Check for updates

Received: 25 December 2023 | Revised: 3 June 2024 | Accepted: 5 June 2024

IET Image Processing

DOI: 10.1049/ipr2.13161

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A lightweight underwater fish image semantic segmentation model based on U-Net

Zhenkai Zhang[1] | Wanghua Li[1] | Boon-Chong Seet[2]

[1]Ocean College, Jiangsu University of Science and Technology, Zhenjiang, People's Republic of China

[2]Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand

Correspondence
Zhenkai Zhang, Ocean College, Jiangsu University of Science and Technology, Zhenjiang, 212100, People's Republic of China.
Email: zhangzhenkai@just.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61871203

Abstract

Semantic segmentation of underwater fish images is vital for monitoring fish stocks, assessing marine resources, and sustaining fisheries. To tackle challenges such as low segmentation accuracy, inadequate real-time performance, and imprecise location segmentation in current methods, a novel lightweight U-Net model is proposed. The proposed model acquires more segmentation details by applying a multiple-input approach at the first four encoder levels. To achieve both lightweight and high accuracy, a multi-scale residual structure (MRS) module is proposed to reduce parameters and compensate for the accuracy loss caused by the reduction of channels. To improve segmentation accuracy, a multi-scale skip connection (MSC) structure is further proposed, and the convolution block attention mechanism (CBAM) is introduced at the end of each decoder level for weight adjustment. Experimental results demonstrate a notable reduction in model volume, parameters, and floating-point operations by 94.20%, 94.39%, and 51.52% respectively, compared to the original model. The proposed model achieves a high mean intersection over union (mIOU) of 94.44%, mean pixel accuracy (mPA) of 97.03%, and a frame rate of 43.62 frames per second (FPS). With its high precision and minimal parameters, the model strikes a balance between accuracy and speed, making it particularly suitable for underwater image segmentation.

1 | INTRODUCTION

Marine resources are critical to the sustenance and progress of humanity [1]. Fishery plays a significant role in contributing to both economic growth and nutritional well-being for numerous countries globally, serving as a vital source of sustenance and livelihood for millions of individuals. However, the depletion of marine fish stocks is becoming increasingly concerning due to factors such as overfishing and climate change [2].

In recent years, underwater detection technology has developed vigorously, and underwater robots with powerful vision systems have emerged [3–5], which can provide high-precision detection and analysis of marine organisms and marine environment, as well as important information for humans to deeply understand marine resources and environment. At present, underwater vision research mainly focuses on two directions: underwater image enhancement [6–8] and underwater target detection [9–11], aiming at improving the quality of underwater images and achieving accurate positioning and identification of targets of interest. However, there are two main problems. Firstly, most of the existing object detection models have large parameters and size, which are not suitable for deployment on resource-constrained underwater vehicles. Secondly, object detection only provides the rectangular bounding box of the object, which cannot provide more detailed information such as the object contour. In order to cope with the above problems, it is necessary to design an underwater target segmentation model that can achieve a balance between lightweight and segmentation accuracy. The target semantic segmentation technology can effectively distinguish the target from the background. Hence, it is particularly suitable for scenes that require more detailed target description and more accurate target positioning.

Unlike land environment, underwater image features are difficult to extract, and it is difficult to accurately recognize

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2024 The Author(s). *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

and segment underwater images due to the blurring, texture distortion and size change caused by water flow, surge and light in the underwater environment. To cope with this challenge, some researchers have designed artificial lateral line systems (ALLS) by evolving the lateral line systems (LLS) with underwater organisms (such as fish) to achieve underwater sensing and detection capabilities similar to natural LLS [12]. Other researchers have used traditional segmentation techniques including threshold segmentation [13], edge detection [14] and region growth to achieve segmentation of underwater images through hand-designed features. In recent years, with the development of deep learning, image segmentation techniques have been applied in many fields [15–17, 30], particularly in the field of medical imaging. In 2015, Long et al. [18] proposed the renowned fully convolutional network (FCN) architecture, which revolutionized the field of semantic-level image segmentation by allowing pixel-level classification and end-to-end training. This approach simplified the segmentation process and effectively addressed the challenge of accurately segmenting images at the semantic level. In the same year, Ronneberger et al. [19] enhanced FCN by proposing a U-Net network with a symmetric encoding and decoding structure that achieved remarkable results in medical image segmentation tasks. Building upon the success of U-Net, several subsequent researchers have introduced various network architectures, including R2U-Net [20], MultiResUNet [21], U-Net+ [22], TransUNet+ [23], and Attention UNet [24], etc. In 2017, SegNet [25] further advanced the symmetric encoding and decoding structure by utilizing minimal data to preserve index values while mapping low-resolution features to input resolution for accurate boundary feature localization. From 2016 to 2018, Chen et al. sequentially introduced DeepLab v1, v2, v3 and v3+ [26–29]. Among them, DeepLab v1 utilized atrous convolution and fully connected conditional random field (CRF) operations to improve the accuracy of segmentation. However, this approach posed a potential drawback of sacrificing the preservation of detailed information present in the images. DeepLab v2 further advanced the previous research by incorporating an atrous spatial pyramid pooling (ASPP) module, which enables the processing of objects at multiple scales. DeepLab v3 integrates global average pooling and 1×1 convolutional layers into the ASPP module, allowing for the effective handling of multi-scale segmentation targets and enhancing the performance of semantic image segmentation. DeepLab v3+ has introduced an ASPP module that captures multi-scale contextual semantic information, in addition to a decoding module that refines boundary segment levels. In 2023, Zhou et al. [30] proposed a cross-level feature aggregation network (CFA-Net), which uses techniques such as boundary aggregation to capture context information and cross-layer feature fusion to improve segmentation performance. Although these image segmentation techniques have made numerous advances in the medical field, they cannot be directly applied to underwater environment image segmentation.

In recent years, more researchers have begun to shift their attention to the field of fish image semantic segmentation in underwater environment. By learning from the existing segmentation network ideas and combining the characteristics of underwater images, many network models suitable for underwater fish video or image segmentation have been developed. Labao et al. [31] proposed a ResNet-FCN semantic segmentation network for underwater videos, which can segment fish targets only through colour-based input features without motion cues. However, due to the excessive number of fully convolutional residual network layers used, the model parameters are huge. Nezla et al. [32] also tried to apply image semantic segmentation to underwater target exploration. Based on the U-Net network, the model was trained and fine-tuned with optimal hyperparameters, and achieved an average IoU score of 0.8583 segmentation accuracy, without considering the issue of making the model lightweight. Garcia et al. [33] proposed a single fish detection system that directly placed the vision system in the trawl, aiming at reducing the amount of small and medium-sized fish caught. They achieved good segmentation performance by using the Mask R-CNN model for localization and segmentation and adopting gradient optimization to improve the fish boundary contour, but it still needs to be optimized in terms of segmentation speed. Zhang et al. [34] proposed a dual pooling-aggregated attention network (DPANet). Through pooling-aggregation position and channel attention modules, the context semantic features are adaptively fused, and good segmentation performance is achieved. However, the backbone network parameters chosen by the authors are large. Abe et al. [35] used SegNet network to detect fish at a pixel level. Although it can distinguish foreground and background well, it has the problem of gradient disappearance and needs to train a large number of parameters. Yang et al. [36] proposed a fish-feeding segmentation method (FFSS-Net) to achieve a balance between speed and accuracy performance. However, due to the high fish density in the aquatic environment, the segmentation edge is not accurate enough. In addition, extremely imbalanced datasets also lead to limited segmentation performance [37]. Although these methods have made some progress, they have some limitations and do not consider making the model lightweight and improving segmentation accuracy at the same time. Therefore, there is still considerable room for improvement in the field of underwater fish image semantic segmentation.

