WaveCodNet: A Frequency-Aware Framework for Underwater Camouflaged Object Detection

WaveCodNet: A Frequency-Aware Framework for Underwater Camouflaged Object Detection

Jiajun LIN[1], Yuanbo CHEN[2], Yi LE[3], Shangjing SUN[3], Hongyang BAI[1,4*], Shuai GUO[1], Tianyu DENG[1]

[1] Nanjing University of Science and Technology, Nanjing, Jiangsu 210094 China
[2] Xi'an Modern Control Technology Research Institute, Xi'an, Shaanxi 710065 China
[3] Nanjing Institute of Electronic Engineering, Nanjing, Jiangsu 210000 China
[4] This work was supported by the National Natural Science Foundation of China (U21B2028), the Fundamental Research Funds for the Central Universities (30925010521), and the Graduate Education and Teaching Reform Project, Nanjing University of Science and Technology (KT2024_C17).

Email: linjiajun@njust.edu.cn

Abstract. The environmental perception of Unmanned Surface Vehicles (USVs) is often challenged by complex backgrounds, target shape variability, and camouflage, limiting the effectiveness of conventional detection methods. To overcome these issues, we propose WaveCODNet, a novel camouflaged object detection (COD) framework that integrates frequency-domain analysis with multi-scale feature learning. The model utilizes an enhanced Pyramid Vision Transformer (PVTv2) as its backbone to capture hierarchical global features. We introduce a Frequency-enhanced Dense Interaction Decoder (F-DID) to simulate coarse-to-fine human visual perception by decomposing and fusing high- and low-frequency information for initial target localization. A Spatial-Frequency Attention Block (SFAB) is designed to jointly model spatial and frequency dependencies, thereby enhancing feature discrimination. Furthermore, the Frequency-Aware Multi-Scale Fusion (FAMSF) module incorporates frequency-domain transformations to effectively integrate multi-level features with localization cues from F-DID, improving detection across varying object scales. Extensive evaluations on four benchmark COD datasets demonstrate that WaveCodNet surpasses current state-of-the-art techniques, providing a robust and accurate solution for detecting camouflaged objects.

1. Introduction

With the rapid development of technologies such as artificial intelligence and unmanned control, unmanned systems will serve as important complementary forces to manned systems, performing various complex tasks in a more covert, efficient, and cost-effective manner. As a key component of unmanned systems, the Unmanned Surface Vessel (USV) plays a significant role in fields such as fishery monitoring, maritime security, and oceanographic research, leveraging its unique advantages. USVs employ various sensors as information perception systems to achieve environmental awareness or target detection and identification. However, with continuous advancements in underwater vehicle propulsion and bionic technology, it is now possible to simulate fish-like underwater movement using sinusoidal wave propulsion methods, enabling operation in a near-silent state over hundreds of miles. During surface operations, the self-noise of such systems increasingly approximates the level of ambient ocean noise, which significantly challenges detection by radar and sonar sensors. Therefore, underwater image target detection via electro-optical devices has become a crucial component of environmental perception for acquiring features of camouflaged targets.

Currently, research on underwater image target detection primarily focuses on image enhancement and Generic Object Detection (GOD). Within this framework, studies on underwater generic object detection are mainly concentrated on algorithm improvement. Among these, research on underwater generic object detection primarily focuses on algorithm improvement. Lin et al.[1] proposed the RoiMix algorithm to solve overlap, occlusion, and blurring problems by fusing multiple images, thereby improving the detection accuracy of underwater images. Zeng et al.[2] enhanced the standard Faster R-CNN detection algorithm by integrating an adversarial occlusion network, which learns to create occlusions that challenge the detection network's ability to correctly classify objects. Boosting R-CNN[3] utilizes a reweighting strategy to adjust the classification loss based on the errors in the region proposal network, effectively addressing the challenges specific to underwater environments, such as blurriness, low contrast, and mimicry. However, these methods generally require considerable computational power, which may limit their practical applicability in resource-constrained scenarios.

Single-stage underwater object detection algorithms operate at high speed, making them suitable for real-time detection applications. Muksit et al.[4] introduced YOLO-Fish, an enhanced version of YOLOv3, for efficient underwater fish detection, they utilized two new datasets (i.e., DeepFish and OzFish) for testing. Zhang et al.[5] proposed a quantitative detection method for marine benthos, which solves problems such as low light and blurred images in underwater environments. U-YOLOv7[6], which integrates new features such as the cross-efficient squeeze-excitation structure, content-aware reassembly of features (CARAFE) upsampling, and a 3D attention mechanism, is specifically developed to detect underwater organisms. FishNet[7] combines the strengths of pyramid vision transformers and CNNs in a dual-path approach to effectively extract both global and local features from underwater images, which significantly improves the network performance.

To enhance the detection capability of Unmanned Surface Vessels (USVs) for camouflaged underwater targets within complex backgrounds, we propose a novel architecture called WaveCODNet, which consists of three key modules: a Frequency-enhanced Dense Interaction Decoder (F-DID), a Spatial-Frequency Attention Block (SFAB), and a Frequency-Aware Multi-Scale Fusion (FAMSF) module. Specifically, the F-DID module addresses the imbalance of semantic information across backbone layers by decomposing features into high- and low-frequency components, fusing them to form a complete frequency representation. These features are further enhanced via Spatial-Channel Synergistic Attention (SCSA), enabling more accurate coarse localization to guide the FAMSF module.

The SFAB module jointly learns spatial and frequency representations to enhance target perception under low contrast between objects and background. Finally, the FAMSF module aggregates multi-level features and coarse localization maps via frequency transformation to generate more accurate predictions by enriching contextual diversity and improving feature fusion.

Extensive experiments on benchmark COD datasets such as COD10K[8] and CAMO[9] demonstrate that WaveCODNet achieves superior performance compared to state-of-the-art methods, particularly in accurately detecting camouflaged objects with complex boundaries.

To summarize, the key contributions of this paper are as follows:

We present a comprehensive camouflaged object detection framework that employs multi-scale context-aware feature learning and seamlessly integrates frequency-domain information, inspired by the human visual system's coarse-to-fine perceptual mechanism. Extensive evaluations on benchmark datasets show that our approach outperforms existing state-of-the-art methods in terms of both accuracy and boundary preservation.

