Journal of *Marine Science and Engineering*

MDPI

Article

Neural Network for Underwater Fish Image Segmentation Using an Enhanced Feature Pyramid Convolutional Architecture

Guang Yang [1], Junyi Yang [1], Wenyao Fan [1] and Donghe Yang [2,*]

[1] School of Mechanical Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
[2] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
* Correspondence: yangdonghe@zstu.edu.cn

Abstract: Underwater fish image segmentation is a crucial technique in marine fish monitoring. However, typical underwater fish images often suffer from issues such as color distortion, low contrast, and blurriness, primarily due to the complex and dynamic nature of the marine environment. To enhance the accuracy of underwater fish image segmentation, this paper introduces an innovative neural network model that combines the attention mechanism with a feature pyramid module. After the backbone network processes the input image through convolution, the data pass through the enhanced feature pyramid module, where it is iteratively processed by multiple weighted branches. Unlike conventional methods, the multi-scale feature extraction module that we designed not only improves the extraction of high-level semantic features but also optimizes the distribution of low-level shape feature weights through the synergistic interactions of the branches, all while preserving the inherent properties of the image. This novel architecture significantly boosts segmentation accuracy, offering a new solution for fish image segmentation tasks. To further enhance the model's robustness, the Mix-up and CutMix data augmentation techniques were employed. The model was validated using the Fish4Knowledge dataset, and the experimental results demonstrate that the model achieves a Mean Intersection over Union (MIoU) of 95.1%, with improvements of 1.3%, 1.5%, and 1.7% in the MIoU, Mean Pixel Accuracy (PA), and F1 score, respectively, compared to traditional segmentation methods. Additionally, a real fish image dataset captured in deep-sea environments was constructed to verify the practical applicability of the proposed algorithm.

Keywords: fish segmentation; attention mechanism; pyramid architecture; deep learning

check for updates

Academic Editor: Weicheng Cui

Received: 26 December 2024
Revised: 22 January 2025
Accepted: 24 January 2025
Published: 26 January 2025

Citation: Yang, G.; Yang, J.; Fan, W.; Yang, D. Neural Network for Underwater Fish Image Segmentation Using an Enhanced Feature Pyramid Convolutional Architecture. *J. Mar. Sci. Eng.* **2025**, *13*, 238. https://doi.org/10.3390/jmse13020238

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fish are a major source of protein in human diets, playing a crucial role in human health [1]. As fish research expands, monitoring and studying the living habits, health, and growth of fish in deep-sea environments have become increasingly important. To track the condition of fish over the long term, many researchers have begun using fixed underwater observation devices to capture fish activity, generating large amounts of data [2]. However, the raw images captured are often noisy and difficult to process. To extract key information from these images, image segmentation has become a focal task for researchers. Traditional segmentation methods, however, suffer from low efficiency and poor accuracy. With the use of advanced neural network models, researchers can now automate the segmentation of fish images. This approach allows for the precise extraction of key information such as the size, quantity, and species of fish. Such accurate biomass data are invaluable for guiding aquaculture practices and deep-sea research [3].

As algorithms advance, deep learning has become increasingly important in image segmentation tasks, particularly with convolutional neural network models like U-Net [4], SegNet [5], and Mask R-CNN [6]. In the context of fish analysis, pre-trained models can be fine-tuned to recognize and segment fish in underwater images [7]. These models extract key biological features, such as the size, shape, and species, which are crucial for biomass estimation, population health monitoring, and aquaculture management [8]. By applying deep learning, researchers can automate the processing of large image datasets, improving efficiency while providing more precise and consistent measurements compared to traditional manual methods.

As machine learning techniques evolve, the advantages of convolutional neural networks (CNNs) in feature extraction have been widely recognized. CNNs outperform many traditional methods, such as morphology-based recognition techniques [9]. Currently, CNN-driven object-detection algorithms have gradually replaced traditional models, becoming the new industry standard. However, despite the significant performance improvements brought by automated feature extraction, detail loss during pixel-level analysis remains a challenge [10]. This is because convolution and pooling operations can reduce feature maps, degrading image quality. To address this, several strategies have been proposed to improve CNN performance in image segmentation tasks.

After analyzing large datasets, Konovalov's research team developed an efficient convolutional neural network (CNN) classifier [6], which was fine-tuned in practice to improve its recognition performance. They observed that, in certain scenarios, omitting traditional image preprocessing steps could enhance model performance, helping to address overfitting caused by a limited dataset size. Majumder and his team explored and evaluated six CNN-based transfer learning architectures [11]. In tests, the InceptionResNetV2 [12] and Xception models showed outstanding performance, achieving 98.81% accuracy. N.S. Abinaya and colleagues used YOLOv4 to segment and identify fish in concealed underwater environments [13], applying multi-segment processing to track a fish's movement. Mohammad Jahanbakht and his team employed weakly supervised learning models to segment fish in murky waters, significantly increasing the processing speed at the cost of a 4% reduction in accuracy [14].

Convolutional neural networks (CNNs) learn abstract representations of data through deep convolutional operations, enabling the network to remain robust to local transformations in the image. Building on this foundation, PSPNet [15] employs dilated convolutions to expand the receptive field and perform multi-scale feature extraction. The DeepLab series has also made significant contributions to the field of image segmentation, with DeepLabV3+ [16], developed by Chen's team in collaboration with Google, being an important version. DeepLabV3+ improves segmentation accuracy by introducing enhanced skip connections (also referred to as atrous or dilated convolution) and an encoder–decoder structure, which effectively capture multi-scale information in images. To further expand the receptive field and integrate multi-scale information, Chen et al. proposed a hierarchical dilated pyramid pooling structure [17]. Fu et al. integrated a dual-attention network model, incorporating both spatial and channel attention, which performs better in complex environments [18]. By learning dependencies between different image locations and between channels, this model improves segmentation performance. In semantic segmentation, Liu et al. optimized model architecture through hierarchical neural architecture search [19]. Their approach uses a detail branch to capture high-resolution information, while a semantic branch extracts semantic features, which are then fused to generate the final segmentation result. Alshdaifat et al. proposed a segmentation method based on an enhanced dilated convolutional pyramid pooling approach [20], addressing the challenges of multi-scale target segmentation and category boundary prediction. This method en-

hances segmentation accuracy by fusing multi-level features, particularly for segmenting targets of varying sizes in complex scenes.

Current underwater fish datasets fail to accurately represent the diversity of underwater environments and the various challenges they pose [21]. In these ecosystems, fish often exhibit appearances that closely resemble the background (such as rocks), and occlusion issues between individuals are common [22,23].

To address these challenges and improve underwater fish image segmentation accuracy, this paper proposes a novel model based on attention mechanisms and feature pyramids. The model focuses on multi-scale feature extraction, with enhancements made to the attention module and pyramid feature architecture. These modifications improve the receptive field and extend the range of contextual information extraction. As a result, the model can capture more relevant information in complex environments, increase the weight of low-level features, and enhance segmentation accuracy.