In order to better adapt to resource-constrained underwater robots and other detection devices, this paper proposes an image segmentation method based on a lightweight U-Net network to segment fish targets in underwater images. Specifically, we first design a multiple-input method in the first four encoder levels to obtain the optimal number of down-sampling through experiments to obtain more segmentation details. Next, in order to make the model achieve a balance between lightweight and high accuracy, we propose an MRS module with strong feature extraction ability, which not only reduces the number of parameters and calculations but also makes up for the accuracy loss caused by channel reduction. In addition, in order to better utilize the feature information obtained by the encoder, we introduce an MSC structure module to fuse the features of different scales. At the same time, the CBAM attention mechanism is introduced at the end of each decoder level for weight adjustment to improve the segmentation accuracy. Finally, the

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

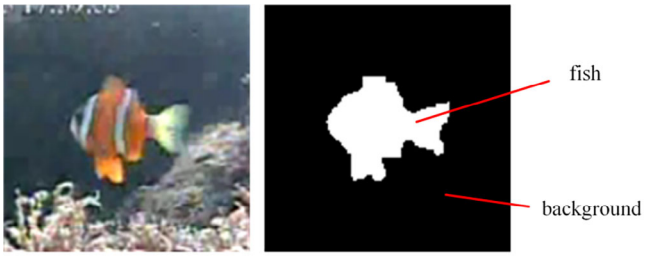

FIGURE 1 Example original and label images from the dataset.

proposed model is verified using a dataset of marine fish images in a real environment. Our contributions in this paper are as follows:

(1) We propose a U-Net based lightweight semantic segmentation model for underwater fish image segmentation, which can achieve a balance between the model weight and segmentation accuracy while ensuring the real-time performance of the network.
(2) We design a multiple-input method in the first four encoder levels, which can be used to compensate for the details lost by multiple down-sampling.
(3) We propose a novel MRS feature extraction block, which can not only make the model lightweight, but also extract more detailed features.
(4) We propose an MSC structure to intelligently combine different levels of semantic information and avoid redundancy, which can segment the object more accurately.

The rest of this paper is organized as follows. Section 2 describes the dataset used in this paper and the proposed lightweight segmentation model in detail. Section 3 presents the experimental verification and analysis. A discussion of failure cases and limitations is given in Section 4. Finally, Section 5 concludes the work presented in this paper.

2 | DATASETS AND PROPOSED MODEL

2.1 | Datasets

To evaluate the proposed model in this paper, the Fish4Knowledge dataset [38] is selected as the experimental data. This is a dataset of fish images collected between 1 October 2010, and 30 September 2013, by the Taiwan Power Corporation, Taiwan Institute of Oceanography, and Kiting National Park at underwater viewing platforms in the Taiwan South Bay Strait, Lanyu Island, and Hubi Lake. It has a total of 23 underwater fish species and 27,370 fish images. Figure 1 shows a dataset example. Each fish image in the dataset corresponds to a Label image, where the white area is the mask of the corresponding fish target. There are two semantic categories in the Label image, namely background and fish target, with pixel values of 0 and 255, respectively. Herein, all fish images are uniformly resized to 128 × 128 pixels, and the dataset is randomly partitioned into three parts of ratio 8:1:1 for the training set, validation set, and test set, respectively.

2.2 | Proposed model

Here, this paper first gives the overall framework of the improved U-Net network, and then introduces four specific improvement made to the original network in sequence: MRS block structure, multiple-input structure, MSC structure and inserted CBAM attention mechanism.

2.2.1 | Improved U-Net network framework

Here, the U-Net network structure is improved as shown in Figure 2. Here, ×1 and ×2 mean that there are 1 and 2 identical modules connected to the hierarchy, respectively. Achieving the balance between lightweight and accuracy is an urgent problem to be solved for underwater exploration robots. At present, semantic segmentation algorithms have too many parameters, low efficiency and low accuracy, which greatly limits their application in the field of underwater exploration. Considering that the underwater fish target segmentation task is relatively simple compared to classification tasks such as COCO dataset and VOC dataset, we first cut the network by reducing the number of convolution kernels, reducing the number of down-sampling channels from {64,128,256,512,1024} to {16,32,64,128,256}. Then, in order to enhance features while minimizing the number of parameters and computational complexity, a lightweight feature extraction structure MRS block based on Inception-ResNet-C is designed. To further reduce the impact of detail loss caused by down-sampling, we design a multiple-input approach in the first four levels of the encoder. This is followed by combining the feature maps of different scales of the encoder and decoder to realize the feature fusion of different scales to accurately segment the fish target and solve the problem of unclear fish boundary segmentation. Finally, the CBAM attention mechanism is introduced at each level of the decoder in order to make the network model focus on important feature information and reduce the interference of irrelevant information.

2.2.2 | MRS block structure

The conventional belief is that deeper stacking of CNNs leads to improved performance and increased knowledge acquisition. However, as the network becomes deeper, the number of model parameters and its size significantly increases, impacting the training speed. Moreover, the actual accuracy does not necessarily improve and may even slightly decrease. This is primarily due to the fact that increasing the depth of CNNs not only substantially increases computational complexity but also results in the "gradient explosion" [39] phenomenon during model training, making it more challenging. At present, it has been shown that the use of decomposed convolution can alleviate the "gradient

