frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 11 November 2024
DOI 10.3389/fmars.2024.1471312



OPEN ACCESS

EDITED BY
Chunlei Xia,
Chinese Academy of Sciences (CAS), China

REVIEWED BY
Chao Zhou,
Beijing Research Center for Information Technology in Agriculture, China
Suja Cherukullapurath Mana,
PES University, India

*CORRESPONDENCE
Hong Yu
yuhong@dlou.edu.cn

RECEIVED 27 July 2024
ACCEPTED 23 October 2024
PUBLISHED 11 November 2024

CITATION
Zhang P, Yang Z, Yu H, Tu W, Gao C and Wang Y (2024) RUSNet: Robust fish segmentation in underwater videos based on adaptive selection of optical flow. *Front. Mar. Sci.* 11:1471312. doi: 10.3389/fmars.2024.1471312



# RUSNet: Robust fish segmentation in underwater videos based on adaptive selection of optical flow

Peng Zhang[1,2,3,4], Zongyi Yang[1,2,3,4], Hong Yu[1,2,3,4]*, Wan Tu[1,2,3,4], Chencheng Gao[1,2,3,4] and Yue Wang[1,2,3,4]

[1]College of Information Engineering, Dalian Ocean University, Dalian, China, [2]Dalian Key Laboratory of Smart Fisheries, Dalian Ocean University, Dalian, China, [3]Key Laboratory of Facility Fisheries, Ministry of Education (Dalian Ocean University), Dalian, China, [4]Liaoning Provincial Key Laboratory of Marine Information Technology, Dalian, China

Fish segmentation in underwater videos can be used to accurately determine the silhouette size of fish objects, which provides key information for fish population monitoring and fishery resources survey. Some researchers have utilized underwater optical flow to improve the fish segmentation accuracy of underwater videos. However, the underwater optical flow is not evaluated and screen in existing works, and its predictions are easily disturbed by motion of non-fish. Therefore, in this paper, by analyzing underwater optical flow data, we propose a robust underwater segmentation network, RUSNet, with adaptive screening and fusion of input information. First, to enhance the robustness of the segmentation model to low-quality optical flow inputs, a global optical flow quality evaluation module is proposed for evaluating and aligning the underwater optical flow. Second, a decoder is designed by roughly localizing the fish object and then applying the proposed multidimension attention (MDA) module to iteratively recover the rough localization map from the spatial and edge dimensions of the fish. Finally, a multioutput selective fusion method is proposed in the testing stage, in which the mean absolute error (MAE) of the prediction using a single input is compared with that obtained using multisource input. Then, the information with the highest confidence is selected for predictive fusion, which facilitates the acquisition of the ultimate underwater fish segmentation results. To verify the effectiveness of the proposed model, we trained and evaluated it using a publicly available joint underwater video dataset and a separate DeepFish public dataset. Compared with the advanced underwater fish segmentation model, the proposed model has greater robustness to low-quality background optical flow in the DeepFish dataset, with the mean pixel accuracy (mPA) and mean intersection over union (mIoU) values reaching 98.77% and 97.65%, respectively. On the joint dataset, the mPA

and mIoU of the proposed model are 92.61% and 90.12%, respectively, which are 0.72% and 1.21% higher than those of the advanced underwater video object segmentation model MSGNet. The results indicate that the proposed model can adaptively select the input and accurately segment fish in complex underwater scenes, which provides an effective solution for investigating fishery resources.

KEYWORDS

**underwater video processing, motion evaluation, adaptive output selection, robust segmentation, deep learning**

## 1 Introduction

Fish are essential marine resources, providing approximately 20% of daily high-quality animal protein for more than 3.3 billion people worldwide (Food and Agriculture Organization of the United Nations, 2022), and the consumption demand for fish and other seafood is still growing; therefore, fishery resource surveys have naturally become a focus of attention. To effectively conduct fishery resource surveys and simultaneously control the impact of production on the environment, the growth status of fish populations in natural habitats, as well as that of artificially cultured fish, must be monitored (Saleh et al., 2023). Currently, the commonly used fish monitoring method involves mounting a camera on underwater equipment for automatic fish segmentation or detection (Chatzievangelou et al., 2022). Through accurate segmentation of underwater fish, information such as fish outline morphology, body length, and size (Zhao et al., 2022) can be obtained; this information can provide key support for the prediction of the long-term production capacity of fish populations (Hall et al., 2023), among other applications. In addition, the size profile information obtained from fish segmentation can be applied to marine environmental protection through monitoring the size and length of juvenile fish. This monitoring can help with formulating a reasonable fish grading feeding and fishing strategy and mitigate marine pollution caused by the irrational discharge of breeding pollutants, such as bait and fish excreta (Cheng et al., 2023). Therefore, accurately segmenting fish objects in underwater environments is important.

In the past, fish length was typically measured by catching and then measuring the fish, which may cause damage to the fish and requires considerable labor and time (Petrell et al., 1997). With the development of computer vision technology, automatically segmenting underwater fish to obtain their size has attracted increasing attention from related researchers. Underwater fish segmentation is a binary semantic segmentation task that is used to separate foreground fish objects from complex backgrounds in underwater scenes. As shown in Figure 1, underwater scenes may exhibit many complex conditions, such as turbid water, insufficient light, and camouflage, so accurately segmenting fish in underwater environments is challenging. In early work, binary masks of underwater fish were typically obtained using hand-selected features such as color, texture, or image morphology methods for segmentation. For example, Yao (Yao et al., 2013) et al. proposed a K-means clustering segmentation algorithm combined with mathematical morphology for fish image segmentation, which separates fish from the background. To address the poor robustness of fish segmentation under low-light conditions, Chuang et al. (Chuang et al., 2011) proposed a method using histogram back-projection to ensure fish segmentation accuracy. However, the particularity of the underwater environment was not considered in these works. Furthermore, the robustness of fish segmentation is poor, and the segmentation accuracy for nongray images and complex background conditions is limited. Since underwater fish images are usually characterized by color deviation and blurring, some researchers have considered enhancing or preprocessing underwater images before segmenting

FIGURE 1
Complex underwater example with turbid water, insufficient light, small target, and camouflaged target scenarios from left to right.

the fish objects. Banerjee et al. (Banerjee et al., 2023) first used vertical flipping and gamma correction to enhance images and then applied two deep learning-based segmentation networks, U-Net and PSPNet, to automatically segment the heads of fish. However, the model's segmentation mIoU was only 76%, and its generalizability in complex underwater situations was not shown. Similarly, Li et al. (Li et al., 2023) proposed a multifeature fusion model-based segmentation method for fish images in aquaculture environments. In this method, the threshold value is first redefined by using the minimum Euclidean distance between the peaks of the original image, and the thresholded image is fused with the original image to augment the fish features in the underwater image; then, a multiscale attention module is proposed for the fusion of the different scale features to obtain the final prediction. Such methods tend to depend on data preprocessing, and the datasets used are small. These methods are thus limited and cannot be effectively applied to real marine environments. Because the attention mechanism is simple and effective, in some works, the attention mechanism from different dimensions is used in underwater scenes to improve the generalizability of the model, quickly locate blurred fish in complex scenes, and ignore background interference. To address the inability of general semantic segmentation methods to accurately segment fish objects in underwater images, Zhang (Zhang et al., 2022) and others designed a novel dual-pooling aggregated attention mechanism, which utilizes maximum pooling and average pooling to address and aggregate the target information from the spatial and channel dimensions while remaining lightweight. To distinguish foreground underwater targets from cluttered low-contrast backgrounds, Kim (Kim and Park, 2022) proposed a parallel semantic segmentation model that simultaneously segments the foreground fish and background of a complex underwater scene. This method utilized different attentional attention and loss functions for the foreground and background, and achieved better segmentation results than previous methods. To handle underwater foreground targets with different scales, Chen (Chen et al., 2022) first used style adaptation to enhance underwater images and then applied multiscale attention to fuse the information of different types of features, effectively improving the segmentation accuracy of small fish in complex underwater scenes.

