frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 21 September 2023
DOI 10.3389/fmars.2023.1256594

OPEN ACCESS

EDITED BY
Haiyong Zheng,
Ocean University of China, China

REVIEWED BY
Huanda Lu,
Zhejiang University, China
Lars Christian Gansel,
Norwegian University of Science and Technology, Norway

*CORRESPONDENCE
Hong Yu
yuhong@dlou.edu.cn

RECEIVED 11 July 2023
ACCEPTED 04 September 2023
PUBLISHED 21 September 2023

CITATION
Zhang P, Yu H, Li H, Zhang X, Wei S, Tu W, Yang Z, Wu J and Lin Y (2023) MSGNet: multi-source guidance network for fish segmentation in underwater videos. *Front. Mar. Sci.* 10:1256594. doi: 10.3389/fmars.2023.1256594

MSGNet: multi-source guidance network for fish segmentation in underwater videos

Peng Zhang[1,2,3,4], Hong Yu[1,2,3,4]*, Haiqing Li[1,2,3,4], Xin Zhang[1,2,3,4], Sixue Wei[1,2,3,4], Wan Tu[1,2,3,4], Zongyi Yang[1,2,3,4], Junfeng Wu[1,2,3,4] and Yuanshan Lin[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 Ocean University, Dalian, China

Fish segmentation in underwater videos provides basic data for fish measurements, which is vital information that supports fish habitat monitoring and fishery resources survey. However, because of water turbidity and insufficient lighting, fish segmentation in underwater videos has low accuracy and poor robustness. Most previous work has utilized static fish appearance information while ignoring fish motion in underwater videos. Considering that motion contains more detail, this paper proposes a method that simultaneously combines appearance and motion information to guide fish segmentation in underwater videos. First, underwater videos are preprocessed to highlight fish in motion, and obtain high-quality underwater optical flow. Then, a multi-source guidance network (MSGNet) is presented to segment fish in complex underwater videos with degraded visual features. To enhance both fish appearance and motion information, a non-local-based multiple co-attention guidance module (M-CAGM) is applied in the encoder stage, in which the appearance and motion features from the intra-frame salient fish and the moving fish in video sequences are reciprocally enhanced. In addition, a feature adaptive fusion module (FAFM) is introduced in the decoder stage to avoid errors accumulated in the video sequences due to blurred fish or inaccurate optical flow. Experiments based on three publicly available datasets were designed to test the performance of the proposed model. The mean pixel accuracy (*mPA*) and mean intersection over union (*mIoU*) of MSGNet were 91.89% and 88.91% respectively with the mixed dataset. Compared with those of the advanced underwater fish segmentation and video object segmentation models, the *mPA* and *mIoU* of the proposed

model significantly improved. The results showed that MSGNet achieves excellent segmentation performance in complex underwater videos and can provide an effective segmentation solution for fisheries resource assessment and ocean observation. The proposed model and code are exposed via Github[1].

KEYWORDS

computer vision, underwater video processing, MSGNet, fish segmentation, optical flow, coattention

1 Introduction

With over three billion people relying on fish for at least 20% of their daily protein and more than 120 million directly employed in the fishing and aquaculture sectors (Food and Agriculture Organization of the United Nations, 2021), sustainable fisheries and fish habitat monitoring were a natural focus. It has been demonstrated that the shape, size, and body length of fish are essential for monitoring fish habitats (Laradji et al., 2021), which can reflect the long-term sustainable production capacity of populations (Hall et al., 2023). In marine fisheries, accurate information on the size and shape of wild and farmed fish populations can be obtained by segmentation-based methods (Muñoz-Benavent et al., 2022; Zhao et al., 2022), for example, by measuring the centerline of a fish segmentation mask. Such information is the basis of harvest management and is crucial for marine ranching farming, as it contributes to the effective management of feeding regimes, grading times, and ultimately the optimal harvesting time of fish (Beddow et al., 1996; Muñoz-Benavent et al., 2022), thus reducing management and production costs and promoting the sustainable development of marine fisheries. In the past, morphological characteristics of fish were usually obtained using manual methods, which may cause damage to the fish and is inefficient (Petrell et al., 1997). With the development of artificial intelligence, researchers are beginning to use deep learning to analyze the size and shape of fish. Object detection generally use rectangular box labels, which allow for contact-free fish detection, but only the positional information of the fish can be obtained, not details such as shape. By contrast, segmentation of underwater fish can provide more accurate and richer semantic-level details such as shape, size, and edges. As a result, segmentation methods for fish analysis have attracted increasing attention (Garcia et al., 2020).

However, because of water turbidity and insufficient lighting, visual information of underwater fish is not obvious. At the same time, reefs (Zhuang et al., 2020) and dense drifting seagrass (Ditria et al., 2021) constitute a complex underwater background, which interferes with fish segmentation. Furthermore, many marine animals have protective colors (Li et al., 2021b) and blend in with the environment, remaining camouflaged, making it difficult to distinguish from the background. These situations lead to the general semantic segmentation model (Costa et al., 2006; Huang et al., 2015) failing to achieve satisfying results. Thus, it is challenging to segment fish accurately in underwater videos.

To address the problem of water turbidity and insufficient light that results in indistinct features of underwater fish, Chuang et al. (2011) used histogram back-projection procedure to ensure fish segmentation accuracy with insufficient light. Shoffan et al (Shoffan, 2022). applied adaptive histogram equalization for preprocessing, followed by morphological processing using a K-means clustering algorithm and open-close operations to obtain fish contours. This approach has excellent performance for black-and-white scenes but is not practical for color images. To further improve the robustness of fish segmentation in turbid water, Haider et al. (2022) presented a robust segmentation model for underwater fish based on multi-level feature accumulation, which improved the segmentation of obscure fish by using an initial feature refinement and transfer block to refine potential information. Similarly, Zhang et al. (2022) employed dual pooling-aggregated attention with spatial and channel dimensions, greatly reducing the computational effort while providing better segmentation results for fuzzy fish, but the performance in complex scenarios is not known. Natural underwater environments, often with complex backgrounds such as seagrass and reefs, interfere with foreground object localization. To improve the performance of fish segmentation in complex underwater scenes, Kim et al (Kim and Park, 2022). proposed a parallel semantic segmentation network that utilizes model and loss to localize the foreground and background, respectively and achieves efficient detection of marine animals by learning their foreground and background regions separately. To simultaneously segment multiple types of objects, such as underwater fish, reefs, and people, Islam et al. (2020) proposed a fully convolutional underwater semantic segmentation model that uses skip connections between mirror composite layers, realized multi-class semantic segmentation for underwater. In addition, underwater organisms become camouflaged when their contrast with the background is further reduced, as in the case of flounder. Segmentation of camouflaged underwater organisms is challenging because the edges of the object are highly blended with the background, making even coarse localization extremely difficult. To solve the above problems, Li et al. (2021a) propose a novel enhanced cascade decoder network (ECD-Net). By using rich multi-scale features, accurate segmentation of marine animals in

1 https://github.com/zp19990818/MSGNet-main.

complex underwater environments was realized. Similarly, Chen et al. (2022) first used a Random Style Adaptation (RSA) module to enhance underwater images and then achieved accurate segmentation of underwater camouflaged organisms by utilizing multi-scale information with different sizes of receptive fields. However, the above studies used only static appearance information, even though some used video data. After observation and research (Lamdouar et al., 2020), we noticed that moving objects are more likely to attract attention in underwater videos, so our work considers using motion information to help segmentation. By interacting motion information with appearance information, it is possible to enhance inconspicuous underwater features and accurately segment fish in complex underwater scenes.