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

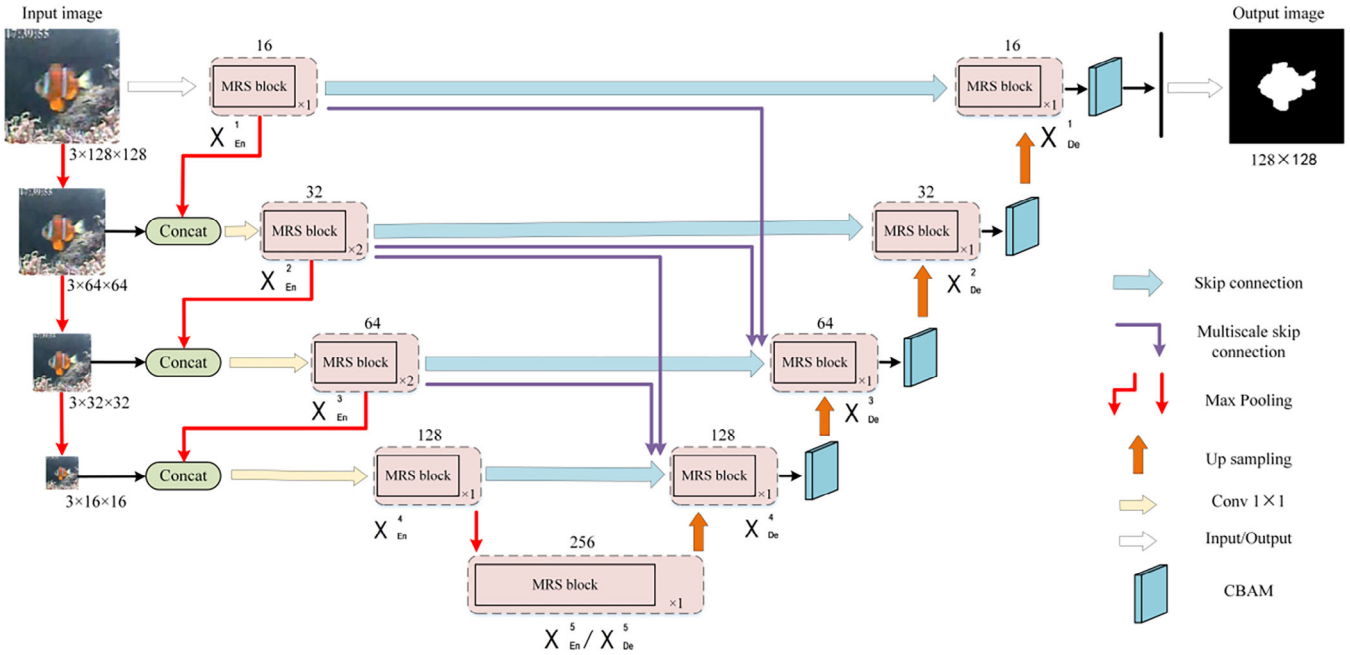

FIGURE 2 Improved U-Net network structure.

explosion" problem while achieving excellent performance at a lower computational cost.

Therefore, this paper draws inspiration from the Inception-ResNet-C architecture and designs a MRS for feature extraction, which is shown in Figure 3. The key to the lightweight nature of this structure lies in the inclusion of decomposed convolutions internally. Additionally, we introduce a 1 × 1 convolution in the leftmost residual connection to alleviate "gradient explosion" and adapt to the number of channels. Moreover, this may allow us to capture additional spatial information. The number of convolutional kernels in the middle and right branches is set to 128 to enhance feature extraction capability while reducing computational complexity. Finally, a channel shuffle operation is performed on the feature map's order of channels to strengthen intercommunication among different information channels.

Decomposed convolution was first proposed in GoogleNet [40] and later used for efficient segmentation in the ERFNet network [41]. The principle is to decompose an ordinary convolution of $k \times k$ size into two one-dimensional convolutions of $k \times 1$ and $1 \times k$. Decomposed convolution has the characteristics of the reduced number of parameters and amount of computation. Without considering the convolution bias, the number P of common convolution parameters for $k \times k$ is calculated as:

$$P = c_{out} k^2 c_{in} \tag{1}$$

where c_{in} and c_{out} are the number of input and output channels, respectively. On the other hand, the ordinary convolution computation Q is calculated as:

$$Q = c_{out} \times \left(2k^2 c_{in} - 1\right) \times H_{out} W_{out} \tag{2}$$

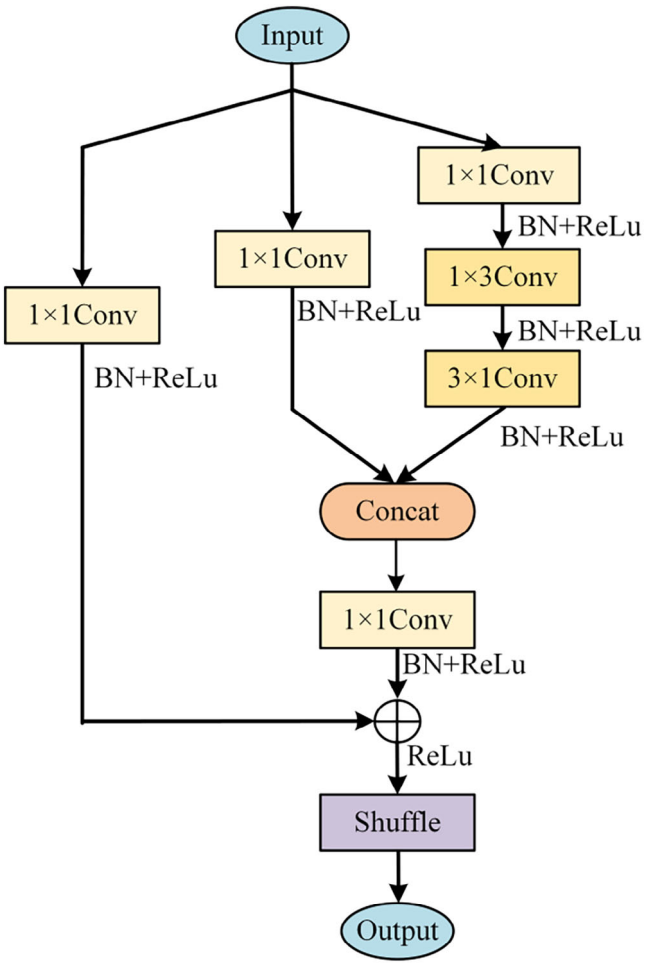

FIGURE 3 MRS block structure.

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where H_{out} and W_{out} are the height and width of the output feature map, respectively. From this, it can be seen that the number of parameters of an ordinary convolution is squared with the convolution kernel size k. The decomposed convolution breaks down an ordinary convolution of size $k \times k$ into two 1D convolutions of $k \times 1$ and $1 \times k$. Therefore, the parameter quantity P_f and computation quantity Q_f can be calculated as follows:

$$P_f = 2c_{out}kc_{in} \tag{3}$$

$$Q_f = 2[c_{out} \times (2kc_{in} - 1) \times H_{out}W_{out}] \tag{4}$$

It can be seen from Equations (3) and (4) that P_f and Q_f have a linear relationship with the convolution kernel size k. Therefore, the larger the size of the convolution kernel, the greater will be the extent of the reduction of parameter number and computation cost by decomposed convolution over ordinary convolution.

2.2.3 | Multiple-input structure

In the field of image segmentation, the feature extraction stage of the encoder is very critical because it directly affects the ability of the model to understand and express the input data. However, regular down-sampling operation will lead to partial loss of part of the detail information, which reduces the network segmentation accuracy, and thus is a challenge for tasks that require high-precision segmentation. To overcome this problem, a new method called multiple-input structure, which makes full use of multi-scale information to improve the network's ability to perceive the target, is designed in this paper, as shown in Figure 4.

In particular, the multiple-input structure first performs three ×2 down-sampling operations on the input features, which is equivalent to ×2, ×4, and ×8 down-sampling of the input features, respectively. Then, the down-sampled features are Concat cascaded with those of the corresponding level to retain more detailed information. Finally, a 1×1 convolution operation is used for feature dimension reduction to speed up the training, so as to obtain the input features of the corresponding level. In this way, the features of multiple down-sampling levels are effectively fused, which improves the network's ability to perceive detailed information and thus the segmentation accuracy. This design can not only effectively deal with the problem of information loss in the down-sampling process, but also accelerate the training speed so that the model has better performance in practical applications.