1. Feature Extraction with a ResNet50 Backbone: The ResNet50 backbone is employed for feature extraction, with the improved pyramid feature extraction architecture integrated after each residual convolution block. This architecture introduces multiple branches that capture the global context using multiple receptive fields. While enhancing high-level semantic feature extraction, it optimizes the weight distribution for low-level features, focusing more on shape or threshold-related aspects. By fine-tuning the weight allocation, this approach addresses the challenges of complex underwater environments and interference from aquatic plants and rocks, improving fish segmentation accuracy.
2. Improved Dilated Convolution with Channel Attention: Compared to the original feature extraction module, the dilated convolution block incorporates an enhanced channel attention mechanism. The original pyramid structure and channel attention mechanism used fully connected layers for global optimization, which caused a dimensionality reduction in the extracted feature maps. In contrast, the proposed model replaces all fully connected layers with 1×1 convolutional kernels, preserving inherent properties while avoiding accuracy degradation due to dimensionality reduction, and enhancing the interaction between different channels.
3. ResNet50 Model Improvement for Underwater Fish Segmentation: To address the blurriness and unclear backgrounds in the fish segmentation dataset, the ResNet50 model is modified. Downsampling is performed in the decoding block to reduce the computational cost, allowing the model to efficiently extract deep semantic information from fish images.
4. Construction of a New Dataset for Deep-Sea Fish and the Use of Dark Channel Dehazing: A new deep-sea fish image dataset is created for validation, and data augmentation techniques are applied to enhance image clarity. The practical applicability of the proposed algorithm is assessed in challenging underwater environments.

2. Related Work

2.1. Attention Mechanism

Attention mechanisms have been widely applied in computer vision tasks. The overall logic is represented by Equation (1) [24]. The attention mechanism is shown in Equation (1).

$$Attention = f(g(x),\ x) \tag{1}$$

In this context, x refers to the input image, and $g(x)$ represents the initial processing of the input features within the attention mechanism. During this stage, the correlation between x and the ground truth is computed, and corresponding weights are derived

through parameters. This correlation is typically calculated using dot product or cosine similarity, followed by normalization, which prepares the data for subsequent processing. $f(g(x),x)$ represents the second step of processing, where the previously computed weights and the normalized x are summed to produce the final output.

The attention mechanism is employed to enhance feature extraction and optimize the model [25]. Different methods emphasize different aspects, with the channel attention mechanism focusing specifically on channel-wise information.

$$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{2}$$

Equation (2) illustrates the compression process of the channel attention mechanism. By inputting the image, the output of size $1 \times 1 \times$ C can be obtained, which theoretically represents the global information. Here, uc represents the input matrix of the feature map along dimension C, and F() denotes the compression process, i.e., the "squeeze" operation. In the actual computation, $\sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j)$ represents the summation of feature values over all spatial locations along dimension c. After summing, it is divided by $H \times W$ to compute the average value. Finally, a 1×1 output is obtained, where Z_c is a scalar.

While the attention mechanism improves local segmentation, it diminishes the contextual relationships, thereby limiting the overall performance of the model [26]. To address this issue, this study leverages a feature pyramid and strengthens generic attention using multi-level dilated convolutions.

2.2. Space Pyramid Pooling

Image segmentation tasks in complex scenes demand extensive information processing, and single-core deep convolutions have limited effectiveness. In contrast, multi-level pyramid dilated convolutions can extract semantic features at different scales through multiple branches [27]. Among these, a dilated convolution plays a key role in the module, as it expands the receptive field without sacrificing resolution. The relationship between the actual output and the original convolution kernel k is illustrated below.

$$Y = k + (k-1)(rate-1) \tag{3}$$

In this equation, Y is the size of the kernel, meaning that if there is no dilation rate, the kernel elements are adjacent. $k-1$ represents the spacing between adjacent elements in the original kernel, while rate -1 indicates how the dilation rate expands the spacing. The larger the dilation rate, the greater the spacing between adjacent elements, which results in a larger receptive field. Expanding the receptive field enhances the information extraction rate, but to reduce computational load, it is important to avoid excessively large feature maps. Dilated convolutions efficiently address this challenge while extracting multi-scale information.

3. Model

The underwater fish image segmentation model presented in this paper is composed of the following modules: the data augmentation module, feature extraction network, and encoder–decoder modules. The encoder module takes preprocessed and fused underwater fish color images as the input. It is based on research from existing segmentation models, balancing computational efficiency and segmentation accuracy, and is inspired by the ResNet architecture. Additionally, an attention mechanism and an optimized pyramid feature extraction module are introduced to enhance the model's performance. After the

improvements, the model not only maintains high segmentation efficiency but these also significantly improve segmentation accuracy. In the decoder, 1 × 1 convolution operations adjust the channel count, fusing low-level shape features with high-level semantic features extracted by the encoder. The fused features are further processed by two depth-wise separable convolution modules, ultimately producing the predicted output. The overall network structure is shown in Figure 1.

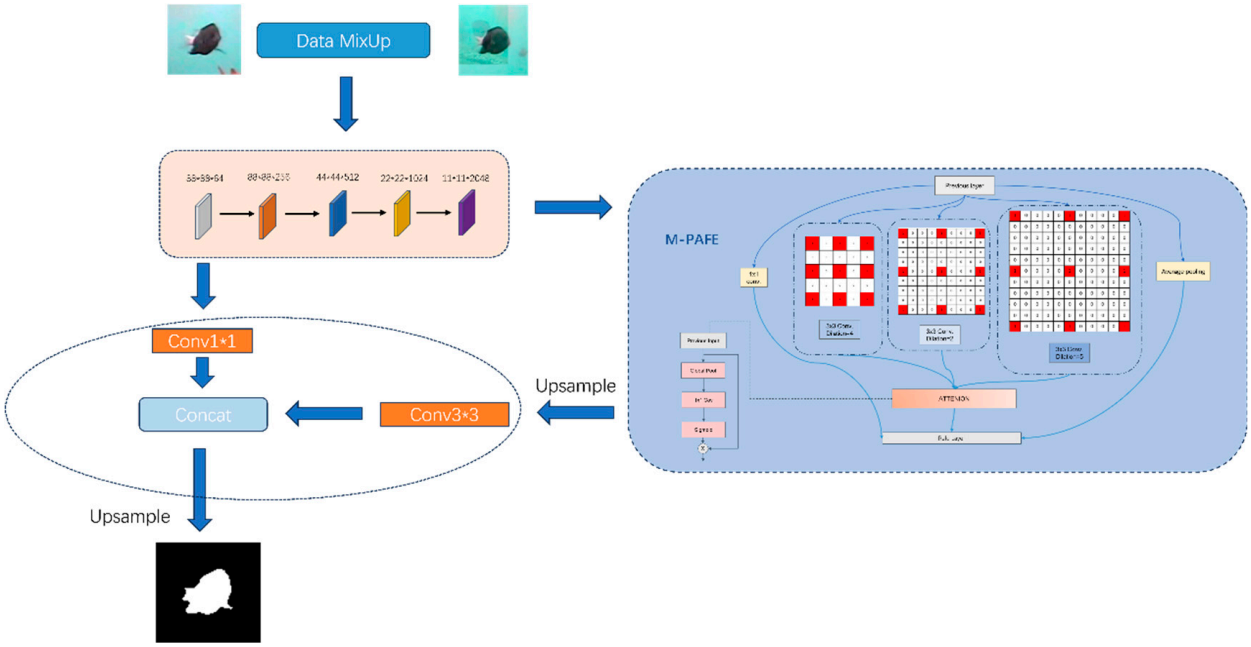


Figure 1. Network Architecture.

3.1. Backbone Network

The advantage of ResNet is its use of residual learning to enhance the performance of deep networks, effectively mitigating the issue of divergence that arises with excessive network depth. In this model, a modified ResNet50 is employed as the backbone for feature extraction. For an input image of size 512 × 512 × 3 (with 3 representing the RGB color channels), traditional ResNet50 extracts high-level features through five convolutional blocks. In this paper, we incorporate a shape attention module after each convolutional block to focus the model on important regions of the image and effectively filter out potential background noise, such as rocks and seaweed.