The F-DID module is designed to simulate the human visual system, producing a coarse localization map from the multi-scale features of the PVTv2 backbone. This coarse map then conditions the fused features within the MASMF module, a process dedicated to enriching scale diversity and refining object boundaries with the aid of frequency-domain information.

To efficiently learn multi-scale context-aware features, we adopt PVTv2 as the backbone and introduce a Spatial-Frequency Attention Block (SFAB) for joint spatial-frequency feature refinement. Furthermore, the Frequency-Aware Multi-Scale Fusion (FAMSF) module enhances multi-level representations through frequency-domain transformations, significantly improving detection precision.

2. Related work

2.1. Camouflaged Object Detection (COD)

Traditional camouflaged object detection (COD) methods mainly rely on low-level image features, where feature extraction is driven by prior knowledge and manual design. These approaches often suffer from time-consuming processing, limited robustness, poor generalization, and subpar detection performance[1]. With the rise of deep learning and the availability of large-scale COD datasets [1,10,11], learning-based methods have gained prominence. Models such as SINet[1], LS [12], and MGL[13] introduce bio-inspired mechanisms, multi-task learning, and graph-based modeling to improve detection through enhanced region localization, segmentation refinement, and spatial-semantic interaction.However, CNN-based methods still struggle to distinguish targets from background in complex scenes due to limited receptive fields and insufficient global context. To address this, Transformer-based architectures have recently been introduced, leveraging self-attention and global modeling to enhance feature discrimination and boost detection accuracy and robustness [14,15].

In parallel, frequency-domain techniques have demonstrated strong potential for COD. Methods based on frequency perception and correction fusion[16], frequency-spatial entanglement learning[17], and frequency-assisted attention[18] effectively integrate multi-scale frequency cues to enhance object representation. Other work incorporates frequency-aware context aggregation to suppress high-frequency background noise and emphasize target-relevant features[19], further improving detection performance across benchmarks.

2.2. Contextual Information Learning

Effective acquisition of contextual information is crucial for camouflaged object detection (COD), as it plays a vital role in enhancing feature representations by capturing spatial-semantic dependencies between camouflaged targets and their often visually similar backgrounds. Precise modeling of these interdependencies allows the network to better distinguish subtle foreground cues from complex and cluttered scenes, which is essential in COD tasks characterized by low contrast and background-foreground confusion.

As a result, considerable research attention has been directed toward developing mechanisms that improve contextual feature learning[20,21]. To address the oversimplified foreground-background settings in earlier COD datasets and the limited context modeling capacity of traditional methods, researchers have constructed more challenging datasets featuring diverse and cluttered backgrounds. In tandem, they proposed semantic context-aware frameworks to facilitate more accurate saliency inference under complex scene conditions[22].Building on this, some works introduced sophisticated architectural modules—such as the combination of Dense Relation Distillation (DRD) and Context-aware Feature Aggregation (CFA)[23]—to strengthen the alignment and interaction between support and query features. These modules enable the network to effectively capture both global and local contextual patterns, thereby improving its robustness to variations in object appearance, scale, and occlusion.Moreover, dual self-attention mechanisms have been proposed[24] to enhance feature fusion by integrating convolutional representations with Transformer-style global attention. This hybrid design leverages the local modeling capability of CNNs and the long-range dependency modeling of self-attention, resulting in stronger context-aware representations and improved detection accuracy, particularly in 3D object detection tasks with complex spatial structures.

3. Proposed method

3.1. Architecture overview

The proposed WaveCodNet framework for camouflaged object detection is illustrated in Figure 1. It leverages the Pyramid Vision Transformer v2 (PVTv2) [25] backbone to extract a hierarchy of features, thereby capturing extensive multi-scale context. Compared with other Transformer variants, PVTv2 offers an efficient trade-off between global semantics and local detail preservation, benefiting from its overlapping patch embedding and pyramid structure.

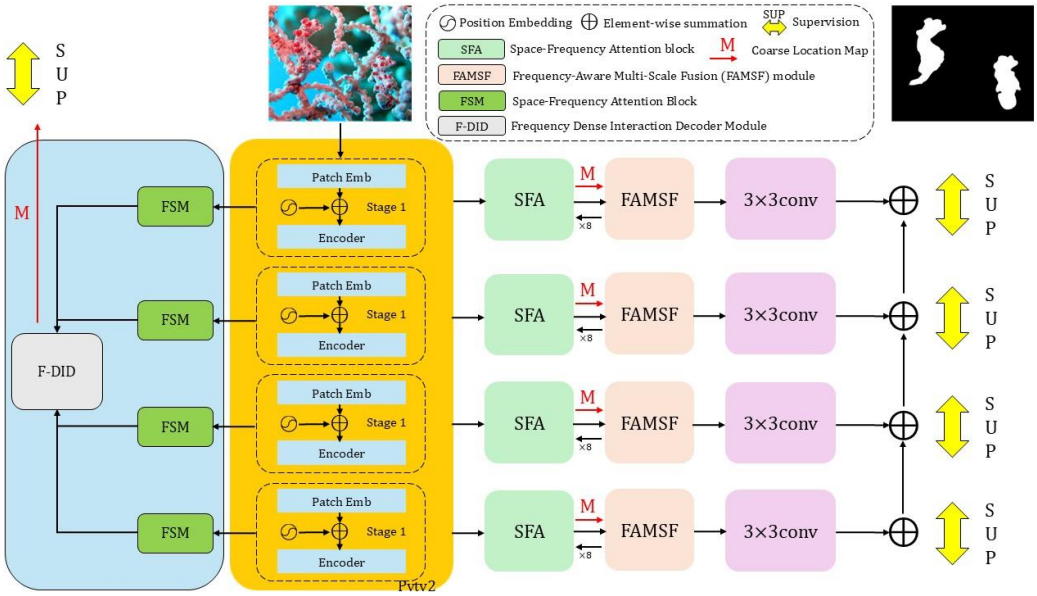

Figure 1. Architecture of the proposed WaveCodNet for camouflaged object detection. The model is built on an enhanced Pyramid Vision Transformer v2 (PVTv2) backbone and consists of three major modules: the Frequency Dense Interaction Decoder (F-DID), the Space-Frequency Attention Block (SFAB), and the Frequency-Aware Multi-Scale Fusion (FAMSF). The symbol "SUP" indicates supervision between each prediction and its corresponding ground truth. During training, both the coarse localization map M produced by the F-DID module and the multi-stage prediction maps are jointly supervised using manually annotated camouflage masks. Additional details are provided in Section III.

