electronics

MDPI

Article

FIRN: A Novel Fish Individual Recognition Method with Accurate Detection and Attention Mechanism

Chunqi Gao [1,2], Junfeng Wu [1,2,*], Hong Yu [1,2], Jianhao Yin [1,2] and Shihao Guo [1,2]

[1] College of Information Engineering, Dalian Ocean University, Dalian 116023, China
[2] Key Laboratory of Environment Controlled Aquaculture, Dalian Ocean University, Ministry of Education, Dalian 116023, China
* Correspondence: wujunfeng@dlou.edu.cn; Tel.: +86-131-9011-1244

Abstract: Fish individual recognition technology is one of the key technologies to realize automated farming. Aiming at the deficiencies in the existing animal individual recognition technology, this paper proposes a method for individual recognition of underwater fish based on deep learning technology, which is divided into two parts: fish individual object detection and fish individual recognition. In the object detection part, the research has improved a new object detection for underwater fish based on the YOLOv4 algorithm, which changed the feature extraction network in YOLOv4 from CSP Darknet53 to Mobilenetv3 and changed the 3 × 3 convolution in the enhanced feature extraction network PANet to depthwise separable convolution. Compared with the original YOLOv4, the mean average precision is improved by 1.97%. For individual recognition, an algorithm called FIRN (Fish Individual Recognition Network) for individual recognition of underwater fish is proposed. The feature extraction network of the algorithm uses the improved ResNext50, and the loss function uses Arcface Loss. The CBAM attention module is introduced in the residual block of ResNext50, the max-pooling layer in the trunk is removed, and dilated convolution is introduced in the residual block, which increases the receptive field and improves the ability of feature extraction. Experiments show that the FIEN algorithm can enhance the compactness within a class while ensuring the separability between classes, and has a better recognition effect than other algorithms.

Keywords: object detection; fish individual recognition; deep learning

Citation: Gao, C.; Wu, J.; Yu, H.; Yin, J.; Guo, S. FIRN: A Novel Fish Individual Recognition Method with Accurate Detection and Attention Mechanism. *Electronics* **2022**, *11*, 3459. https://doi.org/10.3390/electronics11213459

Academic Editor: Byung Cheol Song

Received: 26 September 2022
Accepted: 24 October 2022
Published: 25 October 2022

1. Introduction

With the continuous improvement of people's living standards, aquatic products have become an important source of protein intake, and the production of aquaculture products accounts for more than 60% of the world's total output [1]. Many countries in the world are vigorously developing aquaculture to ensure people's food supply, and fish farming occupies a large proportion of the entire aquaculture industry. With the rapid development of modern technology, the aquaculture industry is changing from traditional artificial farming to intelligent farming [2]. Computer vision technology plays a key role in intelligent aquaculture [3]. Taking fish farming as an example, computer vision technology in fish farming is mainly used to detect, classify, identify, measure, and count by learning the phenotypic characteristics of underwater fish. Among them, achieving accurate recognition of fish individual is of great significance for the development of fish farming.

The traditional fish recognition task mainly realizes the recognition and classification of fish through the machine learning method based on manually selected features [4,5], which can identify what species the fish belongs to. However, the manual method of the selected features is inefficient and the features learned through human experience are not complete enough; thus, the accuracy is low. In addition, with the increase in people's demand for automated farming, it is far from enough to only identify the species of fish. On

the premise of identifying the species of fish, each fish of the species can be given unique identity information and can be identified, which is of more guiding significance to the development of fish farming.

At the present stage, most of the methods for fish individual recognition adopt deep learning models based on the face recognition framework, which mainly include three major processes: fish object detection, fish feature extraction, and fish feature comparison. The process of fish individual recognition is shown in Figure 1.

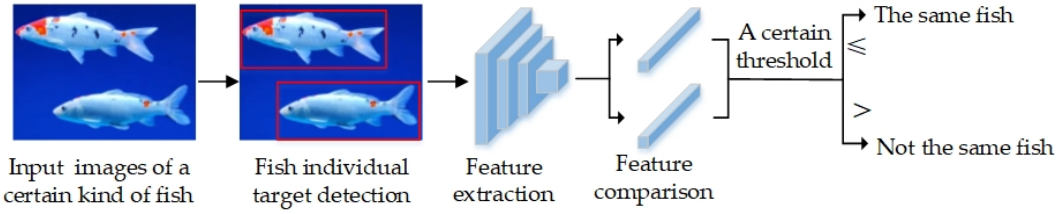


Figure 1. Flow chart of fish individual recognition.

In the fish individual recognition process, the object detection of the fish individual should be carried out first. The object detection algorithm includes three processes: image preprocessing, feature extraction, and classification and recognition. In the traditional object detection algorithm, feature extraction mainly adopts the method of manual selection, and selects important features based on human subjective experience. This method of feature selection is subjective, inefficient, and easily ignores detailed features. The classifiers used in traditional methods mainly include Naive Bayes [6], Decision Tree [7], KNN [8], and Support Vector Machine (SVM) [9]. These classifiers are only suitable for small fish targets with obvious characteristics, and their accuracy is low.

With the continuous development of deep learning, many scholars have applied it to object detection algorithms. Deep learning algorithms can automatically extract features and learn more details with high efficiency. Therefore, the object detection algorithm based on deep learning can be applied to fish targets with large-scale and insignificant features. Object detection algorithms based on deep learning are mainly the R-CNN series algorithm [10], SSD algorithm [11], YOLO series algorithm [12], etc. Among them, the R-CNN series algorithm is a two-stage deep learning algorithm based on candidate boxes, which has a slow detection speed and cannot realize real-time detection. The SSD algorithm and YOLO series algorithm are single-stage deep learning algorithms based on regression. Although the SSD algorithm is fast, it needs to manually set many parameters, and the debugging process is complex. The YOLO series algorithm has the advantages of fast speed, high accuracy, simple debugging, and real-time detection, so it is very suitable for fish individual object detection.

However, in the case of fish individual recognition, the current research is still relatively few and faces great challenges. First, there are few datasets of individual fish, and there is no public large-scale data on underwater individual fish. However, it is relatively difficult to manually collect underwater fish data, the underwater environment is complex, and the quality of the collected images is easily affected by the environment. Secondly, due to the great flexibility of the fish body, it has a wide range of posture changes, which will interfere with the feature extraction of the fish and reduce the recognition accuracy. Based on the existing fish recognition technology, this paper proposes an underwater fish individual recognition method. The main contributions of the proposed work are as follows:

1. A novel fish individual object detection algorithm based on improved YOLOv4 is proposed, which has higher accuracy than the traditional YOLOv4 in fish object detection.
2. A novel fish individual recognition network based on improved ResNext50, named FIRN, is proposed. Dilated convolution is introduced into the residual block of the

feature extraction network, and standard convolution is still used in the trunk. More detailed features can be learned while the receptive field is increased. The distance within the class is reduced while ensuring the separability between classes.

3. A dataset for fish individual recognition is proposed, which includes four different fish species, namely puffer fish, koi, grass goldfish, and clownfish. Each species has 30~50 different individuals, and the total number of images is 8000, which can meet the training and testing of the fish individual recognition method.