Considering the unclear fish features caused by turbid water, insufficient light, camouflage, and other conditions in marine environments, in recent work, motion optical flow has been used to help recover enhanced underwater degraded fish appearance features. Salman et al. (Salman et al., 2020) combined the segmentation results of optical flow with the Gaussian Mixture Model (GMM), and used the predicted pixels as the input of the Convolutional Neural Network (CNN) model to complement each other, and obtained higher segmentation accuracy. Saleh et al. (Saleh et al., 2022) constructed an unsupervised wild fish tracking and segmentation method by combining optical flow, background subtraction, and unsupervised refinement networks with a pseudolabel generation method to achieve more accurate predictions when training a self-supervised segmentation depth model. Zhang et al. (Zhang et al., 2023) proposed multisource guidance segmentation method for fish in underwater video, where the final predictions are obtained using a multiple mutual attention guidance module with a feature adaptive method that fuses the preprocessed optical flow motion information with the RGB appearance information. Ye et al. (Ye et al., 2024) used the Gunnar Farneback optical flow method to quantitatively express the motion characteristics of fish, combined with the surface characteristics of fish, and used convolutional neural networks with different depths to extract the depth feature information of optical flow images to improve the accuracy of high-density fish segmentation in industrial aquaculture. However, the quality of underwater optical flow is often variable and difficult to control. For example, in Figure 2, the images in the first row exhibit less background motion, the scene is relatively static, and the motion of the fish is more significant, whereas the floating seaweed in the second row of the scene exhibit a large amount of motion interference from nonfish objects in the optical flow map, at which time the fish motion information can easily be drowned out by noise. The above method directly fuses the appearance information with the motion information and thus overly relies on the underwater optical flow information, which may result in

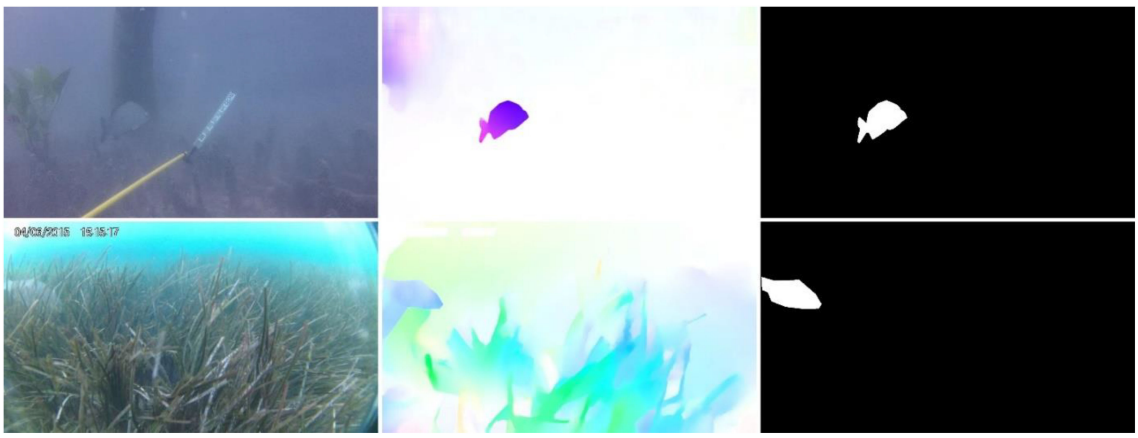

FIGURE 2
Underwater motion data of different qualities. From left to right are the video frames, optical flow, and ground truth.

incorrect segmentation when the network contains low-quality motion information. To solve this problem, in this paper, by rethinking the contribution of optical flow to underwater fish segmentation, we innovatively propose a robust underwater fish segmentation method with adaptive input filtering and fusion, which effectively overcomes the limitations of underwater segmentation models that rely on optical flow. Specifically, the main contributions of this paper are as follows:

(1) To avoid overly relying on optical flow information, which leads to the inflow of low-quality optical flow information into the network and results in final segmentation failure, the robust underwater segmentation network RUSNet is proposed in this paper. A lightweight optical flow quality evaluation module is proposed in the model decoder stage. By channel division and normalization of cross-modal stitching features, the two-dimensional pixel-level global confidence of optical flow features is obtained. This confidence is used to evaluate and correct input motion information.

(2) To segment ambiguous fish and camouflaged fish more accurately, inspired by the human visual perception system for localizing the target before focusing on the details, a decoder structure containing multidimensional attention (MDA) for step-by-step recovery from coarse to fine segmentation is proposed. The rough localization map is obtained through dense pyramid pooling, which is used to capture the coarse map of fish objects, and then this map is used as a guide to iteratively optimize the fish object features in complex underwater scenes from three perspectives—spatial perception, channel separation, and edge detail restoration—to obtain more accurate predictions.

(3) By analyzing the data and rethinking the contribution of optical flow in underwater fish segmentation, an output selective fusion method is designed in the test stage, in which the final underwater fish predictions are obtained by comparing the MAEs of single video frame prediction, optical flow prediction, and multisource input prediction and identifying the modal information with high credibility to be fused with multisource input prediction. The results on two publicly available underwater video datasets and two underwater camouflage datasets show that the proposed robust underwater segmentation network, RUSNet, outperforms other state-of-the-art underwater target segmentation models in terms of mean pixel accuracy and mean intersection over union metrics and achieves excellent performance in terms of model generalizability and robustness.

# 2 Materials and methods

## 2.1 Data preparation

To validate the segmentation effectiveness of the proposed model in complex underwater scenes, four publicly available underwater datasets are used for training and validation: DeepFish (Saleh et al., 2020), Seagrass (Ditria et al., 2021), and MoCA-Mask (Cheng et al., 2022). The DeepFish dataset contains approximately 40K images of fish from 20 different habitats in remote coastal marine environments of tropical Australia. The underwater video fish images were captured using a high-definition digital camera, and these images were divided into three subsets: counting, segmentation, and classification. The segmentation subset contains video clips from 13 different underwater environments, totaling 310 video frames, with a resolution of 1920 × 1080, and there are more scenes of a single fish than of multiple fish. The Seagrass dataset was collected from two estuarine systems in southeastern Queensland, Australia. The raw data were collected using an underwater camera and includes video clips from 18 underwater fishes, totaling 4,280 video frames. The image resolution is 1920 × 1080, and most video clips containing multiple fish objects. Considering that the segmentation set of the DeepFish dataset has only 310 images, to maximize the generalizability of the model, in this paper, the same approach as Zhang et al. (Zhang et al., 2023) is adopted, in which the DeepFish dataset and the Seagrass dataset are jointly trained and tested, and the optical flow data corresponding to video frames are extracted. From the joint dataset, 12 video clips with 3107 images are selected for training, 8 video clips with 693 images are selected for validation, and 11 video clips with 609 images are selected for testing. MoCA-Mask is a subset of the camouflage video dataset MoCA (Lamdouar et al., 2020), from which 32 underwater camouflaged organism video clips are selected in this paper; these videos contain devil scorpion fish, flatfish, and other underwater camouflaged creatures, and a total of 1539 frames are used for testing. In the above datasets, DeepFish and Seagrass have more severe brightness attenuation and water turbidity, as well as complex backgrounds, such as floating dense seaweeds. MoCA-Mask, on the other hand, contains complex scenes of underwater camouflage. These datasets are thus challenging.

## 2.2 RUSNet

The presence of turbidity and low-light conditions in underwater environments is the primarily reason for the low accuracy of video fish segmentation. Earlier work prioritized using methods such as manually selecting features to help recover unremarkable fish object features. However, underwater motion information is often overlooked in video data. From a biological perception perspective, moving targets are more likely to attract attention. This movement thus helps with locating dynamic fish in underwater scenes and improving the effectiveness of fish segmentation in underwater videos. The work of Zhang (Zhang et al., 2023) and Saleh (Saleh et al., 2022) also illustrated that the application of underwater motion information will greatly enhance fish segmentation in complex underwater videos. However, as mentioned in the previous section, underwater optical flow does not positively impact segmentation in all cases. For example, the motion of backgrounds such as floating seaweeds and the haloing of optical flow caused by underwater illumination variations can easily interfere with the motion information of the fish, which can lead to

segmentation failures. In previous work (Zhang et al., 2023), the quality of underwater optical flow information was not effectively assessed; rather, this work aimed to directly reduce the interference of low-quality optical flow on the segmentation model through the cross-modal feature fusion of appearance and motion, which could not effectively improve the robustness of the model to low-quality optical flow input and would lead cause the fish segmentation model to depend on the optical flow for underwater video. The segmentation accuracy loss was thus serious when the optical flow was unavailable or of poor quality. Therefore, evaluating the quality of underwater optical flow and performing effective feature selection and fusion while still maintaining a certain robustness is challenging when the optical flow input is not ideal.

