and M. Dun. "Algorithm for recognition of underwater small target based on shape characteristic." *Ship Sci. Technol* 34, no. 1 (2012).

[4] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).

[5] Redmon, Joseph, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You only look once: Unified, real-time object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779-788. 2016.

[6] Redmon, Joseph, and Ali Farhadi. "YOLO9000: better, faster, stronger." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7263-7271. 2017.

[7] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21-37. Springer International Publishing, 2016.

[8] Dai, Jifeng, Yi Li, Kaiming He, and Jian Sun. "R-fcn: Object detection via region-based fully convolutional networks." *Advances in neural information processing systems* 29 (2016).

[9] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.

[10] Muthana, Mahmoud, and Ahmed R. Nasser. "Using Dynamic Pruning Technique for Efficient Depth Estimation for Autonomous Vehicles." *Mathematical Modelling of Engineering Problems* 9, no. 2 (2022).

[11] Srikrishnan, A., Arun Raaza, Abishek B. Ebenezer, V. Rajendran, M. Anand, and S. Gopalakrishnan. "A Fast and Effective Method for Intrusion Detection using Multi-Layered Deep Learning Networks." *International Journal of Advanced Computer Science and Applications* 13, no. 12 (2022).

[12] Alowaidi, Majed. "Modified Intrusion Detection Tree with Hybrid Deep Learning Framework based Cyber Security Intrusion Detection Model." *International Journal of Advanced Computer Science and Applications* 13, no. 10 (2022).

[13] Ummadisetti, Ganesh Naidu, R. Thiruvengatanadhan, Satyala Narayana, and P. Dhanalakshmi. "Character level vehicle license detection using multi layered feed forward back propagation neural network." *Bulletin of Electrical Engineering and Informatics* 12, no. 1 (2023): 293-302.

[14] Santoso, Albertus Joko, and Raymond Erz Saragih. "Automatic Face Mask Detection Based on MobileNet V2 and Densenet 121 Models." *ICIC Express Letters* 16, no. 4 (2022): 433-440.

[15] Li, Xiu, Min Shang, Jing Hao, and Zhixiong Yang. "Accelerating fish detection and recognition by sharing CNNs with objectness learning." In *OCEANS 2016-Shanghai*, pp. 1-5. IEEE, 2016.

[16] Chen, Long, Feixiang Zhou, Shengke Wang, Junyu Dong, Ning Li, Haiping Ma, Xin Wang, and Huiyu Zhou. "SWIPENET: Object detection in noisy underwater images." *arXiv preprint arXiv:2010.10006* (2020).

[17] Wei, Xiangyu, Long Yu, Shengwei Tian, Pengcheng Feng, and Xin Ning. "Underwater target detection with an attention mechanism and improved scale." *Multimedia Tools and Applications* 80, no. 25 (2021): 33747-33761.

[18] Zeng, Lingcai, Bing Sun, and Daqi Zhu. "Underwater target detection based on Faster R-CNN and adversarial occlusion network." *Engineering Applications of Artificial Intelligence* 100 (2021): 104190.

[19] Chen, Lingyu, Meicheng Zheng, Shunqiang Duan, Weilin Luo, and Ligang Yao. "Underwater target recognition based on improved YOLOv4 neural network." *Electronics* 10, no. 14 (2021): 1634.

[20] Cao, Xiang, Lu Ren, and Changyin Sun. "Dynamic target tracking control of autonomous underwater vehicle based on trajectory prediction." *IEEE Transactions on Cybernetics* 53, no. 3 (2022): 1968-1981.

[21] Huang, Hai, Hao Zhou, Xu Yang, Lu Zhang, Lu Qi, and Ai-Yun Zang. "Faster R-CNN for marine organisms detection and recognition using data augmentation." *Neurocomputing* 337 (2019): 372-384.

[22] Chen, Zhe, Hongmin Gao, Zhen Zhang, Helen Zhou, Xun Wang, and Yan Tian. "Underwater salient object detection by combining 2D and 3D visual features." *Neurocomputing* 391 (2020): 249-259.

[23] Lin, Wei-Hong, Jia-Xing Zhong, Shan Liu, Thomas Li, and Ge 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)*, pp. 2588-2592. IEEE, 2020.

[24] Song, Pinhao, Pengteng Li, Linhui Dai, Tao Wang, and Zhan Chen. "Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection." *Neurocomputing* 530 (2023): 150-164.

[25] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[26] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[27] Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012-10022. 2021.

[28] Hou, Qibin, Daquan Zhou, and Jiashi Feng. "Coordinate attention for efficient mobile network design." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13713-13722. 2021.

[29] Pan, Xuran, Chunjiang Ge, Rui Lu, Shiji Song, Guanfu Chen, Zeyi Huang, and Gao Huang. "On the integration of self-attention and convolution." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815-825. 2022.

[30] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101* (2017).

[31] Zhu, Xizhou, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. "Deformable detr: Deformable transformers for end-to-end object detection." *arXiv preprint arXiv:2010.04159* (2020).

[32] Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection." In *Proceedings of the IEEE international conference on computer vision*, pp. 2980-2988. 2017.

[33] Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature pyramid networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117-2125. 2017.

[34] Qiao, Siyuan, Liang-Chieh Chen, and Alan Yuille. "Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10213-10224. 2021.

[35] Tian, Zhi, Chunhua Shen, Hao Chen, and Tong He. "Fcos: Fully convolutional one-stage object detection." In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9627-9636. 2019.

[36] Zhou, Xingyi, Dequan Wang, and Philipp Krähenbühl. "Objects as points." *arXiv preprint arXiv:1904.07850* (2019).