When monitoring wild fish habitat (Saleh et al., 2020), the data acquisition equipment, usually an underwater camera mounted on an ROV or boat, is in motion and thus need to be able to identify or segment fish when both the camera and the object may be in motion. We used the optical flow method when processing underwater fish movement information. Compared to standard moving object detection methods, such as the background difference method (Zivkovic and van der Heijden, 2006) the optical flow method does not require prior information on the scene. It can detect moving fish in underwater videos independent of camera motion. Benefits from the development of deep learning and improvement in hardware, advanced optical flow extraction models (Teed and Deng, 2020) can meet almost real-time requirements. Some current work also employs optical flow to segment fish in underwater videos. Salman et al. (2020) combined the segmentation results of an optical flow with Gaussian mixture models (GMM) to mutually complement the predicted pixels as input to a convolutional neural network (CNN) model and achieved a higher segment accuracy. However, it is difficult to extract the ideal optical flow due to complex underwater situations such as light changes and background motion. These works used motion information as a complementary method rather than directly optimizing it and thus did not fully consider the vital contribution of motion to segment fish in underwater videos. In contrast, we propose an optical flow data preprocessing scheme for obtaining high-quality underwater optical flows, which highlights the movement of the fish object through a simple overlay operation and significantly reduces the interference of the background movement. In addition, the optical flow method cannot detect stationary fish, so we propose a model to improve fish segmentation in underwater videos using static appearance and motion optical flow information.

Fish segmentation approaches for underwater videos can achieve stronger robustness and higher accuracy by combining both appearance and motion information. The appearance information provides the location of fish, and the motion information provides richer contour details. However, as shown in Figure 1, the appearance of fish becomes fuzzy due to insufficient lighting and water turbidity in underwater environments, and the segmentation result from video frames are incomplete. Figure 1 also illustrates that the additional introduced optical flow can be affected by factors such as lighting changes in the background, resulting in segmentation redundancy. Specially, different colors in Figure 1B indicate different motion directions, and the color depth represents the intensity of the motion. It is challenging to segment fish in complex underwater videos using appearance and motion information reasonably while simultaneously maintaining great generalization and robustness. Therefore, an effective multimodal feature interaction and fusion method is needed to selectively employ appearance and motion information. To address the above challenge, we propose a multi-source guidance network (MSGNet) for fish segmentation in underwater videos. MSGNet contains two key components: the multiple co-attention guidance module (M-CAGM) and the feature adaptive fusion module (FAFM). To address the problem of degradation of visual features of underwater fish due to water turbidity, we designed a multiple co-attention guidance module (M-CAGM) to cross-enhance the appearance and motion information of underwater fish by calculating the similarity between multimodal features and highlighting relatively salient fish in underwater videos. Meanwhile, considering that the optical flow map may still have interference information due to background motion or luminance variation, we employed a feature adaptive fusion module (FAFM) to filter and learn the fused features to avoid accumulating error information. Results on three publicly available datasets, DeepFish (Saleh et al., 2020), Seagrass (Ditria et al., 2021), and MoCA-Mask (Cheng et al., 2022) show that our proposed network effectively solved the low fish segmentation accuracy and poor robustness problems caused by insufficient lighting and water turbidity in underwater videos, and can even segment camouflaged fish. The main contributions of this paper are listed as follows:

1. A video data preprocessing method was designed to obtain underwater optical flow with precise edges. Moving fish in datasets were highlighted by overlaying results of the

FIGURE 1
Single type features and segmentation results. **(A)** the original frame, **(B)** preprocessed optical flow, **(C)** frame segmentation result, **(D)** optical flow segmentation result.

pretrained model and corresponding frames with different pixel values and contrasts.

2. To improve the accuracy of fish segmentation in underwater videos by reasonably using optical flow and video frame information, we proposed a multi-source guidance network MSGNet. By analyzing the characteristics of underwater videos, we found that fish tend to be obscure and in motion. Thus, a multiple co-attention guidance module (M-CAGM) was integrated in the encoder stage to focus more on intra-frame salient fish and moving fish in different frames through nonlocal-based co-attention. The appearance and motion features of fish were bidirectionally enhanced and optimized through M-CAGM.
3. Single type feature errors might accumulate in video sequence, leading to segmentation failure, a feature adaptive fusion module (FAFM) was designed in the decoder of the proposed model. The FAFM applies a mutual gate to filter and fuse the features by evaluating the contribution of different types of features to the final segmentation result, greatly improving the robustness of the proposed model.

The rest of the paper is structured as follows. Section 2 describes the proposed method, Section 3 shows and analyzes the experimental results, Section 4 discusses the model's superiority and potential applications in detail, and Section 5 draws a brief conclusion.

2 Materials and methods

2.1 Experimental dataset

In this work, we used three publicly available underwater video fish datasets, DeepFish (Saleh et al., 2020), Seagrass (Ditria et al., 2021) and MoCA-Mask (Cheng et al., 2022), all three datasets provide video frames and binary ground truth. The videos in the first two datasets suffer from severe water turbidity and insufficient lighting. The MoCA-Mask dataset contains many underwater camouflaged organisms that are highly integrated with their environment. Underwater water turbidity and insufficient lighting can cause fish objects to become blurred, similar to camouflage, so we also considered underwater video camouflage data. Specifically, the DeepFish dataset contains approximately 40k underwater fish frames from 20 different habitats in remote coastal marine environments in tropical Australia. These videos were captured with a high-definition digital camera and divided into three subsets: counting, segmentation, and classification. The segmentation subset contains 13 video clips of different underwater environments with 310 video frames at a resolution of 1920 × 1080 and includes more single fish scenes. The Seagrass dataset was collected in two estuary systems in southeastern Queensland, Australia. The raw data were obtained with submerged action cameras. There are 18 video clips of underwater fish with 4280 video frames. The resolution of the frames is 1920 × 1080, and most of the video clips contain multiple fish. We selected 31 video clips with 4409 images from the DeepFish dataset and Seagrass dataset at a ratio of 6:2:2 for training, validation, and testing. The MoCA-Mask dataset contains 87 videos of camouflaged animals from the MoCA (Lamdouar et al., 2020) dataset, with a total of 22,939 frames with pixel-level ground-truth. The resolution of most frames in MoCA-Mask dataset is 1280 × 720. For this experiment, we selected 32 underwater camouflaged animal video clips in MoCA-Mask dataset, including devil scorpion fish, flatfish, and other underwater camouflaged creatures, with a total of 1539 frames. These images were divided into a training set, validation set, and test set at a ratio of 6:2:2.

2.2 Optical flow data preprocessing

In this study, we used optical flow as motion data. Optical flow can be employed to detect independent motion objects without prior knowledge of the scene and obtain complete information on motion objects and is thus suitable in dynamic backgrounds. In underwater videos, insufficient lighting and water turbidity can easily affect fish appearance features, resulting in less critical information such as fish edges and textures. In contrast, optical flow with precise edges yields more detailed information, which can significantly compensate for the degradation of fish appearance features in underwater videos.