3.2. Pyramidally Attended Feature Extraction

Neuroscientific experiments have confirmed that multi-scale receptive fields in the human visual system help to highlight core regions. After multiple convolutions, fish images generate different input features. Features at different levels contain distinct information. The shape attention module discussed earlier focuses on low-level information to obtain localized optimized parameters, but it still needs to incorporate high-level contour information. Objects in an image have varied scales, and single-scale features cannot capture the multi-scale context of different objects. Thanks to the advantages of pyramid architectures in image segmentation, some segmentation networks utilize pyramid structures. However, due to interference from similar background images, such as rocks and seaweed, high-level features cannot be directly used and require weight optimization. Spatial attention is applied to features at different scales to focus more on visually important regions. To further enhance the feature pyramid extraction module, this model integrates the atten-

tion mechanism into the multi-scale branch network. The detailed structure is illustrated in Figure 2.

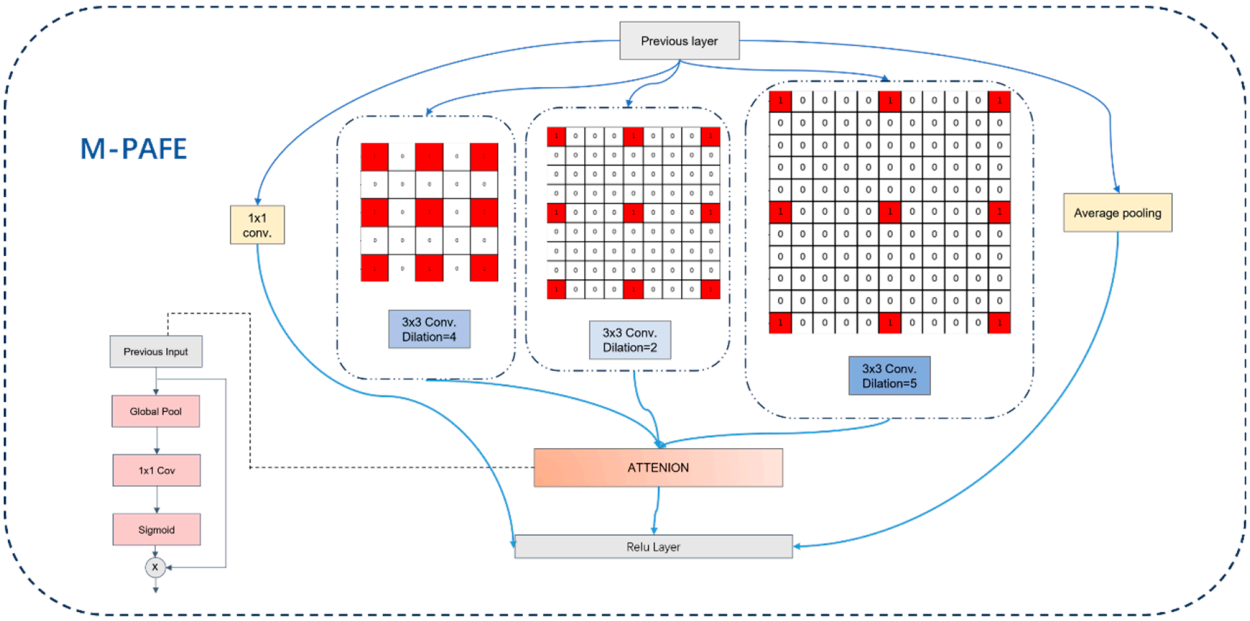


Figure 2. PAFE.

The low-level feature values first pass through three parallel dilated convolution layers to extract multi-scale feature information. The other two branches, consisting of 1×1 convolution layers and average pooling layers, ensure the integrity of the low-level information, reducing the loss of contextual information. The original PAFE convolution kernels and dilation rates are not suitable for fish segmentation. Considering the edge information and target size in fish images, the dilation rates have been modified to 2, 4, and 6. The improved spatial attention mechanism can escape local optima to achieve better results.

Compared to the original pyramid architecture, the dilated convolution block incorporates an improved channel attention mechanism. Both the original pyramid architecture and the channel attention mechanism use fully connected layers, which lead to a dimensionality reduction in the extracted feature maps, causing different features to compute the same result. To avoid the dimensionality reduction, the fully connected layers have been replaced with 1×1 convolution kernels, which maintain the inherent properties while increasing interactions between features.

After entering the PAFE module, the feature values pass through different branches. The 1×1 convolution kernel and average pooling architecture preserve the original dimensions. Feature maps passing through the 2, 4, and 5 dilation kernels then undergo the improved channel attention module. First, global pooling is applied to transpose the features to a $1 \times 1 \times C$ shape. The original channel attention mechanism uses two fully connected layers, which increase the receptive field but also raise the parameter count and ignore spatial architectural features, leading to a dimensionality reduction. Therefore, fully connected layers are not suitable for image segmentation tasks. Instead, 1×1 convolution kernels are used to enhance interactions between channels, further completing the multi-scale feature extraction task. After passing through the 1×1 convolution kernel, the SIGMOID function calculates a weight between 0 and 1, which is then multiplied with the original feature values to obtain the required feature values.

3.3. Loss Function

Ion accuracy, and combining the application of exponential and logarithmic transformations to both Dice loss and cross-entropy loss, can lead to finer segmentation boundaries with the benefits of more accurate data distribution. Its definition is as follows: The L_{Dice} loss is commonly used to evaluate the similarity between predicted results and ground truth in image segmentation tasks [28]. The Dice coefficient measures the similarity between two sets, and Dice loss is 1 minus the Dice coefficient, which helps minimize segmentation errors. Lcross [29], the cross-entropy loss, is used for multi-class classification problems, measuring the difference between the predicted probability distribution and the true labels. Cross-entropy loss is commonly used in training classification models, especially for pixel-level tasks like image segmentation. w is the weight coefficient that controls the relative importance of Dice loss and cross-entropy loss in the combined loss. By combining these losses, this approach optimizes both pixel-level classification accuracy (cross-entropy) and the overlap between the segmented shape and the ground truth (Dice loss). This approach is particularly effective for image segmentation, especially when the target objects are imbalanced in size.

$$L_{Exp} = w_{Dice} L_{Dice} + w_{cross} L_{cross} \tag{4}$$

4. Experiment

4.1. Dataset Acquisition and Processing

This experiment utilized 6056 fish images from the Fish4Knowledge dataset as the focus of the fish target segmentation task. After applying data augmentation, the images were converted to the VOC format. To ensure both the accuracy and generalizability of the experiment, the selected underwater fish images covered a variety of fish species and different viewing angles. Furthermore, to reduce computational complexity and simplify the model, all images are uniformly resized to a fixed resolution of 512 × 512 pixels. The dataset's JPEG images were drawn from 23 fish species within Fish4Knowledge and have been annotated at the pixel level for subsequent experimental validation. A total of 6056 images were selected from the Fish4Knowledge dataset, and they were split into training, validation, and test sets in a 7:2:1 ratio. The distribution was strictly carried out by randomly selecting specific samples.

To further augment the dataset with more images, the following data augmentation methods were applied: Mixup [30], a data augmentation technique, linearly combines features from different samples, enabling the model to capture linear relationships between samples. This form of linear modeling helps to reduce the model's bias when processing data outside the training set; Cutout [31] involves filling a randomly placed square of fixed size with zeros. To minimize the impact of zero-padding on training, some additional data processing is necessary; CutMix [32] places two targets in a local region, enhancing the model's ability to identify complex areas. The effects of these augmentations are illustrated in Figure 3.