To avoid the overreliance of complex underwater fish segmentation models on optical flow information and prevent interference of low-quality optical flow on the segmentation accuracy, the robust underwater segmentation network RUSNet is proposed in this paper. As shown in Figure 3, RUSNet contains three key parts. First, to effectively utilize underwater motion optical flow information, a lightweight global optical flow quality evaluation module, called flow calibration (FC), is proposed for screening and correcting motion information. Second, a decoder structure with step-by-step recovery from coarse to fine segmentation is designed. This structure is guided by a rough localization map and combined with multidimensional attention to iteratively optimize the spatial localization and edge details. Finally, a multioutput adaptive fusion method is proposed to determine which modal information RGB or optical flow, is more reliable by comparing the MAE output of single RGB information and cross-modal branch information and the MAE output of single optical flow information and cross-modal branch information. The more reliable information is fused with the cross-modal branch prediction to obtain the final prediction of complex underwater fish.

## 2.3 Flow calibration module

Introducing motion information in complex underwater scenes is beneficial for localizing underwater fish and reducing segmentation interference caused by blurring underwater. However, as mentioned in the introduction, underwater optical flow does not have a positive impact in all scenes, and when the scene does not contain a foreground target or drastic lighting changes, the quality of the motion optical flow decreases dramatically, which ultimately leads to model segmentation failure. To ensure the inflow of low-quality motion information does not affect the segmentation accuracy, a lightweight global optical flow calibration module (FC) to evaluate and correct the optical flow features is proposed in this paper. As shown in Figure 4, the visual feature Fa and the motion feature Fm are first extracted through two backbone networks with shared weights. Considering that the previous self-attention-based optical flow alignment module (Zhang et al., 2023; Zhou et al., 2020) is computationally large, in this paper, Fa and Fm are directly concatenated along the channel. This channel is used to align the RGB information and the motion information and to perform implicit interactions, which helps capture the relative relationship between multimodal features. Subsequently, a 3×3 convolution operation and a channel split operation are performed on the concatenated features to extract the valid information in the fused features and separate those features. Finally, sigmoid normalization is performed to obtain the global optical flow quality assessment score Gi. The process can be represented by Equation 1:

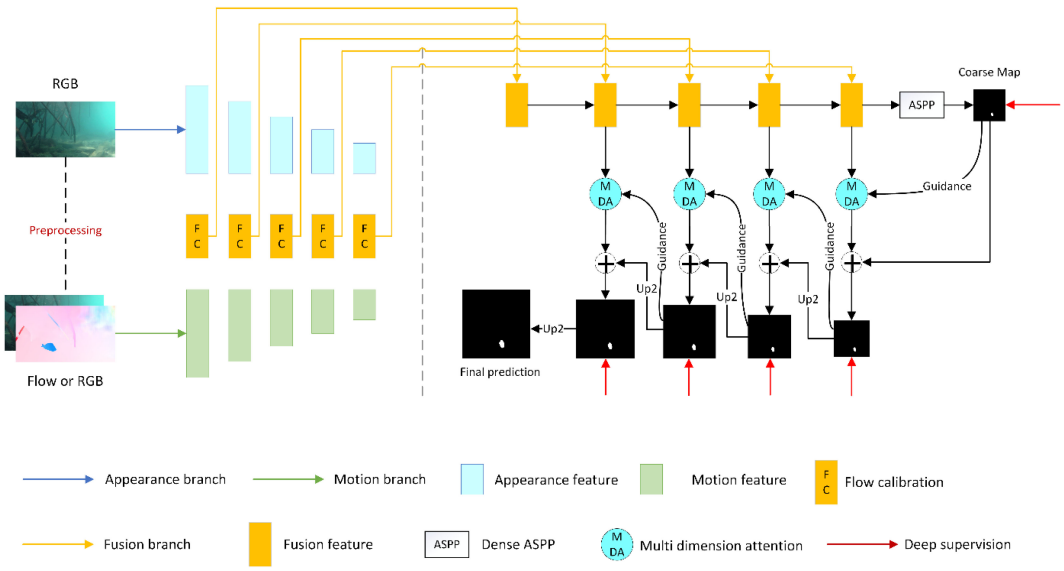


FIGURE 3
RUSNet network. On the left is the encoder, used to extract features, and on the right is the decoder, used to fuse features and generate predictions.

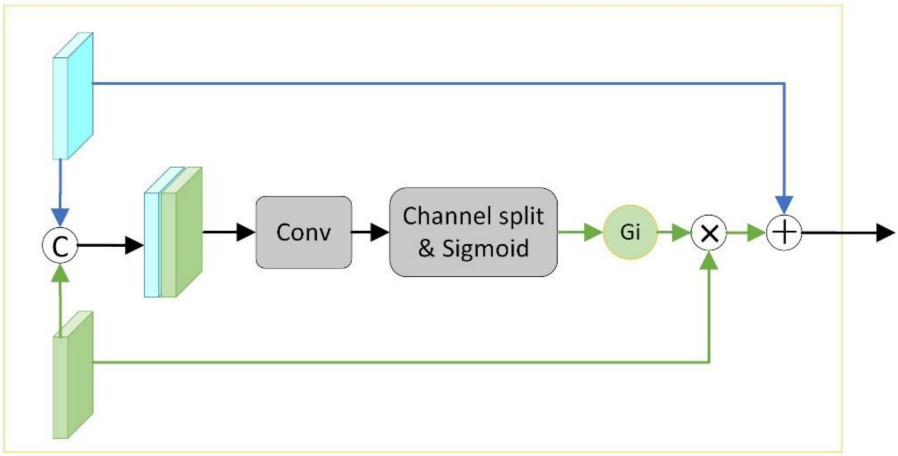


FIGURE 4
Global optical flow calibration module (FC).

$$Gi = \sigma(Split(Conv(Cat(F_i^a, F_i^m)))) \quad (1)$$

In Equation 1, σ denotes the sigmoid normalization function, Split denotes the channel squeeze and split operation, Conv denotes the 3×3 convolution, and Cat denotes the concatenation along the channel. The final fused feature Fi contains both the original appearance and corrected motion information and is computed as follows:

$$Fi = F_i^a + Gi \otimes F_i^m \quad (2)$$

In previous methods for RGB-D quality assessment (Zhang et al., 2021) or optical flow quality assessment (Yang et al., 2021), the fused features are typically pooled and then normalized to obtain a one-dimensional weight. This weight is used as a confidence score for the image-level optical flow information to globally deflate the three-dimensional optical flow features; however, the quality of each localized optical flow of a single image may vary. Furthermore, the one-dimensional weights represent the global feature weights, but this approach may be limited. As a result, instead of calculating the image-level confidence, we remove the pooling layer. In our proposed method, a pixel-level quality assessment is used to correct the global optical flow by two-dimensional weights to ensure that the motion information flowing into the network is relatively reliable.

## 2.4 Coarse-to-fine decoder

The traditional FPN (Lin et al., 2017) decoder structure has achieved relatively strong performance when handling the fusion of multiscale features by using a multiscale pyramid. However, for complex scenes, the FPN cannot effectively use the decoder features of the previous layer to guide the iterative recovery of the subsequent features. Considering water turbidity and camouflage, which result in a lack of detailed information, such as edges in the upsampling recovery of features, a decoder structure that recovers the segmentation step-by-step from coarse to fine is applied in this paper. Coarse-to-fine recovery follows the law of human visual perception, as the rough position of the target is located before detailed observation is conducted. In this paper, we first perform a DenseAspp (Yang et al., 2018) computation on the highest-level fusion feature $F^5$ to obtain the coarse map of the underwater fish at the high-level feature and output the coarse localization map $C_{map}$ as the bootstrap recovery feature for the subsequent decoder. To consider both global information and local details in feature upsampling recovery and to facilitate the perception of fish objects in complex underwater scenes, we design a multidimensional attention module (MDA) in a coarse-to-fine decoder structure. As shown in Figure 5, the MDA consists of two main modules, the spatial channel attention module CBAM (Woo et al., 2018) and the reverse attention module RA. From the three dimensions of spatial localization, channel recovery, and edge guidance, the output of the former stage is used to globally and locally guide the decoder features

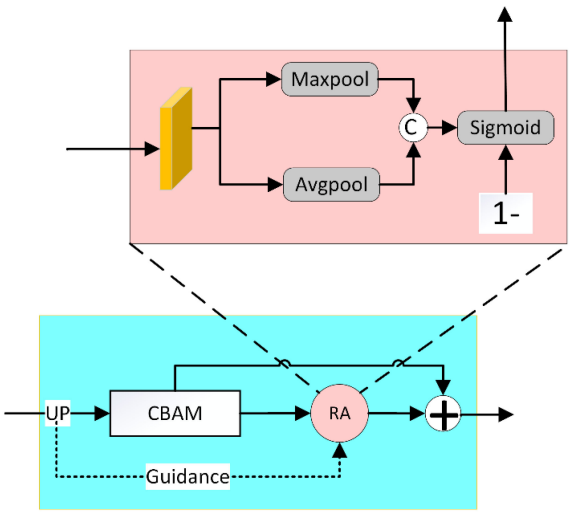


FIGURE 5
Multidimensional attention module (MDA).