Previously, optical flow data were usually synthesized manually or obtained using special equipment such as light detection and ranging (LiDAR). There is no optical flow dataset specifically for fish in underwater videos, so obtaining high-quality optical flow data by preprocessing is necessary. Compared with land scenes, underwater scenes suffer from more light variations and different background motions, and the optical flow of fish in underwater videos usually has inaccurate boundaries. Therefore, generic optical flow extraction models cannot perform satisfactorily in underwater scenes.

Inspired by FlowNet (Dosovitskiy et al., 2015), an optical flow model, and unsupervised fish segmentation work (Saleh et al., 2022), we first used a fully convolutional network (FCN) (Long et al., 2015) segmentation model trained with the ImageNet (Deng et al., 2009) dataset and fine-tuned it with underwater images to obtain a coarse binary segmentation of video frames for optical flow extraction. Different from the background subtraction approach employed in previous unsupervised fish segmentation work (Saleh et al., 2022), to avoid the limitations of a fixed camera, we only used the pretrained model for preprocessing because it has better robustness. After obtaining the coarse segmentation masks O_x and O_{x+1} for the original video frames F_x and F_{x+1}, the masks are overlaid with the corresponding original video frames F_x and F_{x+1} to obtain the inputs M_x and M_{x+1} for the optical flow extraction model. Specifically, we set different three-channel pixel values of coarse segmentation masks and transparency of origin frames to highlight moving fish in underwater videos. We also designed an experiment to test different transparencies for optical flow preprocessing and found that it should be in an appropriate range. By simple overlaying, the model can focus more on foreground motion information and extract a fish optical flow with precise edges and

complete objects in underwater videos. Figure 2 shows some sample images from the two publicly available datasets and the results of optical flow preprocessing and proves the effectiveness of optical flow data preprocessing in this paper.

We applied recurrent all-pairs field transforms (RAFT) (Teed and Deng, 2020) as an optical flow estimation model, in which the similarity between any two points in different images is calculated by constructing a correlation matrix of consecutive frames, and a gated recurrent neural network was also designed for iterative optimization. RAFT can obtain optical flow with clearer edges even under fast movement, occlusion, etc. Compared with other optical flow extraction networks such as FlowNet (Dosovitskiy et al., 2015), RAFT has superior robustness and generalizability. In addition, considering that the correspondence of model inputs, the last frame of each video clip was removed to align the number of video frames and optical flow maps.

2.3 Fish segmentation in underwater videos

2.3.1 MSGNet

Insufficient lighting and water turbidity in underwater videos are the main reasons for poor fish segmentation. Previous works have usually focused on static pictures for underwater semantic segmentation, ignoring the motion information in dynamic video scenes. By visualizing the pixel points of moving objects and focusing more on the motion in a scene rather than the visual information, optical flow can prevent segmentation failure caused by the degradation of underwater visual features. This makes optical flow a preferable choice since it is not impacted by such degradation, which is a significant advantage. From the biological dimension, moving objects can draw more attention and provide more detailed information, thus breaking the low contrast with the background and becoming more obvious. The appearance details can be recovered from the motion information of underwater fish to obtain a completer and more accurate object. Meanwhile, prominent fish can be located with the appearance information, forcing the model to focus more on the motion of fish and ignore the interference of motion information caused by lighting changes and background motion. By employing multi-source information, it is possible to achieve higher accuracy when segmenting fish in underwater videos.However, the selection of appearance and motion features of fish in underwater videos need to be improved. From the appearance perspective, water turbidity and insufficient lighting lead to low contrast between fish and the background. In some scenes, fish even become camouflaged, so the visual features of fish underwater are not significant enough. From the motion perspective, underwater environments have drastic light changes because of light refraction and water surface fluctuations, resulting in vignetting of the extracted underwater optical flow, blurring the edge of fish. In addition, seagrasses and shadows of fish, which are non-fish motion present in the background, disturb the optical flow and capture additional background motion. Thus, it is challenging to effectively utilize the appearance and motion information of fish while suppressing the interference of both information.Thus, we designed MSGNet, a dual-stream network for fish segmentation in underwater videos with multi-source information guidance. For the convenience of variable control, the appearance and motion features of fish are first extracted using two ResNet-101 (He. et al., 2016) with shared weights. Then, the M-CAGM is designed in the encoder stage of MSGNet to enhance the insignificant

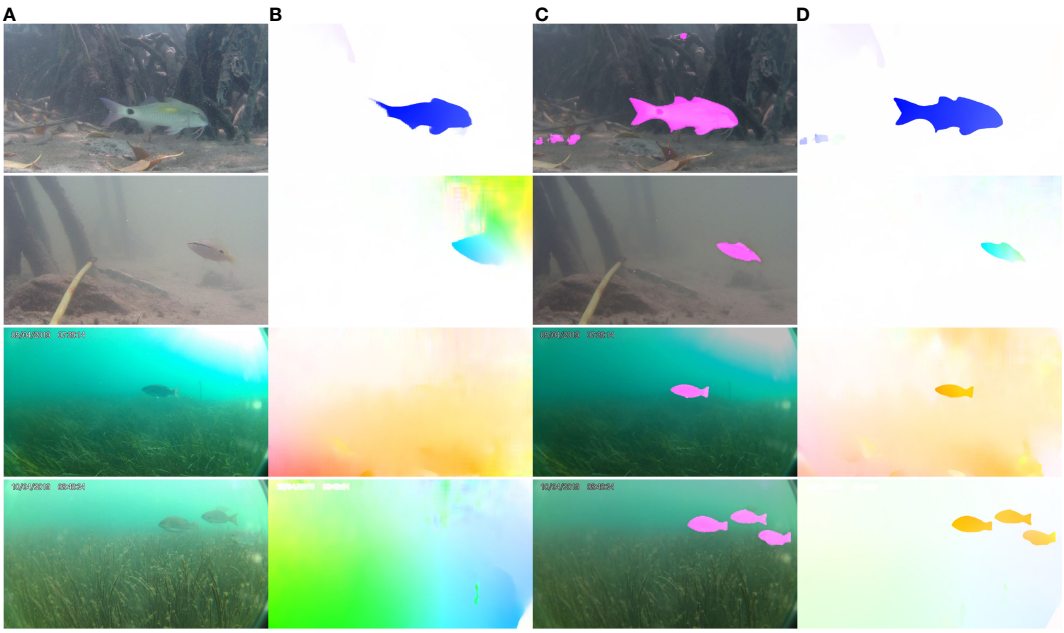

FIGURE 2
Results of optical flow preprocessing. The first two rows of data are from DeepFish and the last two rows of data are from Seagrass. **(A)** the original frame, **(B)** original optical flow, **(C)** overlaid frame, **(D)** overlaid frame optical flow.

appearance features of underwater fish and the motion features with background interference by co-attention. Since fish in underwater videos tend to be obscure and in motion, considering both intra-frame saliency and inter-frame motion, we design two different co-attention parts in M-CAGM. The self co-attention guidance module S-CAGM aims to enhance fish appearance features that are not salient within a single video frame. The flow co-attention module F-CAGM facilitates the interactive enhancement of fish appearance and motion features. By combining S-CAGM and F-CAGM, M-CAGM can focus more on the appearance and motion of foreground fish in underwater videos. Meanwhile, to avoid the accumulation of single-category feature errors throughout the video sequence, such as incorrect object focus or background motion such as seagrass, the FAFM is designed in the decoder stage of MSGNet to filter and fuse different features to improve the robustness of the model. FAFM can learn information at different scales by cascade stacking. We also set a learnable mutual gate in FAFM to measure the contribution of different features to the final segmentation result and selectively fuse fish appearance and motion features. Figure 3 shows the complete MSGNet model.