Given an input image $I \in R^{H\times W\times 3}$, PVTv2 produces four feature maps X_1, X_2, X_3, X_4, with progressively decreasing spatial resolutions and increasing semantic abstraction. These features provide complementary information ranging from fine-grained details to high-level context, which is essential for accurate COD.

To enhance frequency awareness, we propose a Frequency-Enhanced Dense Interaction Decoder (F-DID), which extracts frequency-domain features from high-level backbone outputs and integrates them with spatial cues to perform coarse localization and initial segmentation.

Next, the backbone features are refined through a Spatial-Frequency Attention Block (SFAB), which captures long-range dependencies via cross-domain (spatial–frequency) interaction. We then utilize a series of Frequency-Aware Multi-Scale Fusion (FAMSF) modules, which integrate multi-scale features guided by position priors from F-DID. Fourier transforms are applied throughout to strengthen frequency-domain representation.

Finally, the segmentation map is progressively reconstructed, and a multi-level supervision strategy is employed to optimize the network. The following sections describe the core components and loss function in detail.

3.2. Frequency-Enhanced Dense Interaction Decoder (F-DID)

Frequency Sensitivity Module(FSM). To effectively decompose and fuse multi-frequency components, we adopt an octave convolution-based structure. Given an input feature map, we denote its high-frequency and low-frequency components at the $i-th$ layers as X_i^H and X_i^L, respectively. The corresponding output features Y_i^H and Y_i^L are computed as follows:

$$Y_i^H = F(X_i^H; W^{H\to H}) + Upsample(F(X_i^L; W^{L\to H}), 2) \quad (1)$$

$$Y_i^L = F(X_i^L; W^{L\to L}) + F(Pool(X_i^H, 2)W^{H\to L}) \tag{2}$$

where $F(\cdot; W^{a\to b})$ denotes a convolutional operation with learnable weights mapping from frequency branch a to b $(a, b \in H, L)$.The operator $Pool(\cdot, 2)$ performs average pooling with a stride of 2, and $Upsample(\cdot, 2)$ represents a 2× upsampling operation, typically implemented via nearest-neighbor interpolation.After frequency-specific processing, the fused feature f_i is obtained by:

$$f_i = Resize(Y_i^H) \oplus Resize(Y_i^L) \tag{3}$$

where $Resize(\cdot)$ adjusts the spatial resolution to a common size and $\oplus$ denotes element-wise addition. This formulation enables effective integration of high- and low-frequency information, facilitating robust feature representation for camouflaged object detection.

Dense Interaction Decoder (DID) To fully leverage the multi-scale features, we incorporate the pyramid features from all four stages, denoted as $f_i \in R^{\frac{H}{2i+1}\times\frac{W}{2i+1}\times C_i}$, where $C_i \in \{64, 128, 320, 512\}$ and $i \in \{1, 2, 3, 4\}$ into the F-DID module for the preliminary localization of camouflaged objects. Additionally, to ensure semantic consistency within each layer and maintain contextual connectivity, we adapt the Partial Decoder from [26] by using a nearest-neighbor connection function. The process is mathematically represented as follows:

$$\begin{cases} f_4^l = f_4^{'} \\ f_3^{cl} = f_3^{'} \otimes b[\delta_{\uparrow}^2(f_4^{cl}); W_{DI}^3] \\ f_2^{cl} = f_2^{'} \otimes b[\delta_{\uparrow}^2(f_3^{cl}); W_{DI}^2] \otimes b[\delta_{\uparrow}^2(f_3^{cl}); W_{DI}^3] \\ f_1^{cl} = f_1^{'} \otimes b[\delta_{\uparrow}^2(f_2^{cl}); W_{DI}^4] \otimes b[\delta_{\uparrow}^2(f_2^{cl}); W_{DI}^5] \end{cases} \tag{4}$$

In this process, $f_i^{'}, i \in \{1, 2, 3, 4\}$ denotes the feature maps obtained from f_i by applying a 1×1 convolution to adjust the channel dimension to 64. Furthermore, $b[\cdot, W_{DI}^y], y \in \{1, 2, 3, 4, 5\}$ represents a 3×3 convolutional layer followed by a batch normalization operation. To maintain consistency in spatial dimensions among the candidate features, an upsampling operation, denoted as $\delta_{\uparrow}^2$, is performed prior to the element-wise multiplication.

Moreover, we incorporate the Spatial and Channel Synergistic Attention (SCSA) module, which effectively combines the strengths of both spatial and channel attention mechanisms, enabling the full utilization of multi-semantic information.The overall architecture is shown in figure 2. The process is mathematically represented as follows:

$$\begin{cases} SA(f_1^{cl}) = \sigma(Conv1(AvgPool(f_1^{cl}))) + \\ Conv2(MaxPool(f_1^{cl})) \\ CA(f_1^{cl}) = \sigma(FC_1(GlobalPool(f_1^{cl})) + \\ FC_2(GlobalMaxPool(f_1^{cl}))) \\ SCSA = (f_1^{cl} = SA(f_1^{cl}) \odot CA(f_1^{cl})) \end{cases} \tag{5}$$

where f_1^{cl} denotes the input feature map derived from the input of the F-DID module. $SA(\cdot)$ represents the spatial attention branch, where spatial descriptors are obtained by applying both average pooling and max pooling across the channel dimension, followed by separate convolution layers (Conv1, Conv2) and a sigmoid activation $\sigma(\cdot)$.