of the latter stage to enhance the location and edge details of the underwater fish. The MDA computation process consists of two parts, CBAM and reverse edge attention (RA), in which the *RA* is computed as follows:

$$RA = E - \sigma\left(Cat(Max(CBAM(D_i)),\ Avg(CBAM(D_i)))\right) \quad (3)$$

where $E$ denotes the unit matrix, $\sigma$ denotes the sigmoid normalization function, *Cat* denotes concatenation along the channel, *Max* denotes the maximum pooling operation, *Avg* denotes the average pooling operation, and *CBAM* denotes the channel spatial attention. $D_i$ denotes the decoder feature of the current layer, and the specific computation process for the final decoder feature $D_i$ is as follows:

$$D_i = \begin{cases} CBAM(RA(DenseAspp(F_5))) & i = 4 \\ CBAM(RA((D_{i-1}))) & i \in \{3,2,1\} \end{cases} \quad (4)$$

The DenseAspp-processed high-level features contain preliminary target location information. However, edge details are missing. The ablation results show that guided detail recovery through rough localization maps is very effective in underwater scenes. Moreover, designing multidimensional attention to avoid overly focusing on details and neglecting global information is another issue worth discussing. This topic will be further explored in ablation tests.

## 2.5 Output selective fusion method

In the testing stage, to obtain more accurate binary segmentation masks for underwater fish, a multioutput selective fusion method is designed in this paper. Low-quality motion optical flows usually contain more noise than valid information. To avoid introducing too much interfering information during segmentation, the multioutput selective fusion method is combined with a global optical flow quality assessment module to ensure that the motion information selected by the network is as reliable as possible.

As shown in Figure 1, the direction and extent of foreground target motion are more pronounced in high-quality optical flow maps than in the background and are reflected in optical flow maps because moving fish tend to have clear, continuous, and distinct boundaries. Low-quality optical flow maps, on the other hand, usually contain more background motion, such as cluttered seagrasses, drastically changing lighting, or blurred fish boundaries. To effectively incorporate high-quality fish optical flow and avoid using low-quality motion information or introducing additional background motion interference, in this paper, the optical flow quality is evaluated in the feature extraction stage, and a selective fusion output method is introduced in the test output stage. First, based on the selected input branches, three types of prediction results, $P_r$, $P_m$, and $P_f$, are obtained using the model. These results are calculated as follows:

$$P_r = RUSNet(RGB、\ RGB) \quad (5)$$

$$P_m = RUSNet(Motion、\ Motion) \quad (6)$$

$$P_f = RUSNet(RGB、\ Motion) \quad (7)$$

After obtaining these three prediction types in the testing stage, the mean absolute error (MAE) (Perazzi et al., 2012) of $P_r$, $P_m$ and $P_f$ can be compared to determine which modality contributes more to the final segmentation results. Specifically, if the MAEs of $P_r$ and $P_f$ are smaller than those of $P_m$ and $P_f$, the prediction of $P_r$ is more similar to that of $P_f$, and the RGB information contributes more to the results. Unlike segmentation in nonunderwater scenes, the targets in the RGB inputs are usually not significant enough in underwater scenes, and thus, both appearance and motion information may interfere with the final prediction. Therefore, unlike in (Fan et al., 2020b), a manual threshold is not set for the MAE; only the similarity between the optical flow branch and the fusion branch are compared, and the MAEs predicted by two single-modal branches and the fusion modal branch are calculated simultaneously to mitigate the poor generalization of the model caused by the manual threshold. Finally, the selective output of the test phase is calculated as follows:

$$P_c = \begin{cases} P_r, MAE(P_r, P_f) \le MAE(P_m, P_f) \\ P_m, MAE(P_r, P_f) > MAE(P_m, P_f) \end{cases} \quad (8)$$

$$Output = \alpha \cdot P_f + (1 - \alpha) \cdot P_c \quad (9)$$

In the above formula, $P_c$ denotes the candidate features, *MAE* denotes the mean absolute error, and $\alpha$ denotes the experimental setup of the hyperparameters, which is set to the best value of 0.9; this value is determined through verification. Therefore, the final prediction result contains the segmentation results of the fusion branch and the segmentation results of the candidate branch, effectively improving the model's robustness under different quality inputs.

## 2.6 Loss function

The prediction result of video frame $t$ at different decoder stages is $P_t^i$, where i∈{1,2,3,4,5}. The standard binary cross-entropy loss function $L_{bce}$ is used to measure the difference between the prediction $P_t$ and ground truth $G_t$, where $L_{bce}$ is as follows:

$$L_{bce}(Pt, Gt) = -\sum_{(x,y)}[Gt(x,y)\log(Pt(x,y)) + (1 - Gt(x,y)\log(1 - Pt(x,y)))] \quad (10)$$

where $(x,y)$ denotes the positional coordinates of the pixel in the video frame, and the overall loss $L_{total}$ of the final model is as follows:

$$Ltotal = \sum_{i=1}^{4} Lbce(UP(Pt^i), Gt) + Lbce(Pt^5, Gt) \quad (11)$$

*UP* denotes bilinear interpolation upsampling, which aims to dimensionally align the prediction result $P_t^i$ with the ground truth $G_t$. The loss of the prediction result for each stage of the decoder can be calculated to accurately control the learning of multiscale information at different stages and to facilitate the use of the supervised multidimension attention module MDA for more

accurately performing iterative feature optimization from the perspective of spatial localization and detail recovery.

## 2.7 Evaluation criteria

Fish segmentation in underwater videos is a binary semantic segmentation task in which the foreground pixel value is 255 and the background pixel value is 0. Therefore, two evaluation metrics commonly used in semantic segmentation, the mean pixel accuracy (*mPA*) and the mean intersection over union (*mIoU*), are used to evaluate the gap between the prediction results of the model and the ground truth. *mPA* denotes the mean of the number of correctly categorized pixels of all the classes as a proportion of the number of pixels of that class's predicted number of pixels, and *mIoU* denotes the average of the ratio of intersection and concatenation of the predicted values of pixels of all classes. Thede metrics are calculated as follows:

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{Pii}{\sum_{j=0}^{k}pij} \tag{12}$$

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{Pii}{\sum_{j=0}^{k}pij + \sum_{j=0}^{k}pji - pii} \tag{13}$$

where $k$ denotes the number of categories; in our work, $k$ is taken as 2. $P_{ii}$ denotes the number of pixels correctly predicted as fish, or true-positive *TP*. $P_{ij}$ denotes the number of fish category pixels incorrectly predicted as background or false-positive *FP*. $P_{ji}$ denotes the number of fish category pixels incorrectly predicted as background, false-negative *FN*. $P_{jj}$ denotes the number of pixels correctly predicted as background, true-negative *TN*.

# 3 Experiment and results

In this section, the experiment's relevant environment configuration and the training details are described. Then, ablation tests are conducted to validate the effectiveness of the model's components, including global optical flow calibration, coarse-to-fine decoder structures incorporating multidimensional attention, and output selective fusion methods. Then, the proposed method is compared with other state-of-the-art underwater object segmentation methods and state-of-the-art video object segmentation methods on two publicly available underwater video datasets to validate the model's effectiveness in fish segmentation tasks in underwater videos. Finally, further model robustness tests are conducted on the optical flow unprocessed DeepFish dataset to verify the robustness of the model in the presence of low-quality motion inputs.

## 3.1 Experimental platform and training parameters

The experiments were conducted using GPUs for accelerated training, with the following environment configuration: a GPU model of Geforce RTX3090 with a graphics memory size of 24 GB, a CPU model of Intel(R) Core(TM) i7-9700 CPU (3.00 GHz), a Python 3.8 interpreter, the PyCharm development platform with CUDA version 11.3, and the deep learning framework PyTorch 1.11.0.

In the optical flow data preparation stage, the same method as (Zhang et al., 2023) is adopted for the DeepFish dataset and Seagrass dataset. Considering the water turbidity and other factors, the dataset is first processed using simple presegmentation and label conversion. Optical flow extraction is then conducted by using the RAFT (Teed and Deng, 2020) network. The MoCa-Mask dataset is directly subjected to optical flow extraction. All datasets are divided into training, validation, and test sets at a ratio of 6:2:2. Like the state-of-the-art video target segmentation methods, after the computational speed and accuracy are balanced, MiT-b1 (Xie et al., 2021), which has been pretrained on ImageNet-1K (Deng et al., 2009), is used as the feature extractor for the two-branch network. To facilitate model processing, the RGB images and optical flow resolution used for training and testing are uniformly set to 384 × 384, and the input data are enhanced through horizontal flipping. Video frames and corresponding optical flow images were used as model inputs, 3000 iteration rounds were trained, batch sizes were set to 16, adaptive moment estimation (Adam) was used as the model optimizer, the learning rate was set to 1e-5, the learning rate was kept constant during the training phase, and all the BN layers were frozen to accelerate the model training.