2.3.2 Multiple co-attention guidance module

Co-attention is often employed to construct the connection between muti-modal features by interacting with different types of features for enhancement separately. Co-attention enables reciprocal optimization of different types of information instead of focusing on a single category of information, improving the robustness of the model. For example, a co-attention mechanism is used in the hierarchical feature alignment network (HFAN) (Pei et al., 2022) to assign weights to aligned features. These weights are used to adaptively and selectively fuse motion and appearance information, thus mitigating the interference caused by poor optical flow or obscure appearance features. Similarly, the motion-attentive transition network (MATNet) (Zhou et al., 2020) employs co-attention to interconvert appearance features with motion features, improving the unsupervised video object segmentation accuracy. In underwater videos, appearance information can be used to locate fish, while motion information contains more details of fish edges. Thus, applying co-attention to construct correlations between fish appearance and motion features and guide bidirectional feature enhancement is worth investigating.

Traditional co-attention (Lu et al., 2019) assigns weights to both inputs simultaneously from a channel perspective by calculating the affinity matrix of different inputs. In underwater scenes, because of the degradation of visual features, the relationship between appearance and motion features cannot be captured entirely from the channel perspective only. For this reason, a co-attention guidance module (CAGM) is designed, as shown in Figure 4A. Specifically, the CAGM takes fish appearance and motion features as input, reduces the channel by a 1×1 convolution, then performs global normalization with the softmax function for both types of features, and establishes the relationship between fish appearance and motion features from the spatial dimension by calculating the affinity matrix after global normalization. Compared with simply resizing to obtain the association between appearance and motion features, assigning weights to spatial features before calculating the affinity matrix can better highlight the appearance and motion features of fish. Similar to the co-attention siamese networks (COSNet) (Lu et al., 2019), the affinity matrix M contains the connection between the activated fish appearance and motion features at the channel level. The motion features are weighted by M and the appearance features are weighted with the transpose of M. In summary, the CAGM applies global spatial normalization to enhance the fish appearance and motion features separately from a spatial perspective and establishes the relationship between fish appearance and motion features from the channel dimension by calculating the affinity matrix of the two features, promoting the reciprocal optimization of the two different types of information.

However, by calculating co-attention for appearance and motion features, only the moving fish in underwater videos are focused on, and the best segmentation masks cannot be achieved. In some scenarios, static fish or background motion will further interfere with degraded features and eventually lead to segmentation failure. Thus, as shown in Figure 4B, we designed the M-CAGM. According to different inputs, the M-CAGM is divided into two parts: the self co-attention guidance module (S-

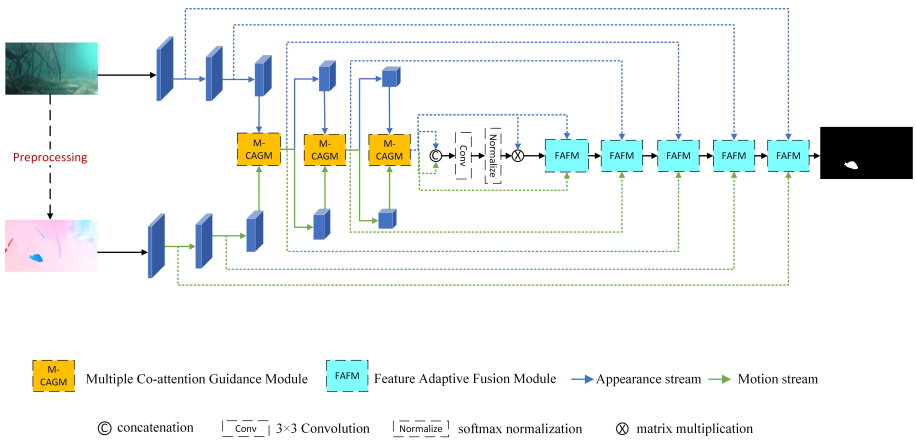

FIGURE 3
Overview of the MSGNet model.

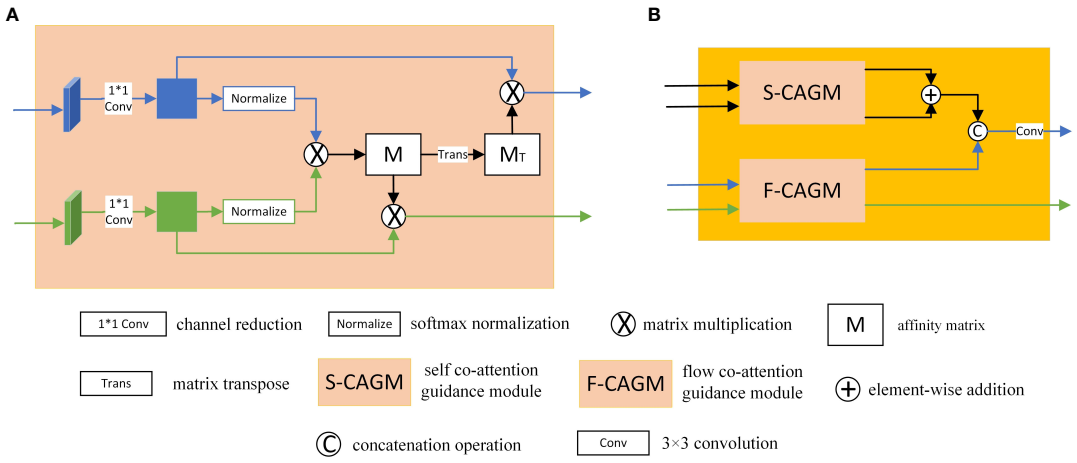

FIGURE 4
Schematic illustration of the CAGM **(A)** and M-CAGM **(B)**.

CAGM) and the flow co-attention guidance module (F-CAGM). The S-CAGM enhances the degraded fish appearance features by inputting the same frame twice and calculating the similarity between the appearance features and itself to obtain the intra-frame salient fish. The F-CAGM uses a frame and the corresponding optical flow as input. The appearance information guides the optical flow branch to focus more on the motion of the fish and ignore the interference of background motion such as seagrass. In addition, the motion information can be employed to recover the edge details of the fish. Finally, we summed the outputs of S-CAGM by element-wise addition and then concatenated them with the appearance outputs of the F-CAGM with a 3×3 convolution to obtain the final enhanced fish appearance features. The motion outputs of the F-CAGM are the enhanced fish motion features. Figure 4 shows the framework of the CAGM and M-CAGM.

In addition, the M-CAGM contains a larger number of parameters, which would be burdensome if used at the first encoder stage because of the high resolution. In practice, low-level features provide rich detailed information, i.e., boundary, texture, and spatial structure information. In contrast, high-level features contain more semantic information. After balancing the inference time and average accuracy, we decided to use the M-CAGM in the last three layers of the encoder.

