

EffiSeaNet: Pioneering Lightweight Network for Underwater Salient Object Detection

Qingyao Wu¹, Zhenqi Fu², Hong Lin³, Chenyu Ma³, Xiaotong Tu^{1,3}*, and Xinghao Ding^{1,3}

¹ Institute of Artificial Intelligence, Xiamen University

² Department of Automation, Tsinghua University

³ School of Informatics, Xiamen University

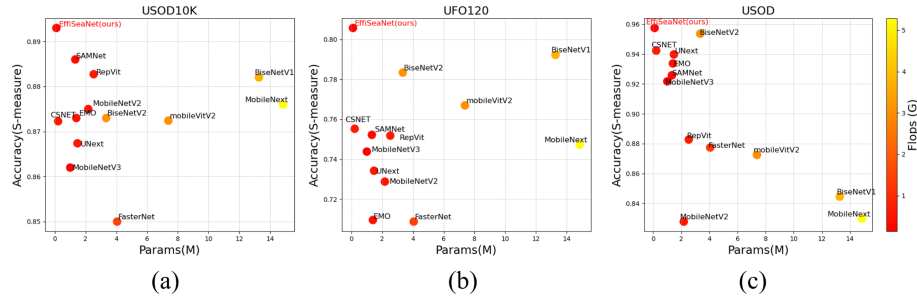
Abstract. Underwater salient object detection seeks to pinpoint the most vital elements in underwater environments, offering considerable promise for underwater exploration. Considering the preference for low-complexity algorithms in underwater applications to maximize overall system efficiency, this paper proposes EffiSeaNet, a lightweight network designed to provide an effective solution for salient object detection in underwater scenarios. On the one hand, EffiSeaNet incorporates a parameter-free image enhancement block to mitigate the effects of image degradation caused by water. This block effectively addresses issues such as color distortion and reduced visibility, which are common challenges in underwater environments. On the other hand, we develop a customized lightweight network structure incorporating a novel cross-layer fusion strategy to efficiently capture and merge features. This enhances the network's ability to handle the variability and complexity of underwater objects and scenes while maintaining a low computational load. Extensive experiments on three public datasets demonstrate that our innovative designs achieve remarkable results while maintaining a low model size and computational complexity. This efficiency and effectiveness make our approach highly suitable for practical underwater applications where resources are limited, yet high precision is essential. Our code and results will be accessible to the public.

Keywords: Underwater Image · Salient Object Detection · Lightweight Model · Feature Fusion

1 Introduction

Underwater salient object detection (USOD) is an emerging subfield of salient object detection (SOD) [21], created to meet the increasing demands of underwater exploration. Despite the heightened focus on this area, USOD research

* This work was supported in part by the Dreams Foundation of Jianghuai Advance Teclology Center (No.2023-ZM01X018), the National Natural Science Foundation of China (Grant Nos. 52105126 and 61971369). (*Corresponding author: Xiaotong Tu, xttu@xmu.edu.cn).



Given the aforementioned challenges, the objective of this paper is to develop an efficient network for real-time salient object detection in underwater environments. To accomplish this objective, we introduce EffiSeaNet, a lightweight encoder-decoder framework via parameter-free image enhancement and multi-scale features fusion. To handle the degraded underwater images, we introduce the parameter-free image enhancement (PEIE) module. This module is based on linear channel normalization, which performs contrast stretching within both the RGB and HSV color spaces of the image. As a result, the enhanced images are more suitable for the USOD task, leading to significant performance improvements. To handle variable underwater objects and scenes while maintaining a low computational cost, we propose the cross-scale feature fusion (CSFF) module, which is composed of three group fusion (GF) blocks. Specifically, the GF conducts channel-wise grouping of feature blocks derived from three adjacent encoding stages. Subsequently, grouped feature blocks are meticulously fused by employing atrous convolutions with corresponding rates. As shown in Fig. 1, by these designs, our EffiSeaNet achieves high detection accuracy while maintaining minimal resource consumption. In summary, our main contributions are as follows:

- We introduce EffiSeaNet, a framework that integrates lightweight encoders and decoders within a multi-scale feature fusion architecture. EffiSeaNet includes a parameter-free image enhancement module aimed at alleviating image degradation effects induced by water.
- We propose a new cross-layer feature fusion module with group interaction and integration mechanisms. This innovative design significantly enhances the network’s capability to represent a wide range of diverse underwater objects and scenes effectively.
- Extensive experiments on three public datasets validate that the proposed network surpasses advanced methods in both computational efficiency and detection accuracy.