## 3.2 Ablation test

To verify the effectiveness of the proposed global optical flow calibration module, the design of the coarse-to-fine decoder structure, and the multioutput selective fusion method, an ablation test is conducted on a joint dataset of DeepFish and Seagrass. The baseline in this test is an ordinary two-stream network. Specifically, the encoder stage of the model is a Mit-b1 network with shared weights, which is used to extract the input features of different modalities, and the decoder obtains the final prediction by concatenating the fish appearance features and motion features obtained from the encoder along the channel using a convolution with a kernel size of 3*3. The model proposed in this paper, RUSNet, is obtained by adding the designed global optical flow calibration module FC, the decoder structure including MDA, and the multioutput selective fusion

TABLE 1 Ablation test results for RUSNet.

| Method | FC | MDA | Select | mPA/% | mIoU/% |
|---|---|---|---|---|---|
| Base | | | | 91.86 | 86.89 |
| Base + FC | √ | | | 93.73 | 88.07 |
| Base + MDA | | √ | | **93.11** | 89.64 |
| Base + Select | | | √ | 91.32 | 87.63 |
| RUSNet | √ | √ | √ | 92.61 | **90.12** |

Bold values are the value that gives the best results under the evaluation metrics.

method to the baseline. The experimental results are shown in Table 1.

The test results show that compared to Baseline, the proposed three modules exhibit different enhancements. Introducing the decoder embedded with MDA yields more obvious enhancements than introducing the other modules, with the mPA and mIoU increasing by 1.25% and 2.75%, respectively, compared to those of the Baseline, Furthermore, the obtained mPA is greater than that of the full RUSNet module because the Seagrass dataset used contains more small targets. Due to the lack of strong target positioning ability and attention to detailed information, the additional global optical flow quality calibration module FC misidentifies small fish objects as background, resulting in incorrect pixel prediction and reducing the mPA. Moreover, using only the multioutput selective fusion method in the test stage results in the least significant improvement because the lack of reasonable optical flow quality assessment and target feature recovery ability in the training stage causes the initial prediction results to deviate, and the segmentation mask before fusion is not sufficiently accurate; therefore, improving the accuracy and the mean intersection over union of segmentation via multioutput fusion is difficult. By combining these three modules and methods, the proposed RUSNet achieves a more significant improvement, with mPA and mIoU improved by 0.75% and 3.23%, respectively, compared to Baseline. Considering that the decoder MDA better improves the performance, to further investigate whether the sequence of the use of channel spatial attention and edge attention in MDA impact the final prediction results and because segmenting underwater camouflaged fish usually relies on edge details, more profound ablation tests are conducted on MDA using the underwater camouflaged organism video dataset MoCA-Mask. As mentioned above, following the laws of human visual perception, localization is performed before recognition during fish segmentation. First, channel spatial perception attention is used to capture the coarse map of the fish, and reverse edge attention is then used to iteratively recover the edge details of the ambiguous object, whereas in the control group, the opposite order is used. As shown in Figure 6, when edge correction is performed before spatially aware recovery, due to the lack of global location information for guiding the prediction, the prediction results tend to focus excessively on the edges and ignore the spatial consistency within the target, causing a hollow phenomenon to appear within the prediction results of the fish object. The attention heatmap visualization results for reverse MDA also show that the attention within the fish exhibits more serious inhomogeneity at this time. In contrast, MDA's heat map visualization illustrates that focusing on space and channel first can be effective in improving the consistency of attention within fish objects. This result is consistent with the findings of other work on saliency detection (Fan et al., 2020a) that incorporate reverse edge attention, further validating the rationality of the module proposed in this paper.

## 3.3 Comparison with other underwater segmentation methods

To validate the effectiveness of the proposed model, we further compare it with the advanced video target segmentation models AMC-Ne$^t$ (Yang et al., 2021) and FSNet (Ji et al., 2021) as well as the underwater fish segmentation models MSGNet (Zhang et al., 2023) and WaterSNet (Chen et al., 2022) on the publicly available complex underwater video dataset DeepFish-Seagrass (Saleh et al., 2020) (Ditria et al., 2021). Considering the running time and computational cost and to facilitate a fair comparison, the experiments conducted in this paper do not include any postprocessing, such as CRF. The segmentation set of the DeepFish dataset has only 310 images; to improve the generalizability of the model and avoid overfitting, following the same training principle as the SOTA underwater video fish segmentation model MSGNet, DeepFish is compared with another complex underwater video public dataset, Seagrass, for joint training. The specific results are shown in Table 1. FSNet effectively facilitates the information interaction and enhancement of the RGB branch and the optical flow branch in the video target segmentation model through bidirectional cross-attention. However, the optical flow information is not evaluated and filtered during feature refinement. Instead, the optical flow branches are reused, which leads to unsatisfactory prediction results in the presence of low-quality optical flow. AMC-Net is used for to segment fish in underwater videos by constructing a set of coattention mechanisms for the joint evaluation of cross-modal features and incorporating the channel space attention mechanism into both the encoder and decoder stages. The mean pixel angle and mean intersection over union obtained using AMC-Net are substantially better than those obtained using FSNet. However, as analyzed in the previous section, using a one-dimensional scalar as the global quality evaluation score limits the optical flow quality assessment when the regional optical flow quality is inconsistent. WaterSNet effectively improves the segmentation accuracy of nonsalient and camouflaged fish through random style adaptation (RSA) and multiscale fusion of the input image. However, when the group size for RSA is set, the batch size is directly used as the number of blended images in the group, significantly decrease the robustness

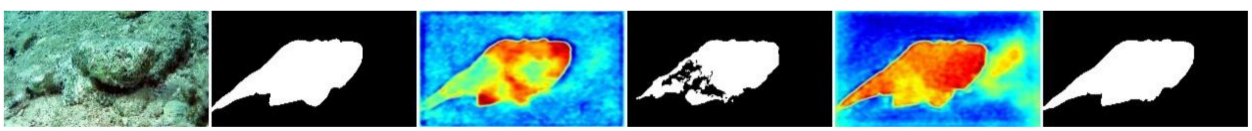

FIGURE 6
MDA ablation visualization. From left to right are the video frames, ground truth, reverse MDA attention heatmap, reverse MDA prediction, MDA heatmap, and MDA prediction.

