applied sciences

MDPI

Article

An FSFS-Net Method for Occluded and Aggregated Fish Segmentation from Fish School Feeding Images

Ling Yang [1,2], Yingyi Chen [3,4,*], Tao Shen [1,2,*] and Daoliang Li [3,4]

1 Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming 650500, China; yangling@cau.edu.cn
2 Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
3 National Innovation Center for Digital Fishery, China Agricultural University, Beijing 100083, China; dliangl@cau.edu.cn
4 College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
* Correspondence: chenyingyi@cau.edu.cn (Y.C.); shentao@kust.edu.cn (T.S.)

Abstract: Smart feeding is essential for maximizing resource utilization, enhancing fish growth and welfare, and reducing environmental impact in intensive aquaculture. The image segmentation technique facilitates fish feeding behavior analysis to achieve quantitative decision making in smart feeding. Existing studies have largely focused on single-category object segmentation, ignoring issues like occlusion, overlap, and aggregation amongst individual fish in the fish feeding process. To address the above challenges, this paper presents research on fish school feeding behavior quantification and analysis using a semantic segmentation algorithm. We propose the use of the fish school feeding segmentation method (FSFS-Net), together with the shuffle polarized self-attention (SPSA) and lightweight multi-scale module (LMSM), to achieve two-class pixel-wise classification in fish feeding images. Specifically, the SPSA method proposed is designed to extract long-range dependencies between features in an image. Moreover, the use of LMSM techniques is proposed in order to learn contextual semantic information by expanding the receptive field to extract multi-scale features. The extensive experimental results demonstrate that the proposed method outperforms several state-of-the-art semantic segmentation methods such as U-Net, SegNet, FCN, DeepLab v3 plus, GCN, HRNet-w48, DDRNet, LinkNet, BiSeNet v2, DANet, and CCNet, achieving competitive performance and computational efficiency without data augmentation. It has a 79.62% mIoU score on annotated fish feeding datasets. Finally, a feeding video with 3 min clip is tested, and two index parameters are extracted to analyze the feeding intensity of the fish. Therefore, our proposed method and dataset provide promising opportunities for the urther analysis of fish school feeding behavior.

Keywords: fish feeding behavior; semantic segmentation; attention mechanism; multi-scale feature; intensive aquaculture

check for updates

Citation: Yang, L.; Chen, Y.; Shen, T.; Li, D. An FSFS-Net Method for Occluded and Aggregated Fish Segmentation from Fish School Feeding Images. *Appl. Sci.* **2023**, *13*, 6235. https://doi.org/10.3390/app13106235

Academic Editor: Yu-Dong Zhang

Received: 18 April 2023
Revised: 7 May 2023
Accepted: 11 May 2023
Published: 19 May 2023

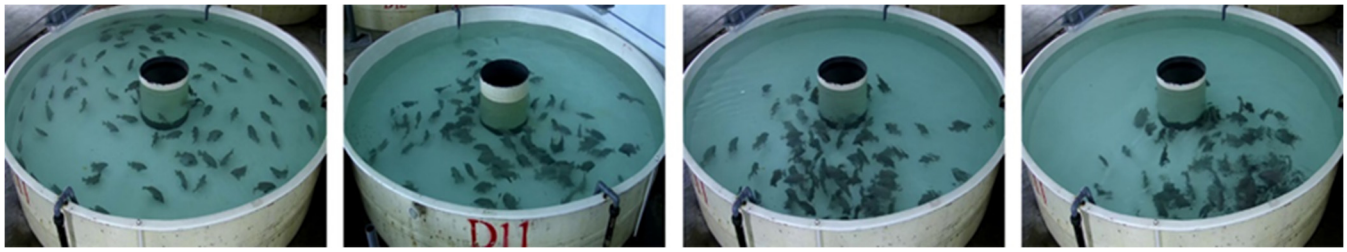
Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Intelligent feeding is crucial in intensive aquaculture systems for optimizing resource usage and promoting fish development and well-being while minimizing environmental consequences. Image analysis technologies have emerged as powerful tools in achieving precision aquaculture and have shown potential in image-based tasks such as fish identification [1], detection [2], tracking and behavior analysis [3]. Using image technologies to analyze fish behavior enables us to provide important information to farmers for their use in making feeding decisions. In general, image techniques based on tracking are used to extract fish trajectories and then utilize extracted characteristic from trajectories to quantify individual fish feeding behavior [4–7]. However, these methods are not suitable for intensive aquaculture since fishes are generally managed at the group level, and motion of fish school is complex [8]. Recently, fish school feeding behavior analysis has

appeared as a competitive alternative to individual behavior analysis [9–14]. For the quantification analysis of feeding behavior, one of the most challenging problems is automatic fish segmentation. Thus, it is essential to obtain a precise, reliable, and standardized fish segmentation approach, surpassing a trained human-level performance, which can provide valuable assistance to farmers.

Several segmentation methods have been widely applied in fish size measurement, counting, and species identification. Traditional segmentation methods, including background models [15] and contour-based [16] and color-based [17,18] models, focus on segmenting single-category objects and only segment the objects in an image into foreground and background. However, the traditional methods do not accurately segment the object into multiple categories in the case of images in complex scenes. Moreover, there are still significant challenges to the segmentation of individual fish from fish school feeding images due to motion blur, adhesion, occlusion, overlap, and aggregation between individual fish, as shown in Figure 1b–d, ultimately leading to incorrect behavior analysis. The characteristics vary considerably between feeding image and non-feeding image, resulting in the algorithm designed on the non-feeding image not being well generalized to the feeding image.



(**a**) (**b**) (**c**) (**d**)

Figure 1. Existing dataset examples with different feeding states. These images show difference between feeding and non-feeding image states. There is motion blur, adhesion, occlusion, overlap, and clustering in the feeding image. (**a**) Non-feeding; (**b**) moderate feeding; (**c**) weak feeding; (**d**) strong feeding.

The semantic segmentation approach proved to be excellent for addressing multi-category object segmentation problems in an image [19]. It is a pixel-wise classification technique, with differences from other classification and detection methods. Its main task is assigning each pixel from the input image with a semantic label and aims to segment the instances of interest. With the popularization of data-driven deep learning techniques, the use of semantic segmentation has increased dramatically in image analysis tasks and demonstrated promising results in empirical tasks such as autonomous driving [20], and medical diagnosis [21,22]. Despite the significant progress achieved, the most existing SOTA semantic methods are not directly applicable since the size of objects and categories in an image are diverse. Therefore, the algorithm needs to be redesigned for application to different segmentation tasks. On the other hand, deep learning algorithms have shown great potential for automatic fish identification, detection, and individual tracking, but there are few studies to have quantified fish school feeding behavior by using a semantic segmentation algorithm.

Fish semantic segmentation is a challenging endeavor that faces two major obstacles: it is multi-scale and there are similarities between fish. (1) In terms of multiple scales between the same categories and different categories: as shown in Figure 2, for different categories, one is small in scale, and the other has a large size difference. In addition, there are existing scale differences between the same categories. Therefore, multi-scale objects require the model to have multi-scale capture capabilities. (2) The similarity between fish causes category confusion due to occlusion and overlapping between the individual fish, which may lead to low segmentation accuracy. Moreover, the similarity of colors between

individual fish and the blurring of object edges make it difficult to depict fish boundaries and classify them into correct categories for the pixels close to or at the boundaries of fish due to the intrinsic weakness of convolutional networks. The above factors bring serious segmentation challenges and make it difficult to generalize the model.

Figure 2. Multi-scale in the fish feeding image. (**A**) multi-scale between different categories; (**B**,**C**) multi-scale between the fish2 semantic category; (**D**,**E**) multi-scale between the fish1 semantic category. The white font indicates the semantic category of fish objects.

To solve the above problems, we use a semantic segmentation algorithm to quantify fish school feeding behavior in a high-density scene. In this paper, we propose the use of an FSFS-Net method to achieve the segmentation of two-category objects (labeled rules for semantic categories are detailed in Section 4.1) in the feeding process. To solve the problems of similarity, we propose the use of shuffle polarized self-attention (SPSA) methods to learn dependencies between features in images. Furthermore, considering the trade-off between speed and accuracy performance, we also propose a lightweight multi-scale module (LMSM) to extract multi-scale features and more high-level semantic features by using a different the receptive field which does not introduce computation overhead for inference. Finally, we conduct comprehensive experiments to evaluate the effectiveness of the algorithm. The experimental results show that our proposed method achieves the competitive performance with a 79.62 mean IoU without dataset augmentation of fish feeding dataset, and that it outperforms other segmentation algorithms, such as U-Net, SegNet, FCN, DeepLab v3 plus, GCN, HRNet-w48, DDRNet, LinkNet, BiSeNet v2, DANet and CCNet. To the best of our knowledge, the superior performance of our proposed method over the state-of-the-art method is mainly due to the proposed SPSA and LMSM module. The result and analysis of comparative experiments prove that our method is straightforward and effective, and that it can be used to quantify the feeding intensity of a fish school.

Our contributions can be summarized as follows:

- An annotated dataset is presented for occluded and aggregated fish segmentation from feeding images. The dataset contains 1038 training and 323 testing images, as well as their ground truth semantic labels. The 1361 images marked have 72,537 instances with the fish1 category and 15,925 instances with fish2 category, and the background pixel accounts for a relatively large amount in an image.
- This study demonstrates that semantic segmentation is employed in fish school feeding behavior quantification. Particularly, we propose using a FSFS-Net method, equipped with a simple and effective module, to extract feature dependencies and multi-scale features from an image.
- A SPSA module is proposed to learn dependencies between features, which can alleviate color and size similarity. Additionally, a novel LMSM is developed to fuse multi-scale features by using different receptive fields to address the multi-scale problem, which is helpful in the task of segmentation.
- The proposed approach achieves a competitive performance with a mean IoU of 79.62% while ensuring computational efficiency over other semantic segmentation algorithms for building fish feeding datasets. Further, the proposed approach is tested on a video with 3 min clip to analyze the feeding intensity of fish. The promising results demonstrate the superiority of our proposed algorithm, the use of which would promote fish segmentation and behavior analysis research in aquaculture.