2 Related Work

2.1 Salient Object Detection

SOD has made significant progress in the past few decades [2, 19, 48]. MLM-Net [38] takes advantage of the supervision of the foreground boundary and edges for SOD. AADFNet [50] generates local saliency cues by dilated convolutions with a small rate and global saliency cues by dilated convolutions with a large rate. GateNet [46] designs a gated dual branch structure to build cooperation between different levels of characteristics and improve the discriminability of the entire network. U²Net [26] obtains intrastate multiresolution features without reducing feature map resolution. CANet [29] presents a context-aware attention module that detects salient objects by simultaneously constructing connections between each image pixel and its local and global contextual pixels.

EDN [39] uses an extreme down-sampling method to effectively learn global features and Scale-Correlated Pyramid Convolution in the decoder to recover local details. Despite substantial progress in SOD, most current methods are tailored for high-quality terrestrial images. However, underwater scenes suffer from significant image degradation and have specific requirements for model complexity. Consequently, existing SOD methods cannot achieve satisfactory performance when applied directly to underwater environments.

2.2 Underwater Salient Object Detection

Recently, USOD has attracted more and more attention due to its importance and challenges [28]. Existing USOD methods can be divided into traditional method and learning-based methods. The former explores low-level visual features such as color, texture, and contours to obtain saliency maps of underwater images. Such features are commonly manually designed. For example, Cui *et al.* [6] proposes an USOD model based on an enhanced histogram equalization algorithm to address the issue of low visual quality in underwater images. Chen *et al.* [5] proposes a novel USOD method by combining hand-crafted 2D features (color and intensity) and hand-crafted 3D features (depth map). Recently, learning-based methods have become a research hotspot. Islam *et al.* [18] develops an effective solution for saliency-guided visual attention modeling by integrating bottom-up and top-down learning within an encoder-decoder architecture. Islam *et al.* [17] design an end-to-end training pipeline to jointly learn the saliency prediction on a shared hierarchical feature space and present a dataset of UFO-120 with great contribution. Hong *et al.* [12] constructs a large-scale USOD dataset named USOD10K. Meanwhile, the authors design a straight-forward hybrid architecture for USOD. The entire framework is built upon an encoder-decoder structure incorporating transformer and convolutional blocks.

2.3 Efficient Neural Network

In recent years, there has been a surge of research focused on developing lightweight and hardware-efficient convolutional neural networks specifically designed for mobile vision tasks. This trend reflects the growing demand for efficient neural network architectures that can operate effectively on resource-constrained mobile devices. [14, 16]. For instance, MobileNet [13, 14, 32] propose separable convolution and inverse residual bottlenecks. The former acts as a substitute for traditional convolution, significantly lowering computational demands. Meanwhile, the latter addresses the problem of vanishing gradients. Together, these approaches effectively reduce the parameter count while enhancing model efficiency. ShuffleNet [23, 45] which introduces techniques such as pointwise group convolution and channel shuffle, aimed at reducing network capacity and enhancing lightweightness. BiSeNet [40, 41] leverages dual-branch paths to capture low-level details and high-level context information to reduce model complexity while maintaining model performance. MixNet [33] proposes a mixed depth-wise convolution approach that incorporates multiple-sized convolution kernels

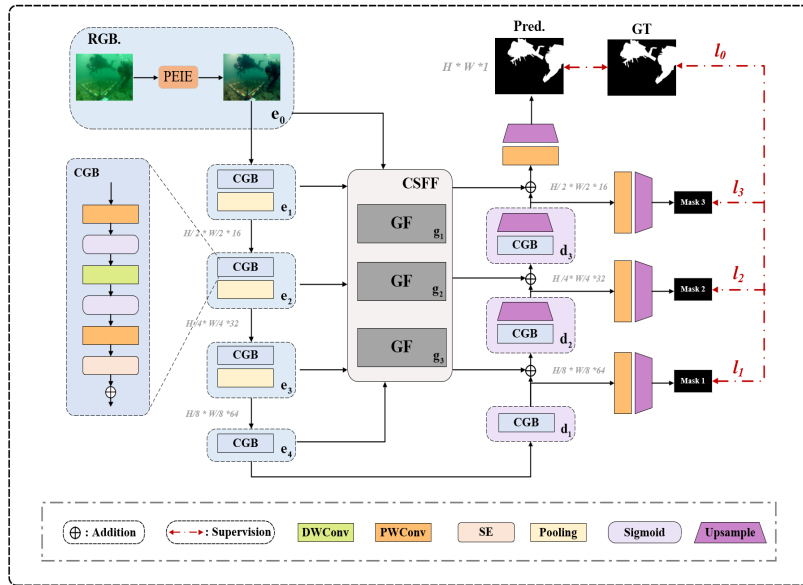


Fig. 2: The detailed architecture of EffiSeaNet. First, the parameter-free image enhancement module (PFIE) is employed to enhance the input image quality, making it more suitable for the task at hand. Then, four encoders ($\{e_0 \rightarrow e_4\}$) are utilized to extract multi-scale representations. Following this, the cross-layer feature fusion Module (CSFF) refines the learned features. Finally, the decoder ($\{d_1 \rightarrow d_3\}$) gradually upsamples the hierarchical features. DWConv refers to Depth-Wise Convolution. PWConv denotes Point-Wise Convolution. SE refers to squeeze-and-excitation module [15].

to enhance accuracy and reduce model complexity. EGE-UNet [31] groups input features and performs Hadamard Product Attention mechanism (HPA) on different axes to extract pathological information from diverse perspectives. In our work, we draw inspiration from the above-mentioned methods, but our model employs these methods uniquely.