2.3.3 Feature adaptive fusion module

The M-CAGM promoting the network focuses more on the appearance and motion of the fish to be segmented. However, there are still some failures, such as the incorrect optical flow optimization due to obscure fish appearance information and inaccurate appearance optimization caused by background motion such as seagrass in underwater optical flow. To avoid the accumulation of errors in single-type features within video sequences, it is necessary to filter and fuse features adaptively.

Inspired by the work of Yang et al. (2021), we proposed the FAFM to filter and adaptively fuse the enhanced fish appearance and motion features to obtain the final segmentation masks of fish in underwater videos. The FAFM framework is shown in Figure 5.

The FAFM has three inputs: the output of the upper-level decoder and the appearance and motion features obtained from the corresponding encoder stage. First, the FAFM follows a cascade structure, transferring fish appearance and motion features of different stages to learning multi-scale fish information in the network. Specifically, the decoder output D_{x-1} carries the information from the previous layer. The output is concatenated with the appearance features F_{xa} and motion features F_{xm} obtained from the next encoder and summed along the channel to obtain the fused features F,

$$F = (Cat(UP(D_{x-1}), F_{xa}) + (Cat(UP(D_{x-1}), F_{xm}) \quad (1)$$

UP is the upsampling operation with stride 2. *Cat* is the concatenation operation along the channel axis. Fused features F are enhanced from both channel and spatial dimensions using the convolutional block attention module (CBAM) (Woo et al., 2018) to obtain F', which contains both appearance and motion information of the main fish. To fuse the appearance and motion information selectively and avoid the accumulation of errors of single-category features in the sequence, such as fuzzy appearance or background motion, we design a learnable mutual gate G to balance the contribution of different features instead of using CBAM results directly. Specifically, the fused features obtained by the CBAM are normalized as an appearance input weight, while the motion input is weighted using *1-G*, as in Eqs. (2-4):

$$G = \sigma(F') \quad (2)$$

$$F_{xa'} = G \odot F_{xa} \quad (3)$$

$$F_{xm'} = (1 - G) \odot F_{xm} \quad (4)$$

σ indicates the sigmoid normalization function, $\odot$ denotes element-wise multiplication, $F_{xa'}$ is the appearance features weighted by G, and $F_{xm'}$ is the motion features weighted by *1-G*.

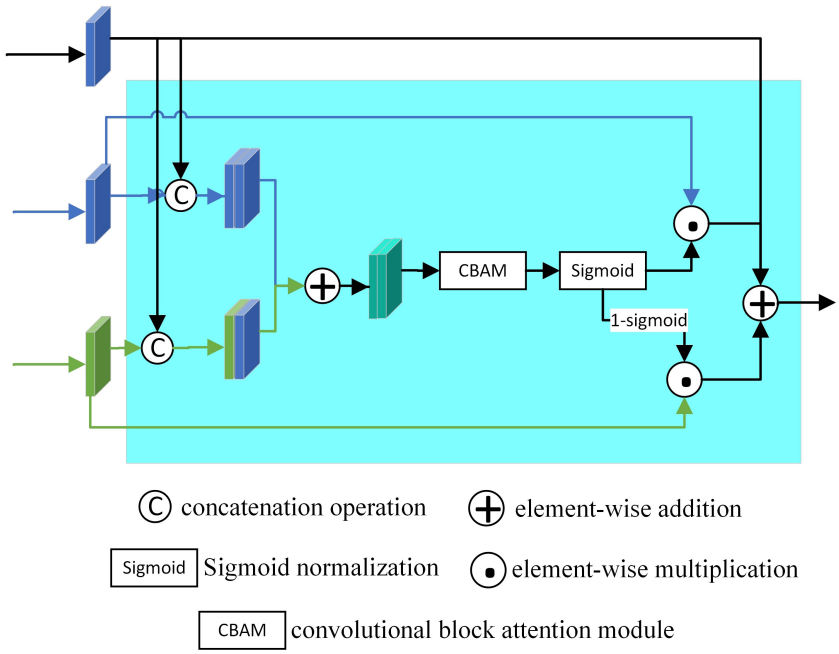

FIGURE 5
The FAFM (Feature Adaptive Fusion Module) framework.

Finally, the decoder output D_x of stage X is as follows:

$$D_x = F_{xa'} + F_{xm'} + D_{x-1} \tag{5}$$

In contrast to Yang (Yang et al., 2021), in the decoder stage, considering that the optical flow may not be accurate because of background motion or light changes, we fuse appearance features with motion features to perform feature filtering and enhancement rather than converting motion features into weights to optimize appearance features alone. By fusing multilevel information with both fish appearance and motion information, the segmentation failures caused by low-quality optical flow can be suppressed. In addition, the mutual gate places the importance of different types of features for the final segmentation mask in a learnable state. When G approaches 1, all appearance features contribute to the final segmentation mask. In contrast, when G approaches 0, motion features contribute to the final segmentation mask. The mutual gate increases the robustness of the model and avoids the errors of blurred appearance or low-quality optical flow on the segmentation results with single information.

2.4 Loss function

The prediction mask for the video frame t at different decoder stages is P_t^i, where $i \in \{1,2,3,4\}$. The gap between the prediction mask P_t and the ground-truth G_t is measured by the standard cross-entropy loss L_{bce} (De Boer et al., 2005). L_{bce} is calculated as follows:

$$L_{bce}(P_t, G_t) = -\sum_{(x,y)} [G_t(x,y)log(P_t(x,y)) + 1 - G_t(x,y)log(1 - P_t(x,y)))] \tag{6}$$

(x, y) are the location coordinates of the pixel points in video frames, and the final loss L_{total} is as follows:

$$L_{total} = \sum_{i=1}^{3} L_{bce}(UP(P_t^i), Gt) + L_{bce}(P_t^4, Gt) \tag{7}$$

UP is the upsampling operation with stride 2, which aims to align the prediction mask P_t^i with the ground-truth G_t in the spatial dimension. Calculating the loss for each decoder allows for precisely controlling the learning of multi-scale information at different stages. Additionally, it facilitates the accurate fusion of appearance and motion features by the mutual gate in the FAFM.

2.5 Evaluation metrics

Fish segmentation in underwater videos is a binary semantic segmentation task with an object pixel value of 255 and a background pixel value of 0 for the prediction and ground-truth. Thus, we employed two semantic segmentation evaluation methods to evaluate the performance, the mean pixel accuracy (mPA) and the mean intersection over union($mIoU$). The mPA is the ratio of correctly classified pixels to the total number of pixels averaged over all classes. The $mIoU$ denotes the average of the ratio of the intersection and union of the pixel predictions for all classes, and they are calculated as follows.

$$mPA = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{8}$$

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

Where k indicates the number of categories, and k is 2 in this study. p_{ii} is the number of pixels correctly predicted as the fish category, i.e. true positive (*TP*). p_{ij} is the number of fish category pixels incorrectly predicted as background, i.e. false positive (*FP*). p_{ji} is the number of background pixels incorrectly predicted as the fish category, i.e. false negative (*FN*). p_{jj} is the number of pixels correctly predicted as background, i.e. true negative (*TN*).