2.2.4 | MSC structure

Due to the direct connection between the encoder and decoder in the same layer of the U-Net network, there exists a significant semantic difference among the feature maps, which

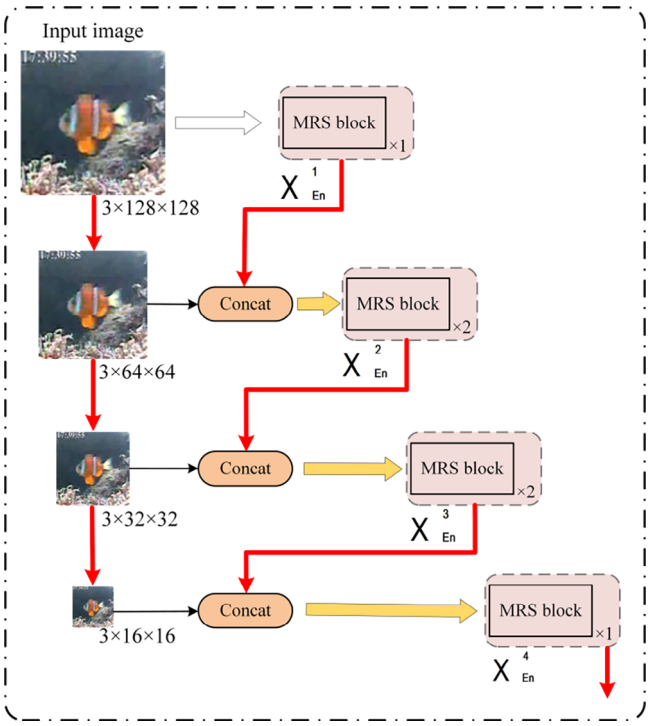

FIGURE 4 Multiple-input structure.

inevitably increases the complexity of network learning. To address this issue, this paper introduces the MSC module to make the feature information of network fusion more abundant and comprehensive.

Unlike U-Net's skip connection structure, our proposed MSC structure not only modifies the simple skip connection between the encoder and decoder but also alters connections within the decoder. In the improved U-Net decoder in Figure 2, X^3_{De} fuses the smaller-scale feature maps of X^1_{En} and X^2_{En} with the same scale feature maps of X^3_{En} and the larger-scale feature maps of X^4_{De}. Similarly, X^4_{De} fuses the smaller-scale feature maps of X^2_{En} and X^3_{En} with the feature maps on the same scale of X^4_{En} and the larger-scale feature maps of X^5_{De}. In order to avoid information redundancy, this paper only sets the fusion of multi-scale feature maps in decoders X^3_{De} and X^4_{De}. The specific process of constructing the feature map, taking the decoder X^3_{De} as an example, is illustrated in Figure 5.

As depicted in Figure 5, the aforementioned two skip connections reduce the feature map resolution to 1/4 and 1/2 through max pooling operations, respectively, thereby conveying low-level semantic information. Conversely, the lower skip connection expands the resolution by a factor of 2 using transposed convolution to convey high-level semantic information. In terms of channel numbers, both X^1_{En} and X^3_{En} convert them to 32 channels via a 1 × 1 convolution operation while reducing the number of up-sampling channels to 1/4 of their original value. The design of the MSC structure makes the decoder X^3_{De} fuse small-scale, same-scale, and large-scale feature maps effectively, thus obtaining fine-grained and coarse-grained semantic information for clearer localization and edge detection

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

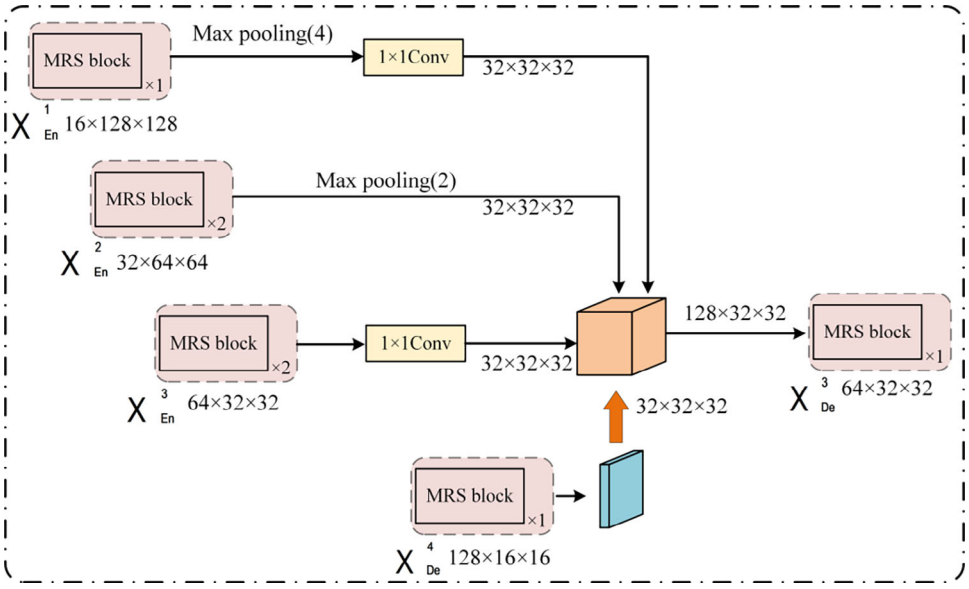

FIGURE 5 Example of constructing a feature map using MSC structure.

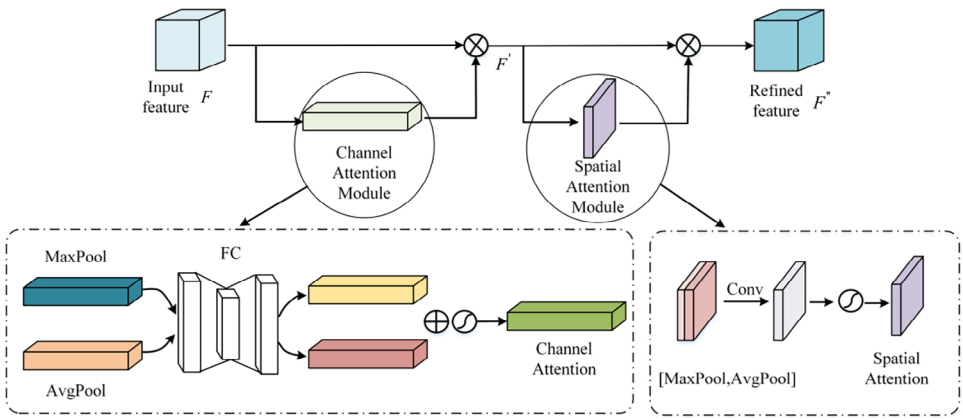

FIGURE 6 CBAM structure.

of fish targets, even under challenging underwater conditions characterized by fuzzy boundaries.

2.2.5 | Attention mechanism

In order to improve the network model's attention to important feature information, suppress irrelevant features, and improve the extraction ability and segmentation accuracy of target features, we introduce the CBAM [42, 43]. We incorporate the CBAM attention module into each level of the decoder, adjusting the feature weights through CBAM to strengthen meaningful features while suppressing irrelevant ones. This enhancement improves the decoder's image recovery capability.

The CBAM attention module consists of a channel attention module and a spatial attention module, which can refine the feature map and integrate it into two dimensions to optimize the intermediate feature map adaptively. The structure diagram of CBAM is illustrated in Figure 6.