2. Related Work

The detection and recognition of fish has always been a research hotspot in the field of computer vision. In the early stage, many scholars studied fish recognition by manually selecting features based on the image content of fish such as color, shape, and texture, and made some progress. Larsen et al. [13] used Latent Dirichlet Allocation (LDA) to classify the shape and texture features of different species of fish. Mehdi et al. [14] used the Haar classifier to classify shape features modeled by principal component analysis. Hsiao et al. [15] designed and implemented a detection system for underwater fish video using methods such as maximum likelihood estimation. Wu Yiquan et al. [4] proposed a recognition method based on least squares support vector machine (LSSVM), which has high accuracy.

Although the method based on image content has also made some achievements, its disadvantages appear with the continuous increase of data volume. The recognition method based on image content is computationally complex and manual feature selection is inefficient, which can no longer meet the needs of big data.

With the continuous development of convolutional neural networks, more and more scholars are applying them to the object detection and recognition tasks of fish. Villon et al. [16] used the GoogLeNet to extract features and adopted the Softmax classification method to detect reef fish. Rauf et al. [17] deepened the number of convolutional layers based on the CNN framework as a way to improve the accuracy of fish recognition and classification tasks. Labao et al. [18] used the cascade structure of R-CNN and LSTM models to identify fish, and their precision and recall rates were relatively high. Aiming at the problem of the low quality of underwater videos, Zhang Minghua et al. [19] proposed a object detection method for underwater fish based on background removal using the partial least squares (PLS) classifier, which had good results. Li Chongchong et al. [20] used the YOLOv3 network to detect and identify the targets of underwater fish, which had a good detection effect. Cai et al. [21] improved the YOLOv3 model by replacing the feature extraction network in YOLOv3 with MobileNet to improve the feature extraction rate. This model can achieve high-precision fish detection. Xue Yongjie et al. [22] added an item-based flexible attention layer on the basis of AlexNet and used transfer learning for classification training, which greatly improved the classification effect of fish.

Based on the research above, this paper proposes a method for individual recognition of underwater fish using a deep learning model.

3. The Proposed Method

The model structure of the method for individual recognition of underwater fish proposed in this paper is shown in Figure 2. The method is divided into two modules: fish individual object detection and fish individual recognition. The object detection module adopts the improved YOLOv4 algorithm, which has better detection performance after improvement. The recognition module adopts the FIRN algorithm proposed in this paper, including feature extraction network and loss calculation. The feature extraction network is the improved ResNext50, and the loss function adopts Arcface Loss.

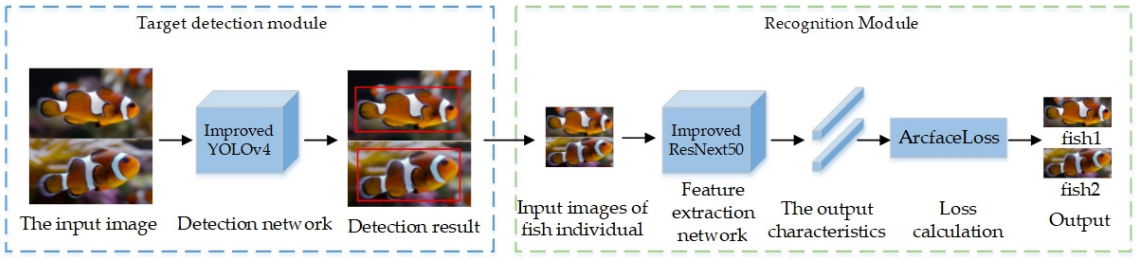


Figure 2. The diagram of the model's structure.

3.1. Object detection Module

Based on the above introduction of the advantages of YOLO series algorithms, this paper adopts the YOLOv4 [23] algorithm for fish individual object detection. We made improvements on the basis of YOLOv4, including the replacement of the feature extraction network and the enhancement of the feature extraction network, so as to detect the target of individual fish. Figure 3 is the network structure diagram of YOLOv4 after improvement.

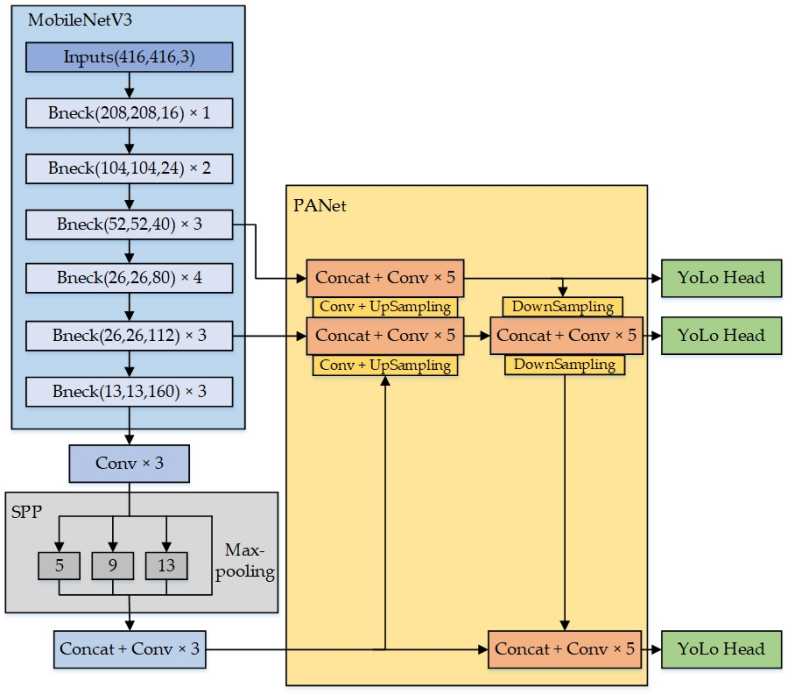


Figure 3. Improved YOLOv4.

3.1.1. Introduction of YOLOv4

YOLOv4 consists of three major components, namely the backbone feature extraction network Backbone, the enhanced feature extraction networks SPP and PANet, and the prediction network YoloHead. YOLOv4 is further optimized on the basis of the original YOLO network in terms of feature extraction network, activation function, and loss function, etc. The feature extraction network is changed from Darknet53 used in YOLOv3 [24] to CSP Darknet53, which reduces the model size while ensuring accuracy, and the activation function is changed from Leaky_relu to Mish. Compared with the previous version of the YOLO algorithm, YOLOv4 has higher accuracy and faster speed.

However, after YOLOv4 was proposed, YOLOv5 [25] was proposed. We delved into YOLOv5 and YOLOv4 and found that YOLOv5 and YOLOv4 have extremely similar network structures. Compared with YOLOv4, YOLOv5 has no substantial innovation but integrates a large number of state-of-the-art methods in the field of computer vision. In

addition, WongKinYiu, the second author of YOLO V4, used V100 GPU to provide comparable benchmarks [26]. From the data, it can be found that although YOLOv5 is superior in flexibility and speed, it is slightly inferior to YOLOv4 in performance. Moreover, YOLOv4 has a high degree of customization, and it is still an excellent object detection framework at present. Therefore, this paper adopts YOLOv4 instead of YOLOv5.

3.1.2. Improvements to YOLOv4

In order to make the object detection network more lightweight and applicable to mobile devices, the feature extraction network CSP Darknet53 in the YOLOv4 network is replaced by MobilenetV3 [27]. MobilenetV3 can be used for classification and has a good effect on feature extraction, which was proposed by Andrew Howard et al. [28] on the basis of MobilenetV2. It adds the SE module and changes the complex tail structure of MobilenetV2, which is more lightweight than MobilenetV2 and at the same time has higher accuracy. Using MobilenetV3 as the backbone feature extraction network for YOLOv4 can achieve good detection results with fewer parameters, and it is more lightweight than CSP Darknet53.