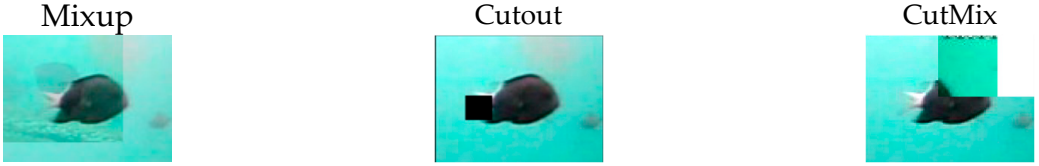


Figure 3. Fish dataset after data augmentation.

4.2. Evaluation Criteria

Due to the minimal differences between each segmented image, it is challenging to evaluate them accurately with the naked eye. This study selects three commonly used evaluation metrics, namely F1-score, Global Accuracy, and the Mean Intersection over Union.

F1-Score:

$$\text{F1-}Score = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

F1 is a commonly used metric for imbalanced datasets. It is the harmonic mean of precision and recall, providing a balanced evaluation of a model's performance on both positive and negative samples. Precision, or positive predictive value, is the proportion of predicted positive samples that are actually positive. Recall is the proportion of true positives correctly identified by the model. The F1 score is the harmonic mean of these two metrics, aiming to balance precision and recall. Unlike their arithmetic mean, the harmonic mean is more sensitive to lower values, so when there is a large discrepancy between precision and recall, the F1 score will be lower.

Pixel Accuracy (PA):

$$PA = \frac{\sum_{i=0}^{n} p_{ii}}{\sum_{i=0}^{n} \sum_{j=0}^{n} p_{ij}} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

In this context, *TP* (True Positive) refers to the number of pixels correctly predicted as the positive class. *TN* (True Negative) refers to the number of pixels correctly predicted as the negative class. *FP* (False Positive) refers to the number of negative pixels incorrectly predicted as positive, and *FN* (False Negative) refers to the number of positive pixels incorrectly predicted as negative. Therefore, pixel accuracy measures the proportion of correctly predicted pixels among all pixels. The higher the value, the better the model's overall performance.

The Mean Intersection over Union (MIoU):

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{7}$$

IoU (Intersection over Union) measures the overlap between the model's prediction and the ground truth for each class. The higher the value, the more accurate the model's prediction. The MIoU (Mean Intersection over Union) is the average of IoU values, providing a comprehensive evaluation of the model's performance across multiple classes. It is especially useful for multi-class segmentation tasks.

4.3. Comparison with Other Models

This study evaluated 7 different network architectures under the same training conditions on the same dataset, with the results presented in Table 1. The following models are commonly used in recent image segmentation research, and the results were obtained using the fish dataset.

The system configuration consisted of an Intel 12th Gen Core™ i5-12400F processor, 3060 GPU, and 16 GB (8 GB × 2) of RAM. All comparison groups presented in this paper were trained using this machine with identical parameter settings, and all models were pre-trained on the ImageNet dataset to ensure the initial parameters were not excessively biased. The model was trained using Stochastic Gradient Descent (SGD) with a learning rate of 7×10^{-3}, a minimum learning rate of $(7 \times 10^{-3}) \times 0.01$, and a cosine learning rate decay. The training was carried out for 100 epochs, with a batch size of 4. The downsample_factor

was set to 16, as a value of 8 was not feasible due to GPU memory limitations. The input image size was 512 × 512, and Weight decay was set to 1×10^{-4} to prevent overfitting.

Table 1. Comparison of Network Segmentation Performance.

	F1-Score	MIoU	PA
UNet [4]	0.891	0.923	0.912
PSPNet [10]	0.879	0.902	0.881
Deepv3 (VGG16) [12,33]	0.895	0.911	0.900
Deepv3+ (VGG16) [12,33]	0.881	0.922	0.913
Deepv3 (ResNet50) [12,34]	0.895	0.915	0.910
Deepv3+ (ResNet50) [12,34]	0.897	0.925	0.901
Proposed model	0.901	0.932	0.921

All of the algorithm models used in this paper are based on publicly available research papers. No modifications were made except for the addition of backbone networks. The models include UNet [4], PSPNet [10], DeepV3, DeepV3+ [12], VGG16 [33], and ResNet50 [34].

Table 1 represents the average accuracy of different models for the same test set under the same training environment, starting from three indicators respectively. Figure 4 is the visual display and comparison, Figure 5 is the segmentation effect comparison.

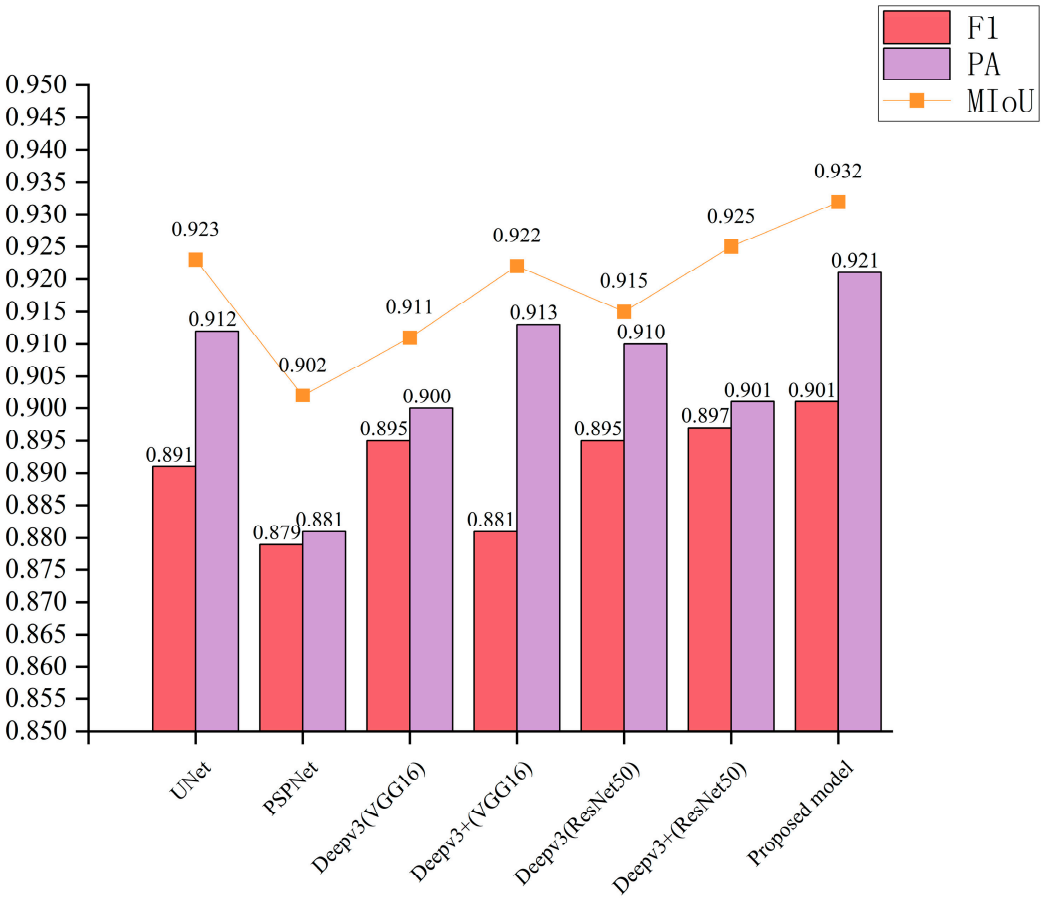


Figure 4. Data Comparison Chart.

To further validate the stability of the model proposed in this paper, the Black Sea Spart dataset was selected as an additional validation set, providing a more intuitive understanding of the improvement in segmentation performance. The results are shown in Figure 6.