2. Related Work

2.1. Fish Segmentation

Fish segmentation is a prerequisite for fish identification, size measurement, counting and behavior analysis. Toward these goals, numerous researchers are carrying out extensive exploration in this field. However, fish segmentation is a very challenging task in the complex background due to the similarities in color, shape, and fish overlapping. To overcome the aforementioned issues, various segmentation methods are proposed by researchers according to different visual tasks with different scenes. There are three common methods of fish segmentation: traditional method, semantic segmentation method, and instance segmentation method. These methods and their application have been described in Table 1.

Table 1. Summary of fish segmentation method and its application. Category and Number, respectively, represent the segmentation category and the number of fishes in an image.

	Method	Category	Number	Application
Traditional method	Background models [15], clustering [23].	One-class	Fish school	Counting, biomass and behavior.
	Contour-based [16], improved thinning [24], color-based [17,18].	One-class	Single fish	Mass and size measurement, identification.
Semantic segmentation	SegNet [25]	Two-class	Single fish	Size measurement.
	ResNet-FCN [26], U-Net [27], DPANet [28], SegNet [29].	One-class	Fish school, multi-fish.	Counting, size measurement.
Instance segmentation	Self-proposed [30]	Multi-category	Multi-fish	Identification
	Mask R-CNN [31] and other [32]	One-class	Single fish	Size measurement, identification.
	Mask R-CNN [33,34]	One-class	Multi-fish	Tracking, size measurement.
	Mask R-CNN [35]	Four-class	Single fish	Morphological features.

Traditional fish segmentation methods are based on unsupervised techniques, and data do not need to be manually labeled as in color-based the GrabCut, Otsu threshold, edge detection, mean-shift and region growing algorithm [17,18]. These methods are beneficial for determining the difference between the color of fish object and background. However, they are manual segmentation methods and require a white or uniform background. Background subtraction and background modeling have proved to be useful for station underwater scenes to achieve counting in fish school segmentation because the difference between fish objects and background is obvious. If a background changes rapidly, background subtraction may lose its precision. Moreover, background subtract cannot handle well several underwater imaging problems, such as varying or sudden illumination caused by wave movement or color changes. To address these problems, an approach is used to eliminate reliance on background features by using motion-robust features such as edges or shapes [16].

Nonetheless, the traditional methods mentioned above are only proposed for segmenting single-category objects in fish school or in images of single fish. The deep learning method, as a powerful tool, enables researchers to extract complex global and local features. With the popularization of deep learning, the majority of existing works have utilized deep learning to precisely segment fish objects. Semantic and instance segmentation methods based on deep learning can perform pixel-wise classification without relying on a static background, effectively distinguishing the foreground fish from the background. More importantly, their use has been explored for size measurement, counting and species identification, and they are employed to segment multi-category objects. Specially, the SegNet [25] method is used to segment fish body and fins for size measurement in a single-fish image. Some semantic-based methods are utilized to segment single-category in multi-fish image, such as ResNet-FCN [26], U-Net [27], DPANet [28], and SegNet [29]. In particular, SegNet [29] has achieved a segmentation accuracy of 72% mIoU for counting in

aquaculture net cages. The instance-based method has achieved multi-class segmentation in the multi-fish image [30] and multi-class segmentation in the single-fish image in species identification from underwater [35].

Although semantic and instance segmentation have achieved an excellent performance, all fish objects are labeled as foreground categories for counting in the high-density fish school image segmentation, as shown in Figure 3. However, there are few reports on fish feeding behavior quantification using a segmentation method. Unlike previous works, we introduce a novel semantic segmentation method for fish school feeding behavior quantification in industrial recycling aquaculture. Our proposed method can achieve two-class pixel-wise classification in the fish school feeding process.

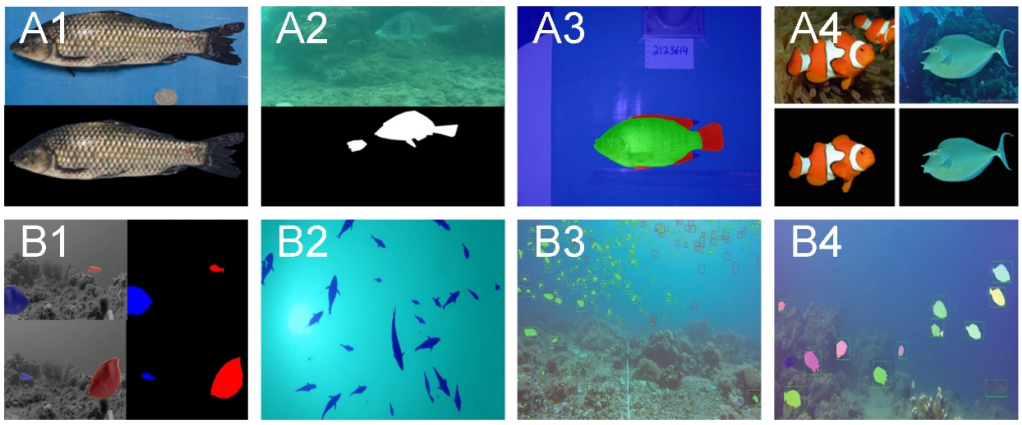


Figure 3. Examples image of fish segmentation in previous works. (**A1**): single fish and single category in the simple background; (**A2**): multi-fish and single category in the underwater environment; (**A3**): single fish and multi-category in the simple background; (**A4**): single fish and multi-category, but only single category in an image; (**B1**): multi-fish and multi-category in underwater; (**B2**,**B3**): single fish in underwater fish school image; (**B4**): multi-fish and multi-category in the complex underwater environment.

2.2. Semantic Segmentation

Deep learning methods have made tremendous progress and have achieved very impressive results in image classification and detection. Despite the success, these methods still face existing challenges in image segmentation, including multi-scale features, receptive field, and low resolution. Among them, the occurrence of multi-scale features is caused by the different objects size between the same or different categories in an image. A small convolution kernel has a small receptive field, which is not conducive to extracting contextual information. While a large convolution kernel can solve the problem of the receptive field, it increases model parameters, meaning that the real-time requirements cannot be met. For resolution issues, low-stage features have more detailed information, while the high-stage features have higher semantic information. Using a downsampling operator increases the receptive field to extract high-level (global/contextual) semantic information, but it reduces the image resolution, making the algorithm unable to accurately locate the individual and extract the object spatial information (object contour and edge information).

To address the abovementioned challenges, various advanced semantic segmentation algorithms are proposed. SegNet [36] and U-Net [37] adopted encoder–decoder architecture to recover low-level spatial information for keeping a sharper edge. SegNet utilized saved pool indices to recover the reduced spatial information in downsampling. U-Net utilized skip connections to fuse features from different layers. Some straightforward ways to extract multi-scale features include resizing the feature map to different scales using the proposed PSP [38], ASPP [39], and multi-receptive field module (MRFM) [40], as shown in Figure 4. PSPNet and DeeplabV3 use the PSP and ASPP modules, which utilize several pooling with different sizes and multiple parallel atrous convolutions to obtain multi-scale

receptive fields, and these modules are located at the end of the backbone. Additionally, the multi-receptive field module (MRFM) [40] is introduced by re-designing the backbone to obtain multi-scale features. Nonetheless, different scales yield features with varying degrees of discrimination, resulting in inconsistency.

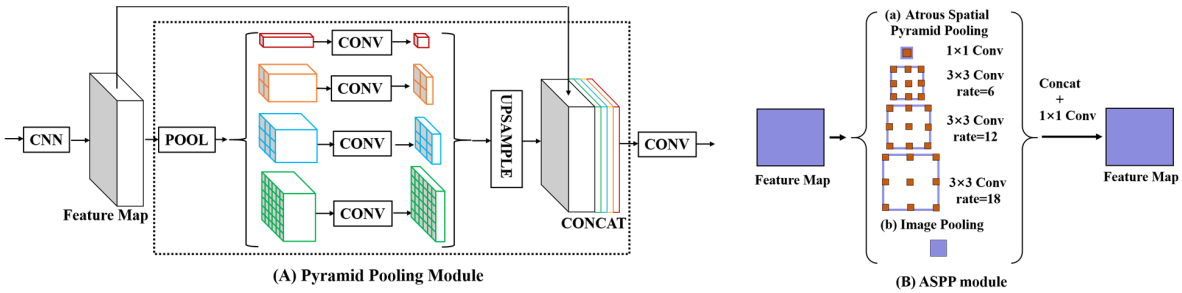


Figure 4. Multi-scale module for extracting contextual semantic information.

The attention mechanism can capture global dependencies. To distinguish between the boundaries of objects (classification of pixels at edges) and extract key features for feature fusion from different stages, some studies utilized attention mechanisms to extract spatial information and build context dependencies by learning the relationship between pixels. Examples of this approach include DFN [41], DANet [42], OCNet [43], OCRet [44] and other related studies, and they have achieved outstanding results on public datasets. Inspired by them, we propose a multi-scale module to extract features of different scales of target. Subsequently, we propose a shuffle polarized self-attention to extract key features and global dependencies between features.

3. Method

We rethink the UNet and propose a shuffle polarized self-attention U-Net architecture (FSFS-Net). The overall architecture is shown in Figure 5B. The proposed method consists of U-Net, a shuffle polarized self-attention module (SPSA), and a global context module (LMSM). Specifically, an SPSA can learn the relationship between pixels in an image. LMSMs are used to generate high-level semantic global maps. We introduce each component in the following subsections.