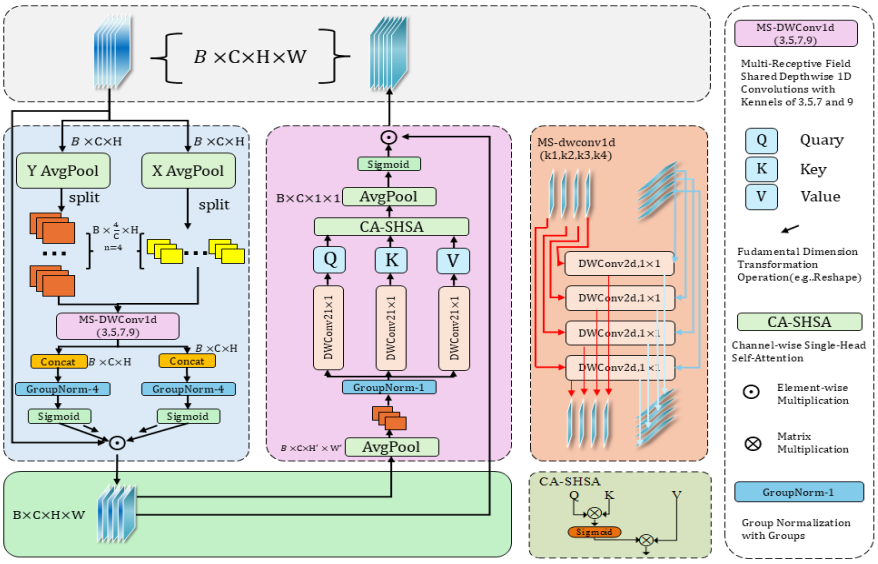

Figure 2. An illustration of SCSA, which uses multi-semantic spatial information to guide the learning of channel-wise self-attention. B denotes the batch size, C signifits the number of channels, and H and W correspond to the height and width of the feature maps, respectively. The variable n represents the number of groups into which sub-features are divided, and $1P$ denotes a single pixel.

$CA(\cdot)$ denotes the channel attention branch, where channel-wise descriptors are extracted by global average pooling and global max pooling, followed by fully connected layers (FC_1, FC_2) and a sigmoid activation. $\odot$ element-wise multiplication, used here to synergistically fuse the spatial and channel attention maps.

By jointly leveraging spatial and channel cues, the SCSA module enhances the DID module to form the F-DID architecture, which effectively suppresses background clutter and strengthens target-related features, leading to improved accuracy in camouflaged object detection.

3.3. Space-Frequency Attention block (SFAB)

Frequency Self-Attention. To model frequency-domain dependencies, we introduce a Frequency Self-Attention block(FSAB) as depicted in figure 3.

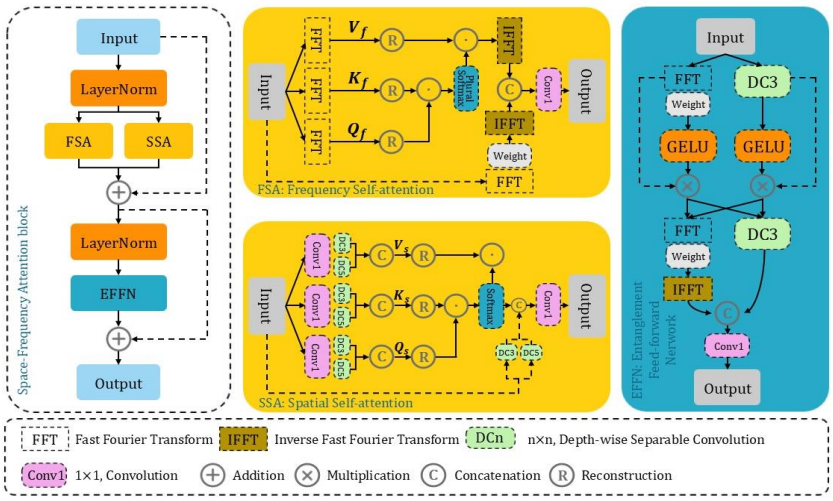

Figure 3. Details of the space-frequency attention block

Given input $x \in R^{C\times H\times W}$,we apply 2D FFT to obtain frequency features. Query, key, and value are computed as:

$$q_f = FFT^q(x), k_f = FFT^k(x), v_f = FFT^v(x) \tag{6}$$

The attention map is obtained via scaled dot-product with a learnable temperature parameter:

$$attn_f = \frac{q_f \cdot k_f}{\sqrt{d_k}} \cdot temperature \tag{7}$$

where d_k is the dimensionality of the key and $temperature$ is a learnable scaling parameter. The attended output is transformed back to the spatial domain:

$$out_f = \left| IFFT(attn_f, v_f) \right| \tag{8}$$

To complement this with low-frequency information, a convolution is applied to the real FFT component followed by IFFT. The final FSAB output is:

$$out_{FSA} = project_{out}(cal(out_f, out_l), 1) \tag{9}$$

Spatial Self-Attention. To capture spatial context, we apply convolutions with different receptive fields to x and construct multi-scale $query(q)$, $key(k)$, and $value(v)$ features. After normalization:

$$q, k = normalize(q, \dim = -1), normalize(k, \dim = -1) \tag{10}$$

The attention map is computed and applied:

$$attn_s = softmax(\frac{(q \cdot k^T)}{\sqrt{d_k}}, 1), out_s = attn_s \cdot v \tag{11}$$

Then reshaped to the original spatial format:

$$out_s = rearrange(out_s, b, h_n, h_d, h, w \to b, h_n \cdot h_d, h, w) \tag{12}$$

To enrich local details, we fuse out_s with multi-scale convolutions:

$$out_{SSA} = project_{out}(concat(out_s, out_{st}), 1) \tag{13}$$

This mechanism integrates global and local spatial features, enhancing the detection of camouflaged objects.