2.6 Experimental design

To verify the effectiveness and robustness of the proposed method, the following experiments were designed for validation. (1) A series of ablation experiments were designed to verify the effectiveness of the proposed module. (2) Experiments with different optical flow contrast were designed to compare the visualization results and select the most suitable contrast for data preprocessing. (3) Experiments with different numbers of M-CAGMs were designed to compare the inference speed and accuracy and select the appropriate positions and numbers of M-CAGMs. (4) Comparison experiments with other advanced underwater fish segmentation models and video object segmentation models using additional optical flow information were designed to validate the advancement of the model. (5) The model was tested with video datasets of underwater camouflaged organisms to verify the generalizability.

3 Results

3.1 Implementation and training detail

We utilized a graphics processing unit (GPU) to accelerate training, and the environmental configurations are as follows: GeForce RTX3090 with 24 GB of video memory, an Intel(R) Core(TM) i7-9700 (3.00 GHz) central processing unit (CPU), a Python 3.8 interpreter, a PyCharm development platform with CUDA (version 11.3), the PyTorch 1.11.0 deep learning framework, and MMSegmentation (Contributors, 2020), an open source semantic segmentation toolkit based on PyTorch.

First, in the optical flow preprocessing stage, the underwater optical flow dataset is acquired according to the method described in Section 2.1, the FCN here is provided by MMSegmentation and pre-trained on ImageNet1k, fine-tuned using 3576 underwater fish images publicly available online, and the optical flow extraction model is the RAFT model integrated into PyTorch 1.11.0. Then, after balancing the inference time and accuracy, we use two ResNet-101 branches with shared weights to extract the appearance feature and the motion feature of underwater fish. The ReNet-101 used here was pre-trained on the ImageNet1K dataset, and the final average pooling and fully connected layers were removed. The input frame and optical flow resolution used for training and testing were uniformly set to 384 × 384 to facilitate model processing and enhanced by random flipping. The model inputs are video frames and their corresponding optical flow. We trained our model for 100 epochs with a batch size of 4. Stochastic gradient descent (SGD) was employed as the model optimizer, where the initial learning rate, momentum, and weight decay were set to 1e-3, 0.9, and 5e-4, respectively. Figure 6 visualizes the *loss* and *IoU* changes in the training process, which indicates that as the training epochs increase, the *loss* gradually converges, and the *IoU* increases.

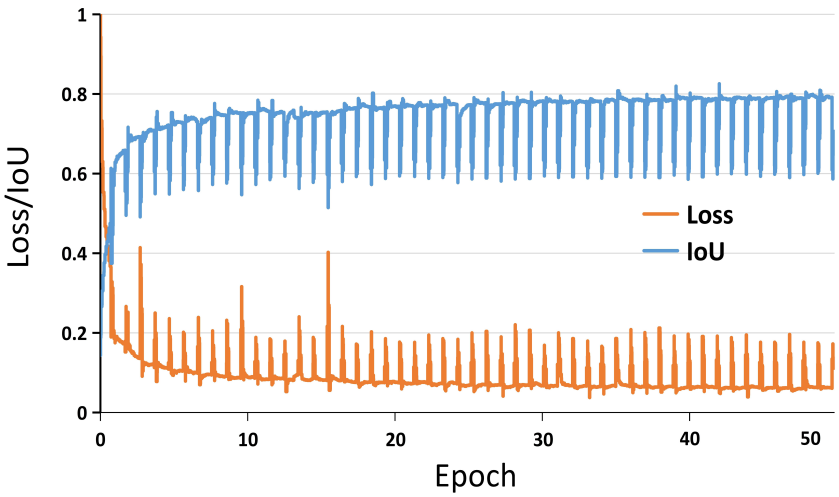

FIGURE 6
Training loss and IoU plot.

3.2 Ablation test

To verify the effectiveness of the proposed model as well as the modules, an ablation test was designed. The baseline in this paper was a two-stream network. Specifically, the encoder stage of the model is a shared weight ResNet-101 network, which was used to extract the features of different inputs. The decoder obtains the final segmentation masks by concatenating the fish appearance and motion features along the channels, applying a convolution with a kernel size of 3 x 3. Combining the baseline with the M-CAGM and FAFM, we obtain the MSGNet model.

The test results in Table 1 indicated that both the M-CAGM and FAFM improved the *mPA* and *mIoU* of the baseline in different ways. The improvement of the baseline by the FAFM was more significant than that by the M-CAGM because the M-CAGM only enhanced fish appearance and motion features using co-attention but did not filter the features to guide the enhancement, causing additional interference. The FAFM achieved better results by filtering the enhanced features through the mutual gate to avoid error accumulation leading to false segmentation. In addition, the *mPA* of the baseline with only the FAFM added was slightly higher than that of MSGNet because there are more small fish in the datasets, especially the Seagrass dataset. Meanwhile, the M-CAGM in MSGNet segments part of the background as a target, causing a high percentage of incorrectly segmented pixels. We visualized the features in the last stage of the decoder. As shown in Figure 7, the baseline features are not obvious enough to obtain only the coarse location of the fish. By adding the M-CAGM to the baseline, the fish features are enhanced, but interference is introduced in the background. By adding the FAFM, the features are filtered, and the background interference is suppressed, but this leads to missing fish object edges. Combining the M-CAGM and FAFM, MSGNet enhances the fish object features, suppresses the noise in the background, and obtains clear edge information. The *mPA* is improved by 1.77% and the *mIoU* is improved by 2.81% in MSGNet compared with those of the baseline, which verifies the effectiveness of using both the M-CAGM and FAFM.

3.3 Contrast selection test for optical flow preprocessing

In the optical flow data preprocessing stage, using different contrast ratios for the superimposed object and background frames will result in different optical flows, which will have an impact on the experimental effect. Contrast refers to the ratio of the

TABLE 1 Ablation analysis of MSGNet where the metrics with the highest rankings are shown in bold.

Method	M-CAGM	FAFM	*mPA*/%	*mIoU*/%
Baseline			90.12	86.10
Baseline + M-CAGM	√		91.31	87.23
Baseline + FAFM		√	**92.11**	88.06
MSGNet	√	√	91.89	**88.91**