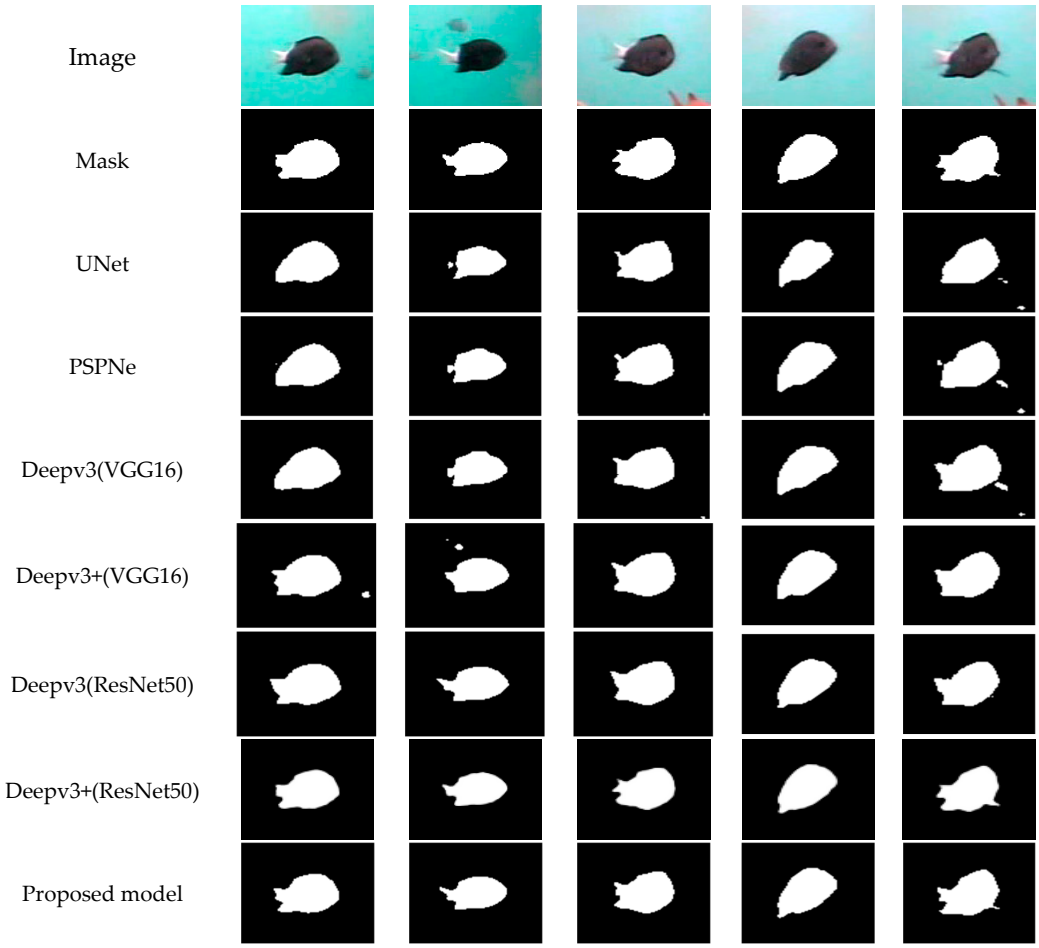


Figure 5. Results Display.

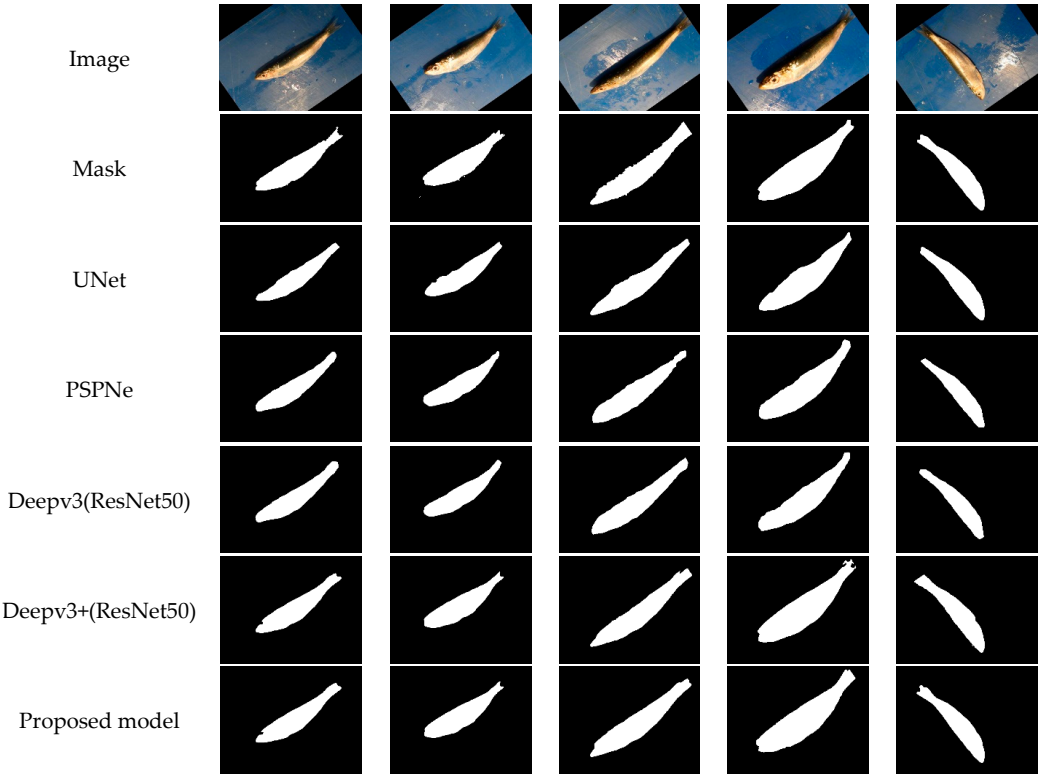


Figure 6. Black Sea Spart.

4.4. Discussion

The fish segmentation model proposed in this study achieved the highest F1 score among all of the models. In comparison with other algorithms, it consistently demonstrated state-of-the-art detection metrics. Furthermore, the ResNet50 backbone outperformed VGG16 in the fish segmentation tasks. Compared to the baseline algorithms, the algorithm incorporating the pyramid feature extraction architecture showed a significant improvement in segmentation performance, sacrificing some computational speed in exchange for a higher detection accuracy.

The Table 2 were obtained after testing the model post-training, where the time taken was averaged over ten repeated tests to reduce errors. As is shown in the chart, VGG16 has the largest number of parameters due to its three fully connected layers. DeepV3 has the advantage in terms of size while maintaining a certain level of accuracy. The model proposed in this paper, however, involves multiple iterations through the feature pyramid and attention modules, leading to some parameter redundancy and a disadvantage in speed. This is one of the areas for future improvement.

Table 2. Time and Parameters.

	Parameters	Time
UNet	73.52 M	42
PSPNet	60.89 M	39
Deepv 3 (VGG16)	153.25 M	53
Deepv 3+ (VGG16)	123.11 M	47
Deepv 3 (ResNet50)	97.65 M	43
Deepv 3+ (ResNet50)	83.65 M	42
Proposed model	145.9 M	56

4.5. Ablation Experiment

We use the ResNet50 backbone as the baseline to highlight the contribution of different modules. The first step was to evaluate how much the accuracy of the new network improved compared to the original model when various modules were added. The segmentation results after modifying the pyramid architecture and shape attention module in the network are presented in Table 3. The experimental results demonstrate that the pyramid feature extraction architecture improved segmentation accuracy by approximately 4% in underwater fish segmentation. Additionally, adding the shape attention module after the convolutional blocks of ResNet50 led to an improvement in detection accuracy.

Table 3. Ablation Experiment I.