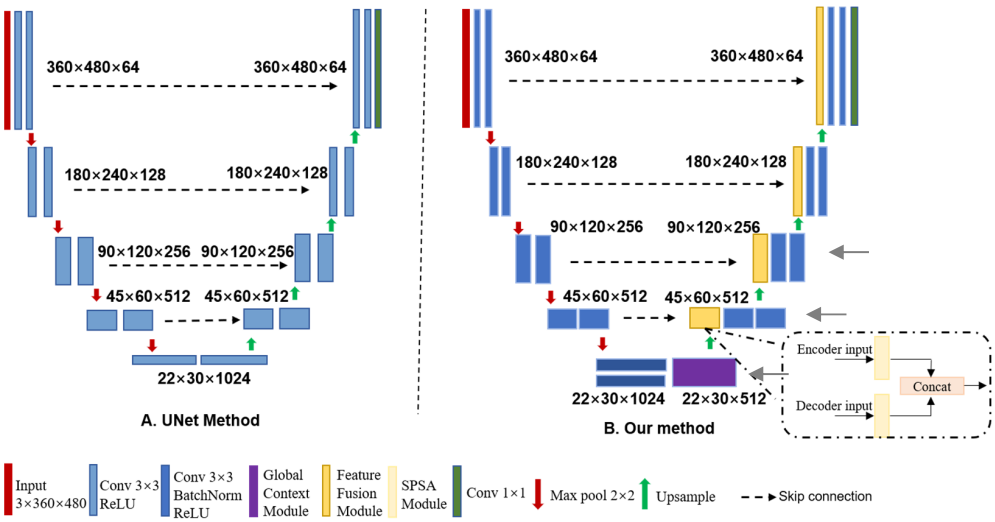


Figure 5. The overall architecture of our proposed FSFS-Net. (**A**) is original U-Net network. (**B**) is our method proposed. The encoder downsamples the input image with 360 × 480 until it reaches a size 22 × 30 in the bottleneck. Then, the decoder upsamples it to the original input. An additional three output heads are used for deep supervision loss.

3.1. Overall Network Architecture

This achievement proves that a well-designed U-Net architecture has the ability to perform very well on segmentation tasks. We proposed a solution for fish school feeding behavior quantification, intending to make it accurate, robust, and efficient. In this paper, we use the U-Net method [37] as the baseline for our network, mainly because it is a simple and effective network, and the result is proved in the comparison experiment section. The U-Net method consists of an encoder and decoder. Figure 5A depicts the network structure. The encoder with a depth of 5 extracts features continuously by using two 3×3 convolutions $((3 \times 3\,\mathrm{Conv}) \rightarrow (\mathrm{ReLU}))$ and expands the receptive field to capture contextual information by using four downsampling operations (2×2 max pooling). The decoder recovers spatial information through upsampling and uses skip connection to perform feature fusion to compensate for the details lost during the downsampling process.

The idea of our proposed method is to combine SPSA and the LMSM module to improve algorithm accuracy without data augmentation based on U-Net architecture. Compared with the U-Net baseline, we use $(3 \times 3\ \mathrm{Conv}) \rightarrow (\mathrm{BN}) \rightarrow (\mathrm{ReLU})$ to extract features while keeping the encoder/decoder depth and down-sampling/up-sampling times the same as the U-Net structure. To extract more semantic information without reducing the spatial resolution, we propose an LMSM module, which is added to the last encoder convolution block. The purpose of the proposed SPSA module is to lessen the reliance on external information, while also better capturing internal data/feature correlation. Our SPSA module is added to the decoder section, where encoder and upsampling features are transferred to the SPSA module, which then performs the feature fusion process.

The final architecture is shown in Figure 5B. The input of shape $360 \times 480 \times 64$ is first processed by an input block with two $3 \times 3 \times 3$ convolution layers, batch normalization, and ReLU activation. Then, the feature map is transformed by 4 blocks, each of which reduces the spatial dimension of the feature map by firstly downsampling with 2×2 max pooling. The reduced feature maps are then refined by two convolution layers with a kernel size of $3 \times 3 \times 3$. Downsampled feature maps have a size of $22 \times 30 \times 1024$ at the bottleneck. Then, their feature map is sent to the LMSM module, and the output size remains the same as the input of LMSM. Finally, spatial dimensions are increased by 4 upsampling blocks which consist of the feature fusion module and two convolution layers with a kernel size of $3 \times 3 \times 3$. Especially, the first feature fusion module concatenates the output of the encoder and LMSM module. The remaining three feature fusions module concatenate the output of the encoder and decoder. The final output feature map with input size has 3 output channels, where each channel corresponds to a different class.

3.2. Shuffle Polarized Self-Attention

An attention mechanism allows a neural network to precisely focus on all relevant components of the input. Although numerous visual tasks use advanced algorithms together with attention mechanisms to improve algorithmic performance, the computational overhead still inevitably increases. Inspired by shuffle attention [45] and polarized self-attention [46], we propose a shuffle polarized self-attention (SPSA) to address this problem. The general architecture of the SPSA module is depicted in Figure 6. SPSA consists of the following components: feature group, polarized self-attention, and aggregation. Firstly, the feature map of the input is separated into groups, and two branches of each group are fed into the PSA module. Following that, all sub-features are aggregated, and ultimately the "channel shuffle" operator is employed to interchange information amongst distinct sub-features.

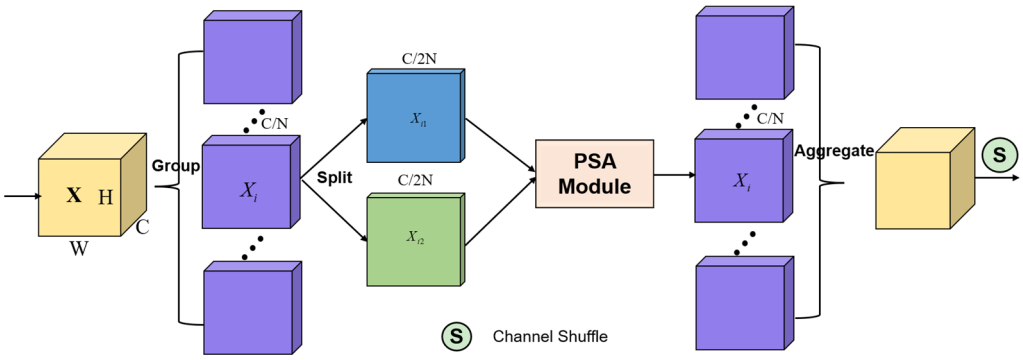


Figure 6. Shuffle Polarized Self-Attention architecture.

3.2.1. Feature Group

A feature map, X, with $C \times H \times W$ is divided into N groups (default parameters is 8 groups) along with the channel dimension. C, H, W indicate the channel number, height and width of feature map, respectively. For each group divided, it has the same the number of channels C/N, and each feature map i^{th} is $X_i \in R^{\frac{C}{N} \times H \times W}$. Then, X_i is split into 2 branches along the channel dimension, and each branch has $C/2N$ the number of channels. Finally, these two branches are fed into the SPA module, and its output is a feature map with $\frac{C}{N} \times H \times W$.

3.2.2. Polarized Self-Attention

The attention mechanism learns specific weights along the channel, as well as spatial dimensions, to estimate category score and identify the spatial location information of the same semantic algorithm, respectively. The self-attention mechanism enables us to further highlight the key features of the channel and its spatial dimensions. In this paper, we applied polarized self-attention [46] to extract the interdependence between features in an image. It is suggested that high-quality pixel-wise regression is achieved by integrating two key photographic factors: filtering and high dynamic range (HDR). Filtering is used to maintain high internal resolution in terms of both spatial dimensions and channel direction, while completely collapsing features along their opposite direction. By using Softmax normalization, HDR with compositional non-linearity can directly fit the output distribution of typical fine-grained regression. The structure of the PSA mechanism is shown in Figure 7.

The PSA module is made up of two parts: channel-only self-attention and spatial-only self-attention. The input $X_1 \in R^{C \times H \times W}$ and $X_2 \in R^{C \times H \times W}$ from divided features $X_i \in R^{2C \times H \times W}$ are sent to channel-only self-attention and spatial-only self-attention of the PSA module. The output of the PSA module is a feature map with $2C \times H \times W$.

Channel-only self-attention output $Z^{ch} \in R^{C \times H \times W}$:

$$Z^{ch} = B^{ch}(X_1) \odot^{ch} X_1 \tag{1}$$

$$B^{ch}(X_1) = F_{sigmoid}\left[W_{z|\theta_1}\left((\sigma_1(W_v(X_1)) \times F_{SoftMax}(\sigma_2(W_q(X_1)))\right)\right] \tag{2}$$

Spatial-only self-attention output $Z^{sp} \in R^{C \times H \times W}$:

$$Z^{sp} = B^{sp}(X_2) \odot^{sp} X_2 \tag{3}$$

$$B^{sp}(X_2) = F_{sigmoid}\left[\sigma_3\left(F_{SoftMax}(\sigma_1(F_{GP}(W_q(X_2)))) \times \sigma_2(W_v(X_2))\right)\right] \tag{4}$$

PSA module output:

$$PSA(X_1, X_2) = Z^{ch} © Z^{sp} \tag{5}$$

where $B^{ch}(X_1) \in R^{C\times1\times1}$ and $B^{sp}(X_2) \in R^{1\times H\times W}$ represent the channel-only branch and spatial-only branch. W_v, W_q, and W_z are 1×1 convolution layers. σ_1, σ_2, σ_3 are three tensors reshaping operators. $F_{SoftMax}(\cdot)$, $F_{sigmoid}(\cdot)$, $F_{GP}(\cdot)$ and "$\times$" represent SoftMax function, sigmoid function, a global pooling operator, and matrix dot-product operator, respectively. $\odot^{ch}$, and $\odot^{sp}$ represent a channel-wise and spatial-wise multiplication operator. © is concatenation operator.