For the feature map F (H×W×C) input to CBAM module, the mathematical derivation formula is as follows:

$$\begin{aligned} M_c(F) &= \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_0\left(F_{avg}^c\right)\right) + W_1\left(W_0\left(F_{max}^c\right)\right)\right) \end{aligned} \tag{5}$$

$$F' = M_c(F) \otimes F \tag{6}$$

$$\begin{aligned} M_s(F') &= \sigma\left(f^{7\times7}\left(\left[AvgPool\left(F'\right); MaxPool\left(F'\right)\right]\right)\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F_{avg}^s; F_{max}^s\right]\right)\right) \end{aligned} \tag{7}$$

$$F'' = M_s(F') \otimes F' \tag{8}$$

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $M_c(F)$ is the output channel attention feature map, σ is the Sigmoid activation function, *MLP* is a multi-layer perceptron [44], *AvgPool* and *MaxPool* are average pooling, and max pooling respectively, W_0 and W_1 are the weights of *MLP* ($W_0 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_1 \in \mathbb{R}^{C \times \frac{C}{r}}$ where $r = 16$ is the attenuation factor), F^c_{avg} and F^c_{max} are the output of feature map F through average pooling, and max pooling, respectively, F' is the input of the spatial attention module, $\otimes$ is the element-wise multiplication operation, $M_s(F')$ is the output spatial attention feature map, F^s_{avg} and F^s_{max} are the output of the feature map through average pooling, and max pooling, respectively, $f^{7\times7}$ is a convolution layer with a convolution kernel size of 7×7, and F'' is the final output feature map of the CBAM module.

2.3 | Experiment configuration

The experiment was conducted on a computer platform with Intel(R) Core(TM) i7-13700KF 3.40 GHz processor, NVIDIA GeForce RTX 4090 graphics card, 24GB video memory, and Windows 10 operating system. The experiment utilized the Python 3.8 programming language, a deep learning framework based on PyTorch, and the VS Code compiler. The batch size was set to 16; the cross entropy (CE) loss function was employed; Adam optimizer was used to update the network training weight; the learning rate was set to 1e-4; and the number of epochs was set to 30.

3 | EXPERIMENT AND ANALYSIS

3.1 | Performance metrics

To assess the performance of the network model in a quantitative manner, the mean intersection over union (mIOU), mean pixel accuracy (mPA), and frames per second (FPS) are utilized as performance metrics for the algorithm's target identification. The mean intersection over union (mIOU) is a metric that calculates the average ratio of the intersection and union of real and predicted values across all categories. On the other hand, the mean pixel accuracy (mPA) is a metric that calculates the average accuracy of pixel classification for all object classes in the image. They are calculated as follows:

$$mIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} (p_{ji} - p_{ii})} \tag{9}$$

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{10}$$

where k denotes the number of categories excluding the background, i and j represent the true value and predicted value, respectively, p_{ij} represents the count of pixels that predict i as j, p_{ii} represents the count of pixels that predict i as i, and p_{ji} represents the count of pixels that predict j as i.

TABLE 1 Network performance under different down-sampling times.

Number of down-sampling	mIOU (%)	mPA (%)	Params (10^6)	FLOPs (10^9)
2	94.37	96.98	6.09	8.40
3	94.63	97.21	6.16	8.41
4	94.49	97.04	6.43	8.43

3.2 | Model verification

3.2.1 | Effect of different down-sampling times in the multiple-input module

In order to verify the effect of different down-sampling times for input features in the multiple-input module on network performance, this paper conducts experiments on the basis of the improved U-Net network when the number of channels is {64, 128, 256, 512, 1024}. The results are shown in Table 1.

As evident from the results in Table 1, the best mIOU and mPA performances are achieved when the input features are down-sampled three times in sequence in the multiple-input module, while the number of parameters and FLOPs are between those of two and four down-sampling. When down-sampling two times, although the parameter number and FLOPs will decrease, mIOU and mPA will also decrease. When down-sampling four times, not only will mIOU and mPA decrease, but the number of parameters and FLOPs will also increase. This shows that only the appropriate down-sampling times can retain more useful details. More detailed information is lost when the sample is down-sampled four times. Even if Concat is concatenated with the sampled features at the corresponding level, it cannot be replaced. However, the network performance is affected by the decrease in parameter number when the sample is down-sampled two times. Therefore, in this paper, the down-sampling frequency was set to 3 times in the multiple-input module.

3.2.2 | Effect of different channel reduction factor

In order to verify the effect of a different channel reduction factor on network performance, this paper conducts experiments similar to those in the previous section. Experiment 1 indicates that the number of improved U-Net channels is {64, 128, 256, 512, 1024}. Experiment 2 indicates that the number of channels in the improved U-Net network is reduced to 1/2 of the number of channels in experiment 1, namely {32, 64, 128, 256, 512}. Experiment 3 indicates that the number of channels in the improved U-Net network is reduced to 1/4 of the number of channels in experiment 1, that is, {16, 32, 64, 128, 256}. The results are shown in Table 2.

As evident from the results in Table 2, with the reduction of channel number, both the parameter number and FLOPs decrease significantly, while mIOU and mPA decrease slightly.

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Network performance under different channel reduction factors.

Experiment	Channel reduction factor	mIOU (%)	mPA (%)	Params (10^6)	FLOPs (10^9)
1	1	94.63	97.21	6.16	8.41
2	1/2	94.51	97.09	2.79	6.41
3	1/4	94.44	97.03	1.74	5.59

TABLE 3 Network performance under different attention mechanisms.

Model	mIOU (%)	mPA (%)	Params (10^6)	FLOPs (10^9)
Improved U-Net	94.37	96.98	1.73	5.58
+SE	94.42	97.03	1.74	5.58
+CA	94.40	97.00	1.73	5.58
+ECA	94.38	96.99	1.73	5.58
+CBAM	94.44	97.03	1.74	5.59

When the number of channels is reduced to 1/4 of the original number, mIOU and mPA decrease by 0.19% and 0.18%, respectively, and the number of parameters and FLOPs decrease by 71.75% and 33.53%, respectively, relative to that before the reduction. Therefore, based on the lightweight index of the model and the recognition performance index of the algorithm, the number of channels in the improved U-Net network is reduced to 1/4 of the number of channels in experiment 1 in this paper {16, 32, 64, 128, 256}.

3.2.3 | Effect of attention mechanism

This section further verifies whether the introduction of the CBAM attention mechanism can help the network model improve the extraction ability of target features and segmentation accuracy. After a series of operations of channel reduction and the introduction of MRS block, multiple-input and MSC structure, we obtain an improved lightweight network model, which is called improved U-Net. Here, several commonly used attention mechanisms SE (squeeze-and-excitation), CA (coordinate attention), ECA (efficient channel attention) and CBAM are fused into the improved U-Net network model, and comparative experiments are carried out on the same data set. The results are shown in Table 3.

It can be seen from Table 3 that the improved U-Net network without the attention mechanism has the lowest mIOU and mPA, which represent the segmentation accuracy indicators. After integrating the attention mechanisms of SE, CA, ECA and CBAM, the number of parameters and FLOPs are hardly improved, while the mIOU is increased by 0.05%, 0.03%, 0.01% and 0.07% respectively, and the mPA is increased by 0.05%, 0.02%, 0.01% and 0.05%, respectively. This also shows that the introduction of an attention mechanism can improve the ability to extract target features. Therefore, considering the experimental results, this paper chooses to fuse the CBAM attention mechanism.

3.3 | Performance comparison between the improved model and the basic network

The proposed algorithm (this work) and the original U-Net model are trained and tested on the same marine fish image dataset, and the results are shown in Table 4.