	F1	MIoU	Time (ms)
Resnet50 [34]	0.830	0.841	46
Resnet50 + XA	0.885	0.875	45
Resnet50 + XA + PAFE	0.922	0.928	54

Next, we investigated specific parameters in the architecture, beginning with the role of the 1×1 COV and whether it can replace the fully connected layer. From Table 4, we observe that incorporating the 1×1 COV into the PAFE architecture resulted in noticeable improvements in image segmentation performance. The new network efficiently extracts higher-quality multi-scale features from images. Therefore, the effects of dilation and the 1×1 COV are significant. Figure 7 also confirms this point from the perspective of implementation effect.

Table 4. Ablation Experiment II.

	F1	MIoU	Time (ms)
1 × 1COV	0.922	0.922	46
FC	0.910	0.908	65
D = 2 4 6	0.905	0.915	54
D = 2 4 5	0.924	0.923	55

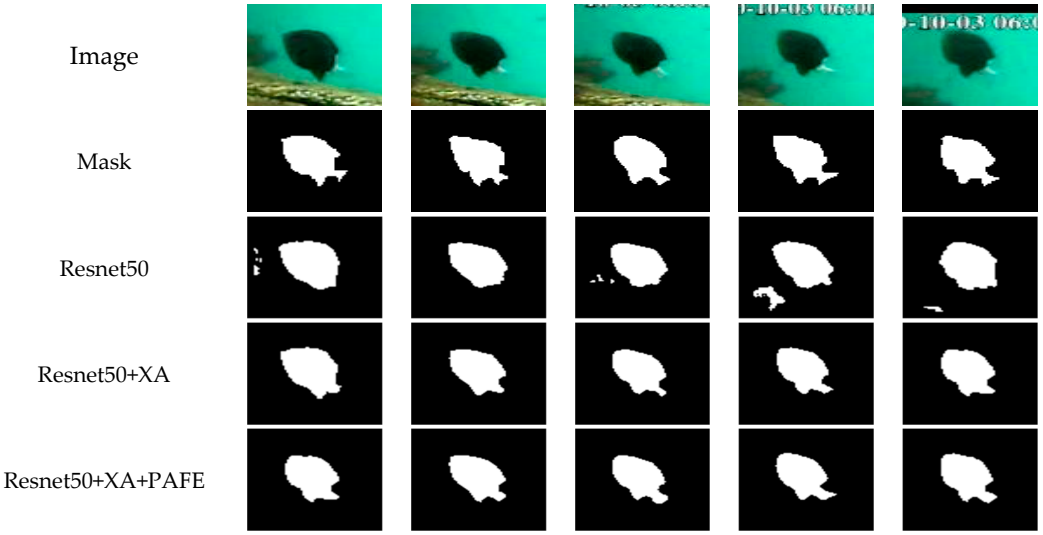


Figure 7. Ablation Experiment III.

4.6. Validation on Real Deep-Sea Fish Images

We constructed a real underwater image dataset consisting of over a hundred high-definition images captured at a depth of 3000 m using specialized underwater imaging equipment. The dataset includes underwater species such as fish, crabs, and shrimp, where the targets closely resemble the background. After preprocessing the images, segmentation experiments were conducted using various algorithms. The results show that classic neural networks, prior to any improvements, performed poorly in segmentation. However, the model proposed in this paper, inspired by camouflaged object segmentation techniques, optimizes the loss function and successfully segments the target to some extent. This approach presents a promising direction for future research regarding how to effectively segment fish in real-world environments. Figure 8 Model segmentation effect, the middle column represents the original model effect.

Obtaining deep-sea fish images is challenging due to issues such as uneven lighting and the high similarity between the target and the background. The model presented in this paper improves upon the combination of neural networks and feature pyramids by focusing on the specific regions where the targets are located for detailed segmentation. Traditional segmentation algorithms struggle to distinguish between fish shadows and fish bodies, especially in areas with similar brightness. To enhance the model's performance in deep-sea fish image segmentation, specific optimizations and image preprocessing steps are necessary.

In deep-sea fish image segmentation, two main challenges arise: first, the fish shadow and the fish body are similar in shape; second, the RGB values of the bait and soil reflections are close to those of the fish body. To address these challenges, both the images and the model underwent further processing, as shown in Figure 9. During segmentation, the algorithm first optimizes the image by enhancing contrast and improving readability. The model then prioritizes color features, comparing them with the pre-trained model to identify key color values. This step filters out fish shadows and soil color blocks, focusing the attention on

the fish itself. Next, the PAFE module uses optimized parameters to identify the fish's shape features. Thanks to pre-training, the model performs secondary identification, directly detecting the fish and generating a segmented image. However, since the shape and reflective RGB values of the bait are similar to those of the fish, the bait is segmented alongside the fish. This is an area that requires further improvement in future research.