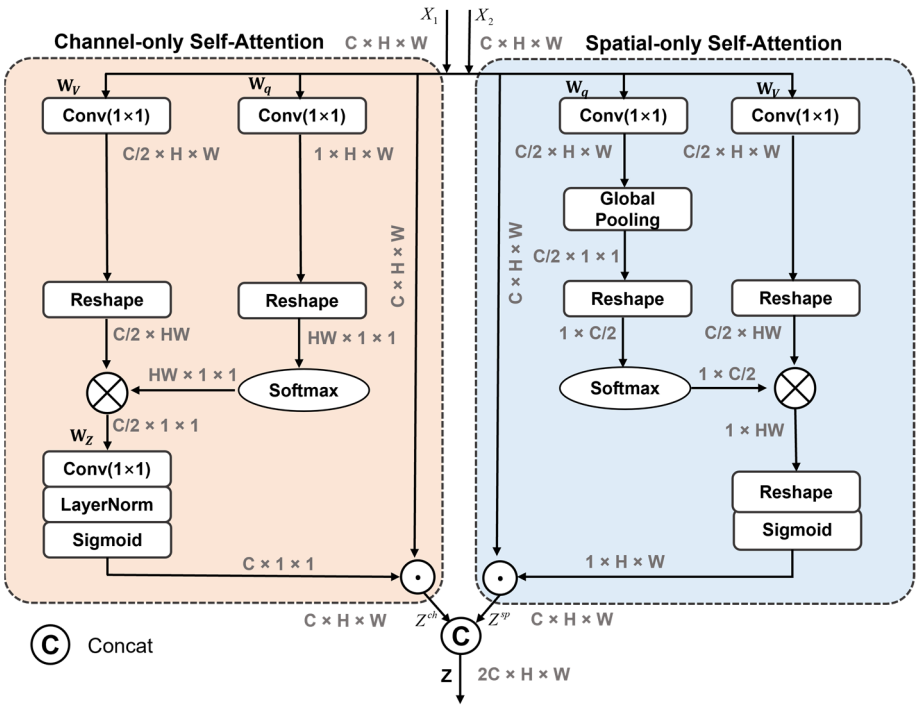


Figure 7. Polarized Self-Attention Module.

3.2.3. Aggregation

All the sub-features are aggregated in the last part of the network. A "channel shuffle" operator is employed to allow cross-group information flow along the channel dimension. Our SPSA module's final output is the same size as the input. As a result, this module is easy to integrate with other semantic segmentation algorithms.

3.3. Lightweight Multi-Scale Module

The goal of semantic segmentation is to attempt to strike a balance between classification and location. To obtain accurate classification, the model needs to expand the receptive to extract contextual information by a using downsampling operation. Unfortunately, the spatial localization information of the feature map is lost during the downsampling step, resulting in low positioning accuracy.

To overcome the abovementioned drawbacks, we design a lightweight multi-scale module (LMSM) which generates multi-scale features to extract global contextual semantic information while maintaining the feature resolution constant. The proposed LMSM consists of two parallel 5×5 and 3×3 deep separable convolutions, as shown in Figure 8B. Specifically, the feature map with w $\times$ h $\times$ 1024 is fed into two branches, each of which uses deep convolution and point convolution with batch normalization and ReLU activation. Among them, 5×5 convolution is used to extract high semantic information from large objects by expanding the receptive field rather than using a downsampling operation. Point convolution is used to reduce the amounts of parameters. Then, the output features of the two branches are fused together. Finally, the fused features are input to 3×3 convolution to produce a feature map with w $\times$ h $\times$ 512.

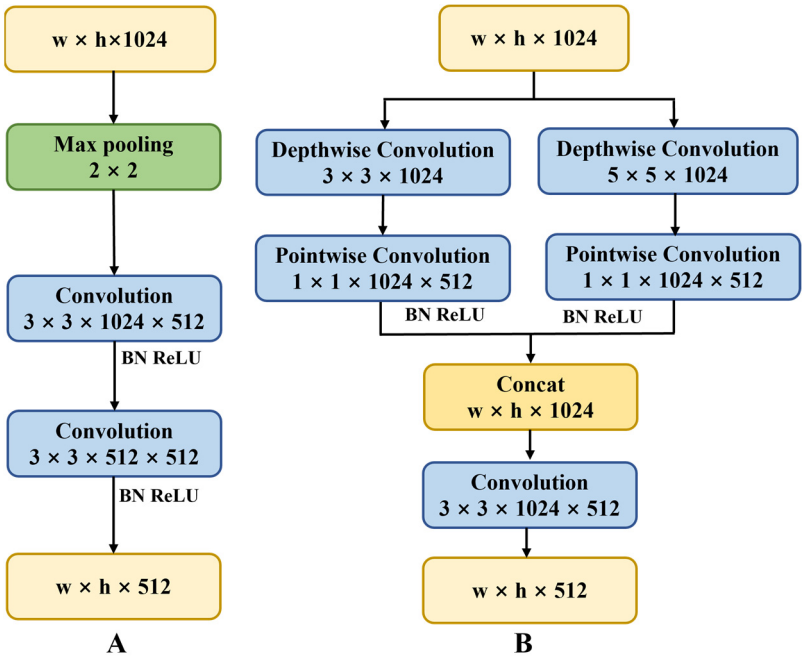


Figure 8. (**A**) Stack of original 3 × 3 convolutions. (**B**) Our proposed LMSM.

3.4. Deep Supervision

Deep supervision (namely auxiliary training) is beneficial to enhancing the semantic information representation for each stage. In contrast to the previous method following BiSeNet V2 [47] and DFN [41], we propose the use of a segmentation head for booster training, as shown in Figure 9. Additionally, this study aims to match the shape of the additional predictions with the ground truth of (360, 480). Three additional segmentation heads at the decoder stage are inserted for the training stage, and this does not increase computational complexity since it is discarded in the inference stage. We apply region mutual information loss (RMI) [48] to calculate the final loss and auxiliary loss. The total loss function is denoted as:

$$l_{total} = l(g, y_{final}) + \sum_{i=3}^{5} l(g, y^{i}_{\text{auxiliary}}) \tag{6}$$

where g represents ground truth. y_{final} is the final prediction output. $y^{i}_{\text{auxiliary}}$ represents the i deep supervision loss.

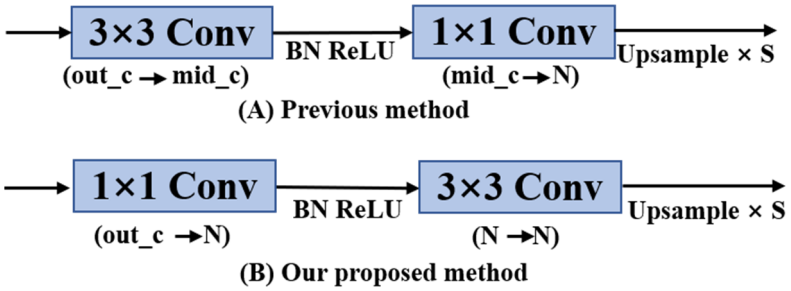


Figure 9. Segmentation head of deep supervision. Notation: Con is the convolution operation. BN denotes batch normalization. Upsample means bilinear interpolation. N denotes the number of segmentation objects categories. C denotes the number of the output channels.

4. Experiments

In this section, the data acquisition and data annotation rules are described in detail. To validate the effectiveness of our method proposed, we perform a series of ablation

experiments and compare with the use of other state-of-the-art methods on datasets of fish feeding behavior. Finally, the experimental results are analyzed and discussed.

4.1. Datasets

The videos/images data from the surveillance camera were obtained from LaiZhou Mingbo Aquaculture Co, Ltd., Yantai City, China. All videos/image were taken from real industrialized recirculating aquaculture, and there were approximately 60 fish in the pond, as shown in Figure 10. Each video obtained, with a frame rate of 24 fps and a resolution of 1440 × 2560 pixels, lasted 10 to 15 min long and included various feeding stages such as before, during, and after feeding. In addition, images of different feeding intensities (strong, medium, weak, and non-feeding) were acquired to increase the richness of data. The majority of the obtained images clearly distinguished between foreground and background. However, the similarity in the color and appearance between fish objects, as well as the blurry images caused by the fish's fast movement during the feeding process, brought serious challenges to efforts to accurately segmenting fish.

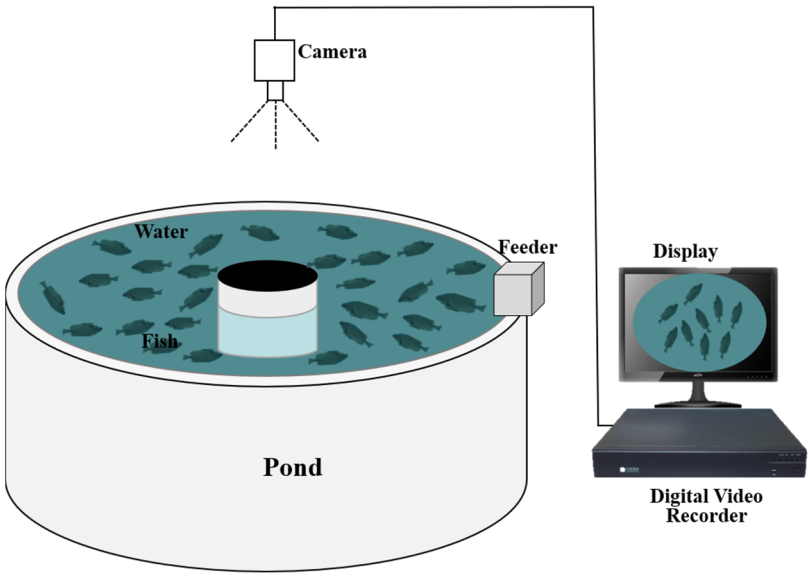


Figure 10. An image acquisition station. Images were taken from real industrialized recirculating aquaculture, and there were approximately 60 fish in the pond.

The data obtention required manual labeling because deep learning is a supervised approach. Firstly, a frame image was captured every second, for a total of 23 videos obtained. After removing the frame images that were quite similar, we obtained 1361 frame images. Then, fish objects for each frame were labeled manually into two target categories (fish1 and fish2) by our team members, and the marking rules were as follows (see Figure 11): (1) the individual fish with clearly separated and border adhesion is named fish1; (2) for overlapping fish, when the occluded area of fish is less than 1/3, it is marked as fish1; (3) fish are occluded by other fish, and occlusion area exceeds 1/3, in which case the two fish are marked as fish2 as a whole; and (4) the fish gathered during the feeding process are marked as fish2. Finally, 1361 instance-level images with two object categories and corresponding masks were obtained. Among them, 1038 images with 67,518 annotations instances of the fish region were used for training. A total of 323 images, along with 20,943 segmentation annotations of fish regions, were used for testing.