3 Method

In this section, we begin by introducing the overall architecture of EffiSeaNet. Following this, we present two key components: the parameter-free image enhancement module and the cross-layer feature fusion module. Lastly, we provide a detailed explanation of the loss functions used to train our pipeline.

3.1 Overall Structure

As shown in Fig. 2, EffiSeaNet is built on a U-shaped architecture. The encoder and decoder are composed of lightweight modules called CGB, which consists of

several PWConv, DWConv, SE, and Sigmoid layers. In the encoder part, each CGB downsamples (Maxpooling) the feature by a factor of 1/2. Correspondingly, each CGB upsamples (Bilinear Interpolation) the feature by a factor of 2 in the decoder part. The core design of EffiSeaNet is the PFIE and CSFF modules. We use PEIE to enhance the original underwater image (regarded as e_0). The purpose of this process is to reduce the impact of image degradation. The multi-scale features learned from $\{e_0\}$ and three CGB $\{e_1 \rightarrow e_4\}$ are inputted into CSFF for feature interaction and fusion. The processing of the proposed CSFF can be expressed as:

$$g_1, g_2, g_3 = CSFF(\{e_0 \rightarrow e_4\}), \quad (1)$$

where, g_1, g_2, g_3 represent the outputs of CSFF. The decoder is mirror-aligned with the encoder, creating a symmetrical structure that facilitates the generation of multi-scale masks. Specifically, we use PWConv and Upsample layers to produce the salient mask at each scale. Finally, we calculate the loss functions and optimize the entire network based on these multi-scale predictions and the ground-truth map.

3.2 Parameter-Free Image Enhancement

In order to reduce the impact of underwater image degradation on salient object detection, we propose a parameter-free pre-processing method based on the channel linear normalization principle. First, we normalize and stretch the contrast of the image color in channel dimensions of R, G, and B. Then, the optimized *RGB* image is converted into the *HSV* space. We perform a same normalization in the three dimensions of hue, saturation, and value. Stretching saturation and value can generate elements with a wider color range, while hue can increase the true color of the image and solve lighting problems. Channel linear normalization in RGB and HSV color spaces enhances the contrast and color of the degraded underwater image. Mathematically, the above normalization can be defined as:

$$f_n = \frac{p^i - p^{min}}{p^{max} - p^{min}}, \quad (2)$$

where i refers to the index of image pixels, p^{max} and p^{min} refer to the maximum and minimum values in each channel, respectively. Finally, we convert the enhanced image to RGB space:

$$e_0 = RGB(f_n(HSV(f_n(x)))), \quad (3)$$

where e_0 refers to the final output of PEIE which we regard as first encoder block, and $RGB(x)$ and $HSV(x)$ are the corresponding image space conversion projection. PEIE is simple yet effective. With PEIE, our method improves the S-measure [8], mean E-measure [9], weighted F-measure [24], and mean absolute error with large margins (see Tab. 3).

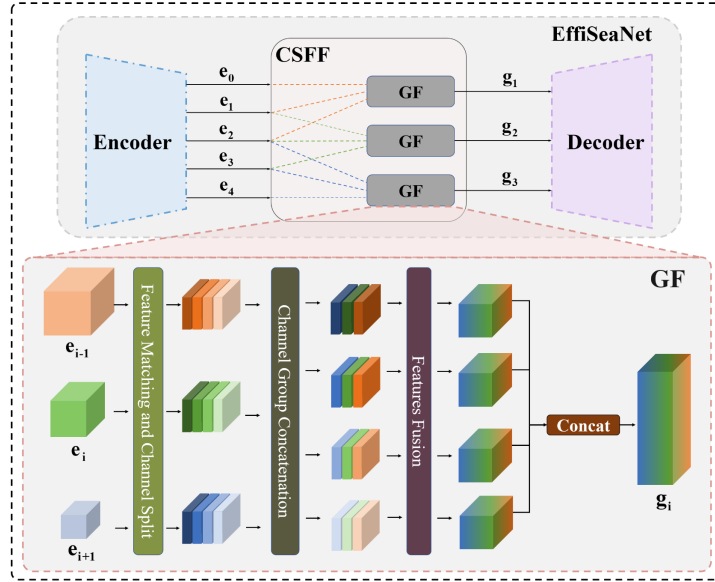


Fig. 3: Overview of cross-scale feature fusion (CSFF) and group fusion (GF) block.