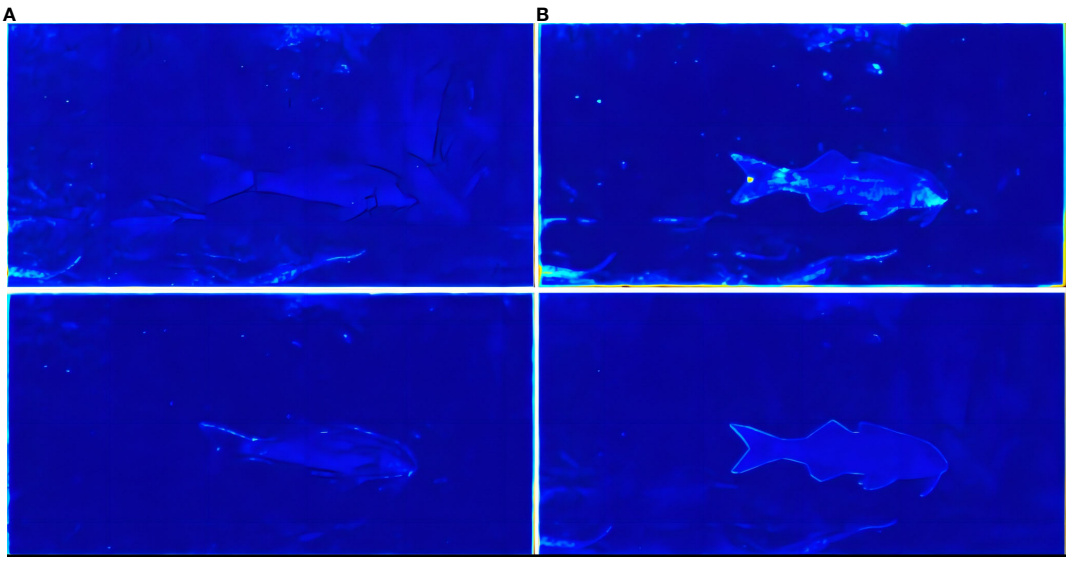

FIGURE 7
Illustration of the ablation test. **(A1)** means the first row of column A, and so on. **(A1)** Visualization features of the baseline; **(B1)** visualization features of the baseline + M-CAGM; **(A2)** visualization features of the baseline + FAFM; **(B2)** visualization features of MSGNet.

background frame luminance to the original frame luminance, so the contrast selection test for optical flow preprocessing is designed to select and verify the optimal front- and back-view contrast ratios. The standard evaluation metric for optical flow is the endpoint error (EPE), which is the mean of the Euclidean distance between the predicted optical flow vector and the ground- truth over all pixels. However, there is no available optical flow labeling method to compare the predicted optical flow and the ground-truth. To research the effect of different optical flows on segmentation, we used preprocessed optical flows with contrasts of 1, 0.75, 0.5, 0.25, and 0 as input to the proposed model. Figure 8 demonstrates the segmentation masks of the preprocessed optical flow with different contrast ratios. It is evident that as the contrast ratio decreases, the optical flow is gradually disturbed by light changes and background motion. Especially when the contrast ratio is 0, there is no background in the overlay frame, and the moving object can no longer be detected by the optical flow. Optical flow can be used to detect similar texture regions in the image and calculate each point's distance and direction of motion by dense matching. If the background disappears completely, the textures are identical, and the random motion of any point in the background satisfies the matching condition, which is not conducive to calculating optical flow. In Figure 8, when the contrast is 0.75, the interference caused by the change in illumination appears in the upper right part of the fish, but the outline of the fish is still clear, and the proposed model can effectively remove this small part of the interference. When the contrast is 0.5 or 0.25, the outline below the fish becomes blurred, the predicted mask contains redundant incorrectly predicted pixels, and the fish and background cannot be accurately distinguished. After experimenting with the complete datasets, the predicted masks had the highest *mIoU* and *mPA* when the contrast was 0.75. Thus, we selected 0.75 as the background contrast in the optical flow data preprocessing stage. The specific results are shown in Table 2. Notably, the *mPA* and *mIoU* of the model decrease significantly when the contrast is 0. Under this condition, the optical flow cannot bring any benefits and will only cause model interference.

3.4 Test and analysis of the number of M-CAGM

The M-CAGM is used to facilitate reciprocal enhancement of different types of features. The number of M-CAGMs affects the accuracy of model segmentation and the inference speed, where the number setting refers to the number starting from the deep encoder level. Specifically, the low-level encoder captures the main fish edges and location information, while the high-level encoder contains more advanced semantic information. Theoretically, the more M-CAGMs applied, the higher the segmentation accuracy of the model. However, considering the real-time demand, the inference speed is also an essential performance measure. Thus, we designed this test to balance the model accuracy and inference speed and selected the most suitable number of M-CAGMs. Table 3 shows the *mIoU* and inference speed for different numbers of M-CAGM settings. The experiments are based on a GeForce RTX3090, and *fps* is the speed of inference, which indicates the number of frames per second that can be processed. When the number of M-CAGMs

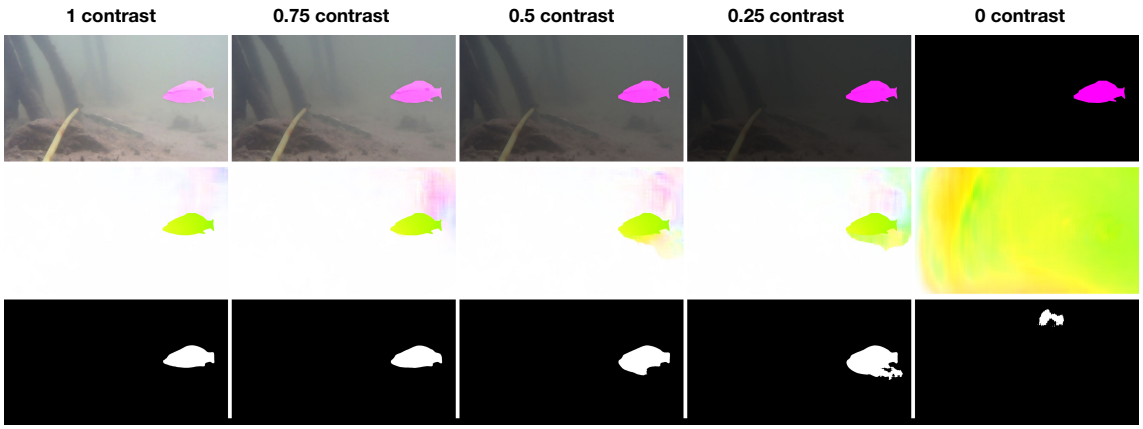

FIGURE 8
Illustration of the optical flow contrast ratio test. From top to bottom: the object overlay frames, optical flow of the overlay frames, and predicted masks of the overlay frames.

TABLE 2 Evaluation of optical flow with different contrast where the metrics with the highest rankings are shown in bold.

Contrast ratio	1	0.75	0.5	0.25	0
mPA/%	91.55	**91.89**	87.32	84.74	71.33
mIoU/%	88.16	**88.91**	84.46	79.60	61.52

TABLE 3 Inference speed and *mIoU* for different numbers of M-CAGMs where the metrics with the highest rankings are shown in bold.

Number	5	4	3	2	1
fps/(frame/s)	8.64	23.33	26.86	28.08	**28.90**
mIoU/%	–	**88.94**	88.91	88.15	87.71

is 5, the network cannot be trained because the M-CAGM has high-resolution input. The processing burden is too large, which is also the reason for the low *fps*. When the number of M-CAGMs is 3, the *mIoU* is very close to the best, and the *fps* is reduced within a reasonable range compared to that using fewer M-CAGMs. Therefore, we applied 3 M-CAGMs in the proposed model.