The technique of applying Mobilenet to YOLO has been studied by some scholars. Zhang, Xiaxia et al. [29] replaced Darknet53 in YOLOv3 with the improved MobileNetv3 for feature extraction to reduce the algorithm complexity and simplify the model. Moreover, a new attention module SESAM is constructed by channel attention and spatial attention in MobileNetv3. Chen, YunFei et al. [30] proposed a structure based on the YOLOV5-Mobilenetv3Small network model and applied MobileNetv3Smal to YOLOv5, which improved the BackBone network structure to solve the problem of inference high-pixel images taking up too much memory for low power edge computing nodes. In the literature [31], the backbone of YOLOv4 adopts Mobilenetv3, which is improved by CBAM and modified SENet. As a result, the effect of high-light background interference is eliminated and the complexity of the model is reduced.

Different from the three papers above, this paper changed CSP Darknet53 in YOLOv4 to Mobilenetv3 for the problem of fish individual object detection, adjusted the input feature size to 416 × 416, and the output channels of the three effective feature layers were 40,112,160, respectively. Then the three effective feature layers are connected into PANet and SPP.

In addition, depthwise separable convolution [32] is considered to be applied to YOLOv4 in this paper. Depthwise separable convolution divides the convolution into the spatial dimension and the channel dimension for convolution, respectively. It has the same input and output as the standard convolution but reduces the number of parameters and calculations compared to the standard convolution. Therefore, applying it to the network can greatly reduce the number of parameters and amount of computation.

The application of depthwise separable convolution has also been studied by some scholars. In [33], the standard convolution is replaced by depthwise separable convolution in the feature extraction network, at the same time, the attention mechanism is introduced in the channel and spatial dimensions in each residual block of the feature extraction network to focus on small targets. In [34], the classic res bottleneck block is improved to compact res bottleneck block by removing the last 1 × 1 convolution layer and using a 3 × 3 depthwise separable convolution. In [35], the author proposed Reverse Depthwise Separable Convolution (RDSC) and applied it to the backbone network and feature fusion network of YOLO v4.

Different from the three studies above, this paper applies the depthwise separable convolution to the enhanced feature extraction network PANet, and replaces all 3 × 3 convolutions in PANet with depthwise separable convolutions in order to further reduce the numbers of parameters and speed up the calculation.

3.2. Recognition Module

As for the fish individual recognition module, this paper proposes an individual recognition algorithm for underwater fish, named FIRN (Fish individual recognition network). It includes two parts: the backbone feature extraction network and the loss calculation. The backbone feature extraction network is improved on the basis of ResNext50 [36] and the loss function adopts Arcface Loss. The main improvements to Resnext50 are as follows.

1. The CBAM attention module is embedded in the residual block of Resnext50;
2. Batch normalization is performed in the residual block before convolution;
3. The max-pooling layer is removed and the dilated convolution is introduced into the residual block, while the standard convolution is still used in the backbone network;
4. We use the Hard-Swish activation function instead of ReLu;
5. The BN–dropout–FC–BN structure is adopted.

The residual block of the improved ResNext50 is named C-Bottleneck, and its structure and the structure of the improved backbone network are shown in Figure 4 and Figure 5, respectively.

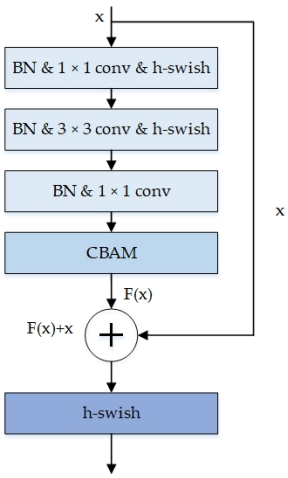


Figure 4. C-Bottleneck: residual block after embedding CBAM.

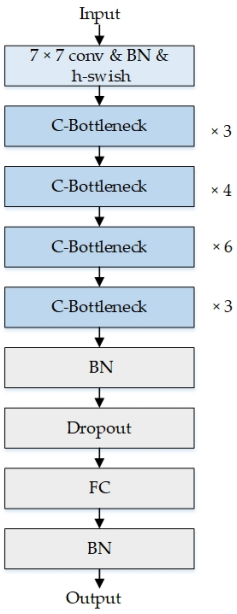


Figure 5. The backbone network of the improved ResNext50.

3.2.1. ResNext50

The reason why this paper adopts the ResNext50 network and improves it is that ResNext has better performance than Resnet. ResNext adds a hyperparameter to Resnet, called cardinality. Reference [36] wrote that increasing the cardinality is a more effective way to obtain accuracy than increasing the depth or increasing the width. Increasing the cardinality is to divide the input channels into multiple groups for convolution so that the output channels are widened and the obtained features are more abundant. The block structure comparison between Resnet and ResNext is shown in Figure 6. The left image is Resnet and the right image is ResNext.

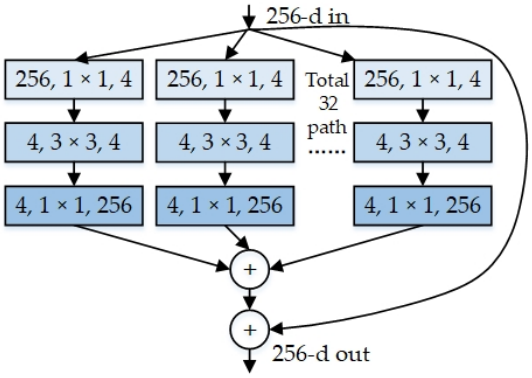


Figure 6. The block structures of Resnet and ResNext.

As seen in Figure 6, the input of the block of ResNext is a feature map of 256 channels, and then it is divided into 32 branches, the number of input channels for the first convolutional layer of each branch is 256, the size of the convolution kernel is 1 × 1, and the number of output channels is 4. The number of input channels of the second convolutional layer of each branch is four, the kernel size is 3 × 3, and the number of output channels is four. The number of input channels of the third convolutional layer of each branch is 4, the kernel size is 1 × 1, and the number of output channels is 256. Then, the output feature maps of the 32 branches are added point by point. Finally, the final output is obtained by summing the result of the summation with the input part through a short connection. Figure 7 is the completely equivalent structure obtained by simplifying the right figure in Figure 6, which is the most widely used structure. The block structure shown in Figure 7 is also adopted in this paper.

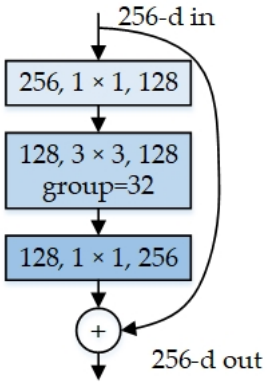


Figure 7. The equivalent block of ResNext.

3.2.2. CBAM Attention Module

In order to improve the ability of the network to extract features of individual fish, the Convolutional Block Attention Module (CBAM) [37] is embedded in the residual

blocks of ResNext50. It is a lightweight attention module proposed by Sanghyun Woo et al., which is composed of the Channel Attention Module (CAM) and Spatial Attention Module (SAM), as shown in Figure 8.