Feed-Forward Network (FFN). Frequency and spatial features capture complementary information—frequency reflects global signal variations, while spatial features retain local structural details. To effectively merge them, we treat these modalities as separate states and apply entangled learning via a series of convolutional layers. The fused output is computed as:

$$x_{output} = x + out_{SSA} + out_{FSA} + x_{norm2} \tag{14}$$

where x_{norm2} is a layer-normalized version of the input x, ensuring numerical stability. The final representation is obtained by concatenating x_{output} with the initial projection $conv_{1\times1}$ along the channel dimension, followed by dimensionality reduction and residual learning:

$$output = conv_{1\times1}(concat(conv_{1\times1}(x), x_{output})) + conv_{1\times1}(x) \tag{15}$$

This structure enhances feature expressiveness while stabilizing training via residual connections.

3.4. Frequency-Aware Multi-Scale Fusion (FAMSF) Module

To address scale variation in object detection, the proposed FAMSF module fuses multi-stage features and integrates position and frequency information, the architecture of our FAMSF module is shown in figure 4.

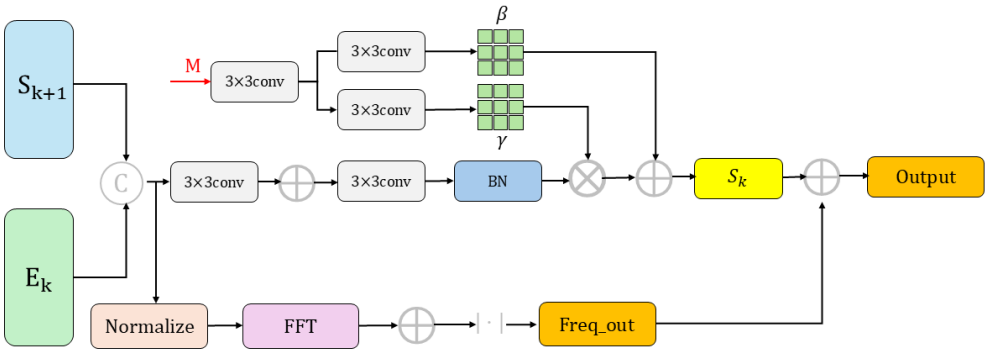

Figure 4. The architecture of our frequency-aware multi-scale fusion (FAMSF) module.

At the $k-th$ stage $(k \in \{1,2,3\})$, features from the next stage S_{k+1} and the current stage E_k are fused:

$$F_k = CB(concat(S_{k+1}, E_k) \oplus S_{k+1}) \otimes \gamma(M) \oplus \beta(M) \tag{16}$$

where $CB(\cdot)$ is a 3×3 convolution followed by conditional batch normalization, $\gamma(M)$, $\beta(M)$ are scale and shift parameters generated from coarse localization map M, $\oplus$ and $\otimes$ denote element-wise addition and multiplication.

Frequency-Domain Enhancement. To further enrich the features, a frequency-domain branch is employed. Given input features x and y ,we apply normalization and perform 2D Fast Fourier Transform (FFT):

$$x_f = \left|FFT(x)\right|, y_f = \left|FFT(y)\right| \tag{17}$$

These frequency-domain features are enhanced using a learnable convolution:

$$freq\_out = \left|IFFT(freq\_conv(x_f, y_f) + normalized)\right| \tag{18}$$

Finally, the fused output of the module is:

$$S_k = F_k + freq\_out \tag{19}$$

This formulation allows the model to jointly leverage multi-scale, positional, and frequency-domain information for robust feature representation.

3.5. Loss Function

To address foreground–background imbalance in pixel-wise binary classification, we employ a weighted binary cross-entropy(BCE) and weighted IOU loss:

$$L = L_{\omega BCE} + L_{\omega IOU} \tag{20}$$

where $L_{\omega BCE}$ and $L_{\omega IOU}$ denote the weighted BCE and IOU loss, with pixel-wise weights computed based on local intensity disparities [27]. To further enhance model attention to ambiguous regions, an Uncertainty-Aware Loss (UAL) is introduced:

$$UAL = cal\_ual(seg\_\log its = map_4, seg\_gts = G) \times ual\_codf \tag{21}$$

where G is the ground truth mask, and ual_coef is a scaling factor.

The model products five outputs: one coarse localization map M and four stage-wise prediction maps $R_1 - R_4$. The total training loss is defined as:

$$\begin{aligned} L_{total} &= L_{BCE}(M,G)\times 0.00625 + L_{BCE}(R_1,G)\times 0.125 \\ &+ L_{BCE}(R_2,G)\times 0.25 + L_{BCE}(R_3,G)\times 0.25 + L_{BCE}(R_4,G) \\ &+ 2\times UAL \end{aligned} \tag{22}$$

This loss formulation ensures multi-scale supervision while adaptively emphasizing uncertain regions for improved segmentation accuracy.

4. Experiments

This section first introduces the implementation details and benchmark settings, including the datasets and evaluation metrics used in our experiments. Next, we present both quantitative and qualitative comparisons with existing camouflaged object detection (COD) methods. Finally, an ablation study is conducted to demonstrate the effectiveness of the key components in the proposed COD framework.

4.1. Implementation Details and Benchmark Protocols

Our method is implemented in PyTorch and optimized using Adam. Following common practice[8,28,29,30], input images are resized to 352×352 via bilinear interpolation. The model is trained for 100 epochs with a batch size of 36, an initial learning rate of 8e-5 (decayed by 10× every 50 epochs), and a weight decay of 0.1. Training on two NVIDIA RTX 3090 GPUs takes approximately 8 hours.

We evaluate the proposed approach on four widely adopted camouflaged object detection (COD) datasets: COD10K [8], CAMO [15], NC4K [31], and CHAMELEON [14]. COD10K is the largest benchmark, containing 3040 training and 2026 testing images. CAMO includes 1250 samples (1000 for training and 250 for testing), while NC4K (4121 images) and CHAMELEON (76 images) are used solely for testing to assess the generalization performance of the model. Following prior works, we train on the combined COD10K and CAMO training sets. Evaluation is conducted using four standard metrics: Structure-measure (S_α), Enhanced-measure (E_ϕ), Weighted F-measure (F_β^ω), and Mean Absolute Error (MAE). Higher values of the first three and a lower MAE indicate better performance.