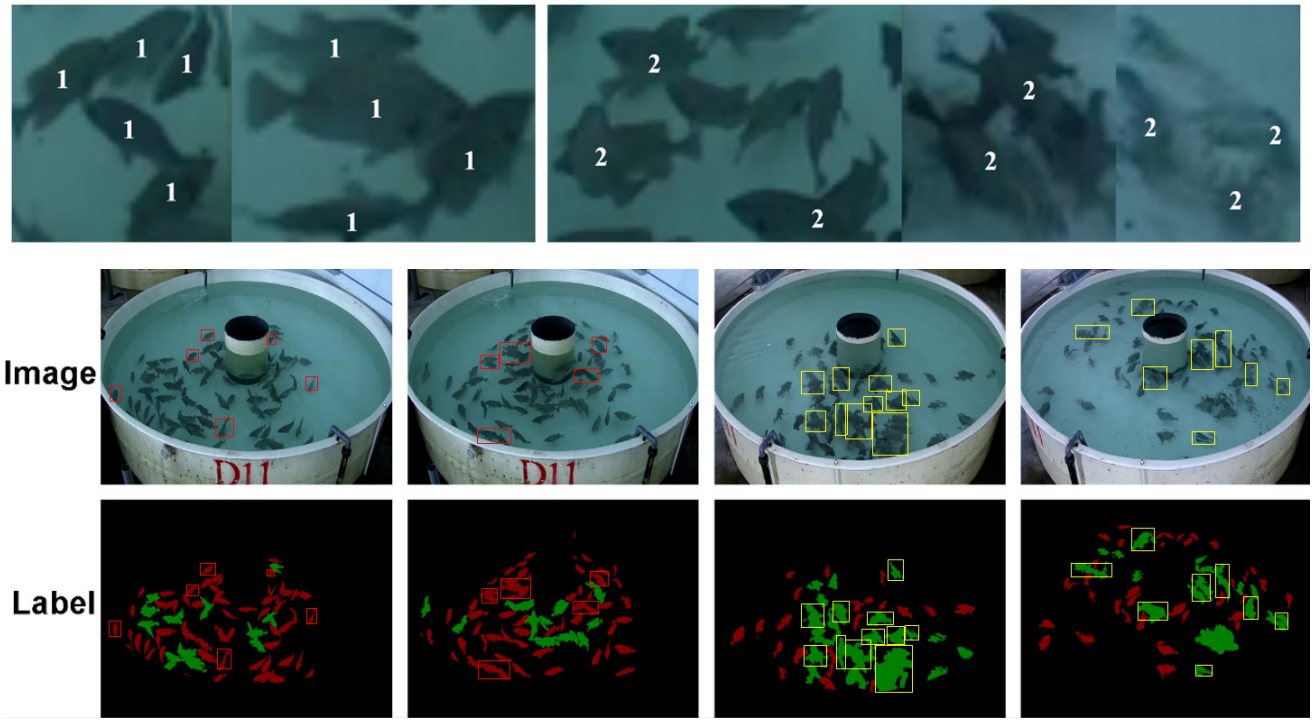


Figure 11. Sample Images Labeled. The three rows top-bottom presents the marking rules for two target categories, the original image, and the ground truth. The red box and the yellow box, respectively, represent the fish1 and fish2 category.

4.2. Implementation Details

Training schedule: To train our method proposed, the input image was resized to 360 × 480 pixels. The approaches proposed were trained from scratch, with no pre-training strategies used. The epochs and batch size were set to 50 and 4, respectively. We trained the network using an Adam optimizer with a 0.0005 initial learning rate and a 0.0004 weight decay. We also adopted warmup and the cosine annealing learning rate decay with 10 epochs. Moreover, DropBlock was used to alleviate overfitting in the training stage. All the experiments was executed with the PyTorch framework and ran on a single NVIDIA 2080Ti GPU. Finally, we employed the mean intersection over union (mIoU) as the metric with which to evaluate the performance of algorithm.

Data augmentation: Data augmentation is a technique that alleviates the overfitting problem by extending a dataset sample. The following data augmentation methods were used, including mean subtraction, randomly horizontal flipping, gaussian blur, contrast, brightness, and saturation.

4.3. Comparison with the State-of-the-Art

In this section, we compare our result with 15 state-of-the-art segmentation models on test dataset, including SegNet [36], U-Net [37], HRNet, DFN [41], DANet [42], OCNet [43], PSNet [38], DeepLab v3 plus [39], GCN [49], BiSeNet V2 [47], and DDRNet [50], et al. To achieve a fair comparison, the same iterative training strategies were employed to evaluate the effect of the various models to a certain extent. The experimental results show that our proposed approach outperformed other methods and achieved the highest score with 79.62% mIoU.

Table 2 presents the result. Namely, our FSFS-Net method with 79.62% mIoU surpassed baseline U-Net and other state-of-the-art methods, which demonstrates the effectiveness of the proposed model. The original U-Net method achieved 77.98% mIoU over the HRNet method, which was in line with our expectations, especially since the U-Net method has been widely used in medical image segmentation and has achieved remarkable results.

The HRNet, GCN, DeepLab v3 plus, and real-time DDRNet39 methods had similar scores because they use multi-scale fusion to extract contextual semantic information. The HRNet maintained high-resolution representation from all the high- to low-resolution parallel convolutions, which enhance the high-resolution and low-resolution features and achieve multi-scale feature fusion by repeatedly exchanging information between parallel multi-resolution subnets. DeepLab v3 plus utilized encoder–decoder methods and the ASPP module to recover spatial information and to extract multi-scale contextual information. GCN utilized a large convolution kernel to increase the receptive field for addressing classification and locations task simultaneously. DDRNet39 used a deep aggregation pyramid pooling module (DAPPM) to extract contextual information. Despite this, the method we proposed surpassed the abovementioned method since our method combines the advantages of both methods.

Table 2. MIoU score comparison with other methods on fish feeding behavior dataset. Our proposed method achieves state-of-the-art performance.

Method	Backbone	Image Size	MIoU (%)
LinkNet [51]	ResNet18	704 × 1280	59.06
ENet [52]	__	720 × 1280	65.03
BiSeNet v2 [47]	__	720 × 1280	72.49
DDRNet [50]	39	704 × 1280	75.68
OCNet [43]	ResNet101	360 × 480	42.39
DANet [42]	ResNet101	360 × 480	43.29
FCN-8s [53]	VGG16	352 × 480	69.44
SegNet [36]	VGG16	360 × 480	69.52
DFN [41]	ResNet101	352 × 480	69.77
DFN [41]	ResNet50	352 × 480	55.26
ExFuse [54]	ResNet50	352 × 480	70.46
ExFuse [54]	ResNet101	352 × 480	70.05
PSNet [38]	ResNet50	473 × 473	71.18
HRNet V2 [55]	W48	352 × 480	75.36
DeepLab v3 plus [39]	VGG16	360 × 480	75.83
GCN [49]	VGG16	352 × 480	75.81
GCN [49]	ResNet152	352 × 480	76.14
U-Net [37]	__	360 × 480	77.98
SPSA-Net (ours)	__	360 × 480	**79.62**

From Table 2, we can observe that our FSFS-Net method outperforms the method-attention mechanism method (DFN, DANet, and OCNet) and other methods, such as BiSeNet V2, ExFuse, PSPNet and FCN-8s. It is worth mentioning that although the PSNet method achieves a performance that is closer to ExFuse and DFN, its backbone network uses ResNet50 rather than ResNet101. To validate the effect of the backbone network on the algorithm, we used ResNet 50 and ResNet101 as the backbone network of ExFuse and DFN. The experimental results showed that DFN using deeper ResNet101 was favorable to fish segmentation, with a gap of at least 10% mIoU compared to that obtained with a shallow ResNet50 network; however, ExFuse was unaffected by the depth of the backbone network due to the multi-scale fusion module. Additionally, when VGG16 and ResNet152 were applied to DeepLab v3 plus, we discovered that their results were quite similar, proving that the multi-scale fusion could achieve very competitive results without relying on the depth of the feature extraction network.

The comprehensive experimental result shows that our method is superior to other methods and can produce more accurate results. The sophisticated HRNet approach has poor generalization performance as compared to our method proposed. The poor performance is primarily due to the fact that it is designed on multi-category feature and large sample datasets. However, our dataset of fish school feeding has a small sample, with

only three categories (contains background category). Therefore, the method we proposed is effective in segmenting fish school images.

4.4. Ablation Study

To further verify our contributions, we conducted an extensive ablation study on the fish feeding test dataset. We took the basic U-Net network as the baseline, and gradually designed the strategies in this paper, such as loss function, PSPA, LMSM, and deep supervision. Our experiments demonstrated that increasing LMSM and SPSA further improved the score over the baseline model. To validate the performance of our proposed module, our method was compared with previous ASPP, PPM, and improved U-Net benchmarks. In addition, the various loss functions were also used for comparison, such as focal loss, weighted cross-entropy, LDAM loss, OHEM, Lovas loss, and RMI loss.

Baseline. Based on comparative experiments (the detailed results are shown in Section 4.3), we observed that baseline U-Net achieves the best results compared with other methods. Therefore, we selected U-Net architecture for further exploration on the fish school feeding dataset.

The result in Table 3 presents the influence of data augmentation, the number of downsamplings, and the number of skip connections. We find that increasing the number of downsamplings to 5 improves mIoU from 77.98% to 78.97% without dataset augmentation, demonstrating that expanding the receptive is beneficial to our fish segmentation task. However, the number of skip connections is 5, and mIoU decreases from 78.97% to 78.8%. We also find that using data augmentation does not lead to significant improvements in the U-Net model (only improving by 0.17). U-Net+ (the number of downsamplings is 5) with dataset augmentation achieves 79.29 mIoU, but our proposed FSFS-Net method achieves a high score with 79.62% mIoU and 2.45 M parameters.

Table 3. Performance comparisons of our proposed and baseline U-Net method. U-Net+ shows that the U-Net model uses five down-sampling. DA: data augmentation; NSC: number of skip connections.