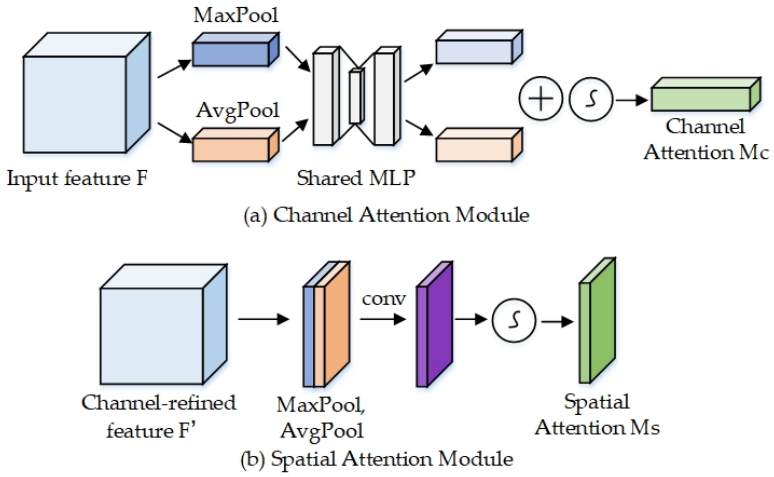


Figure 8. Structure of the CBAM attention module. (**a**) Structure of the Channel Attention Module; (**b**) Structure of the Spatial Attention Module.

In the CAM, the input feature maps are first pooled through max pooling and average pooling, respectively, and then input into the Shared MLP. After the Shared MLP, the output feature elements are summed to merge the output features. Finally, activated by sigmoid, the output features of the CAM can be obtained.

In the SAM, the features output by the CAM are used as input. Max pooling and average pooling are also performed first, then the two layers are spliced, and then the convolution operation is performed to reduce the channel to 1. Finally, the output features of the SAM can be obtained through sigmoid activation.

In [38], the CBAM attention module is applied to the output of ResNext50, which aims to conduct CBAM processing on each group of the detailed features to get informative features, suppress unnecessary features, and improve the information utilization.

However, for the fish individual recognition problem, we applied the CBAM attention module to the residual block of ResNext50, named the residual block after embedding the CBAM as C-Bottleneck, and then replaced Bottleneck in ResNext50 with C-Bottleneck. The purpose of this is to improve the ability of the network to extract features of individual fish and extract more detailed features.

3.2.3. Batch Normalization

Data distribution is particularly important in the process of neural network training. When the training data and test data do not satisfy the same distribution, the generalization ability of the network will be greatly reduced. In addition, during training, if each batch of data does not satisfy the same distribution, the network has to be re-adapted every time, which will greatly reduce the training speed. The batch normalization (BN) [39] normalizes the input data to meet the normal distribution with a mean of 0 and variance of 1. In this way, the convergence speed of the network can be accelerated, a larger initial learning rate can be set, and the generalization ability of the network can be improved.

In the residual blocks of the original ResNext50 network, the convolution operation is carried out first, and then the BN operation is carried out, which is effective. However, it is easy to make the network unstable in the training process, and the loss decrease is easy to produce large fluctuations, which will affect the training effect. In order to make the training more stable and further speed up the training, this paper takes the form of

advancing the BN layer and putting it before the convolutional layer. The data input from the upper layer is first subjected to batch normalization operation to make it satisfy the normal distribution, and then convolution operation is performed on it, which can speed up the training speed and make the training process more stable.

3.2.4. Pooling

In the original ResNext, the input feature map is subjected to max pooling and average pooling. Max pooling is to take the maximum value of the feature values in the neighborhood and conduct the down-sampling operation on the feature map. After the max-pooling layer, the size of the feature map is halved, which can play the role of dimension reduction and reduce the network parameters. However, in the process of dimension reduction, some details and some smaller targets will be lost, and the lost information is irreducible, so the effect is not ideal. Average pooling is to take the average value of the feature values in the neighborhood, which can preserve the background well, but easily make the image blurred. In addition, the images corresponding to the perceptual fields of different points are different, so the weight of each point should be different. However, average pooling considers them as the same weight, so the network performance will decrease.

Due to the problems caused by the above-pooling layer, the pooling layer in the ResNext50 backbone network is modified in this paper. Firstly, the max pooling is removed, so that the feature map size will not change and a lot of information will not be lost. Second, the average-pooling layer is removed, and the BN–dropped–FC–BN structure is then used.

After removing the average pooling, the information of the original feature map can be retained without blurring the image. At this time, connecting a BN layer first can normalize the input data above, and then enter the dropout layer to randomly inactivate some neurons to prevent overfitting. Then it enters the fully connected layer to extract features to achieve classification. Finally, a BN layer is passed so that the output data is uniformly regularized to obtain the final output features.

For max pooling, if it is directly removed, a new problem will arise, that is, the receptive field of the original image corresponding to the obtained feature map becomes smaller because the max-pooling layer can increase the receptive field of the image. This will have a certain impact on the subsequent series of convolution operations. In order to increase the receptive field, this paper introduces dilated convolution into a series of residual structures of ResNext50, while the backbone network, except for the residual structure, still uses standard convolution.

3.2.5. Dilated Convolution

Dilated convolution [40] injects holes on the basis of standard convolution to increase the receptive field and keep the size of the original input feature map unchanged. Based on the standard convolution, dilation convolution adds a new hyperparameter, called dilation rate, which means the number of kernels spaced. As shown in Figure 9, the left figure shows the standard convolution with a convolution kernel size of 3 × 3, which can also be regarded as the dilated convolution with an expansion rate of 1. The figure on the right shows the dilated convolution, the dilation rate is 2, and the size of the convolution kernel is expanded to 5. It can be seen that dilated convolution expands the size of the convolution kernel on the basis of ordinary convolution, but the number of elements involved in the operation does not change. Therefore, dilated convolution increases the receptive field by expanding the size of the convolution kernel. As long as the dilated convolution uses different dilated rates and they superimpose each other, its receptive field will grow exponentially, and the size of the original feature map will not change.

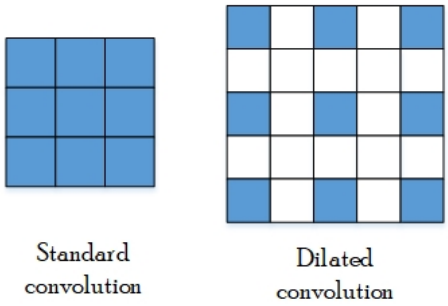


Figure 9. Schematic diagram of standard convolution and dilated convolution.

In reference [40], the author mentioned that the gridding effect will occur when dilated convolution is used. The reason is that when the same dilated rate is adopted and multiple convolutions are continuously superimposed, some pixel values are not used, and some information will be lost, which will affect the effect of feature extraction. In order to avoid such problems, this paper adopts the Hybrid Dilated Convolution (HDC) proposed in reference [40] and adopts three different expansion rates of 1, 2, and 3 for each convolutional layer. In this way, the information of pixels will not be missed, and the receptive field is increased.

3.2.6. The Activation Function

The ReLu activation function is used in the original ResNext. Although the ReLu activation function is widely used, there are certain drawbacks. When the input is close to zero or negative, the gradient of the ReLu function becomes 0, the network cannot perform backpropagation, and the problem of neuron deactivation occurs. In order to avoid such problems, this paper adopts the Hard-Swish activation function.

The Swish activation function was proposed by Prajit Ramachandran et al. [41], which is an upgrade of the Sigmoid and ReLu activation functions. Swish has the advantages of two activation functions of Sigmoid and ReLu, and its effect is better than both in the deep model. Its expression is:

$$f(x) = x \cdot \text{sigmoid}(\beta x) \tag{1}$$

where β is a constant or learnable parameter.