3.3 Cross-Scale Feature Fusion

To address the variability and complexity of underwater objects and scenes while maintaining a low computational load, we propose a cross-scale feature fusion (CSFF) module and group fusion (GF) blocks. As depicted in Fig. 3, CSFF is responsible for receiving multi-scale features from $\{e_0 \rightarrow e_4\}$. The internal GF is responsible for fusing the neighbor features. More specifically, GF initially selects three neighboring scale features. It then uses upsampling for high-level features and downsampling for low-level features to match their sizes with the middle-level features.

Then, GF splits each matched feature into four groups along the channel dimension:

$$f_1^k, f_2^k, f_3^k, f_4^k = \text{split}(f^k), \quad k \in \{e_{i-1}, e_i, e_{i+1}\}, \quad (4)$$

where e_i denotes the encoder and f^k refers to the matched feature map of the encoder. Next, it concatenate the adjacent three levels of splited feature blocks in each group as:

$$s_j = \text{Concat}(f_j^k), \quad (5)$$

where $j \in \{1, 2, 3, 4\}$ is the group index. Subsequently, we perform feature fusion by layer normalization (LN) [1] and dilated convolutions (DConv) [42] with a kernel size of 3 and dilation rates of $\{1, 2, 5, 7\}$ with corresponding to $\{s_1 \rightarrow s_4\}$ to obtain four fused feature maps:

$$F_j^i = \text{DConv}_j(s_j), \quad (6)$$

where $j \in \{1, 2, 3, 4\}$ denotes the group index. To further facilitate information interaction between features of diverse scales, we employ point-wise convolution (PWConv) [14] to process the four feature maps concatenated along the channel dimension, yielding an output feature map that incorporates multi-scale interaction information. This process is represented in the Fig. 2 as Feature Fusion, so there is:

$$g_i = PWConv(Concat(F_j^i)), \quad (7)$$

where g_i denotes the output of the GF module.

3.4 Loss Function

We employ binary cross-entropy and dice loss [20] to optimize our network. Binary cross-entropy evaluates the performance by calculating the difference between the predicted and actual binary labels, penalizing incorrect predictions. Dice loss, on the other hand, compares the similarity of two binary classifications of an image by measuring the overlap between the predicted and ground-truth masks. In this work, binary cross-entropy loss is defined as:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_i^N y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)), \quad (8)$$

where i represents the index of the pixel, $p(y_i)$ denotes the predicted value of the pixel, and y_i represents the corresponding ground-truth. Dice loss is expressed as:

$$\mathcal{L}_{Dice} = 1 - \frac{2 \sum_i^N y_i p(y_i) + \epsilon}{\sum_i^N (y_i)^2 + \sum_i^N (p(y_i))^2 + \epsilon}, \quad (9)$$

where ϵ is a small positive value. Its purpose is to avoid division by zero in the calculation process. Here we choose a value of e^{-7} . To generate an accurate mask, the overall loss function \mathcal{L} can be expressed as:

$$\mathcal{L} = \sum_{s=0}^3 \lambda_s \times \ell_s, \quad (10)$$

where $\ell_s = \mathcal{L}_{BCE}^s + \mathcal{L}_{Dice}^s$, s denotes the scale index, λ_s represents the weight. We set $\lambda_0 = 1$, $\lambda_1 = 0.5$, $\lambda_2 = 0.4$, and $\lambda_3 = 0.3$, empirically.

4 Experiments

4.1 Experimental Settings

Datasets. We train the proposed model on the USOD10K training set [12], which includes 7,178 underwater images with pixel-level annotations. These images are sourced from 12 distinct underwater scenes. The salient objects in the USOD10K dataset are categorized into 70 classes, encompassing fish, ruins,

Table 1: Accuracy comparisons on three datasets. The best results are highlighted in red, and the second-best results are indicated in blue.