Method	DA	NSC	MIoU (%)	Params	GFLOPs
U-Net	W/O	4	77.98	3.45 M	166.2
	W	4	78.15	3.45 M	166.2
U-Net+	W/O	5	78.80	5.35 M	99.1
	W/O	4	78.97	2.99 M	89.3
	W	4	79.29	2.99 M	**89.3**
FSFS-Net	W	4	79.37	**2.45 M**	89.6
	W/O	4	**79.62**	**2.45 M**	89.6

Thus, our approach outperforms the UNet and U-Net+ methods in terms of both segmentation performance (the mIoU score) and computation costs (the number of parameters). According to the ablation study based on U-Net+, increasing the number of skip connection not only increases model calculations but also reduces algorithmic performance. Furthermore, data augmentation makes no discernible contribution to the improvement of algorithmic performance. Therefore, we use four skip connections without data augmentation in the following experiments.

Multi-scale module. Table 4 shows the performance comparison and reports the parameters contained by the LMSM, PPM, and ASPP methods. Comparing FSFS-Net with the ASPP and PPM method, we observe that the LMSM method we proposed brings a considerable improvements and obtains the best result with fewer parameters and GFLOPs. Since the proposed LMSM approach is a straightforward implementation for learning contextual semantic information by expanding receptive field and extracting multi-scale features, the FSFS-Net with PPM and ASPP achieves 79.14% and 79.29% mIoU, respectively. When depth separable convolution of our proposed method proposed is replaced with general convolution, the algorithm achieves 79.38 mIoU, while their overall learnable

parameters are higher than those of the original LMSM method. It is worth noting that we alter the ASPP module to adapt to our dataset since only two categories need to be segmented for our segmentation task, as shown in Figure 12. The PPM method is consistent with the literature.

Table 4. Ablation study of adding LMSM on proposed methods. LMSM *: depth separable convolution is replaced with general convolution; LMSM: proposed method.

Context Module	MIoU (%)	Params	GFLOPs
PPM	79.14	3.33 M	94.7
ASPP	79.29	4.51 M	102.9
LMSM *	79.38	4.54 M	103.4
LMSM	**79.62**	**2.45 M**	**89.6**

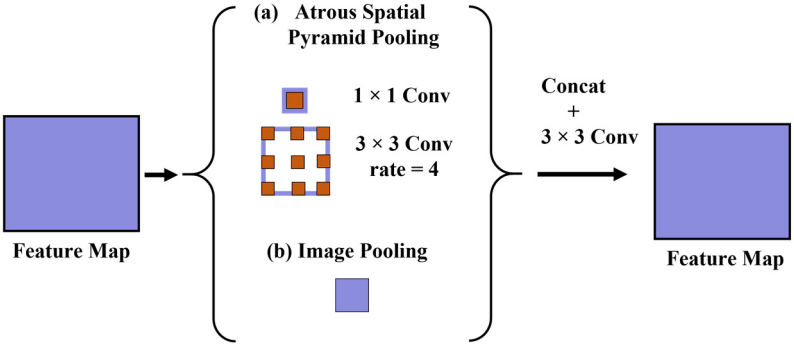


Figure 12. Improved ASPP module. The goal is to adapt to our fish segmentation task.

We further conduct an experiment to validate how the number of multi-scale feature extraction branches affects algorithmic performance. For instance, a parallel 7×7 depth separable convolution is added to LMSM method, which reduces mIoU on fish feeding behavior dataset. Similar experiments are performed on the ASPP module. When the original ASPP module is applied in our segmentation task, the algorithm is unable to segment small target objects. Therefore, it is critical to present alternative multi-scale modules for different segmentation tasks, measure which also proves the effectiveness of our proposed method. Moreover, we find that using 3×3 convolution rather than 1×1 convolution in the last layer of the LMSM method is more conducive to feature extraction and target segmentation.

SPSA module. Table 5 shows the performance comparison between our proposed and the original attention mechanism module. Comparing the original PSA module, we find that the SPSA we proposed obtains competitive performance while ensuring fewer parameters and GFLOPs. This is because the proposed SPSA approach divides groups in the channel dimension, which is very helpful for efforts to reduce the number of parameters.

Table 5. Ablation study of adding SPSA on proposed methods.

Attention Module	MIoU (%)	Params	GFLOPs
PSA	79.49	2.52 M	97.9
SPSA	**79.62**	**2.45 M**	**89.6**

Ablation study of gradually increasing various modules. The experimental results presented in Table 6 have shown that applying each strategy separately can improve the score over baseline U-Net. Using the SPSA module can significantly boost the score over the baseline U-Net (improving it from 77.98% to 79.14%). Notably, we also discovered that only adding the LMSM module was able to dramatically improve the results by 1.48% mIoU (improving from 77.98% to 79.46%) when compared to using the SPSA module alone.

Our proposed method, together with SPSA and LMSM methods, yielded 79.52% mIoU. Finally, by adding deep supervision, the score was further slightly improved to 79.62% mIoU. The experimental results demonstrated that the various modules we proposed had obvious performance gains and showed reliable improvements.

Table 6. Ablation study of gradually increasing LMSM, SPSA, and the deep supervision (DS) based on baseline U-Net. The experimental results show that the performance can be further improved by using the LMSM and the SPSA strategy.

U-Net	SPSA	LMSM	DS	MIoU (%)	Gain
✓				77.98	
✓	✓			79.14	↑ 1.16
✓		✓		79.46	↑ 1.48
✓	✓	✓		79.52	↑ 1.54
✓	✓	✓	✓	**79.62**	↑ 0.1

Loss function. Table 7 shows the result of different loss functions on our proposed method. The results show that changing the loss function improvs the segmentation score slightly. We can see from the table that RMI achieves the highest score since it utilizes correlation between the pixels. However, the fact that the pixels in the image are interdependent is ignored by cross entropy loss. Thus, RMI is used as the loss function for fish school feeding image segmentation based on this experiment.

Table 7. Different loss functions bring different improvements.

Method	MIoU (%)
Focal loss	76.75
Weighted cross-entropy	77.70
LDAM loss	77.72
OHEM	77.94
Lovas loss	79.01
RMI loss	**79.62**

4.5. Visualizations and Discussion

In this section, we present visualized segmentation results to evaluate the effectiveness of our method. Additionally, we also discuss the possible limitations of our proposed method in some special cases. The images used for analysis are selected from test samples with simple (feed-before) and hard samples (strong feeding intensity). Finally, a video is tested on the method we proposed to analyze the changes of feeding intensity during the fish school feeding process.

Experimental results show that our proposed method can distinguish key features and extract multi-scale features from an image, resulting in the highest score. As shown in Figure 13a–f, the method we proposed is effective in segmenting multiple instances in an image. Especially, our method can distinguish two adhesion fish from simple samples. Additionally, our algorithm can correctly segment some challenging instances, such as category confusion, motion blur, and multi-scale features between the same category and different categories. Furthermore, our proposed method can classify some mislabeled instances (label error) into true categories.

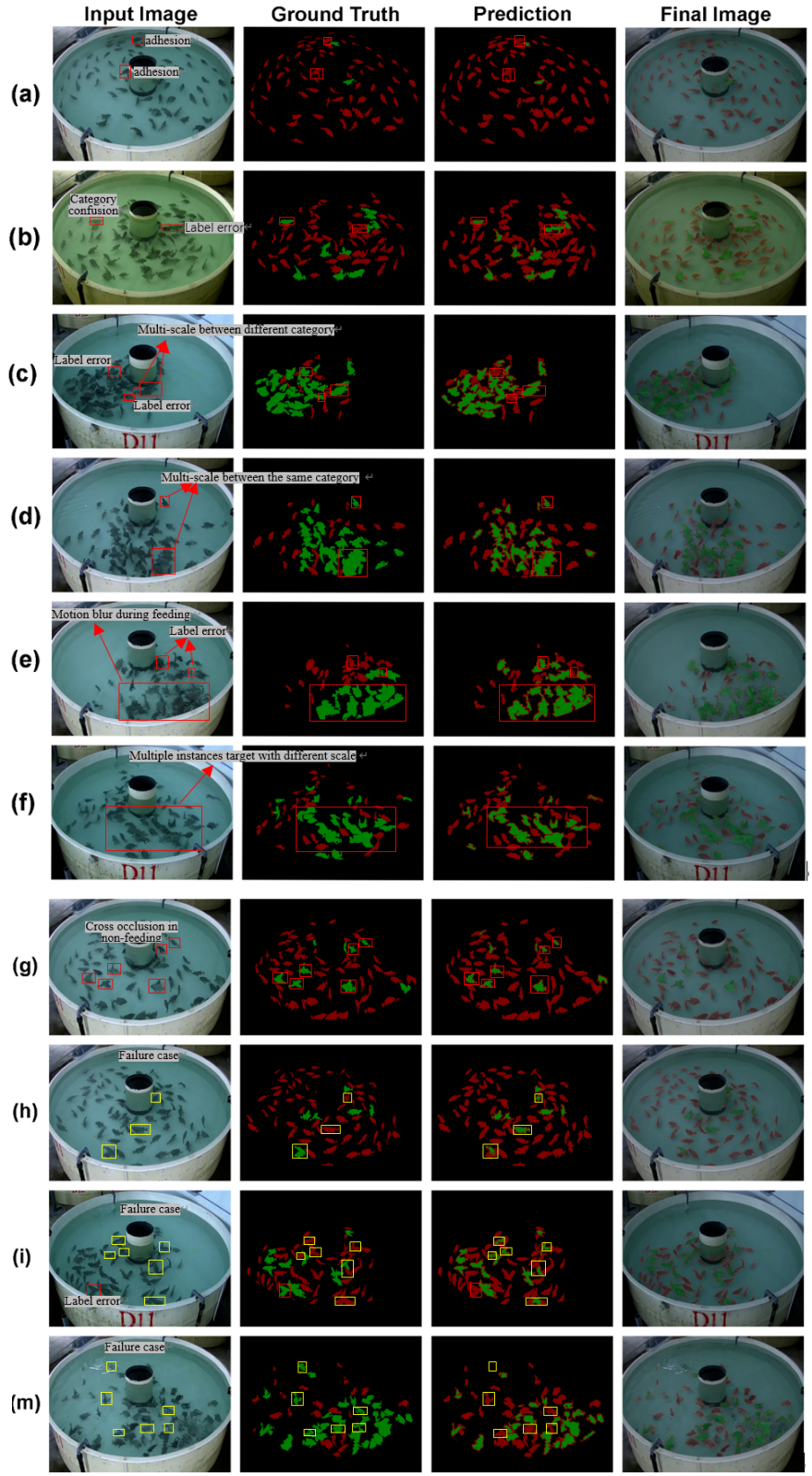
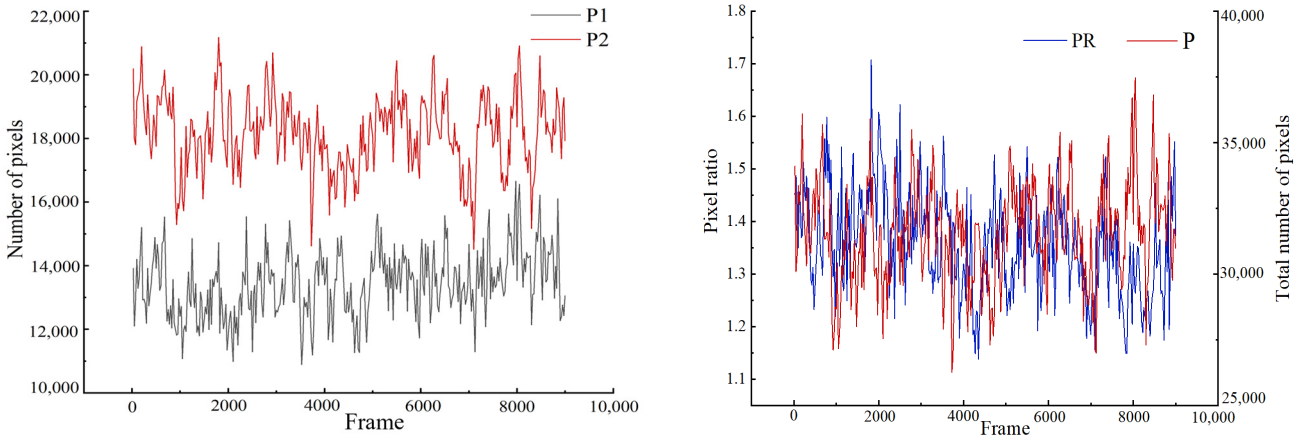