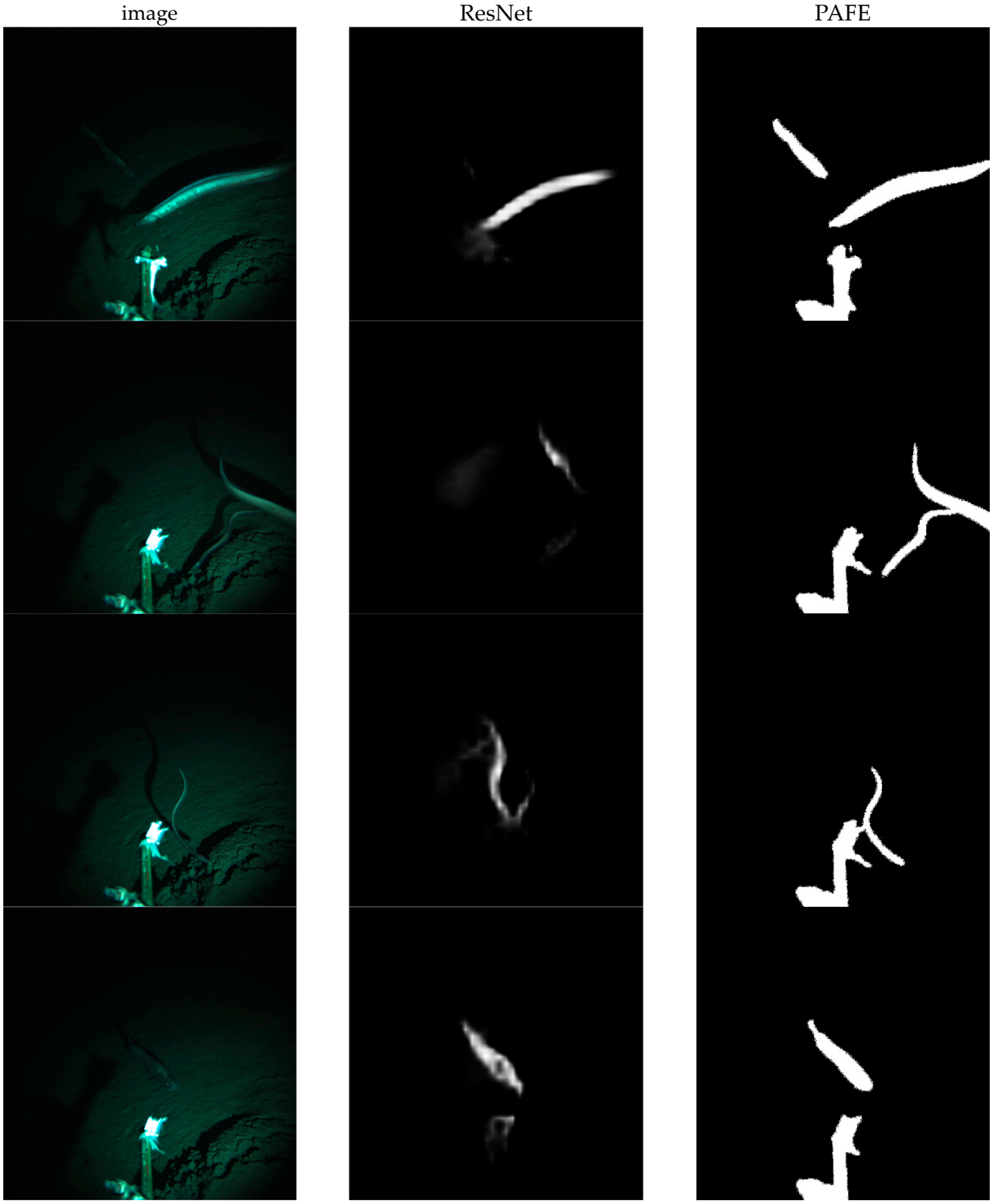


Figure 8. *Cont.*

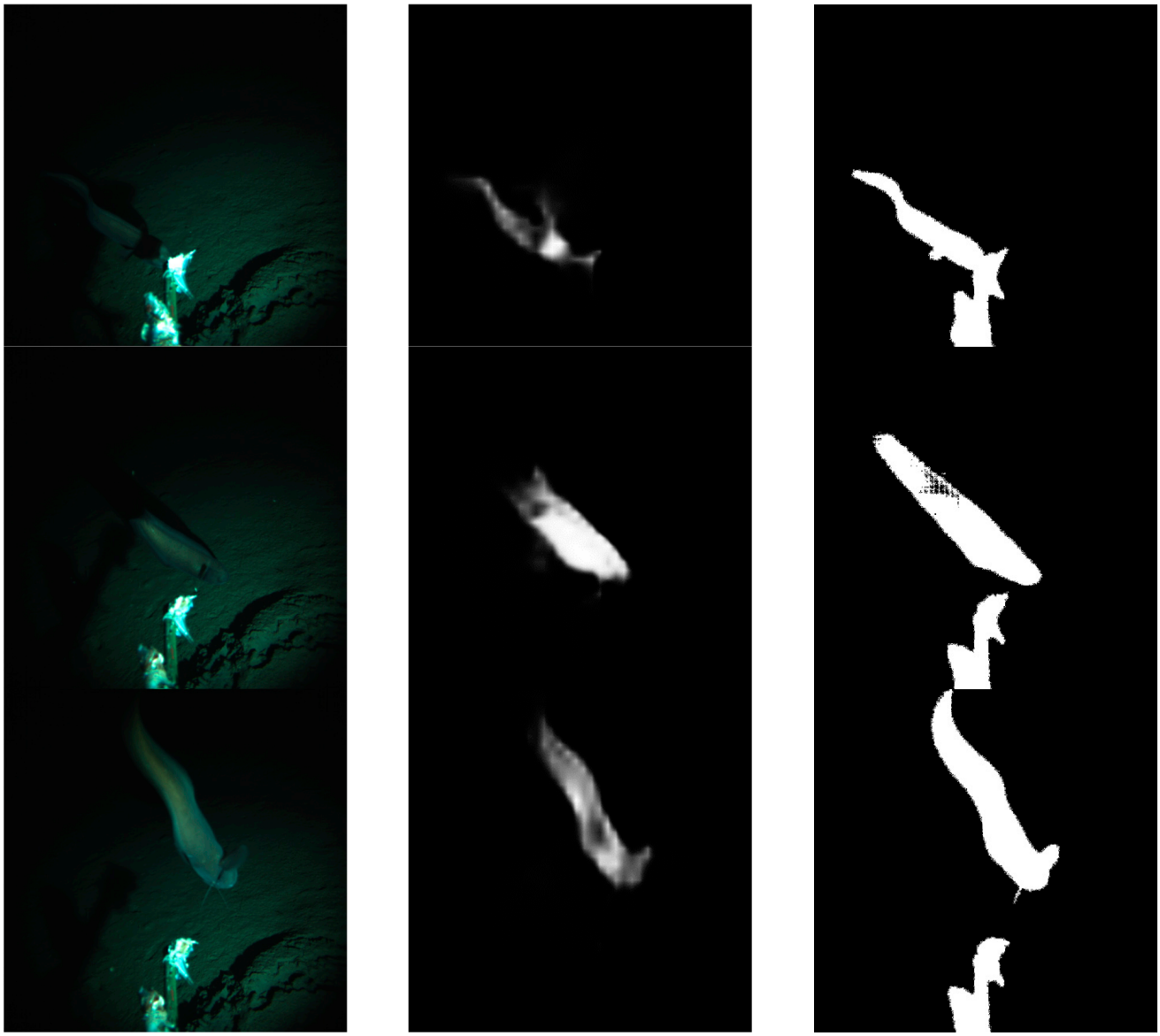


Figure 8. Presentation of Experimental Results for Deep-Sea Fish Images.

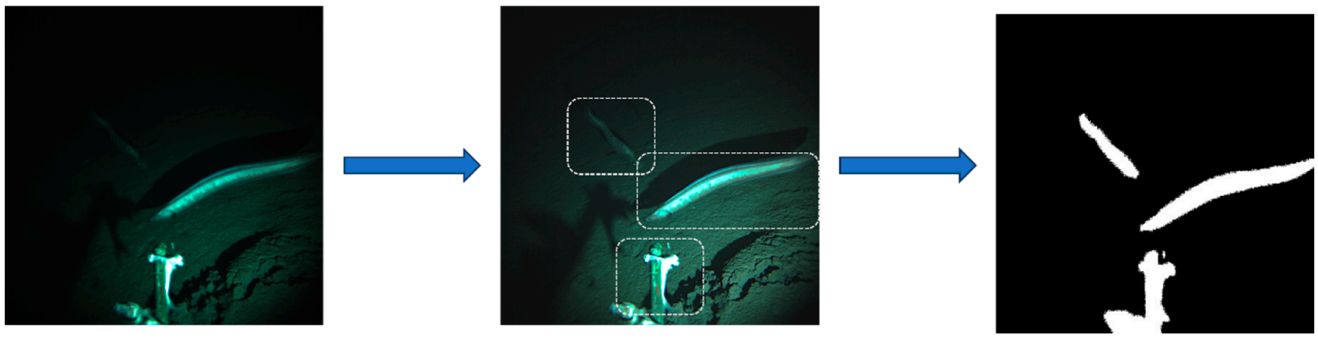


Figure 9. Image enhancement effect.

5. Conclusions

To address the challenges in underwater fish image segmentation, this study proposes an innovative model based on a combination of the attention mechanism and feature pyramid module. The model's uniqueness lies in its ability to adaptively assign weights through the attention mechanism, enhancing feature selectivity and sensitivity, while also extracting

features from multiple levels using the feature pyramid module to strengthen the correlation of contextual information. This multi-level feature extraction strategy significantly boosts segmentation accuracy. In particular, the feature pyramid structure effectively integrates high-level semantic information with low-level detail features, enabling the model to better capture fish boundaries and shape features in complex scenes, thereby improving the overall segmentation performance. The model achieves high-precision segmentation results through pixel-level classification. Comprehensive evaluation of an underwater fish image dataset, along with comparison with other mainstream models, demonstrates the significant advantage of the proposed model in terms of segmentation accuracy. This study not only provides a novel and effective tool for underwater image processing but also lays the groundwork for future research, highlighting its broad application potential, particularly in target recognition and segmentation tasks in complex environments.