4.2. Comparison With SOTA Methods

We evaluate the performance of the proposed method by comparing it with 4 state-of-the-art (SOTA) COD methods, including MSCAF-Net[32], F-a Net[19], FMNet[18], AdaptCOD[33].

Quantitative Comparison: Table 1. presents the quantitative performance of various COD methods across four benchmark datasets. From the results, it is evident that our WaveCodNet consistently outperforms the other 4 methods on all four evaluation metrics for each dataset. In particular, NC4K large-scale dataset, our WaveCodNet improves the S_α , E_ϕ, F_β^ω by 1.1%, 2%, and 3.4%, respectively, compared to the second-best model, AdaptCOD, while reducing the MAE by 24.1%. On the COD10K dataset, our method outperforms the AdaptCOD model, achieving improvements of 2.5%, 1%, 3.5% in terms of S_α, E_ϕ, F_β^ω respectively. Additionally, MAE is reduced by 14.3%, demonstrating the strong generalization capability of our approach. On the CAMO dataset,our WaveCodNet outperforms the AdaptCOD model, achieving improvements of 2.6%, 2.9%, 5.3%, respectively, compared to the second-best model, AdaptCOD, while reducing the MAE by 30.2%. In summary, the proposed method significantly advances the state-of-the-art performance. Furthermore, we present the number of parameters (#Params, in million) for various COD models in table 1.

Our WaveCodNet exhibits a comparatively lower number of parameters than the majority of the models considered. In fact, only one of the four models included in the comparison has fewer parameters than ours.

Table 1(a). Quantitative comparison of various COD methods on four benchmark datasets using the metrics S_α, E_ϕ, F_β^ω, and MAE. The best-performing results are highlighted in bold.

Method	Year	#params	COD10K				CAMO-Test			
			$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
MSCAF-Net	2023	29.68	0.865	0.927	0.775	0.037	0.814	0.870	0.781	0.076
F-a Net	2023	38.87	0.809	0.889	0.684	0.035	0.783	0.839	0.702	0.081
FMNet	2025	68.89	0.895	0.940	0.798	0.020	0.890	0.946	0.863	0.039
Adapt COD	2025	66.40	0.892	0.938	0.819	0.021	0.886	0.932	0.884	0.043
Wave CODNet	**-**	**37.84**	**0.914**	**0.948**	**0.848**	**0.018**	**0.909**	**0.959**	**0.889**	**0.030**

Table 1(b). Quantitative comparison of various COD methods on four benchmark datasets using the metrics S_α, E_ϕ, F_β^ω, and MAE. The best-performing results are highlighted in bold.

Method	Year	#params	NC4K				CHAMELEON			
			$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
MSCAF-Net	2023	29.68	0.887	0.935	0.839	0.032	0.912	0.958	0.865	0.022
F-a Net	2023	38.87	0.862	0.910	0.828	0.040	0.888	0.939	0.828	0.032
FMNet	2025	68.89	0.890	0.944	0.871	0.029	0.916	0.950	0.880	0.022
Adapt COD	2025	66.40	0.906	0.942	0.860	0.029	0.922	0.960	0.879	0.020
Wave CODNet	**-**	**37.84**	**0.916**	**0.957**	**0.889**	**0.022**	**0.931**	**0.966**	**0.886**	**0.019**

Visual Comparison: Figure 5 presents a visual comparison of detection results on four test samples from the COD10K-Test and NC4K datasets. The detection outcomes of AdaptCOD, FMNet, MSCAF-Net, F-a COD, and our WaveCodNet are shown. These samples cover a range of challenging scenarios, including occlusion (rows 1), intricate edge details (row 2), high luminance (row 3), small targets (row 4). In all these scenarios, the proposed method consistently produces much more accurate results compared to the other methods when compared to the ground truth. Notably, for the last three challenges—dense fine details, small targets—other methods show a decline in performance, whereas our approach still achieves highly accurate results (refer to rows 4).

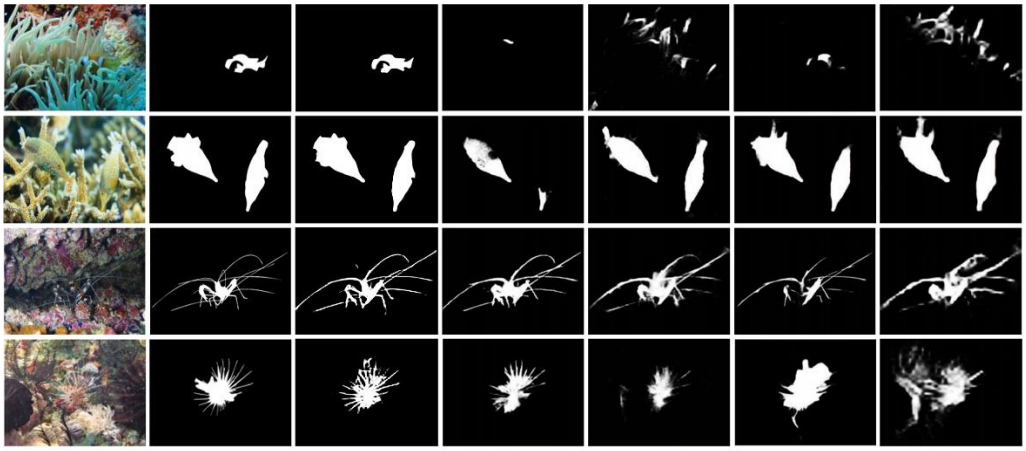

Figure 5: Visual comparison of detection results produced by different COD methods. (a) Input images; (b) Ground-truth (GT) maps; (c)–(g) Results generated by (c) the proposed method, (d) AdaptCOD, (e) FMNet, (f) MSCAF-Net, and (g) F-a COD.

Ablation Analysis: We conduct a series of ablation studies to evaluate the effectiveness of the SFAB, F-DID, and FAMSF modules in our WaveCodNet framework. Specifically, the ablation analysis centers on the following six model variants:

Basic (M1): This baseline model represents a simplified version of our network in which the SFAB, F-DID, and all FAMSF modules are omitted.