However, Andrew Howard et al. [27] found that the Swish function can only play a better role in the deep model, and the calculation is complex, so they proposed Hard-Swish. The Hard-Swish function has the advantages of the Swish function and is not affected by the depth of the model, so it has a wider range of applications. Its expression is:

$$\text{Hardswish}(x) = \begin{cases} 0, & x \le -3 \\ x, & x \ge 3 \\ x \cdot (x+3)/6, & \text{otherwise} \end{cases} \tag{2}$$

This paper replaces all the ReLu activation functions in the original ResNext network with Hard-Swish, which has a good effect.

3.2.7. Loss Function

The loss function is the key to ensuring model training effect and prediction accuracy. Common loss functions in the field of biometric technology include Softmax Loss [42], Triplet Loss [43], Arcface Loss [44], etc. The equation of Softmax Loss is shown in Equation (3). It can ensure that categories are separable in face recognition but does not require intra-class compactness and inter-class separation, so it is not suitable for fish individual recognition tasks.

$$L_{soft\max} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} \tag{3}$$

The expression of Triplet Loss is Equation (4). Unlike Softmax Loss, Triplet Loss needs to compare the distances between three feature vectors: two face image features of the same class and one face image feature of a different class. By training, the distance between classes is larger and the distance within classes is smaller. However, when the dataset is large, the number of Triplets explodes, resulting in a significant increase in the number of iterations.

$$L_{triplet} = \sum_{i}^{N}[\| x_i^a - x_i^p \|_2^2 - \| x_i^a - x_i^n \|_2^2 + \alpha]_+ \tag{4}$$

Later, many scholars proposed variants of Softmax Loss to enhance the discriminative ability of Softmax Loss. Jiankang Deng et al. proposed Arcface Loss on the basis of Softmax Loss, see Equation (5).

Arcface Loss sets the bias term b_j in Softmax Loss to 0 and normalizes the input feature x_i and weight W_j^T, so $W_j^T x_i = \| W_j \| \| x_i \| \cos\theta_j$. The arccosine function is used to calculate the angle θ_j between the current feature and the target weight. Then the additive angular margin m is added to the angle θ_j to enhance the intra-class compactness and inter-class difference. The angle after increasing m is obtained by the cosine function to obtain the target logit, and then all logits are rescaled by the feature norm s. The newly obtained logit is then calculated by Softmax Loss to obtain Arcface Loss.

$$L_{arcface} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1, j\neq y_i}^{n} e^{s\cos\theta_j}} \tag{5}$$

After Arcface Loss training, larger inter-class distance and smaller intra-class distance can be obtained, stable performance can be obtained without combining with other loss functions, and it can be easily converged. Therefore, Arcface Loss is used as the loss function of the fish individual recognition network in this paper.

4. The Simulation Experiment

4.1. Experimental Environment and Evaluation Index

The computer system used in the whole experiment is ubuntu20.04, the CPU is Intel i7-11700K@3.6 GHz, the memory is 32 G, the graphics card is GeForce RTX3090, accelerated by CUDA11.1, the language is Python3.6, and the version of Pytorch is 1.8, implemented in a virtual environment.

In the object detection module, the mean average precision (mAP) is used as the evaluation index of network performance. In the recognition module, intra-class and inter-class distances are used to verify the effectiveness of the algorithms in this paper by comparing them with different algorithms.

4.2. Dataset

Since there is no publicly available individual dataset of large underwater fish at this stage, this paper collects and constructs an underwater fish dataset, which includes four different species of fish, such as puffer fish, koi, grass goldfish, and clownfish. Figure 10 is a sample image of the dataset. Among them, the images of puffer fish and grass goldfish were taken on-site by underwater cameras in the fish farm, and the images of koi and clownfish were collected from various websites [45,46]. Each type of fish has 30 to 50 different individuals, and the total number of images is 8000. This paper takes koi as an

example to verify the proposed algorithm. A total of 3500 koi images were selected as the dataset of fish individual object detection, of which 2800 were used for training and 700 were used for testing.

Figure 10. Sample graph of the dataset.

After the detection of fish individual targets, the image is trimmed according to the coordinates of the detected prediction box, and every fish detected in the image is trimmed out. After screening, a new dataset is formed for fish individual recognition. A total of 3600 images of koi individuals were selected, of which 2600 were used for training and 1000 were used for testing, including 50 fish individuals with different identities.

4.3. Specific Implementation

This paper is divided into two modules: fish individual object detection and fish individual recognition. Therefore, the two tasks are trained and tested, respectively, and then the two parts are combined to carry out the final effect verification by putting the two weights of the trained object detection and recognition.

Firstly, the image object detection dataset of koi was manually annotated. The labeling tool uses "labelimg" to label each fish in the image, the label is fish, and the label file format is "xml". After labeling, the format of the dataset should be converted to the format of the VOC dataset, and "train.txt" and "val.txt" are generated for training and testing, respectively.

The processed dataset is sent to the improved YOLOv4 network for the training of fish individual object detection. The initial training learning rate is set to 0.001, using the SGD optimizer, setting the batch size to 32 and the number of epochs to 300. The weight is saved every 20 epochs, and the optimal weights are selected for network testing and prediction. Images or videos can be input for prediction.

In the fish individual recognition module, the dataset adopts the cropped koi individual images and put them into 50 folders according to the individual fish with different identities. In order to expand the dataset, the images of each identity category are randomly rotated from −20 degrees to 20 degrees, so that there are 20 to 50 different images of the same fish in the folders of each identity category. We generate "txt" label files for training and testing by reading folder data.

The processed data is sent to FIRN for the training of fish individual recognition. The feature extraction network is the improved ResNext50, and the loss function is Arcface Loss. The training sets the initial learning rate to 0.0001, using the Adam optimizer, the batch size is 16, the number of epochs is 300, and the value of momentum is 0.9. After the training, the optimal weight is selected for the test and prediction of fish individual recognition.

4.4. Experimental Results

4.4.1. Fish Individual Object detection

In order to test the effectiveness of the method proposed in this paper, it is verified on the same test dataset with other algorithms, and the performance index results are shown in Table 1.

Table 1. Performance comparison of different object detection algorithms.

Algorithms	mAP/%
YOLOv4	86.89
YOLOv4-tiny	85.65
The proposed method	88.86

It can be seen from Table 1 that the proposed method in this paper improves mAP by 1.97% compared with the original YOLOv4 algorithm and 3.21% compared with YOLOv4-tiny, which is enough to verify the effectiveness of the improved algorithm in this paper. Figure 11 is a schematic diagram of the detection results of fish individual targets by the proposed method in this paper. It can be seen from Figure 11 that all fish individual targets can be detected with an accuracy of more than 90%.

Figure 11. Example of object detection results.

4.4.2. Fish Individual Recognition

In the fish individual recognition module, after the training of the FIRN algorithm, an accuracy rate of 98.7% can be obtained on the test dataset, and the obtained recognition results have a smaller intra-class distance while ensuring inter-class separability.

In order to verify the effectiveness of the improvement of each part in this paper, several experiments were carried out. Firstly, the original ResNext50 was used for the experiment and compared with the ResNext50 embedded in the CBAM attention module. The comparison results are shown in Table 2.