Method	USOD10k [12]				UFO120 [17]				USOD [18]			
	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE
BiseNetV1 [41]	0.892	0.929	0.855	0.031	0.792	0.859	0.766	0.118	0.844	0.910	0.814	0.064
BiseNetV2 [40]	0.883	0.913	0.829	0.038	0.783	0.843	0.764	0.122	0.954	0.983	0.957	0.013
MobileNetV1 [14]	0.822	0.899	0.771	0.052	0.701	0.811	0.677	0.149	0.825	0.887	0.803	0.098
MobileNetV2 [32]	0.875	0.934	0.848	0.036	0.729	0.819	0.695	0.139	0.828	0.908	0.786	0.077
MobileNetV3 [13]	0.862	0.929	0.828	0.037	0.744	0.824	0.704	0.136	0.922	0.971	0.922	0.024
MobileNeXt [47]	0.876	0.940	0.854	0.034	0.747	0.818	0.725	0.132	0.830	0.920	0.800	0.062
CSNet [10]	0.872	0.926	0.852	0.937	0.755	0.783	0.663	0.148	0.924	0.977	0.943	0.019
SAMNet [22]	0.886	0.938	0.861	0.032	0.752	0.813	0.719	0.126	0.926	0.975	0.920	0.024
UNeXt [34]	0.801	0.872	0.712	0.057	0.734	0.803	0.692	0.142	0.940	0.971	0.933	0.021
MobileVitV2 [25]	0.872	0.915	0.798	0.043	0.767	0.766	0.671	0.160	0.948	0.971	0.939	0.022
EMO [43]	0.873	0.933	0.842	0.035	0.709	0.783	0.663	0.148	0.934	0.976	0.934	0.021
FasterNet [3]	0.850	0.917	0.809	0.043	0.709	0.785	0.670	0.156	0.877	0.931	0.919	0.036
FastVit [35]	0.889	0.934	0.840	0.034	0.762	0.802	0.714	0.134	0.958	0.977	0.954	0.017
RepVit [36]	0.883	0.937	0.842	0.032	0.752	0.818	0.726	0.135	0.944	0.979	0.947	0.017
EffiSeaNet	0.893	0.935	0.872	0.030	0.806	0.839	0.787	0.129	0.967	0.977	0.960	0.014

divers, underwater robots, and more. We evaluate the proposed method on three datasets including test set of USOD10K [12], UFO120 [17] and USOD [22]. These datasets consist of 1,026, 1500 and 300 underwater images with corresponding pixel-level labels, respectively.

Implementation Details. Our model is developed with the PyTorch framework, and all experiments are performed on a single NVIDIA RTX 3080Ti GPU. During the training phase, the input image is resized to 256×256 . We apply various data augmentation methods, such as horizontal flip, vertical flip, and random rotation to increase the diversity. We use Adam as the optimizer and Multi-Step-LR as the learning rate scheduler. The model is trained for a total of 800 epochs. After the first 400 epochs, the learning rate is reduced by 50%. The batch size is fixed at 16, and no pretraining strategies involving additional datasets like ImageNet [7] are employed.

Evaluation Metrics. We use S-measure (S_m) [8], mean E-measure (E_ϕ) [9], weighted F-measure (F_m^w) [24], and mean absolute error (MAE) to objectively evaluate each method. Among them, S-measure calculates the structural similarity between the predicted saliency map and the ground-truth map. Mean E-measure computes the similarity for the binarized predicted map and the binary

Table 2: Efficiency comparisons. The best results are highlighted in red, and the second-best results are indicated in blue.

Method	Params (M)	FLOPs (G)
BiseNetV1 [41]	13.27	3.716
BiseNetV2 [40]	3.340	3.083
MobileNetV1 [14]	13.68	28.901
MobileNetV2 [32]	2.173	0.519
MobileNetV3 [13]	1.002	0.136
MobileNeXt [47]	14.84	5.289
CSNet [10]	0.780	0.610
SAMNet [22]	1.330	0.329
UNeXt [34]	1.472	0.573
MobileVitV2 [25]	7.380	3.009
EMO [43]	1.397	0.493
FasterNet [3]	4.052	1.324
FastVit [35]	7.551	2.201
RepVit [36]	2.527	0.645
EffiSeaNet	0.101	0.237

ground-truth map. Weighted F-measure solves the problems of F-measure that may cause interpolation flaw, dependency flaw, and equal-importance flaw [24]. Since underwater applications favor lightweight models, we calculate model parameters (M) and FLOPs (G) to show the effectiveness of our solution.

4.2 Comparison with Other Methods

Quantitative Comparison. Tab. 1 presents the accuracy comparisons on three datasets. 14 representative algorithms are selected as the competitors. It’s important to note that there are few publicly available USOD methods, and we do not include RGBD-related approaches [12]. As observed, EffiSeaNet achieves promising performance with the best S_m , F_m^w , and MAE scores in USOD10K, the best S_m and F_m^w scores in UFO120, and the best S_m and F_m^w scores in USOD. Among the competitors, BiseNetV1 and BiseNetV2 achieve relatively high performance due to their specific fusion mechanisms. These models effectively learn rich spatial information with a sizable receptive field. We further compare the efficiency of all methods in Tab. 2. From the table, it is evident that the proposed EffiSeaNet is highly efficient. EffiSeaNet comprises approximately 0.1 million parameters and 0.237 billion FLOPs, which are significantly lower than those of the other methods. In summary, Tab. 1 and Tab. 2 demonstrate that our method effectively balances computational complexity and segmentation performance.