Figure 13. Visualized segmentation results on fish school feeding datasets. The result shows that our proposed method can segment multiple instance objects. Examples of fish images are randomly selected from the test dataset. The four column left-to-right represents the input image, the ground truth (Label), the output predicted of FSFS-Net, and the final image. The red rectangles in (**a**–**f**) rows show that the instances are successfully segmented. The yellow rectangles in (**g**–**i**,**m**) rows present some segmentation failures.

Although our proposed method can recover the spatial information to capture sharper object boundaries (fish with label 1), it still has three limitations. We find that our method designed is ineffective at distinguishing object borders, particularly for fish with the fish2 label, due to the fact that edge pixels tend to be incorrectly classified. In other cases, as seen in Figure 13g–m, there is considerable category confusion. When there is a border adhesion or slight overlap between two fish in a non-feeding image, the algorithm may be unable to clearly distinguish the border of the adhesion area, causing it to misclassify the instances (for example, two fish1 are misclassified as one fish2). In addition, there are clusters and severe occlusions between individual fish in the fish school feeding image (Figure 13m), but the algorithm misclassifies the fish2 category into the fish1 category since boundaries can be distinguished.

Finally, to verify the effectiveness of the proposed semantic segmentation method in fish feeding behavior analysis, the FSFS method proposed is tested on a video with 3 min clip. Two indicators (number of pixels and pixel ratio of two semantic categories) are used to distinguish the feeding intensity of fish school. As shown in Figure 14a, we can observe that the number of fish2 semantic pixels is greater than that of fish1, which is mainly due to the large amount of aggregation and occlusion between fish during feeding. From Figure 14b, we find that the curve tends to be highly oscillatory in the first 4000 frame sequence, which proves that the feeding intensity of fish is very intense. In 4354 frame, PR value (pixel ratio of two semantic categories) reaches the minimum, which indicates that the feeding intensity of fish has changed from strong to medium. After 8000 frames, the PR value becomes smaller and the p value (total number of pixels) becomes larger, indicating the number of clustered fish that are sheltered becomes less and the fish tends to disperse. At this time, the feeding intensity of the fish school will gradually weaken. Therefore, the FSFS method we proposed can well quantify fish school behavior and analyze the fish feeding intensity.



(**a**) (**b**)

Figure 14. Visualization of the number of pixels that was extracted from a 3 min video clip with fish feeding. (**a**) Trends of P1 and P2 values across 3-minute video clip; (**b**) Trends of P and PR values across 3-minute video clip. P1, P2, P and PR represent the number pixels of fish1, fish2, total and pixel ratio of fish2 to fish1, respectively.

It is worth mentioning that our algorithm is limited by the fish species, number of fish, and fish maturity in aquaculture ponds. When these factors change, our model needs to be retrained. Therefore, our proposed method only provides a new perspective for quantifying feeding intensity. In the future, we will design a network that can obtain accurate edge information from the low stage while simultaneously obtaining semantic information from the high stage, thereby eliminating some original edge's lack of semantic information. Moreover, we will develop a more lightweight network structure to achieve

real-time segmentation, using stronger backbones or more sophisticated architectures. Further, we need to conduct further research to validate our findings across diverse fish species, number of fish, fish maturity, and tank designs.

5. Conclusions

In this paper, we explored and demonstrated the importance of fish semantic segmentation in fish school feeding behavior quantification. A FSFS-Net algorithm was proposed to achieve two-class pixel-wise classification in fish feeding image. Especially, the SPSA module designed was able to capture long-range dependencies from the feature in an image. Moreover, we raised an effective LMSM that could extract multi-scale features to learn contextual information. The experimental results show that the proposed method achieved a performance of 79.62% mIoU score on annotated fish feeding dataset, surpassing other semantic segmentation algorithms such as U-Net, SegNet, FCN, DeepLab v3 plus, GCN, HRNet-w48, DDRNet, LinkNet, BiSeNet v2, DANet, and CCNet. The competitive performance on fish feeding datasets shows that our method proposed can contribute to quantifying fish school feeding intensity. In the future, we will design a network that can obtain accurate edge information from the low stage while simultaneously obtaining semantic information from the high stage, thereby eliminating some of the original edge's lack of semantic information. Moreover, we need to conduct further research to validate our algorithm against diverse fish species, number of fish, fish maturity, and tank design.

Author Contributions: Conceptualization, L.Y.; Methodology, L.Y.; Validation, L.Y.; Writing—Original draft preparation, L.Y.; Supervision, Y.C. and T.S.; Writing—Reviewing and Editing, D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Natural Science Foundation of China 'Analysis and feature recognition on feeding behavior of fish school in facility farming based on machine vision' (No. 62076244), the Beijing Digital Agriculture Innovation Consortium Project (No. BAIC10-2022), Yunnan Fundamental Research Projects (No. 202301AV070003), the Yunnan Reserve Talents of Young and Middle-aged Academic and Technical Leaders (No. 2019HB005), the Yunnan Young Top Talents of Ten Thousands Plan (No. 201873), and the Major Science and Technology Projects in Yunnan Province (No. 202002AB080001-8).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Xu, X.; Li, W.; Duan, Q. Transfer Learning and SE-ResNet152 Networks-Based for Small-Scale Unbalanced Fish Species Identification. *Comput. Electron. Agric.* **2021**, *180*, 105878. [CrossRef]
2. Wang, W.; He, B.; Zhang, L. High-Accuracy Real-Time Fish Detection Based on Self-Build Dataset and RIRD-YOLOv3. *Complexity* **2021**, *2021*, 4761670. [CrossRef]
3. Hu, X.; Liu, Y.; Zhao, Z.; Liu, J.; Yang, X.; Sun, C.; Chen, S.; Li, B.; Zhou, C. Real-Time Detection of Uneaten Feed Pellets in Underwater Images for Aquaculture Using an Improved YOLO-V4 Network. *Comput. Electron. Agric.* **2021**, *185*, 106135. [CrossRef]
4. Wageeh, Y.; Mohamed, H.E.D.; Fadl, A.; Anas, O.; ElMasry, N.; Nabil, A.; Atia, A. YOLO Fish Detection with Euclidean Tracking in Fish Farms. *J. Ambient Intell. Humaniz. Comput.* **2021**, *12*, 5–12. [CrossRef]
5. Huang, R.; Lai, Y.; Tsao, C.; Kuo, Y.; Wang, J.; Chang, C. Applying Convolutional Networks to Underwater Tracking without Training. In Proceedings of the 2018 IEEE International Conference on Applied System Invention (ICASI), Chiba, Japan, 13–17 April 2018; pp. 342–345. [CrossRef]
6. Cheng, X.E.; Du, S.S.; Li, H.Y.; Hu, J.F.; Chen, M.L. Obtaining Three-Dimensional Trajectory of Multiple Fish in Water Tank via Video Tracking. *Multimed. Tools Appl.* **2018**, *77*, 24499–24519. [CrossRef]