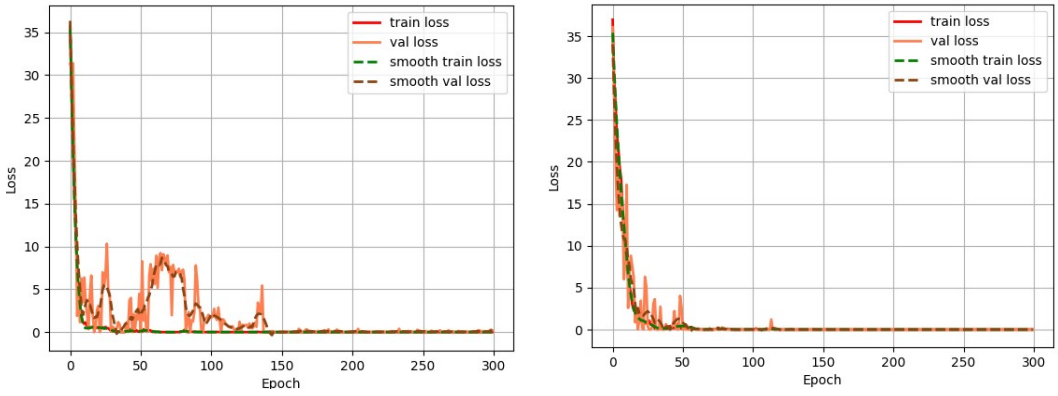
Table 2. Comparison of network performance before and after embedding CBAM.

Category	Image	ResNext50	CBAM- ResNext50
intra-class distance	101 102	0.351	0.337
	201 202	0.595	0.419
inter-class distance	101 301	1.413	1.439
	101 202	1.442	1.534

The algorithm verifies whether they belong to the same individual by calculating the Euclidean distance between the two images. We set the threshold to 1.0. When the Euclidean distance between the two images is less than 1.0, it is judged to be the same individual, otherwise, it is a different individual. In Table 2, images "101" and "102" and images "201" and "202" are two different images of the same individual; images "101" and "301" and images "101" and "202" are different individuals, and other numbers in the table represent the Euclidean distance between the two images.

It can be seen in Table 2 that the intra-class distance of the image after embedding CBAM becomes smaller because the feature extraction ability of the network after embedding CBAM is enhanced, which shows that embedding CBAM is helpful for the improvement of the network training effect.

In the training process of CBAM-ResNext50, we found that the decreasing process of loss function fluctuated greatly, which was guessed because the data did not meet the same distribution and the randomness was large. Therefore, in this paper, the BN layer is placed before the convolution layer. The data input from the upper layer is first subjected to BN operation to make it satisfy the normal distribution, and then convolution operation is performed on it. The loss function decline curve before and after the change is shown in Figure 12. Experiments show that placing the BN layer before the convolutional layer does make the training more stable, and the drop of the loss function no longer fluctuates greatly.



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

Figure 12. (**a**) Decline of the loss function before adjusting the order of BN layers; (**b**) decline of the loss function after adjusting the order of BN layers.

On the basis of the above improvements, the pooling layer in the network is removed for the experiment. The experiment shows that the effect becomes worse after the pooling layer is removed. According to the above analysis, the receptive field becomes smaller after the pooling layer is removed, so it can be seen that the change in the receptive field has a certain impact on the training effect of the network. In view of this, dilated

convolution is introduced into the residual block in this paper to increase the receptive field. The experimental data are shown in Table 3.

As can be seen in Table 3, after the introduction of dilated convolution, the intra-class distance decreases greatly, and the inter-class distance also increases. Therefore, after using dilated convolution, the network learns more detailed information while increasing the receptive field, which is of great help to the effect of fish individual recognition.

Table 3. Comparison of network performance before and after the introduction of dilated convolution.

Category	Image	After Directly Removing the Pooling Layer	After Introducing Dilated Convolution
intra-class distance	101 102	0.374	0.182
	201 202	0.490	0.237
inter-class distance	101 301	1.336	1.411
	101 202	1.357	1.426

The activation functions used in the above experiments are all ReLu. It is known from the above that the ReLu activation function has certain drawbacks. Therefore, on the basis of all the above improvements, this paper replaces all activation functions in the network with Hard-Swish. After using the Hard-Swish activation function, it is the final improved algorithm in this paper. In order to verify the effectiveness of the algorithm in this paper, compared with other classical feature extraction networks, the loss function in all experiments is Arcface Loss, and the results are shown in Table 4.

Table 4. Performance comparison of different algorithms.

Category	Image	Resnet50	Mobile-Facenet	The Proposed Method
intra-class distance	101 102	0.315	0.249	0.150
	201 202	0.544	0.442	0.192
inter-class distance	101 301	1.427	1.504	1.430
	101 202	1.338	1.467	1.469

Figure 13 shows the visualization results of the experiment. The left side of the dotted line shows the intra-class distances obtained by using three feature extraction networks for two different images of the same fish "201" and "202", while the right side of the dotted line shows the inter-class distances obtained by using three feature extraction networks for two images of different fish "101" and "201". The three feature extraction networks are Resnet50, Mobilefacenet, and the algorithm in this paper is from top to bottom.

It can be seen from the chart that under the premise of using the same loss function, the algorithm in this paper has a smaller intra-class distance than using Resnet50 and Mobilefacenet as the feature extraction network. Additionally, the inter-class distance is slightly larger than that of Resnet50, but there is no obvious improvement compared with Mobilefacenet.

Compared with Mobilefacenet, although the inter-class distance obtained by the proposed algorithm is not significantly improved, it can also obtain a large inter-class distance, which can meet the requirements of fish individual recognition. As long as the threshold value is greater than 1 in the whole fish individual recognition, it will be judged as different individuals, and the system will abandon the image to find the next image for feature comparison. Therefore, as long as the separability between classes can be guaranteed, a smaller intra-class distance between the same individuals is more meaningful.

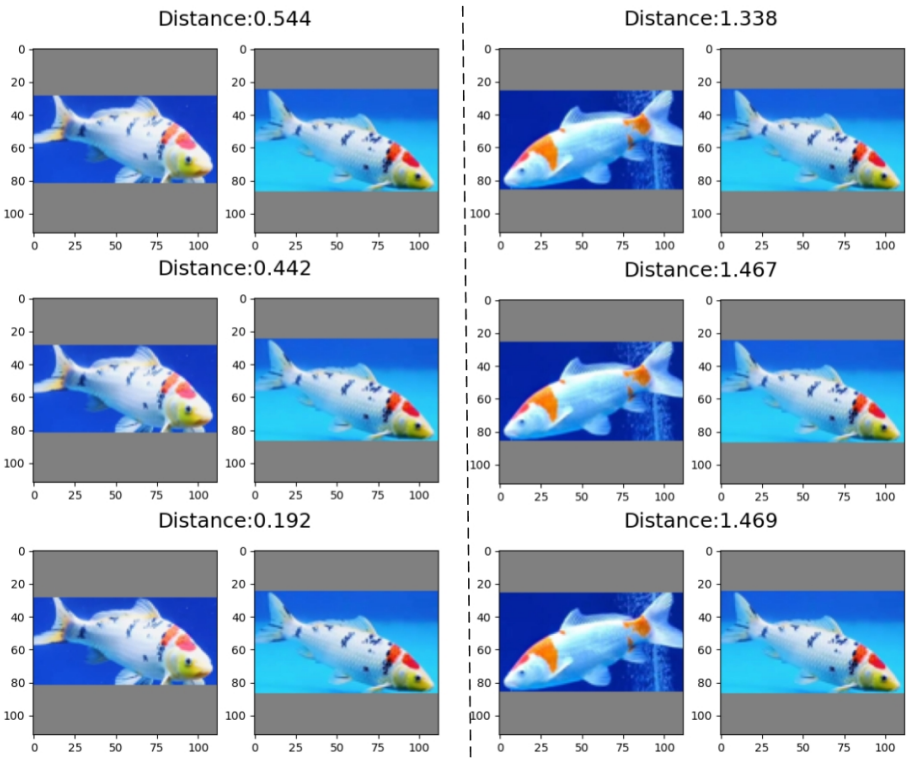


Figure 13. Examples of fish individual recognition visualization results.