Qualitative Comparison. Fig. 4 presents qualitative evaluation results. For a comprehensive comparison, we chose three samples from each dataset, each

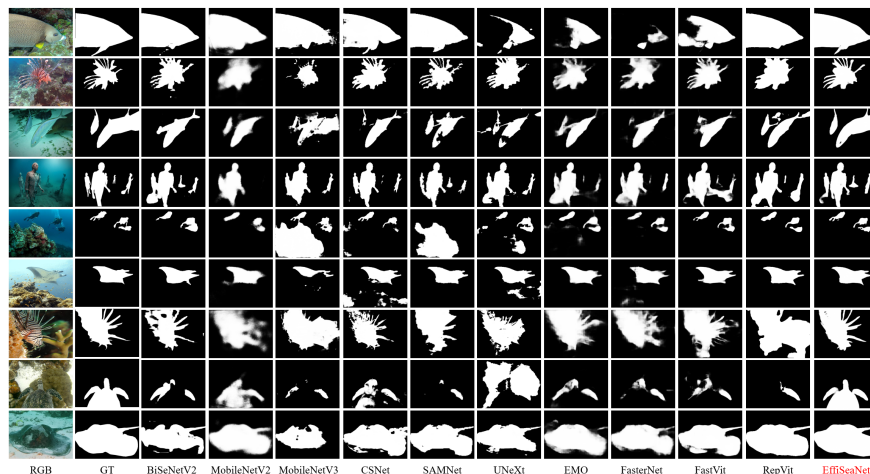


Fig. 4: Qualitative comparisons on USOD10K (first three rows), UFO120 (middle three rows) and USOD (last three rows).

Table 3: The role of CSFF and PEIE. The the best results are highlighted.

Setting		USOD10k [12]				UFO120 [17]				USOD [18]			
CSFF	PEIE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE
✗	✗	0.853	0.914	0.793	0.046	0.700	0.760	0.644	0.163	0.891	0.904	0.894	0.032
✗	✓	0.862	0.842	0.802	0.044	0.712	0.766	0.688	0.156	0.918	0.923	0.933	0.029
✓	✗	0.875	0.929	0.845	0.037	0.787	0.814	0.734	0.139	0.947	0.966	0.933	0.021
✓	✓	0.893	0.935	0.872	0.030	0.806	0.839	0.766	0.129	0.967	0.977	0.960	0.014

with different objects and distortions. EffiSeaNet consistently demonstrates superior performance across all selected samples. Specifically, none of the methods accurately detect salient objects in the image in the eighth row. UNeXt, EMO, and FasterNet fail to handle the image in the first row adequately, while MobileNetV3 and SAMNet struggle with the image in the fifth row. Compared to ground truth (GT), competing methods exhibit inaccurate boundaries and incomplete object representations, whereas our method’s saliency maps are more accurate and complete.

4.3 Ablation Studies

The role of CSFF and PEIE. CSFF and PEIE are to core modules in our EffiSeaNet. In this subsection, we conduct ablation studies to understand the role of these two modules. The results are listed in Tab. 3. Observationally, when both CSFF and PEIE are simultaneously removed (leaving the network similar to UNet [30]), there is a noticeable decline in performance, underscoring

Table 4: Impact of different scale features. The the best results are highlighted. ‘Low’, ‘Mid’, and ‘High’ denotes three neighbor scale features.

Setting			USOD10k [12]				UFO120 [17]				USOD [18]			
Low	Mid	High	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE
✓	✗	✗	0.850	0.904	0.784	0.046	0.728	0.752	0.679	0.148	0.894	0.914	0.874	0.030
✗	✓	✗	0.851	0.905	0.785	0.048	0.710	0.778	0.660	0.153	0.873	0.917	0.892	0.026
✗	✗	✓	0.861	0.904	0.778	0.048	0.730	0.780	0.680	0.144	0.892	0.897	0.916	0.027
✗	✓	✓	0.880	0.928	0.839	0.039	0.772	0.790	0.701	0.134	0.922	0.927	0.921	0.020
✓	✗	✓	0.871	0.924	0.818	0.038	0.760	0.791	0.698	0.145	0.923	0.947	0.922	0.021
✓	✓	✗	0.881	0.921	0.818	0.036	0.770	0.805	0.729	0.132	0.927	0.942	0.926	0.020
✓	✓	✓	0.893	0.935	0.872	0.030	0.806	0.839	0.766	0.129	0.967	0.977	0.960	0.014

Table 5: Impact of the fusion mechanism. The bold indicates the best performance.