It can be seen from Table 4 that the proposed algorithm has a significant decrease in the number of parameters, model size and calculation amount, and a small increase in mIOU and mPA. Compared with the original U-Net model, the number of parameters of the proposed algorithm is reduced to 1.74 M, a reduction of 94.39%. The model size is reduced to 6.87 MB, a reduction of 94.20%. The FLOPs are reduced to 5.59G, a reduction of 51.52%. The mIOU reaches 94.44%, an increase of 0.57%. The mPA reaches 97.03%, an increase of 0.62%. Therefore, through the comparison of various evaluation indicators, it can be concluded that the proposed algorithm can achieve model lightweight while maintaining high segmentation accuracy.

3.4 | Ablation experiment

In order to verify the effectiveness of the improved network and each improvement module of the network, we set up an ablation experiment as follows: Scheme 0—original U-Net network; Scheme 1—add MRS block to the original U-Net network with channel numbers {16, 32, 64, 128, 256}; Scheme 2—introduces multiple-input structure based on Scheme 1; Scheme 3—introduces MSC structure on the basis of Scheme 2; and Scheme 4—introduces CBAM attention mechanism based on Scheme 3. Scheme 4 is the final model proposed in this paper. Table 5 illustrates the designs of several schemes, and Table 6 shows the outcomes of ablation experiments.

As evident from the results in Table 6, the inclusion of each module into the U-Net network leads to an enhancement in performance, accompanied by a notable reduction in the number of parameters and FLOPs. Although the mIOU and mPA of Scheme 1–4 show only marginal improvement compared to Scheme 0, the network model demonstrates enhanced lightweight characteristics. The number of parameters and FLOPs of Scheme 1–4 exhibits a significant decrease, primarily attributed to the reduction in the number of channels within

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Performance comparison before and after improvement.

Model	mIOU (%)	mPA (%)	Params (10^6)	Model size (MB)	FLOPs (10^9)
U-Net	93.87	96.41	31.04	118.49	11.53
This work	94.44	97.03	1.74	6.87	5.59

TABLE 5 Different schemes for ablation experiments.

Scheme	+MRS block	+Multiple-input	+MSC	+CBAM
0				
1	✓			
2	✓	✓		
3	✓	✓	✓	
4	✓	✓	✓	✓

"√" denotes the method of introducing the corresponding column.

TABLE 6 Model comparison in the ablation experiment.

Model	mIOU (%)	mPA (%)	Params (10^6)	FLOPs (10^9)
Scheme 0	93.87	96.41	31.04	11.53
Scheme 1	94.18	96.81	1.79	5.57
Scheme 2	94.26	96.85	1.80	5.58
Scheme 3	94.37	96.98	1.73	5.58
Scheme 4	94.44	97.03	1.74	5.59

the model and the utilization of a substantial amount of decomposed convolution. Compared to Scheme 0, the mIOU and mPA of Scheme 1–4 exhibited improvements of 0.31%, 0.39%, 0.50%, and 0.57%, respectively. Additionally, the number of parameters decreased by 94.23%, 94.20%, 94.43%, and 94.39%, respectively; and the FLOPs decreased by 51.69%, 51.60%, 51.60%, and 51.52%, respectively. These results suggest that each module of the design is effective.

3.5 | Performance comparison of different algorithms

In order to verify the advantages of the proposed algorithm, we compare it with the popular semantic segmentation algorithms (U-Net [19], SegNet [25], R2U-Net [20], Mobile_UNet [45], Att U-Net [24], PSPNet [46]) and several existing underwater fish segmentation methods (ARD-PSPNet [47], IST-PSPNet [48]) using the same data set. The results are shown in Table 7. The prediction visualization results of different methods are shown in Figure 7.

As evident from the results in Table 7, when compared to the classic networks U-Net, SegNet, R2U-Net, Mobile_U-Net, Att U-Net, and PSPNet, the proposed algorithm has the least number of parameters (1.74 M), smallest model size (6.87 MB), and FLOPs (5.59 G), while maintaining high segmentation accuracy. The mIOU of our proposed algorithm in this study exhibits improvements of 0.57%, 0.99%, 4.29%, 0.91%, 0.21%, and 4.31% when compared to U-Net, SegNet, R2U-Net, Mobile_UNet, Att U-Net, and PSPNet, respectively. The corresponding mPA is 0.62%, 0.58%, 0.65%, 0.04%, and 1.14% higher than that of U-Net, SegNet, Mobile_UNet, Att U-Net, and PSPNet, respectively, but 0.24% lower than that of R2U-Net. Compared with the existing underwater fish segmentation methods ARD-PSPNet and IST-PSPNet, the proposed algorithm does not have advantages in mIOU, but improves the mPA. Moreover, it also shows clear advantages in the number of parameters, model size and FLOPs. In terms of segmentation speed, the FPS of the proposed method reaches 43.62, which is 3.5 higher than that of the PSPNet network. Although it is lower than other methods except PSPNet, it can still meet the real-time requirements. In summary, our proposed algorithm is less demanding in terms of computing resources and storage space, more lightweight, and can achieve high segmentation accuracy, which is suitable for application in underwater environments with limited resources. As can be seen from Figure 7, compared with the classical network, the segmentation results of the proposed method have higher accuracy, achieve more complete object segmentation, and the resulting segment details are more consistent with the real label images. Compared with the existing underwater fish segmentation methods ARD-PSPNet and IST-PSPNet, the segmentation effect is similar.

4 | FAILURE CASES AND LIMITATIONS

Through the verification of the above experiments, it is shown that the proposed algorithm balance between achieving model lightweight and maintaining segmentation accuracy. However, when dealing with extremely blurry underwater fish images and "incognito" fish images, the proposed algorithm cannot achieve satisfactory segmentation results, as illustrated in Figure 8. The so-called "stealth" means that due to the differences in the living environment and habits of marine fish, the colour of the whole body or part of some fish will be similar to the colour of the environment, so that they seem to blend with the surrounding environment in the image, showing an "incognito" effect.

As can be seen in Figure 8, when dealing with extremely blurry underwater fish images, the boundary of the fish target appears to drift and the details of the fishtail are segmented inaccurately. When dealing with the "incognito" fish image, the fish target and the environment background are similar in colour, and the boundary of the fish target is mixed with that of the environment, resulting in boundary confusion and unsatisfactory segmentation results. In this case, the algorithm proposed