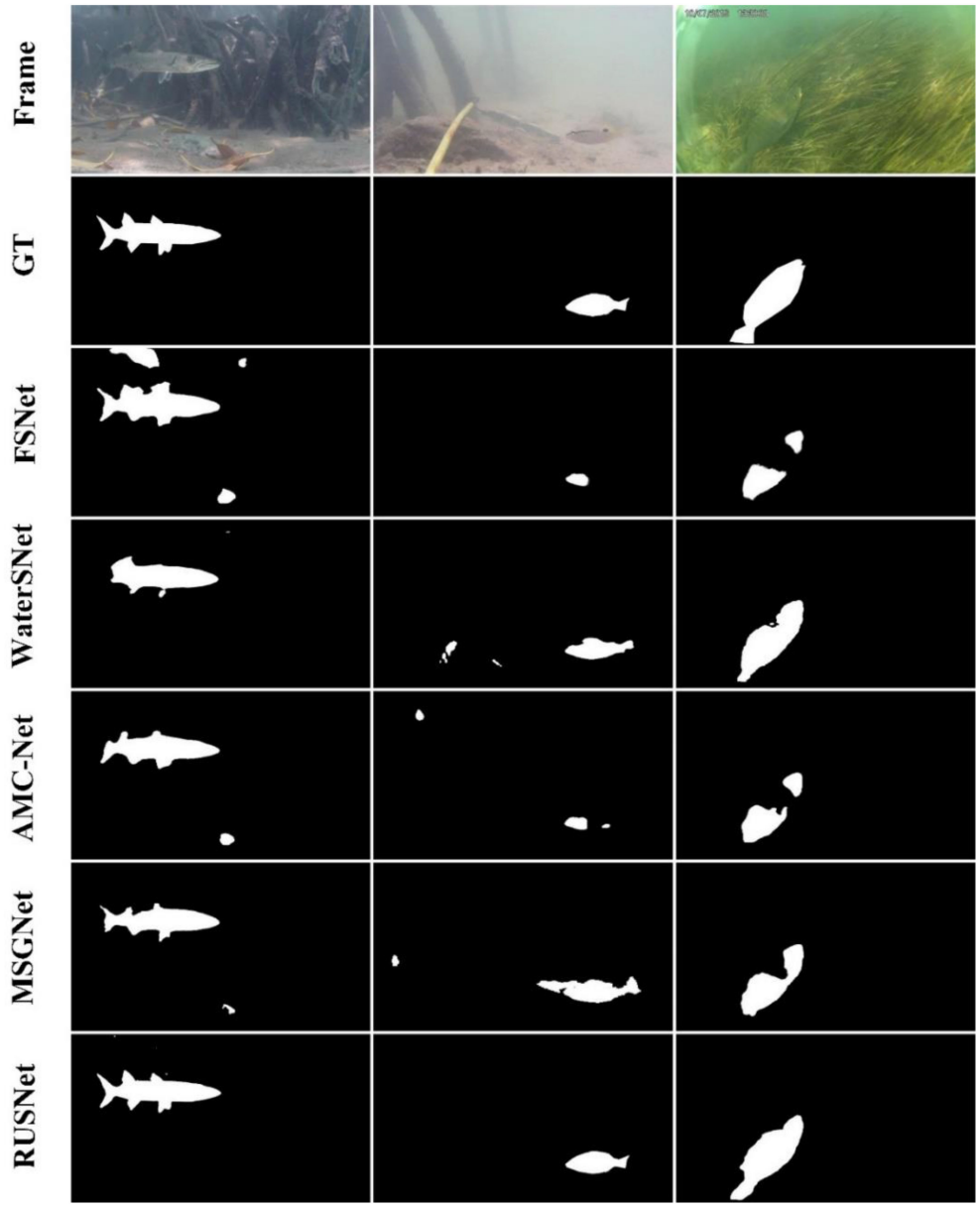


FIGURE 7
Prediction of different models on the DeepFish and Seagrass datasets.

of the model under hardware constraints. Finally, the mPA and mIoU of MSGNet are closest to those of the proposed model because of the use of a self-attention-based multiple interattention interaction mechanism and feature adaptive fusion. However, its dependence on the optical flow is too strong, causing the quality of the input optical flow to greatly impact the segmentation results. Furthermore, the segmentation robustness of MSGNet has room for improvement.

Figure 7 visualizes the experimental results of a real-world model that contains light in marine environments with changing light and complex backgrounds. The prediction results of RUSNet exhibit more complete targets and more accurate edge detail information than do those of other advanced models. RUSNet exhibits an improved robustness by avoiding the introduction of low-quality interfering features through the global quality assessment of the input motion information and the selective fusion of multiple outputs. Moreover, the coarse-to-fine decoder structure embedded with multidimensional attention ensures that the prediction results exhibit relatively complete targets and rich edge details. Compared with those of the advanced underwater fish segmentation model MSGNet, the mPA and mIoU of RUSNet are 0.72% and 1.21% higher, respectively, and compared with those of the advanced video target segmentation model AMC-Net, the mPA and mIoU of RUSNet are 1.06% and 2.51% higher, respectively. The effectiveness of the proposed model in segmenting fish objects in complex underwater scenes is thereby demonstrated. Table 2

TABLE 2 Comparison of the test results with those of advanced segmentation models.

| Model | Image | Flow | mPA/ % | mIoU/ % |
|---|---|---|---|---|
| FSNet (Ji et al., 2021) | √ | √ | 86.77 | 82.84 |
| WaterSNet (Chen et al., 2022) | √ | | 91.14 | 86.83 |
| AMC-Net (Yang et al., 2021) | √ | √ | 91.55 | 87.61 |
| MSGNet (Ye et al., 2024) | √ | √ | 91.89 | 88.91 |
| RUSNet | √ | √ | **92.61** | **90.12** |

Bold values are the value that gives the best results under the evaluation metrics.

To more fairly verify the robustness of the proposed model, the DeepFish (Saleh et al., 2020) dataset is selected for the experiment; this dataset has not undergone any preprocessing of the optical flow and is therefore susceptible to disturbances such as illumination variations, resulting in lower segmentation quality. According to the official segmentation strategy, the DeepFish segmented dataset contains a total of 620 images, and the background and fish objects in these images are labelled. 310 images are used for training, 124 images are used for validation, and 186 images are used for testing. In this paper, we use this approach to train and test the model and compare it with the state-of-the-art underwater fish segmentation model. As shown in Figure 8, there are two fish with different scales in the first row of the scene. However, the optical flow map does not show the motion of the small target on the left side because the target remains stationary for a period, and the advanced underwater multimodal fish segmentation model MSGNet ignores the information of the RGB video frames due to its strong dependence on the input optical flow, which ultimately results in missed detections. The proposed robust underwater segmentation model reduces the dependence on the input optical flow through global optical flow quality assessment and coarse-to-fine iterative recovery; in addition, this model is robust for temporary motionless object segmentation. The second row of Figure 8 shows the last frame of the video clip. Due to drastic scene changes, the optical flow map presents an irregular motion interference state, fuzzy video frames, and interference optical flow input. As a result, MSGNet is unable to segment effective targets from underwater fish, which can be predicted relatively accurately because RUSNet does not pay much attention to underwater optical flow information. The last row of Figure 8 shows the scene without targets when the optical flow is the change of light in the background environment. In this scene, MSGNet identifies a part of the background as a target. Table 3 shows the comparison results on the DeepFish dataset. The proposed segmentation model RUSNet outperforms the current SOTA model RMP-Net (Chen et al., 2022), especially in terms of the intersection union of fish objects. Therefore, RUSNet is more focused on learning fish semantic features and can segment temporary motionless targets while being robust to interference information such as background motion.

# 4 Discussion

## 4.1 Application

Fish habitat monitoring is an important step towards achieving sustainable fisheries, so we need to have access to important fish measurements such as size, shape and weight. These measurements can be used to judge the growth of the fish and serve as a reference for feeding, fishing and conservation (Ying et al., 2000). However, our work focuses on pixel-wise segmentation of fish in underwater videos. By accurately segmenting fish in underwater videos, information such as the length and contour of fish can be obtained, which helps human experts intuitively verify or estimate the size and weight of fish and facilitates the monitoring of fish. Laradji et al. (Laradji et al., 2021) and DeepFish's dataset providers (Saleh et al., 2020) also point out that pixel-wise segmentation is more helpful in assessing fish size and shape, and thus analyzing fish habitat. The segmentation method proposed in this paper can also be combined with tasks such as counting and tracking to integrate

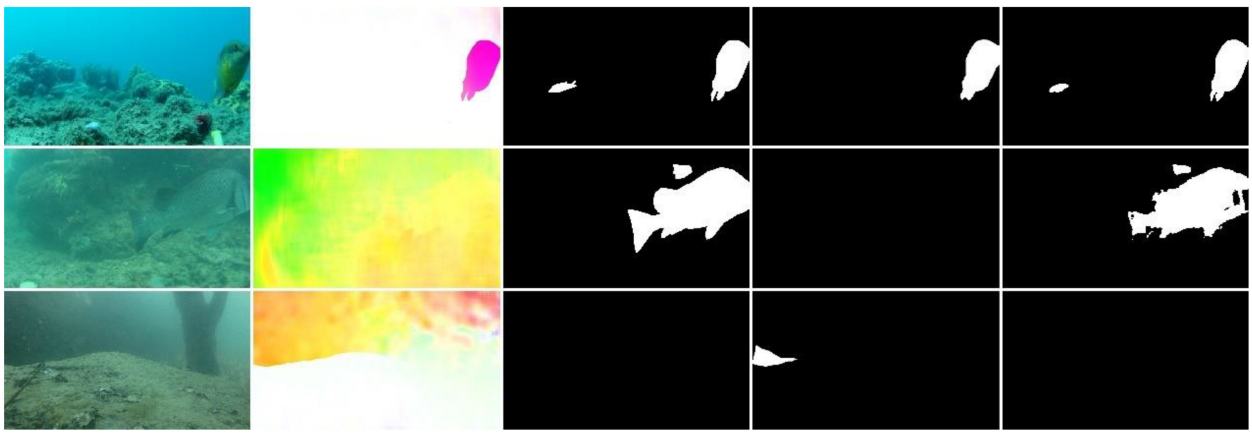

FIGURE 8
Prediction on the DeepFish dataset. From left to right are the video frames, optical flow, ground truth, MSGNet segmentation results, and RUSNet segmentation results.