Method	USOD10k [12]				UFO120 [17]				USOD [18]			
	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE
Baseline	0.851	0.905	0.785	0.048	0.710	0.778	0.660	0.153	0.873	0.917	0.892	0.026
ASPP [4]	0.856	0.915	0.798	0.045	0.696	0.771	0.648	0.163	0.919	0.967	0.915	0.028
CSFF	0.893	0.935	0.872	0.030	0.806	0.839	0.766	0.129	0.967	0.977	0.960	0.014

the effectiveness of CSFF and PEIE. Removing either CSFF or PEIE individually results in the network failing to achieve satisfactory results across all three datasets. This is because PEIE is designed to reduce the impact of image degradation caused by water, while CSFF efficiently captures and merges features. Both of these modules are essential for this task.

Impact of Different Scale Features in CSFF. Tab. 4 presents the ablation studies of different scale features in CSFF. Specifically, we have tested six variants. As shown in Tab. 4, in the first three variants, each GF within CSFF receives only a single scale of features. In the last three variants, each GF within CSFF receives two scale features. From the table, we can make the following observations: 1) Without the feature fusion strategy, our model cannot achieve satisfactory performance due to the complexity of underwater scenes and the diversity of underwater objects; 2) Using two scale features from three neighboring scale features can significantly enhance the network’s ability to capture salient objects; 3) The proposed CSFF achieves the best results as our solution thoroughly incorporates features of different scales, enhancing the network’s representation ability.

Impact of the Fusion Mechanism. In this section, we explore the role of the fusion strategy. Conceretly, we replace our CSFF with vanilla convolution

Table 6: Effectiveness of auxiliary losses. The bold indicates the best performance.

Setting	USOD10k [12]				UFO120 [17]				USOD [18]			
	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE	S_m	E_ϕ	F_m^w	MAE
ℓ_0	.875	.913	.852	.038	0.722	0.781	0.635	0.144	0.927	0.968	0.924	0.025
$\ell_0 + \ell_1$	0.882	0.927	0.865	0.032	0.737	0.802	0.685	0.138	0.933	0.972	0.932	0.023
$\ell_0 + \ell_1 + \ell_2$	0.885	0.929	0.867	0.031	0.746	0.804	0.703	0.134	0.935	0.974	0.936	0.022
$\ell_0 + \ell_1 + \ell_2 + \ell_3$	0.893	0.935	0.872	0.030	0.806	0.839	0.766	0.129	0.967	0.977	0.960	0.014

(baseline) and atrous spatial pyramid pooling (ASPP) [4]. The results are reported in Tab. 5. We can observe that our method significantly outperforms the baseline. This demonstrates that a specific fusion strategy is necessary as it enhances the efficiency of feature utilization. Compared to ASPP, the proposed fusion mechanism introduces a more efficient, adaptive approach to capturing multi-scale contextual information, significantly enhancing the performance on underwater image saliency detection tasks. Furthermore, our method is tailored to address the specific challenges of underwater imagery, such as color distortion and obscured details, ensuring robust performance in challenging conditions.

Effectiveness of Auxiliary Losses. In this paper, auxiliary losses are designed to guide the EffiSeaNet in generating an accurate binary mask of the salient object. Here, we conduct ablation studies to understand the role of these losses. As shown in Tab. 6, the auxiliary losses can improve the USOD performance to a certain extent. Specifically, EffiSeaNet uses three scale auxiliary losses (i.e., $\ell_0 + \ell_1 + \ell_2 + \ell_3$) achieves the best performance on all datasets. The reason is that it offers robust supervision across various scales, which is beneficial for learning discriminative representations.

5 Conclusion

This paper proposes EffiSeaNet, a lightweight network for underwater salient object detection. EffiSeaNet starts by utilizing a parameter-free image enhancement block to mitigate the impact of image degradation in underwater conditions, addressing challenges like color distortion and reduced visibility. To manage the variability and complexity of underwater scenes efficiently and at low computational costs, we introduce a specialized lightweight network structure. This structure incorporates a novel cross-layer fusion strategy designed to effectively capture and integrate features across different layers of the network. Experimental results on three public datasets demonstrate that EffiSeaNet achieves a superior balance between computational cost and detection accuracy and consistently outperforms existing methods. In the future, we plan to design more sophisticated and lightweight network architectures for the detection of salient objects underwater.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Borji, A., Cheng, M.M., Hou, Q., Jiang, H., Li, J.: Salient object detection: A survey. *Computational visual media* **5**, 117–150 (2019)
3. Chen, J., Kao, S.h., He, H., Zhuo, W., Wen, S., Lee, C.H., Chan, S.H.G.: Run, don't walk: Chasing higher flops for faster neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12021–12031 (2023)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 34, pp. 10599–10606 (2020)
6. Cui, Z., Wu, J., Yu, H., Zhou, Y., Liang, L.: Underwater image saliency detection based on improved histogram equalization. In: *Data Science: 5th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2019, Guilin, China, September 20–23, 2019, Proceedings, Part II 5*. pp. 157–165. Springer (2019)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. pp. 248–255. Ieee (2009)
8. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: *Proceedings of the IEEE international conference on computer vision*. pp. 4548–4557 (2017)
9. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
10. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. In: *European Conference on Computer Vision*. pp. 702–721. Springer (2020)
11. Girdhar, Y., Giguere, P., Dudek, G.: Autonomous adaptive exploration using real-time online spatiotemporal topic modeling. *The International Journal of Robotics Research* **33**(4), 645–657 (2014)
12. Hong, L., Wang, X., Zhang, G., Zhao, M.: Usod10k: A new benchmark dataset for underwater salient object detection. *IEEE Transactions on Image Processing* pp. 1–1 (2023). <https://doi.org/10.1109/TIP.2023.3266163>
13. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 1314–1324 (2019)
14. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018)
16. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)