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

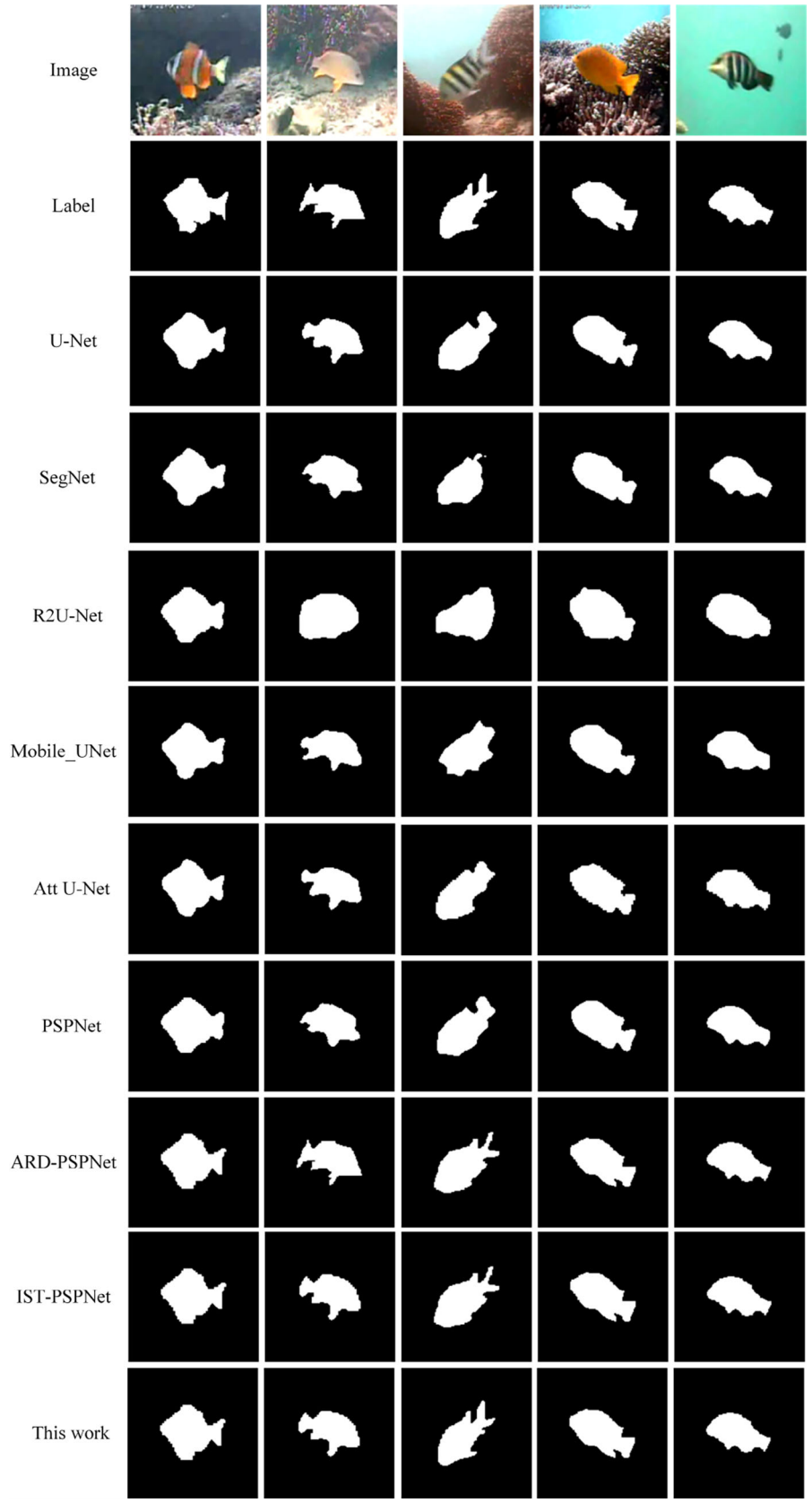

FIGURE 7 Segmentation visualization results of different algorithms.

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7 Comparison of the results of different methods on the test set.

Method	mIOU (%)	mPA (%)	Params (10^6)	Model size (MB)	FLOPs (10^9)	FPS
U-Net	93.87	96.41	31.04	118.49	11.53	61.18
SegNet	93.45	96.45	29.44	112.44	10.04	48.81
R2U-Net	90.15	97.27	39.09	149.23	16.42	49.16
Mobile_UNet	93.53	96.38	15.11	57.83	10.44	61.59
Att U-Net	94.23	96.99	34.88	133.19	16.65	57.42
PSPNet	90.13	95.89	46.71	187.20	11.38	40.12
ARD-PSPNet	94.85	96.12	49.53	47.86	10.71	47.30
IST-PSPNet	94.53	96.76	46.48	168.52	40.27	49.86
This work	94.44	97.03	1.74	6.87	5.59	43.62

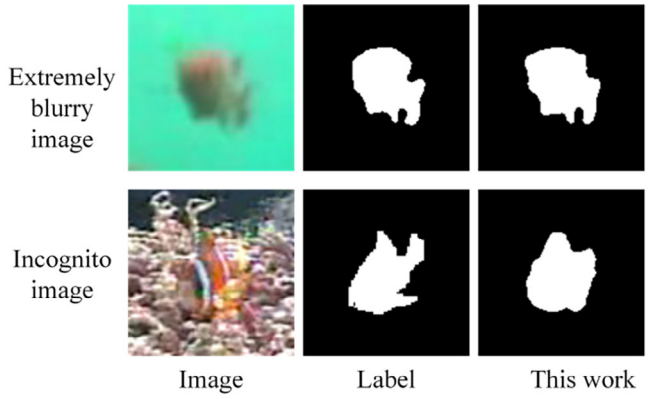

FIGURE 8 Diagram of the two types of failure cases.

in this paper cannot accurately distinguish the fish boundary, but can only roughly judge the position of the fish target.

Therefore, in the future, we can further optimize the deep learning model to systematically analyse and process underwater fish images and improve the segmentation effect by combining techniques such as object detection and image enhancement.

5 | CONCLUSION

Herein, an improved lightweight network for underwater fish image segmentation is proposed based on U-Net. On the basis of retaining the U-Net's U-shape structure, an MRS structure is designed to replace the ordinary convolutional layer, which can greatly reduce the model's volume and parameter number. Meanwhile, a multiple-input structure is designed in the encoder part so that the encoder can obtain more detailed information. The common skip connection between the encoder and decoder is replaced with an MSC structure, which makes full use of deep and shallow layers of semantic information to improve the accuracy of semantic segmentation. It uses an attention mechanism to strengthen the weight of some feature layers and spatial region features, and suppresses the features of the background region, in order to improve the accuracy of underwater fish target segmentation. The mIOU and mPA performances of the model based on the Fish4Knowledge dataset are 94.44% and 97.03%, respectively. The model's 1.74 M parameters and size of 6.87 M are only 5.61% and 5.8% of UNet, respectively. Its average segmentation speed of 43.62 FPS can ensure the real-time performance of the network while maintaining a balance between precision and complexity. Compared with other methods, the performance of the proposed algorithm is verified through simulation experiments.

In the future, we plan to explore relevant techniques for the advancement of unmanned underwater vehicles by enabling intelligent object detection and recognition in diverse underwater environments.

AUTHOR CONTRIBUTIONS

Zhenkai Zhang: Conceptualization; methodology; investigation; supervision; writing—review and editing. **Wanghua Li**: Conceptualization; methodology; software; visualization; data curation; writing—original draft; validation; project administration; investigation. **Boon-Chong Seet**: Methodology; writing—review and editing.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions that have contributed to improve this paper. This work was supported by the National Natural Science Fund (61871203) in China.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Zhenkai Zhang https://orcid.org/0000-0003-2439-0923
Wanghua Li https://orcid.org/0009-0001-0567-4819