In addition, in order to verify the effectiveness of the proposed algorithm under the three conditions of unclear water quality, insignificant difference in fish spots, and rotation of the fish body, three different fish images of these three states were selected for experiments, respectively. Figure14 shows an example of experimental prediction results.

Figure 14a,b are images of puffer fish with unclear water quality. Figure 14a is the same fish, the body of the fish in the left image has been rotated, and the intra-class distance obtained is 0.265. Figure 14b shows two different fish with a small inter-class distance of 1.010. Therefore, it can be seen that the clarity of water quality has a certain impact on the recognition effect of the algorithm. Because the water quality is not clear, the characteristics of fish become blurred, which has a great impact on the inter-class distance. However, the algorithm is still effective and can complete the task of fish individual recognition in the case of unclear water quality.

Figure 14c,d show the recognition result of two grass goldfish with very similar spots. The intra-class distance is 0.271 and the inter-class distance is 1.260. According to the experimental results, when the spots of fish are very similar, the algorithm can still distinguish different individuals. Due to the high degree of similarity of spots, the distance between classes becomes slightly smaller, but the algorithm is still effective.

Figure 14e,f show the recognition results of two koi with larger body rotation angles. Figure 14e shows the same koi with different angles and different lights, and its intra-class distance is 0.623, which is a larger increase than the intra-class distance under other conditions. Figure 14f shows two different koi, and the inter-class distance is 1.455, which is

similar to the result without rotation. It can be seen that when the body rotation angle of the fish is large, the intra-class distance is greatly affected due to the large change of features, but the obtained intra-class distance is still within the threshold of 1.0, so the algorithm is also effective.

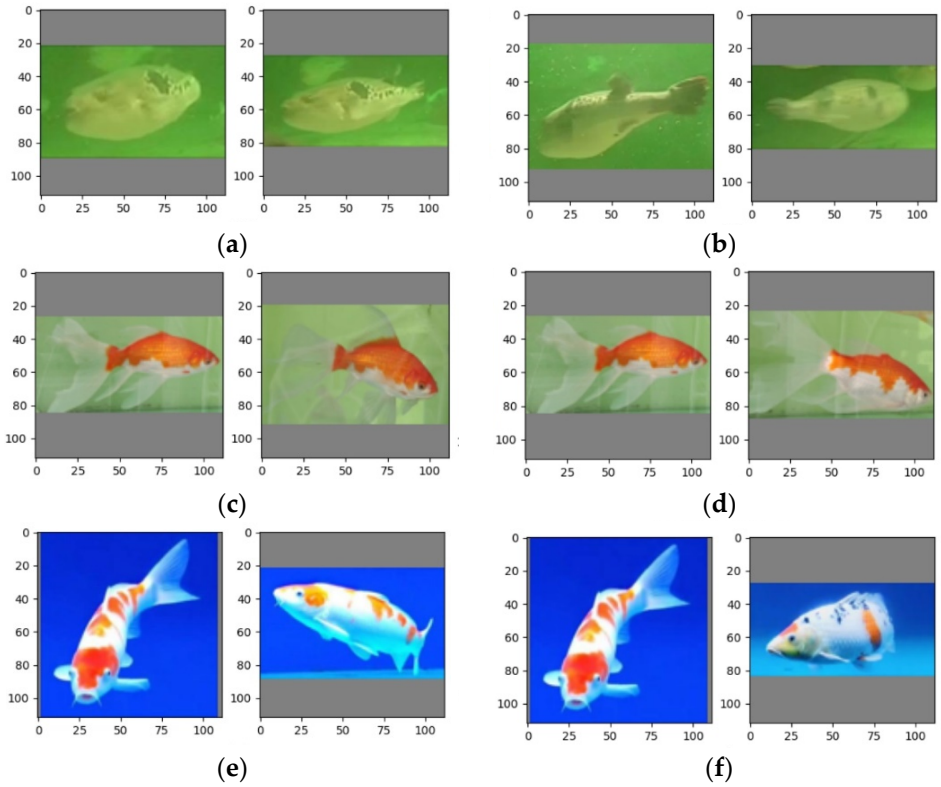


Figure 14. Recognition results of three fish species under three different conditions: (**a**) Distance: 0.265; (**b**) Distance: 1.010; (**c**) Distance: 0.271; (**d**) Distance: 1.260; (**e**) Distance: 0.623; (**f**) Distance 1.455

4.4.3. Visualization Results

The above object detection and individual recognition results are obtained by training alone, but this does not meet practical needs. What we want to achieve is that as long as the video or image of the underwater fish is collected and input into the system, the recognition result of each fish can be obtained. Therefore, we combined the two programs of fish individual object detection and fish individual recognition and put the best weights obtained from the training of the two parts into the synthesis program to verify the final results. As long as an image or video of a certain fish group is put in, the prediction box of each fish can be detected, and its identity information can be marked, so as to realize the detection and recognition of individual fish. Figure 15 shows an example of the final recognition result.

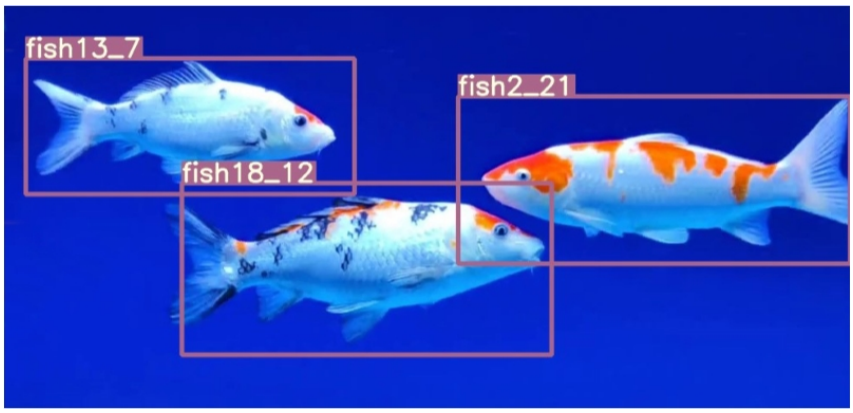


Figure 15. Sample diagram of the final recognition result.

5. Summary

This paper proposes a method for individual recognition of underwater fish, which is divided into two parts: fish individual object detection and fish individual recognition. In the detection module, by replacing the feature extraction network and introducing depthwise separable convolution in PANet, the YOLOv4 object detection algorithm is improved, which is lighter and has higher accuracy than the original YOLOv4. In the recognition module, the FIRN algorithm is proposed, and ResNext50 is improved by adding the CBAM attention module and introducing dilated convolution, etc. The improved ResNext50 is used as a feature extraction network, and Arcface Loss is used as a loss function, which can not only ensure the separability between classes but also enhance the compactness within a class. The combination of detection and recognition algorithms can realize individual recognition of underwater fish and has a good effect.