Equation (12) explains the principle of the attention mechanism, while Equation (1) describes the recalculation and redistribution of weights in the attention mechanism itself. Equation (2) represents channel attention, which compresses global information to expand the receptive field and is one of the cornerstones of the proposed model. Equation (3) refers to dilated convolution, part of the pyramid structure, which increases the receptive field and facilitates multi-scale feature extraction from deep feature maps, improving segmentation accuracy. Equation (4) represents the model's loss function, which combines two loss functions: cross-entropy loss to optimize pixel-level classification accuracy, and Dice loss to ensure the overlap between the segmentation shape and the ground truth label. This combined approach is particularly suitable for image segmentation tasks, especially when the target object is unevenly represented. Equations (5)–(7) assess the utility of the proposed model from three different perspectives. F1 measures the model's ability to maintain a high accuracy across a large sample size, its error rate, and its practical application potential. PA represents pixel accuracy for a single image, testing whether the model can reduce errors to the pixel level in complex environments, rather than just recognizing contours with limited precision. MIoU is an intuitive metric that evaluates the generalizability of the trained model when segmenting images with varying characteristics, given the ground truth.

The complexity of the marine underwater environment and the uncertainty of image clarity require segmentation models to be capable of handling variable and blurry images. While existing models have demonstrated good performance, there remains significant potential for improvement. Future research will primarily focus on optimizing the feature pyramid, particularly in the design and implementation of the feature extraction module. Specifically, exploring more comprehensive and efficient dilated convolution pyramid feature extraction modules will enhance the model's ability to capture multi-scale and blurry features. Improvements should concentrate on strengthening the correlation between features at different levels, thereby improving the model's understanding of complex underwater scenes. Moreover, combining image enhancement techniques with convolutional networks to develop novel segmentation models will provide more effective solutions for monitoring underwater ecosystems. By incorporating advanced image enhancement methods, the quality of input images can be significantly improved, thereby boosting the accuracy of segmentation results.

Finally, future research should also aim to develop learning models with higher interpretability to ensure the transparency and comprehensibility of the decision-making process. This will not only help to increase user's trust in the model but also drive further advancements in the field of underwater image segmentation.

Author Contributions: Resources, J.Y. and D.Y.; Data curation, W.F.; Writing—original draft, G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key R&D Program of China, (No. 2022YFC2803903).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hussain, M.; Saputra, T.; Szabo, E.; Nelan, B. An Overview of Seafood Supply, Food Safety and Regulation in New South Wales, Australia. *Foods* **2017**, *6*, 52. [CrossRef] [PubMed]
2. Allken, V.; Handegard, N.O.; Rosen, S.; Schreyeck, T.; Mahiout, T.; Malde, K. Fish Species Identification Using a Convolutional Neural Network Trained on Synthetic Data. *ICES J. Mar. Sci.* **2019**, *76*, 342–349. [CrossRef]
3. Zhang, Y.; He, S.; Wa, S.; Zong, Z.; Liu, Y. Using Generative Module and Pruning Inference for the Fast and Accurate Detection of Apple Flower in Natural Environments. *Information* **2021**, *12*, 495. [CrossRef]
4. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
5. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
6. Konovalov, D.A.; Saleh, A.; Bradley, M.; Sankupellay, M.; Marini, S.; Sheaves, M. Underwater Fish Detection with Weak Multi-Domain Supervision. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8.
7. Er, M.J.; Chen, J.; Zhang, Y.; Gao, W. Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. *Sensors* **2023**, *23*, 1990. [CrossRef]
8. Cao, F.; Zhao, H. Automatic Lung Segmentation Algorithm on Chest X-Ray Images Based on Fusion Variational Auto-Encoder and Three-Terminal Attention Mechanism. *Symmetry* **2021**, *13*, 814. [CrossRef]
9. Ditria, E.M.; Lopez-Marcano, S.; Sievers, M.; Jinks, E.L.; Brown, C.J.; Connolly, R.M. Automating the Analysis of Fish Abundance Using Object Detection: Optimizing Animal Ecology with Deep Learning. *Front. Mar. Sci.* **2020**, *7*, 429. [CrossRef]
10. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 640–651.
11. Majumder, A.; Rajbongshi, A.; Rahman, M.M.; Biswas, A.A. Local Freshwater Fish Recognition Using Different CNN Architectures with Transfer Learning. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2021**, *11*, 1078–1083. [CrossRef]
12. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:1602.07261. [CrossRef]
13. Abinaya, N.S.; Susan, D.; Sidharthan, R.K. Deep Learning-Based Segmental Analysis of Fish for Biomass Estimation in an Occulted Environment. *Comput. Electron. Agric.* **2022**, *197*, 106985. [CrossRef]
14. Jahanbakht, M.; Rahimi Azghadi, M.; Waltham, N.J. Semi-Supervised and Weakly-Supervised Deep Neural Networks and Dataset for Fish Detection in Turbid Underwater Videos. *Ecol. Inform.* **2023**, *78*, 102303. [CrossRef]
15. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. *arXiv* **2016**, arXiv:1612.01105.
16. 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. *arXiv* **2017**, arXiv:1606.00915. [CrossRef]
17. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
18. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. *arXiv* **2018**, arXiv:1809.02983.
19. Liu, C.; Chen, L.-C.; Schroff, F.; Adam, H.; Hua, W.; Yuille, A.; Fei-Fei, L. Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. *arXiv* **2019**, arXiv:1901.02985.
20. Alshdaifat, N.F.F.; Talib, A.Z.; Osman, M.A. Improved Deep Learning Framework for Fish Segmentation in Underwater Videos. *Ecol. Inform.* **2020**, *59*, 101121. [CrossRef]
21. Zhou, Z.; Yang, X.; Ji, H.; Zhu, Z. Improving the Classification Accuracy of Fishes and Invertebrates Using Residual Convolutional Neural Networks. *ICES J. Mar. Sci.* **2023**, *80*, 1256–1266. [CrossRef]
22. Akram, W.; Hassan, T.; Toubar, H.; Ahmed, M.; Miškovic, N.; Seneviratne, L.; Hussain, I. Aquaculture Defects Recognition via Multi-Scale Semantic Segmentation. *Expert Syst. Appl.* **2024**, *237*, 121197. [CrossRef]
23. Zhou, Z.; Hu, Y.; Yang, X.; Yang, J. YOLO-Based Marine Organism Detection Using Two-Terminal Attention Mechanism and Difficult-Sample Resampling. *Appl. Soft Comput.* **2024**, *153*, 111291. [CrossRef]
24. Li, D.; Yang, Y.; Zhao, S.; Yang, H. A Fish Image Segmentation Methodology in Aquaculture Environment Based on Multi-Feature Fusion Model. *Mar. Environ. Res.* **2023**, *190*, 106085. [CrossRef] [PubMed]

25. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507.
26. Li, D.; Yang, Y.; Zhao, S.; Ding, J. Segmentation of Underwater Fish in Complex Aquaculture Environments Using Enhanced Soft Attention Mechanism. *Environ. Model. Softw.* **2024**, *181*, 106170. [CrossRef]
27. Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; Zhang, L. A Single Stream Network for Robust and Real-Time RGB-D Salient Object Detection. *arXiv* **2020**, arXiv:2007.06811.
28. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *arXiv* **2016**, arXiv:1606.04797.
29. Mao, A.; Mohri, M.; Zhong, Y. Cross-Entropy Loss Functions: Theoretical Analysis and Applications. *arXiv* **2023**, arXiv:2304.07288.
30. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
31. DeVries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
32. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. *arXiv* **2019**, arXiv:1905.04899.
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
34. Beyer, L.; Zhai, X.; Royer, A.; Markeeva, L.; Anil, R.; Kolesnikov, A. Knowledge Distillation: A Good Teacher Is Patient and Consistent. *arXiv* **2022**, arXiv:2106.05237.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.