REFERENCES

1. Lin, M., Yang, C.: Ocean observation technologies: A review. Chin. J. Mech. Eng. 33(2), 33–50 (2020)

17519967, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2. Marc, A.B.A., Lukey, M.J., Cerione, R.A.: The State of World Fisheries and Aquaculture-Meeting the Sustainable Development Goals. Food and Agriculture Organization, Rome, Italy (2018)
3. Bogue, R.: Underwater robots: A review of technologies and applications. Ind. Rob. 42(3), 186–191 (2015)
4. Wynn, R.B., Huvenne, V.A.I., Le Bas, T.P., Murton, B.J., Connelly, D.P., Bett, B.J., Ruhl, H.A., Morris, K.J., Peakall, J., Parsons, D.R., Sumner, E.J., Darby, S.E., Dorrell, R.M., Hunt, J.E.: Autonomous underwater vehicles (AUVs): Their past, present and future contributions to the advancement of marine geoscience. Mar. Geol. 352, 451–468 (2014)
5. Bryson, M., Johnson-Roberson, M., Pizarro, O., Williams, S.B.: True color correction of autonomous underwater vehicle imagery. J. Field Rob. 33, 853–874 (2016)
6. Li, C., Guo, C., Ren, W., Cong, R., et al.: An underwater image enhancement benchmark dataset and beyond. IEEE Trans. Image Process. 29, 4376–4389 (2019)
7. Li, C., Saeed, A., Fatih, P.: Underwater scene prior inspired deep underwater image and video enhancement. Pattern Recognit. 98, 107038 (2020)
8. Li, Y., Chen, R.: UDA-Net: Densely attention network for underwater image enhancement. IET Image Process. 15, 774–785 (2021)
9. Naseer, A., Baro, E.N., Khan, S.D., et al.: A novel detection refinement technique for accurate identification of Nephrops norvegicus burrows in underwater imagery. Sensors 22(12), 4441 (2022)
10. Wei, X., Yu, L., Tian, S., Feng, P., Ning, X.: Underwater target detection with an attention mechanism and improved scale. Multimedia Tools App. 80, 33747–33761 (2021)
11. Zhou, T., Zhou, Y., Gong, C., et al.: Feature aggregation and propagation network for camouflaged object detection. IEEE Trans. Image Process. 31, 7036–7047 (2022)
12. Wang, Z., Wang, S., Wang, X., Luo, X.: Underwater moving object detection using superficial electromagnetic flow velometer array based artificial lateral line system. IEEE Sens. J. 24(8), 12104–12121 (2024)
13. Duan, Y., Stien, L.H., Thorsen, A., Karlsen, Ø., Sandlund, N., Li, D., Fu, Z., Meier, S.: An automatic counting system for transparent pelagic fish eggs based on computer vision. Aquacult. Eng. 67, 8–13 (2015)
14. Abdeldaim, A.M., Houssein, E.H., Hassanien, A.E.: Color image segmentation of fishes with complex background in water. In: The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018), pp. 634–643. Cairo, Egypt (2018)
15. Zhao, Y., Li, W., Li, Y., Qi, Y., Li, Z., Yue, J.: LFCNet: A lightweight fish counting model based on density map regression. Comput. Electron. Agric. 203, 107496 (2022)
16. Saleh, A., Laradji, I.H., Konovalov, D.A., Bradley, M., Vazquez, D., Sheaves, M.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. Sci. Rep. 10(1), 14671 (2020)
17. Zhang, H., Liu, H., Kim, C.: Semantic and instance segmentation in coastal urban spatial perception: a multi-task learning framework with an attention mechanism. Sustainability 16(2), 833 (2024)
18. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. Boston, MA (2015)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Munich, Germany (2015)
20. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. J. Med. Imaging 6, 014006 (2019)
21. Ibtehaz, N., Rahman, M.S.: MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Networks 121, 74–87 (2020)
22. Long, F.: Microscopy cell nuclei segmentation with enhanced U-Net. BMC Bioinf. 21, 1–12 (2020)
23. Liu, Y., Wang, H., Chen, Z., Huangliang, K., Zhang, H.: TransUNet+: Redesigning the skip connection to enhance features in medical image segmentation. Knowledge-Based Syst. 256, 109859 (2022)
24. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. arXiv:1804.03999 (2018)
25. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495 (2017)
26. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv:1412.7062 (2014)
27. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40, 834–848 (2018)
28. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. Computer Science, arXiv:1706.05587 (2017)
29. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 801–818. Munich, Germany (2018)
30. Zhou, T., Zhou, Y., He, K., et al.: Cross-level feature aggregation network for polyp segmentation. Pattern Recognit. 140, 109555 (2023)
31. Labao, A.B., Naval, P.C.: Weakly-labelled semantic segmentation of fish objects in underwater videos using a deep residual network. In: Intelligent Information and Database Systems: 9th Asian Conference, ACIIDS 2017, pp. 255–265. Kanazawa, Japan (2017)
32. Nezla, N.A., Haridas, T.P.M., Supriya, M.H.: Semantic segmentation of underwater images using unet architecture based deep convolutional encoder decoder model. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 28–33 (2021)
33. Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., Vågstøl, H., Løvall, K.: Automatic segmentation of fish using deep learning with application to fish size measurement. ICES J. Mar. Sci. 77(4), 1354–1366 (2020)
34. Zhang, W., Wu, C., Bao, Z.: DPANet: Dual pooling-aggregated attention network for fish segmentation. IET Comput. Vis. 16(1), 67–82 (2022)
35. Abe, S., Takagi, T., Torisawa, S., et al.: Development of fish spatio-temporal identifying technology using SegNet in aquaculture net cages. Aquacult. Eng. 93, 102146 (2021)
36. Yang, L., Chen, Y., Shen, T., et al.: An FSFS-net method for occluded and aggregated fish segmentation from fish school feeding images. Appl. Sci. 13(10), 6235 (2023)
37. Zhao, Y., Chen, S., Liu, S., Hu, Z., Xia, J.: Hierarchical equalization loss for long-tailed instance segmentation. IEEE Trans. Multimedia 26, 6943–6955 (2024)
38. Boom, B.J., Huang, P.X., He, J., Fisher, R.B.: Supporting ground-truth annotation of image datasets using clustering. In: Proceedings of the 21st International Conference on Pattern Recognition, pp. 1542–1545 (2012)
39. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, CA, pp. 4278–4284 (2017)
40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, pp. 2818–2826 (2016)
41. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. 19, 263–272 (2017)
42. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. Munich, Germany, pp. 3–19 (2018)

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

43. Du, L., Lu, Z., Li, D.: Broodstock breeding behaviour recognition based on Resnet50-LSTM with CBAM attention mechanism. Comput. Electron. Agric. 202, 107404 (2022)
44. Riedmiller, M., Lernen, A.: Multi-layer Perceptron. Univ. Freiburg 2014
45. Jing, J., Wang, Z., Rätsch, M., Zhang, H.: Mobile-Unet: An efficient convolutional neural network for fabric defect detection. Text. Res. J. 92(1-2), 30–42 (2022)
46. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)
47. Yue, Y., Geng, L., Zhao, H., Wang, H.: Research on segmentation algorithm of underwater fish image based on ARD-PSPNet network. J. Opt. Laser 33(11), 1173–1182 (2022)
48. Han, Y., Zheng, B., Kong, X., Huang, J., Wang, X., Ding, T., Chen, J.: Underwater fish segmentation algorithm based on improved PSPNet network. Sensors 23(19), 8072 (2023)

How to cite this article: Zhang, Z., Li, W., Seet, B.-C.: A lightweight underwater fish image semantic segmentation model based on U-Net. IET Image Process. 1–13 (2024).
https://doi.org/10.1049/ipr2.13161

17519667, 0, Downloaded from https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/ipr2.13161 by Ministry Of Health, Wiley Online Library on [26/06/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License