Basic+FAMSF (M2): This variant is obtained by adding the FAMSF modules to the Basic model (M1). In this configuration, the FAMSF module operates without the feature modulation component linked to the F-DID module, preserving only the feature fusion functionality.

Basic+F-DID (M3): This model is derived by integrating the F-DID modules into the Basic model (M1).

Basic+FAMSF+F-DID (M4): This variant is formed by incorporating the FAMSF modules into the Basic+F-DID model (M3). In this setting, the FAMSF module is employed in its full configuration, encompassing both the feature fusion and feature modulation components.

Basic+SFAB+F-DID (M5): This model is obtained by introducing the SFAB and F-DID modules into the Basic model (M1).

Basic+SFAB+FAMSF+F-DID (M6): This is the complete version of the network, obtained by incorporating the SFAB modules into the Basic+FAMSF+F-DID model (M5).

Table 2. Quantitative results of the ablation study on two benchmark datasets, evaluated across four metrics (S_α, E_ϕ, F_β^ω, and MAE). The best-performing scores are highlighted in bold.

Method	COD-10K-Test				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
Basic(M1)	0.845	0.914	0.734	0.029	0.875	0.925	0.813	0.037
Basic+FAMSF(M2)	0.868	0.919	0.778	0.027	0.886	0.932	0.853	0.032
Basic+F-DID(M3)	0.875	0.928	0.798	0.025	0.898	0.942	0.863	0.030
Basic+FAMSF+ F-DID(M4)	0.900	0.938	0.823	0.021	0.910	0.949	0.879	0.022
Basic+SFAB+F-DID(M5)	0.875	0.933	0.820	0.022	0.907	0.947	0.879	0.024

Basic+SFAB+FAMSF+ F-DID(M6)	0.914	0.948	0.848	0.018	0.916	0.957	0.889	0.022

Table 2 summarizes the quantitative results of different models in the ablation study on four benchmark datasets. Furthermore, the corresponding visual comparison of these models is illustrated in Figure 6.

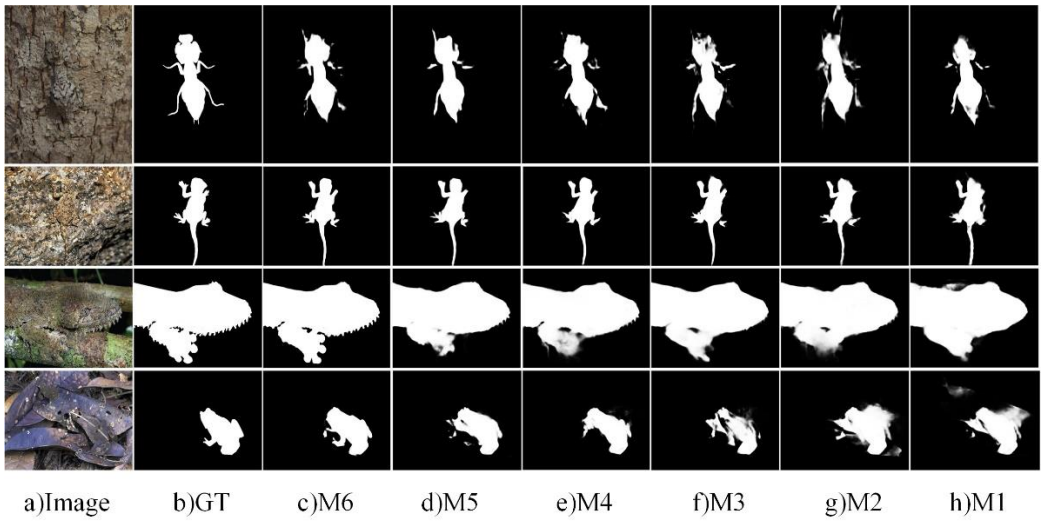

Figure 6. Visual comparison of detection results from different models in the ablation s tudy. (a) Input images; (b) Ground-truth (GT) maps; (c)–(h) Detection outputs from (c) Basic+SFAB+FAMSF+F-DID (M6), (d) Basic+SFAB+F-DID (M5), (e) Basic+FAMS F+F-DID (M4), (f) Basic+F-DID (M3), (g) Basic+FAMSF (M2), and (h) Basic (M1).

Effectiveness of SFAB:The effectiveness of the SFAB module can be confirmed through comparisons between three pairs of models: M5 to M3, M6 to M4. It is evident that incorporating SFAB leads to consistent accuracy gains across all three comparison cases. As shown in Fig.6, the proposed SFAB module refines object boundaries by modeling long-range dependencies and jointly attending to spatial and frequency features. Its cross-domain attention mechanism enhances edge representations and suppresses false responses in non-camouflaged regions, leading to more accurate and focused predictions.

Effectiveness of FAMSF & F-DID: Comparisons between M2 and M1, as well as M6 and M5, the FAMSF module leverages frequency transformation to fuse multi-scale features with the decoder's coarse localization map, enhancing contextual diversity and spatial-frequency integration. This enables more accurate and robust predictions, especially under complex and low-contrast camouflaged scenarios. The F-DID module improves accuracy (M3 to M1, M4 to M2) by using frequency-domain features and the SCSA to capture rich spatial cues, refining the coarse localization map for better subsequent processing. The combined use of both modules leads to consistent gains across all metrics. Visual results in figure 6 confirm that their integration helps the network better distinguish camouflaged objects from the background.

4.3. *Limitation and Discussion*

While the proposed method effectively enhances camouflaged object detection by integrating frequency-domain features with spatial cues, several limitations remain. First, the fusion strategy between frequency and spatial features could be further refined. The current approach may not fully exploit the complementary nature of these domains, limiting the model's potential to adapt to highly complex or diverse camouflage patterns.

Additionally, the use of frequency-domain modules introduces extra computational overhead, which may hinder real-time applications or deployment on resource-constrained devices. Future work will explore lightweight frequency modeling techniques and more adaptive fusion mechanisms that

can dynamically balance the importance of spatial and frequency cues. We also plan to investigate cross-domain generalization to ensure robust performance across different datasets and camouflage scenarios.