7. Lin, K.; Zhou, C.; Xu, D.; Guo, Q.; Yang, X.; Sun, C. Three-Dimensional Location of Target Fish by Monocular Infrared Imaging Sensor Based on a L–z Correlation Model. *Infrared Phys. Technol.* **2018**, *88*, 106–113. [CrossRef]
8. Yang, L.; Liu, Y.; Yu, H.; Fang, X.; Song, L.; Li, D.; Chen, Y. Computer Vision Models in Intelligent Aquaculture with Emphasis on Fish Detection and Behavior Analysis: A Review. *Arch. Comput. Methods Eng.* **2020**, *28*, 2785–2816. [CrossRef]
9. Yang, L.; Yu, H.H.; Cheng, Y.L.; Mei, S.Y.; Duan, Y.Q.; Li, D.L.; Che, Y.Y. A Dual Attention Network Based on EfficientNet-B2 for Short-Term Fish School Feeding Behavior Analysis in Aquaculture. *Comput. Electron. Agric.* **2021**, *187*, 106316. [CrossRef]
10. Zhou, C.; Xu, D.; Chen, L.; Zhang, S.; Sun, C.; Yang, X.; Wang, Y. Evaluation of Fish Feeding Intensity in Aquaculture Using a Convolutional Neural Network and Machine Vision. *Aquaculture* **2019**, *507*, 457–465. [CrossRef]
11. Zhou, C.; Lin, K.; Xu, D.; Chen, L.; Guo, Q.; Sun, C.; Yang, X. Near Infrared Computer Vision and Neuro-Fuzzy Model-Based Feeding Decision System for Fish in Aquaculture. *Comput. Electron. Agric.* **2018**, *146*, 114–124. [CrossRef]
12. Måløy, H.; Aamodt, A.; Misimi, E. A Spatio-Temporal Recurrent Network for Salmon Feeding Action Recognition from Underwater Videos in Aquaculture. *Comput. Electron. Agric.* **2019**, *167*, 105087. [CrossRef]
13. Wei, D.; Bao, E.; Wen, Y.; Zhu, S.; Ye, Z.; Zhao, J. Behavioral Spatial-Temporal Characteristics-Based Appetite Assessment for Fish School in Recirculating Aquaculture Systems. *Aquaculture* **2021**, *545*, 737215. [CrossRef]
14. Ubina, N.; Cheng, S.C.; Chang, C.C.; Chen, H.Y. Evaluating Fish Feeding Intensity in Aquaculture with Convolutional Neural Networks. *Aquac. Eng.* **2021**, *94*, 102178. [CrossRef]
15. Liu, H.; Liu, T.; Gu, Y.; Li, P.; Zhai, F.; Huang, H.; He, S. A High-Density Fish School Segmentation Framework for Biomass Statistics in a Deep-Sea Cage. *Ecol. Inform.* **2021**, *64*, 101367. [CrossRef]
16. Wang, G.; Hwang, J.N.; Wallace, F.; Rose, C. Multi-Scale Fish Segmentation Refinement and Missing Shape Recovery. *IEEE Access* **2019**, *7*, 52836–52845. [CrossRef]
17. Abdeldaim, A.M.; Houssein, E.H.; Hassanien, A.E. Color Image Segmentation of Fishes. In Proceedings of the 3rd International Conference on Advanced Machine Learning Technologies and Applications, Cairo, Egypt, 22–24 February 2018; Volume 4, pp. 634–643.
18. Zhang, L.; Wang, J.; Duan, Q. Estimation for Fish Mass Using Image Analysis and Neural Network. *Comput. Electron. Agric.* **2020**, *173*, 105439. [CrossRef]
19. Sultana, F.; Sufian, A.; Dutta, P. Evolution of Image Segmentation Using Deep Convolutional Neural Network: A Survey. *Knowl. Based Syst.* **2020**, *201–202*, 106062. [CrossRef]
20. Feng, D.; Haase-Schutz, C.; Rosenbaum, L.; Hertlein, H.; Glaser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 1341–1360. [CrossRef]
21. Mei, J.; Cheng, M.M.; Xu, G.; Wan, L.R.; Zhang, H. SANet: A Slice-Aware Network for Pulmonary Nodule Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4374–4387. [CrossRef]
22. Lou, A.; Guan, S.; Loew, M. CaraNet: Context Axial Reverse Attention Network for Segmentation of Small Medical Objects. *J. Med. Imaging* **2023**, *10*, 014005. [CrossRef]
23. Sun, L.; Luo, B.; Liu, T.; Liu, Y.; Wei, Y. Algorithm of Adaptive Fast Clustering for Fish Swarm Color Image Segmentation. *IEEE Access* **2019**, *7*, 178753–178762. [CrossRef]
24. Zhou, C.; Lin, K.; Xu, D.; Liu, J.; Zhang, S.; Sun, C.; Yang, X. Method for Segmentation of Overlapping Fish Images in Aquaculture. *Int. J. Agric. Biol. Eng.* **2019**, *12*, 135–142. [CrossRef]
25. Fernandes, A.F.A.; Turra, E.M.; de Alvarenga, É.R.; Passafaro, T.L.; Lopes, F.B.; Alves, G.F.O.; Singh, V.; Rosa, G.J.M. Deep Learning Image Segmentation for Extraction of Fish Body Measurements and Prediction of Body Weight and Carcass Traits in Nile Tilapia. *Comput. Electron. Agric.* **2020**, *170*, 105274. [CrossRef]
26. Labao, A.B.; Naval, P.C. Weakly-Labelled Semantic Segmentation of Fish Objects in Underwater Videos Using a Deep Residual Network. In Proceedings of the Intelligent Information and Database Systems: 9th Asian Conference, ACIIDS 2017, Kanazawa, Japan, 3–5 April 2017; pp. 255–265. [CrossRef]
27. Christensen, J.H.; Mogensen, L.V.; Ravn, O. Deep Learning Based Segmentation of Fish in Noisy Forward Looking MBES Images. In Proceedings of the 21st IFAC World Congress on Automatic Control—Meeting Societal Challenges, Berlin, Germany, 12–17 July 2020; Elsevier Ltd.: Amsterdam, The Netherlands, 2020; Volume 53, pp. 14546–14551.
28. Zhang, W.; Wu, C.; Bao, Z. DPANet: Dual Pooling-aggregated Attention Network for Fish Segmentation. *IET Comput. Vis.* **2021**, *1*, 67–82. [CrossRef]
29. Abe, S.; Takagi, T.; Torisawa, S.; Abe, K.; Habe, H.; Iguchi, N.; Takehara, K.; Masuma, S.; Yagi, H.; Yamaguchi, T.; et al. Development of Fish Spatio-Temporal Identifying Technology Using SegNet in Aquaculture Net Cages. *Aquac. Eng.* **2021**, *93*, 102146. [CrossRef]
30. 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]
31. 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.* **2020**, *77*, 1354–1366. [CrossRef]
32. Labao, A.B.; Naval, P.C. Simultaneous Localization and Segmentation of Fish Objects Using Multi-Task CNN and Dense CRF. In Proceedings of the 11th Asian Conference on Intelligent Information and Database Systems, Yogyakarta, Indonesia, 8–11 April 2019; Volume 11431, pp. 600–612.

33. Arvind, C.S.; Prajwal, R.; Bhat, P.N.; Sreedevi, A.; Prabhudeva, K.N. Fish Detection and Tracking in Pisciculture Environment Using Deep Instance Segmentation. In Proceedings of the IEEE Region 10 Conference on Technology, Knowledge, and Society, Kochi, India, 17–20 October 2019; Volume 2019, pp. 778–783.
34. Huang, K.; Li, Y.; Suo, F.; Xiang, J. Stereo Vison and Mask-RCNN Segmentation Based 3D Points Cloud Matching for Fish Dimension Measurement. In Proceedings of the Chinese Control Conference, CCC, Shenyang, China, 27–29 July 2020; Volume 2020, pp. 6345–6350.
35. Yu, C.; Fan, X.; Hu, Z.; Xia, X.; Zhao, Y.; Li, R.; Bai, Y. Segmentation and Measurement Scheme for Fish Morphological Features Based on Mask R-CNN. *Inf. Process. Agric.* **2020**, *7*, 523–534. [CrossRef]
36. 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]
37. Weng, W.; Zhu, X. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Strasbourg, France, 27 September–1 October 2021; Volume 9, pp. 16591–16603.
38. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid Scene Parsing Network. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 6230–6239.
39. 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 (ECCV), Munich, Germany, 8–14 September 2018; Volume 11211, pp. 833–851.
40. Yuan, J.; Deng, Z.; Wang, S.; Luo, Z. Multi Receptive Field Network for Semantic Segmentation. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass, CO, USA, 1–5 March 2020; pp. 1883–1892.
41. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Learning a Discriminative Feature Network for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; Volume 1, pp. 1857–1866.
42. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual Attention Network for Scene Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; Volume 2019, pp. 3141–3149.
43. Yuan, Y.; Huang, L.; Guo, J.; Zhang, C.; Chen, X.; Wang, J. OCNet: Object Context Network for Scene Parsing. *arXiv* **2018**, arXiv:1809.00916.
44. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
45. Zhang, Q.L.; Yang, Y. Bin SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; Volume 2021, pp. 2235–2239.
46. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized Self-Attention: Towards High-Quality Pixel-Wise Regression. *arXiv* **2021**, arXiv:2107.00782.
47. Yu, C.; Gao, C.; Wang, J.; Yu, G.; Shen, C.; Sang, N. BiSeNet V2: Bilateral Network with Guided Aggregation for Real-Time Semantic Segmentation. *Int. J. Comput. Vis.* **2021**, *129*, 3051–3068. [CrossRef]
48. Zhao, S.; Wang, Y.; Yang, Z.; Cai, D. Region Mutual Information Loss for Semantic Segmentation. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 1–11.
49. Peng, C.; Zhang, X.; Yu, G.; Luo, G.; Sun, J. Large Kernel Matters—Improve Semantic Segmentation by Global Convolutional Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 2017, pp. 1743–1751.
50. Hong, Y.; Pan, H.; Sun, W.; Jia, Y. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Road Scenes. *arXiv* **2021**, arXiv:2101.06085.
51. Chaurasia, A.; Culurciello, E. LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In Proceedings of the IEEE Visual Communications and Image Processing, St. Petersburg, FL, USA, 10–13 December 2017; Volume 2018, pp. 1–4.
52. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
53. Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 640–651. [CrossRef]
54. Zhang, Z.; Zhang, X.; Peng, C.; Xue, X.; Sun, J. ExFuse: Enhancing Feature Fusion for Semantic Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 11214, pp. 273–288.
55. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-Resolution Representations for Labeling Pixels and Regions. *arXiv* **2019**, arXiv:1904.04514.

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.