Author Contributions: Methodology, C.G. and J.W.; conceptualization, C.G. and J.W.; resources, H.Y. and S.G.; data curation, C.G. and J.Y.; writing—original draft preparation, C.G. and J.W.; writing—review and editing, C.G., J.W., and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Research Projects in Liaoning Province (2020JH2/10100043), National Natural Science Foundation of China (31972846), Key Laboratory of Environment Controlled Aquaculture (Dalian Ocean University) Ministry of Education (202205), and National Key Research and Development Program of China (2021YFB2600200).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to part of the data is provided by the cooperative enterprise.

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

References

1. Mang, Q.;Xu, G.; Zhu,J; Xu, P. Current Situation and Prospects of Aquaculture Development in China. *Fish. Mod.* **2022**, *49*, 1.
2. Zhang, L.; Wang, J.; Duan, Q. Estimation for fish mass using image analysis and neural network. *Computers and Electronics in Agriculture* **2020**, *173*, 105439.
3. Li, D.; Wang, Z.; Wu, S.; Miao, Z.; Du, L.; Duan, Y. Automatic recognition methods of fish feeding behavior in aquaculture: A review. *Aquaculture* **2020**, *528*, 735508.
4. Wu, Y.; Yin, J.; Dai, Y.; Yuan, Y. Recognition of freshwater fish species based on bee colony optimization multi kernel support vector machine. *Trans. Chin. Soc. Agric. Eng.* **2014**, *30*, 312–319.
5. Wan, P.; Pan, H.; Long, C.; Chen, H. Design of online recognition device for freshwater fish species based on machine vision technology. *Food Mach.* **2012**, *6*, 164–167.
6. Webb, G.I.; Keogh, E.; Miikkulainen, R. Naïve Bayes. *Encycl. Mach. Learn.* **2010**, *15*, 713–714.
7. Song, Y.-Y.; Ying, L.U. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **2015**, *27*, 130.
8. Abeywickrama, T.; Cheema, M.A.; Taniar, D. K-nearest neighbors on road networks: A journey in experimentation and in-memory implementation. *arXiv* **2016**, arXiv:1601.01549.

9. Pisner, D.A.; Schnyer, D.M. Support vector machine. In *Machine Learning*; Academic Press: Cambridge, MA, USA, 2020; pp. 101–121.
10. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
12. Joseph Redmon; et al. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016.
13. Larsen, R.; Olafsdottir, H.; Ersbøll, B.K. Shape and texture based classification of fish species. In *Scandinavian Conference on Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2009.
14. Ravanbakhsh, M.; Shortis, M.R.; Shafait, F.; Mian, A.; Harvey, E.S.; Seager, J.W. Automated Fish Detection in Underwater Images Using Shape-Based Level Sets. *Photogramm. Rec.* **2015**, *30*, 46–62.
15. Hsiao, Y.H.; Chen, C.C.; Lin, S.I.; Lin, F.P. Real-world underwater fish recognition and identification, using sparse representation. *Ecol. Inform.* **2014**, *23*, 13–21.
16. Villon, S.; Mouillot, D.; Chaumont, M.; Darling, E.S.; Subsol, G.; Claverie, T.; Villéger, S. A deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecol. Inform.* **2018**, *48*, 238–244.
17. Rauf, H.T.; Lali, M.I.U.; Zahoor, S.; Shah, S.Z.H.; Rehman, A.U.; Bukhari, S.A.C. Visual features based automated identification of fish species using deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *167*, 105075.
18. Labao, A.B.; Naval, P.C.; Jr. Cascaded deep network systems with linked ensemble components for underwater fish detection in the wild. *Ecol. Inform.* **2019**, *52*, 103–121.
19. Zhang, M., Long, T.; Song, W.; Huang, D.; Mei, H.; Tan, X. A moving object detection method for underwater fish dynamic visual sequence. *J. Graph.* **2021**, *42*, 52.
20. Li, C. Detection and recognition of underwater fish targets based on YOLOv3.2020. Master's Thesis, Northwest A&F University: Xianyang, China**.2020**.
21. Cai, K.; Miao, X.; Wang, W.; Pang, H.; Liu, Y.; Song, J. A modified YOLOv3 model for fish detection based on MobileNetv1 as backbone. *Aquac. Eng.* **2020**, *91*, 102117. https://doi.org/10.1016/j.aquaeng.2020.102117.
22. Xue, Y.; Ju, Z. Fish Recognition Algorithm based on improved AlexNet. *Electron. Sci. Technol.* **2021**, *34*, 12–17.
23. Bochkovskiy, A.; Wang, C.Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
24. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Ultralytics/Yolov5. Available online: https://github.com/ultralytics/yolov5 (accessed on 9 October 2022).
26. About Reproduced Results #6. Available online: https://github.com/ultralytics/yoloV5/issues/6 (accessed on 9 October 2022).
27. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 cotober–2 November 2019; IEEE: Piscataway, NJ, USA, 2020.
28. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018.
29. Zhang, X.; Li, N.; Zhang, R. An improved lightweight network MobileNetv3 Based YOLOv3 for pedestrian detection. In Proceedings of the 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 15–17 January 2021; IEEE: Piscataway, NJ, USA, 2021.
30. Chen, Y.; Chen, X.; Chen, L.; He, D.; Zheng, J.; Xu, C.; Lin, Y.; Liu, L. UAV Lightweight Object Detection Based on the Improved YOLO Algorithm. In Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering, Xiamen China, 22–24 October 2021.
31. Ye, X.; Zhang, W.; Li, Y.; Luo, W. Mobilenetv3-YOLOv4-Sonar: Object Detection Model Based on Lightweight Network for Forward-Looking Sonar Image. In *OCEANS 2021*: *San Diego–Porto*; IEEE: Piscataway, NJ, USA, 2021.
32. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
33. Li, Y.; Li, S.; Du, H.; Chen, L.; Zhang, D.; Li, Y. YOLO-ACN: Focusing on small target and occluded object detection. *IEEE Access* **2020**, *8*, 227288–227303.
34. Lu, Y.; Zhang, L.; Xie, W. YOLO-compact: An efficient YOLO network for single category real-time object detection. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 22–24 August 2020; IEEE: Piscataway, NJ, USA, 2020.
35. Liu, T.; Pang, B.; Zhang, L.; Yang, W.; Sun, X. Sea Surface Object Detection Algorithm Based on YOLO v4 Fused with Reverse Depthwise Separable Convolution (RDSC) for USV. *J. Mar. Sci. Eng.* **2021**, *9*, 753.
36. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 24–26 July 2017.
37. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018.

38. Chen, M.; Zhao, C.; Tian, X.; Liu, Y.; Wang, T.; Lei, B. Placental Super Micro-vessels Segmentation Based on ResNeXt with Convolutional Block Attention and U-Net. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Mexico, 1–5 November 2021; IEEE: Piscataway, NJ, USA, 2021.
39. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; PMLR: New York, NY, USA, 2015.
40. Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; Cottrell, G. Understanding convolution for semantic segmentation. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018.
41. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv* **2017**, arXiv:1710.05941.
42. Liu, W.; Wen, Y.; Yu, Z.; Yang, M. Large-margin softmax loss for convolutional neural networks. *arXiv* **2016**, arXiv:1612.02295.
43. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
44. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019.
45. Koi First-Timers. Available online: https://space.bilibili.com/162546647/video?tid=0&page=9&keyword=&order=pubdate (accessed on 21 December 2021).
46. Beautiful Clown Fish Aquarium & Relaxing Music in 4K. Available online: https://www.youtube.com/watch?v=dQfrTmubDzM (accessed on 22 December 2021).