5. Conclusion

We propose WaveCodNet, a novel framework for camouflaged object detection that innovatively integrates frequency- and spatial-domain representations to significantly enhance detection accuracy and robustness. Specifically, we design the F-DID module to extract multi-stage frequency cues and efficiently fuse and decode frequency-aware features, enabling effective coarse-to-fine object localization. To further enrich feature representations, we introduce the SFAB module, which captures long-range dependencies across spatial and frequency domains, resulting in more discriminative and context-aware features. Additionally, the FAMSF module facilitates comprehensive multi-scale feature interaction, allowing the network to better model the complex spatial-frequency structures inherent in camouflaged scenes. To address foreground-background class imbalance, we adopt a weighted binary cross-entropy loss that adaptively assigns pixel-wise weights based on local contrast. Comprehensive experiments on four benchmark datasets show that our WaveCodNet substantially surpasses existing COD methods, achieving a significant improvement over the current state of the art.

6. References

[1] Lin W-H, Zhong J-X, Liu S, Li T and Li G 2020 *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona, Spain) pp 2588-92.
[2] Zeng, L., Sun, B., Zhu, D., 2021. Underwater target detection based on faster R-CNN and adversarial occlusion network. *Eng. Appl. Artif. Intell.* **100** 104190.
[3] Song, P., Li, P., Dai, L., Wang, T., Chen, Z., 2023. Boosting R-CNN: reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **530** 150–164.
[4] Mathias, A., Dhanalakshmi, S., Kumar, R., Narayanamoorthi, R., 2021. Underwater object detection based on bi-dimensional empirical mode decomposition and Gaussian Mixture Model approach. *Ecol. Inform.* **66** 101469.
[5] Zhang, J., Yongpan, W., Xianchong, X., Yong, L., Lyu, L., Wu, Q., 2022. YoloXT: a object detection algorithm for marine benthos. *Ecol. Inform.* **72** 101923.
[6] Yu, G., Cai, R., Su, J., Hou, M., Deng, R., 2023. U-YOLOv7: a network for underwater organism detection. *Ecol. Inform.* **75** 102108.
[7] Liu, Y., An, D., Ren, Y., Zhao, J., Zhang, C., Cheng, J., Liu, J., Wei, Y., 2024b. DP-FishNet: dual-path pyramid vision transformer-based underwater fish detection network. *Expert Syst. Appl.* **238** 122018.
[8] Fan D P, Ji G P, Sun G, Cheng M M, Shen J and Shao L 2020 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (Seattle, WA, USA) pp 2774-84.
[9] Yang F et al 2021 *Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV)* (Montreal, QC,Canada) pp 4126-35.
[10] Le T N, Nguyen T V, Nie Z, et al 2019 *Comput. Vis. Image Underst.* **184** 45-56.
[11] Lv Y, Zhang J, Dai Y, et al 2021 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Nashville, TN, USA) pp 11591-601.
[12] He H, Li X, Cheng G, et al 2021 *Proc. IEEE/CVF Int. Conf. on Computer Vision* (Montreal, QC, Canada) pp 15859-68.
[13] Zhai Q, Li X, Yang F, et al 2021 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Nashville, TN, USA) pp 12997-3007.
[14] Mao Y, Zhang J, Wan Z, et al 2024 *IEEE Trans. Circuits Syst. Video Technol.* [online] Available at: https://doi.org/10.1109/TCSVT.2024.3373642 [Accessed 22 May 2024].
[15] Yang F, Zhai Q, Li X, et al 2021 *Proc. IEEE/CVF Int. Conf. on Computer Vision* (Montreal, QC, Canada) pp 4146-55.
[16] Cong R, Sun M, Zhang S, et al 2023 *Proc. ACM Int. Conf. on Multimedia* (Ottawa, ON, Canada) pp 1179-89.
[17] Sun Y, Xu C, Yang J, et al 2024 *Proc. Eur. Conf. Comput. Vis.* (Milan, Italy) pp 343-60.

[18] Deng M, Sun S, Li Z, et al 2025 *arXiv preprint* [online] Available at: https://arxiv.org/abs/2503.11030 [Accessed 22 May 2024].
[19] Lin J, Tan X, Xu K, et al 2023 *ACM Trans. Multimedia Comput. Commun. Appl.* **19** 1-16.
[20] Huo F, Zhu X, Zhang L, et al 2021 *IEEE Trans. Circuits Syst. Video Technol.* **32** 3111-24.
[21] Mei H, Liu Y, Wei Z, et al 2021 *IEEE Trans. Circuits Syst. Video Technol.* **32** 1378-89.
[22] Siris A, Jiao J, Tam G K L, et al 2021 *Proc. IEEE/CVF Int. Conf. on Computer Vision* (Milan, Italy) pp 4156-66.
[23] Hu H, Bai S, Li A, et al 2021 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Nashville, TN, USA) pp 10185-94.
[24] Bhattacharyya P, Huang C and Czarnecki K 2021 *Proc. IEEE/CVF Int. Conf. on Computer Vision* (Montreal, QC, Canada) pp 3022-31.
[25] Wang W, Xie E, Li X, et al 2022 *Comput. Visual Media* **8** 415-24.
[26] Wu Z, Su L and Huang Q 2019 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Long Beach, CA, USA) pp 3907-16.
[27] Wei J, Wang S and Huang Q 2020 *Proc. AAAI Conf. Artif. Intell.* (New York, NY, USA) pp 12321-8.
[28] Fan D P, Ji G P, Cheng M M, et al 2021 *IEEE Trans. Pattern Anal. Mach. Intell.* **44** 6024-42.
[29] Zhu J, Zhang X, Zhang S, et al 2021 *Proc. AAAI Conf. Artif. Intell.* (Virtual Conference) pp 3599-607.
[30] Li A, Zhang J, Lv Y, et al 2021 *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (Nashville, TN, USA) pp 10071-81.
[31] Chen Z, Xu Q, Cong R, et al 2020 *Proc. AAAI Conf. Artif. Intell.* (New York, NY, USA) pp 10599-606.
[32] Liu Y, Li H, Cheng J, et al 2023 *IEEE Trans. Circuits Syst. Video Technol.* **33** 4934-47.