TABLE 3 Comparison with other advanced underwater segmentation models on the DeepFish dataset.

| Method | Background IoU (%) | Fish IoU (%) | mIoU (%) |
|---|---|---|---|
| SUIM-Net (Islam et al., 2020) | 99.03 | 78.40 | 88.71 |
| DPANet (Zhang et al., 2022) | 99.31 | 82.86 | 91.08 |
| MFAS-Net (Haider et al., 2022) | 99.15 | 84.86 | 92.01 |
| MSGNet (Zhang et al., 2023) | 99.65 | 85.49 | 92.57 |
| RMP-Net (Chen et al., 2022) | 99.61 | 90.9 | 95.26 |
| RUSNet | **99.78** | **95.52** | **97.65** |

Bold values are the value that gives the best results under the evaluation metrics.

into a system that automatically performs comprehensive monitoring to improve efficiency and reduce labor costs. We hope that our work will inspire relevant researchers and continue to contribute to fish habitat monitoring and sustainable fisheries.

## 4.2 Future work

Real marine environments typically include complex conditions such as turbid water and cluttered backgrounds, and locating and segmenting fish by RGB visual information alone is difficult in these environment. In some special environments, such as underwater and camouflaged environments, optical flow, RGB-D, RGB-T, and other information can be combined to understand the visual scene more effectively. However, as mentioned above, the quality of optical flow introduced in real underwater scenes varies, and this flow does not always have a positive impact, so quality assessment of information input from multiple sources is necessary. The quality assessment of additional information such as optical flow is usually categorized into three types depending on how much information flows into the network. The first type is learnable soft weights, which are used to control the proportion of multisource information flowing into the network by obtaining the confidence (Zhang et al., 2021) or common attention score (Yang et al., 2021) of the multisource information through a specific quality assessment module, e.g., edge information learning. This type of method usually incurs less computational overhead, but its confidence learning process is not stable enough and is easily affected by model training and dataset distribution, leading to unreliable evaluation results. In the second category, additional new information flows into the network through information alignment and correction (Wang et al., 2021). This type of method is adequate in terms of theoretical analysis, but the actual process is cumbersome, and manual features may be introduced, resulting in certain limitations. The last method is hard gating-based information selection, in which all the information from multiple sources is selected or discarded through 0–1 gates (Fan et al., 2020b). This type of processing method usually requires a threshold to be manually selected for hard gating, resulting in a lack of generalizability. Compared with these three method types, the proposed global optical flow quality assessment module achieves cross-modal feature alignment through simple channel splicing and separation, and its computational cost is relatively small, which can satisfy the lightweight requirement. Moreover, instead of using a pooling operation to obtain a one-dimensional scalar to measure the optical flow quality, we use pixel-level normalization, which fully considers the global characteristics of the optical flow motion information and avoids uncontrollable correction bias. In addition, in visual tasks that require high-edge details such as processing camouflage, researchers usually design different reverse attentions for focusing on target edges. Reverse attention is a lightweight method that effectively guides the network to learn hard pixels in difficult edge regions through a simple take-inverse subtraction operation. The underwater data used in this paper also include camouflaged scenarios and other difficult scenarios. Following the human process of capturing hidden targets, a multidimensional attention guidance module is proposed that first utilizes DenseASPP and CBAM channel spatial attention to localize underwater fuzzy and camouflaged fish. This module is then combined with reverse attention RA to capture target edge details. In a holistic attention module, target features of complex underwater scenes are simultaneously and iteratively optimized from three perspectives—spatial localization, channel recovery, and edge refinement—and finally, underwater fish are effectively segmented. Experimental results on public datasets show that the segmentation accuracy and robustness of underwater video fish can be effectively improved through assessing the quality of the input optical flow and multidimensional feature recovery.

In terms of deep learning-based visual perception tasks, in April 2023, Meta AI announced a larger and more powerful segment anything model, SAM (Kirillov et al., 2023), which prompted the further development of computer vision tasks through the debugging of large amounts of data and through prompt engineering and inspired many researchers to consider the following question: how can we achieve an open and unified vision model through a unified structural paradigm or engineering for open scene visual perception or open task learning? This is not only a problem for academics but also an issue that needs to be solved urgently for marine fishery resource monitoring. For example, in real marine environments, how can a unified visual model be designed that can simultaneously segment stationary fish and moving fish or even automatically recognize scenes without fish objects? Some recent works (Zhao et al., 2023) (Cho et al., 2023) have attempted to address this issue, but considering the environmental specificity of underwater operations and hardware limitations, these methods can still be improved. For this reason, in this paper, a preliminary attempt to design a low-coupling model structure is also made by separating the encoders and decoders, which is convenient for subsequent expansion and improvement. To verify the effectiveness of the proposed model in segmenting fish in static complex underwater scenes, RGB video frames are used as the input of the second branch instead of optical flow maps. However, the experimental results show that although the rough location of the fish object can be localized, this method is still unsatisfactory for detailed recovery. Therefore, this study still has

limitations, and designing a unified segmentation model that can adaptively segment stationary and moving fish in an open marine scene is the main future research direction.

## 5 Conclusion

To address the strong dependence of fish segmentation models on motion optical flow in complex underwater environments, which yields predictions that are susceptible to the interference of low-quality motion information, a robust underwater fish segmentation method, RUSNet, that adaptively filters and fuses the input optical flow information is proposed by rethinking the contribution of the optical flow to the underwater fish segmentation task. First, a global optical flow quality evaluation module is designed to evaluate and correct the input optical flow information. Second, by embedding the proposed multidimensional attention module into the coarse-to-fine decoder structure, the features are guided iteratively from the spatial, channel, and edge dimensions to fully recover the spatial location and edge details. Finally, a multioutput selective fusion method is proposed in the testing phase. This method is used to determine which modal information contributes more to the final segmentation by comparing the mean absolute errors of the unimodal and cross-modal branch prediction results and predicting the fused output. The experimental results on public datasets show that the proposed method has high accuracy and robustness in complex underwater segmentation and can provide key information for the subsequent sustainability of marine fishery resources. However, this study has limitations. Considering that both moving fish and static fish may exist in an open underwater scene, designing a unified segmentation model that can adaptively segment stationary and moving fish in an open underwater scene is a major research direction for the future.

## Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

## Author contributions

PZ: Writing – review & editing, Writing – original draft. ZY: Writing – review & editing. HY: Writing – review & editing. WT: Writing – review & editing. CG: Writing – review & editing. YW: Writing – review & editing.

## Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work is supported by the Key R&D Projects in Liaoning Province (2023JH26/10200015), the Key Projects of Educational Department of Liaoning Province (LJKZ0729), National Natural Science Foundation of China (31972846), National Natural Science Foundation of China (62406052), Natural Science Foundation of Liaoning Province (2024-BS-214), Basic Research Funding Projects of Liaoning Provincial Department of Education (LJ212410158022) and the special fund for basic scientific research operations of undergraduate universities affiliated to Liaoning Province (2024JBQNZ011).

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note



## References

Banerjee, A., Bhattacharjee, D., Srinivasan, N. T., Behra, S., and Das, N. (2023). "SegFishHead: A Semantic Segmentation Approach for the identification of fish species in a Cluttered Environment," in *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3)*. 1–6 (New York, NY: IEEE).

Chatzievangelou, D., Thomsen, L., Doya, C., Purser, A., and Aguzzi, J. (2022). Transects in the deep: Opportunities with tele-operated resident seafloor robots. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2022.833617

Chen, R., Fu, Z., Huang, Y., Cheng, E., and Ding, X. (2022). "A robust object segmentation network for underwater scenes," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2629–2633 (New York, NY: IEEE).

Chen, J., Tang, J., Lin, S., Liang, W., Su, B., Yan, J., et al. (2022). RMP-Net: A structural reparameterization and subpixel super-resolution-based marine scene segmentation network. *Front. Mar. Sci.* 9, 1032287. doi: 10.3389/fmars.2022.1032287

Cheng, Z., Hong, G., Li, Q., Liu, S., Wang, S., and Ma, Y. (2023). Seasonal dynamics of coastal pollution migration in open waters with intensive marine ranching. *Mar. Environ. Res.* 190, 106101. doi: 10.1016/j.marenvres.2023.106101

Cheng, X., Xiong, H., Fan, D. P., Zhong, Y., Harandi, M., Drummond, T., et al. (2022). "Implicit motion handling for video camouflaged object detection," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. (New York, NY: IEEE Computer Society), 13854–13863. doi: 10.1109/CVPR52688.2022.01349

Cho, S., Lee, M., Lee, J., Cho, M., and Lee, S. (2023). Treating motion as option with output selection for unsupervised video object segmentation. *arXiv preprint arXiv* 2309, 14786. doi: 10.48550/arXiv.2309.14786

Chuang, M. C., Hwang, J. N., Williams, K., and Towler, R. (2011). "Automatic fish segmentation via double local thresholding for trawl-based underwater camera systems," in *//2011 18th IEEE International Conference on Image Processing*. 3145–3148 (New York, NY: IEEE).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. (New York, NY: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Ditria, E. M., Connolly, R. M., Jinks, E. L., and Lopez-Marcano, S. (2021). Annotated video footage for automated identification and counting of fish in unconstrained seagrass habitats. *Front. Mar. Sci.* 8. doi: 10.3389/fmars.2021.629485

Fan, D. P., Ji, G. P., Sun, G., Cheng, M. M., Shen, J., and Shao, L. (2020a). "Camouflaged object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, NY: IEEE), 2777–2787.