17. Islam, M.J., Luo, P., Sattar, J.: Simultaneous enhancement and super-resolution of underwater imagery for improved visual perception. arXiv preprint arXiv:2002.01155 (2020)
18. Islam, M.J., Wang, R., Sattar, J.: Svam: saliency-guided visual attention modeling by autonomous underwater robots. arXiv preprint arXiv:2011.06252 (2020)
19. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* **20**(11), 1254–1259 (1998)
20. Li, X., Sun, X., Meng, Y., Liang, J., Wu, F., Li, J.: Dice loss for data-imbalanced nlp tasks. arXiv preprint arXiv:1911.02855 (2019)
21. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4722–4732 (2021)
22. Liu, Y., Zhang, X.Y., Bian, J.W., Zhang, L., Cheng, M.M.: Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Transactions on Image Processing* **30**, 3804–3814 (2021)
23. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 116–131 (2018)
24. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 248–255 (2014)
25. Mehta, S., Rastegari, M.: Separable self-attention for mobile vision transformers. arXiv preprint arXiv:2206.02680 (2022)
26. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition* **106**, 107404 (2020)
27. Reggiannini, M., Moroni, D.: The use of saliency in underwater computer vision: A review. *Remote Sensing* **13**(1), 22 (2020)
28. Reggiannini, M., Moroni, D.: The use of saliency in underwater computer vision: A review. *Remote Sensing* **13**(1), 22 (2020)
29. Ren, Q., Lu, S., Zhang, J., Hu, R.: Salient object detection by fusing local and global contexts. *IEEE Transactions on multimedia* **23**, 1442–1453 (2020)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. pp. 234–241. Springer (2015)
31. Ruan, J., Xie, M., Gao, J., Liu, T., Fu, Y.: Ege-unet: an efficient group enhanced unet for skin lesion segmentation. arXiv preprint arXiv:2307.08473 (2023)
32. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4510–4520 (2018)
33. Tan, M., Le, Q.V.: Mixnet: Mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595 **43** (2019)
34. Valanarasu, J.M.J., Patel, V.M.: Unext: Mlp-based rapid medical image segmentation network. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 23–33. Springer (2022)
35. Vasu, P.K.A., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: Fastvit: A fast hybrid vision transformer using structural reparameterization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5785–5795 (2023)

36. Wang, A., Chen, H., Lin, Z., Han, J., Ding, G.: Repvit: Revisiting mobile cnn from vit perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15909–15920 (2024)
37. Wei, J., Wang, S., Huang, Q.: F³net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12321–12328 (2020)
38. Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., Ding, E.: A mutual learning method for salient object detection with intertwined multi-supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8150–8159 (2019)
39. Wu, Y.H., Liu, Y., Zhang, L., Cheng, M.M., Ren, B.: Edn: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing* **31**, 3125–3136 (2022)
40. Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision* **129**, 3051–3068 (2021)
41. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
43. Zhang, J., Li, X., Li, J., Liu, L., Xue, Z., Zhang, B., Jiang, Z., Huang, T., Wang, Y., Wang, C.: Rethinking mobile block for efficient attention-based models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1389–1400 (2023)
44. Zhang, W., Jiang, Y., Fu, K., Zhao, Q.: Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
45. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)
46. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: A simple gated network for salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 35–51. Springer (2020)
47. Zhou, D., Hou, Q., Chen, Y., Feng, J., Yan, S.: Rethinking bottleneck structure for efficient mobile network design. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 680–697. Springer (2020)
48. Zhou, T., Fan, D.P., Cheng, M.M., Shen, J., Shao, L.: Rgb-d salient object detection: A survey. *Computational Visual Media* **7**, 37–69 (2021)
49. Zhou, T., Fu, H., Chen, G., Zhou, Y., Fan, D.P., Shao, L.: Specificity-preserving rgb-d saliency detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4681–4691 (2021)
50. Zhu, L., Chen, J., Hu, X., Fu, C.W., Xu, X., Qin, J., Heng, P.A.: Aggregating attentional dilated features for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* **30**(10), 3358–3371 (2019)