Fan, D. P., Lin, Z., Zhang, Z., Zhu, M., and Cheng, M. M. (2020b). "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," in *IEEE Transactions on neural networks and learning systems*, vol. 32. (New York, NY: IEEE), 2075–2089.

Food and Agriculture Organization of the United Nations (2022).The state of world fisheries and aquaculture 2022. In: *Towards blue transformation* (Rome: FAO). https://doi.org/ 10.4060/cc0461en (Accessed July 8, 2023).

Haider, A., Arsalan, M., Choi, J., Sultan, H., and Park, K. R. (2022). Robust segmentation of underwater fish based on multi-level feature accumulation. *Front. Mar. Sci.* 9, 1010565. doi: 10.3389/fmars.2022.1010565

Hall, M., Nordahl, O., Forsman, A., and Tibblin, P. (2023). Maternal size in perch (Perca fluviatilis) influences the capacity of offspring to cope with different temperatures. *Front. Mar. Sci.* 10. doi: 10.3389/fmars.2023.1175176

Islam, M. J., Edge, C., Xiao, Y., Luo, P., Mehtaz, M., Morse, C., et al. (2020). "Semantic segmentation of underwater imagery: dataset and benchmark," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. (New York, NY: IEEE), 1769–1776. doi: 10.1109/IROS45743.2020.9340821

Ji, G. P., Fu, K., Wu, Z., Fan, D. P., Shen, J., and Shao, L. (2021). "Full-duplex strategy for video object segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. (New York, NY: IEEE), 4902–4913. doi: 10.1109/ICCV48922.2021.00488

Kim, Y. H., and Park, K. R. (2022). PSS-net: Parallel semantic segmentation network for detecting marine animals in underwater scene. *Front. Mar. Sci.* 9, 1003568. doi: 10.3389/fmars.2022.1003568

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. *arXiv preprint arXiv* 2304, 02643. doi: 10.1109/ICCV51070.2023.00371

Lamdouar, H., Yang, C., Xie, W., and Zisserman, A. (2020). "Betrayed by motion: camouflaged object discovery via motion segmentation," in *Computer vision – ACCV 2020: 15th asian conference on computer vision(Kyoto, Japan: ACCV) 2020 november 30 – december 4. Revised selected papers, part II* (Springer-Verlag, Berlin, Heidelberg), 488–503. doi: 10.1007/978-3-030-69532-3_30

Laradji, I. H., Saleh, A., Rodriguez, P., Nowrouzezahrai, D., Azghadi, M. R., and Vazquez, D. (2021). Weakly supervised underwater fish segmentation using affinity LCFCN. *Sci. Rep.* 11, 17379. doi: 10.1038/s41598-021-96610-2

Li, D., Yang, Y., Zhao, S., and Yang, H. (2023). A fish image segmentation methodology in aquaculture environment based on multi-feature fusion model. *Mar. Environ. Res.* 190, 106085. doi: 10.1016/j.marenvres.2023.106085

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CPVR)*. (New York, NY: IEEE), 2117–2125.

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). *Saliency filters: Contrast based filtering for salient region detection[C]//2012 IEEE conference on computer vision and pattern recognition* (IEEE, 2012: 733-740).

Petrell, R. J., Shi, X., Ward, R. K., Naiberg, A., and Savage, C. R. (1997). Determining fish size and swimming speed in cages and tanks using simple video techniques. *Aquacultural Eng.* 16, 63–84. doi: 10.1016/S0144-8609(96)01014-X

Saleh, A., Laradji, I. H., Konovalov, D. A., Bradley, M., Vazquez, D., and Sheaves, M. (2020). A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* 10, 1–10. doi: 10.1038/s41598-020-71639-x

Saleh, A., Sheaves, M., Jerry, D., and Azghadi, M. R. (2022). Unsupervised fish trajectory tracking and segmentation. *arxiv preprint arxiv* 2208, 10662. doi: 10.48550/arXiv.2208.10662

Saleh, A., Sheaves, M., Jerry, D., and Azghadi, M. R. (2023). Applications of deep learning in fish habitat monitoring: A tutorial and survey. *Expert Syst. Appl.* 121841. doi: 10.1016/j.eswa.2023.121841

Salman, A., Siddiqui, S. A., Shafait, F., Mian, A., Shortis, M. R., Khurshid, K., et al. (2020). Automatic fish detection in underwater videos by a deep neural network-based hybrid motion learning system. *ICES J. Mar. Sci.* 77, 1295–1307. doi: 10.1093/icesjms/fsz025

Teed, Z., and Deng, J. (2020). "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, 2020 August 23–28, Vol. 2. 402–419 (Switzerland: Springer International Publishing).

Wang, X., Li, S., Chen, C., Hao, A., and Qin, H. (2021). Depth quality-aware selective saliency fusion for RGB-D image salient object detection. *Neurocomputing* 432, 44–56. doi: 10.1016/j.neucom.2020.12.071

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII 1*. (Switzerland: Springer International Publishing), 3–19. doi: 10.1007/978-3-030-01234-2_1

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* 34, 12077–12090. doi: 10.48550/arXiv.2105.15203

Yang, M., Yu, K., Zhang, C., Li, Z., and Yang, K. (2018). "Denseaspp for semantic segmentation in street scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (New York, NY: IEEE), 3684–3692.

Yang, S., Zhang, L., Qi, J., Lu, H., Wang, S., and Zhang, X. (2021). "Learning motion-appearance co-attention for zero-shot video object segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. (New York, NY: IEEE), 1544–1553. doi: 10.1109/ICCV48922.2021.00159

Yao, H., Duan, Q., Li, D., and Wang, J. (2013). An improved K-means clustering algorithm for fish image segmentation. *Math. Comput. Model.* 58, 790–798. doi: 10.1016/j.mcm.2012.12.025

Ye, Z., Zhou, J., Ji, B., Zhang, Y., Peng, Z., Ni, W., et al. (2024). Feature fusion of body surface and motion-based instance segmentation for high-density fish in industrial aquaculture. *Aquaculture Int.* 32, 1–21. doi: 10.1007/s10499-024-01569-2

Ying, Y., Rao, R. X., Zhao, Z. Y., and Jiang, J. Y. (2000). Application of machine vision technique to automatic quality identification of agricultural products (i). *Trans. Chin. Soc Agric. Eng.* 16, 103–108.

Zhang, W., Ji, G. P., Wang, Z., Fu, K., and Zhao, Q. (2021). "Depth quality-inspired feature manipulation for efficient RGB-D salient object detection," in *Proceedings of the 29th ACM international conference on multimedia*. (New York, NY: ACM).

Zhang, W., Wu, C., and Bao, Z. (2022). DPANet: Dual Pooling-aggregated Attention Network for fish segmentation. *IET Comput. Vision* 16, 67–82. doi: 10.1049/cvi2.12065

Zhang, P., Yu, H., Li, H., Zhang, X., Wei, S., Tu, W., et al. (2023). MSGNet: multi-source guidance network for fish segmentation in underwater videos. *Front. Mar. Sci.* 10, 1256594. doi: 10.3389/fmars.2023.1256594

Zhao, X., Chang, S., Pang, Y., Yang, J., Zhang, L., and Lu, H. (2023). Adaptive multi-source predictor for zero-shot video object segmentation. *arXiv preprint arXiv* 2303, 10383. doi: 10.1007/s11263-024-02024-8

Zhao, Y., Sun, Z. Y., Du, H., Bi, C. W., Meng, J., and Cheng, Y. (2022). A novel centerline extraction method for overlapping fish body length measurement in aquaculture images. *Aquacultural Eng.* 99, 102302. doi: 10.1016/j.aquaeng.2022.102302

Zhou, T., Li, J., Wang, S., Tao, R., and Shen, J. (2020). Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.* 29, 8326–8338. doi: 10.1109/TIP.83