3.5 Comparison with other advanced models on the DeepFish and Seagrass dataset

To verify the segmentation effectiveness of the proposed model in the case of underwater visual feature degradation, MSGNet was compared with the robust underwater object segmentation network (WaterSNet) (Chen et al., 2022) and segmentation of underwater imagery network (SUIM-Net) (Islam et al., 2020), advanced underwater fish segmentation models. Considering that optical flow was introduced as an additional input in this study, for a fair comparison, we also compared MSGNet to the full-duplex strategy network (FSNet) (Ji et al., 2021) and attentive multimodality collaboration network (AMC-Net) (Yang et al., 2021), advanced video object segmentation models. The results in Table 4 indicate that the model using additional optical flow data has higher *mPA* and *mIoU* values than those of the model using only pictures. SUIM-Net applies a fully convolutional encoder-decoder with skip connections. It offers competitive performance while ensuring fast end-to-end inference but cannot fully utilize information from different layers by skip connections and deconvolution only. WaterSNet effectively improves the segmentation accuracy of non-significant and camouflaged fish by random style adaptation (RSA) of input images and multi-scale fusion. However, when setting the group size of RSA, the batch size is directly used as the number of mixed images in the group, which leads to a significant decrease in the robustness of the model under hardware-constrained conditions. FSNet effectively enhances the interaction between appearance and motion features with a full-duplex but unselectively reuses motion information, leading to unsatisfactory segmentation results in poor optical flow. AMC-Net suppresses redundant and misleading information through multichannel co-attention gates while designing a motion correction module with a visual motion attention mechanism to highlight features of foreground objects, achieving *mPA* and *mIoU* values close to those of MSGNet. In Figure 9, the results of MSGNet have more accurate bounds compared to the segmentation results of other models while effectively suppressing prediction redundancy. Compared with the above models, the M-CAGM in MSGNet facilitates the reciprocal interactive enhancement of fish appearance and motion information. At the same time, the FAFM effectively suppresses the error accumulation of single-type features through the mutual gate. The proposed model improves the *mIoU* by 2.08% compared with those of the WaterSNet, and compared with those of the AMC-Net, the *mIoU* value is improved by 1.30%. The comparison experiment shows the effectiveness of the MSGNet in segmenting fish in underwater videos.

3.6 Testing results and analysis on camouflage dataset

To verify the generalizability of the proposed model and investigate the segmentation effect of MSGNet in dealing with different degradation conditions of underwater visual features, we conducted validation experiments with the moving camouflaged animals mask dataset MoCA-Mask. The camouflaged object in selected frames has low contrast with the background and blends in with the underwater environment, which is similar to the segmentation difficulties caused by the degradation of underwater visual features. Additionally, considering that there is less water turbidity in the MoCA-Mask dataset, no preprocessing of optical flow data was performed in this experiment. The first two rows of Figure 10 demonstrates that MSGNet can still segment a complete object when the visual features of the fish are degraded and blended

TABLE 4 Comparison test results with advanced models .

Model	Image	Flow	*mPA*/%	*mIoU*/%
SUIM-Net	√		89.89	83.27
WaterSNet	√		91.14	86.83
FSNet	√	√	86.77	82.84
AMC-Net	√	√	91.55	87.61
MSGNet	√	√	**91.89**	**88.91**

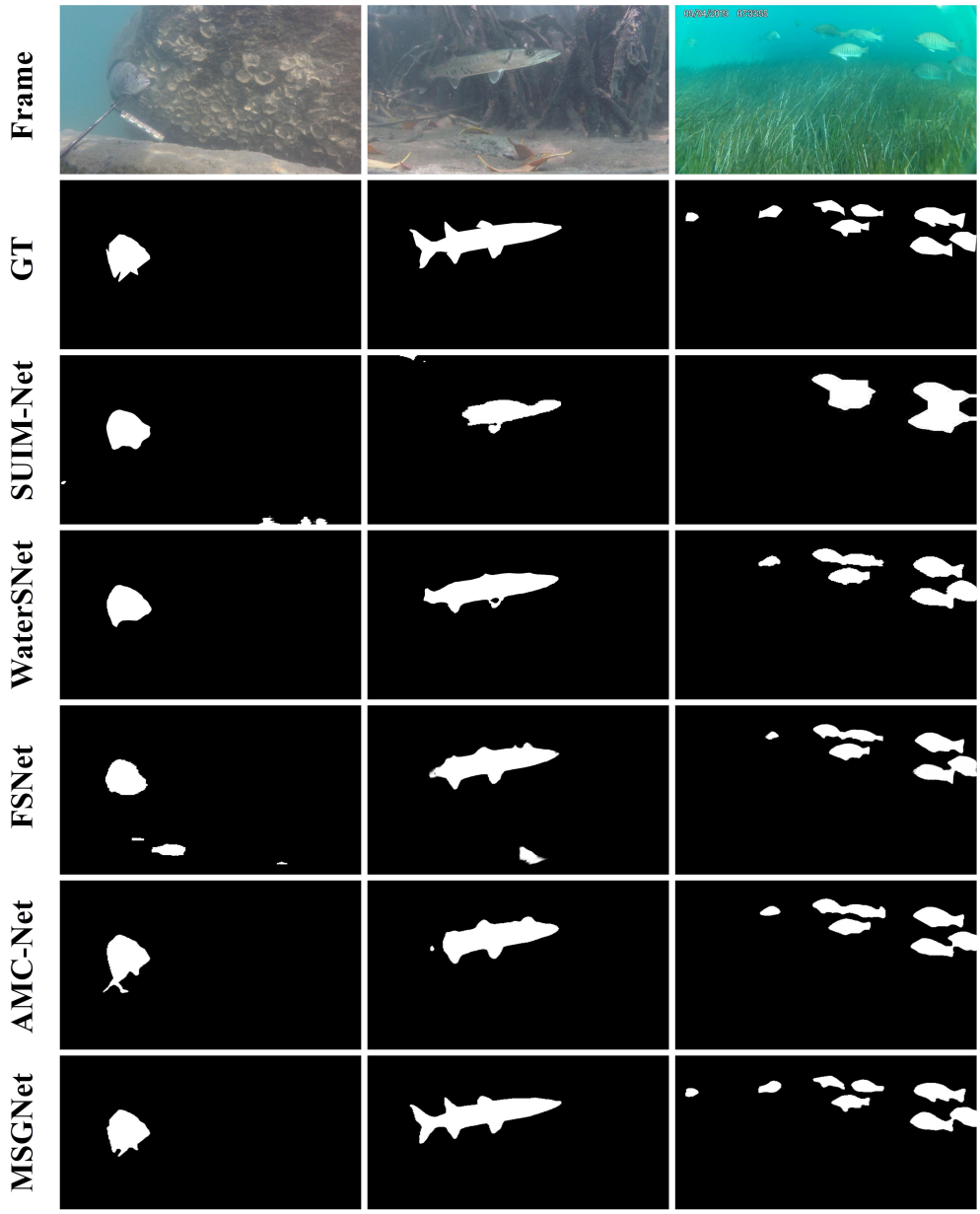

FIGURE 9
Prediction results of different models on DeepFish and Seagrass dataset.

with the background. The evaluation results shown in Table 5 also verify that MSGNet achieves higher mPA and mIoU compared to those of the baseline, indicating that MSGNet can perform well with different datasets, demonstrating its good generalizability. However, as demonstrated in the third row of Figure 10, the model proposed in this paper still has failure cases when segmenting underwater camouflaged creatures. On the one hand, the third row of Figure 10 belongs to a highly camouflaged state, where the camouflaged object is not only unremarkable at the edges but even difficult to be distinguished from the background texturally. Another reason is that the underwater object is small and hard to be roughly localized. These failure cases indicate that the proposed model still has space for improvement, such as using multi-scale information to improve the accuracy of segmenting small underwater objects.

4 Discussion

4.1 Model superiority discussion

Accurately segmenting fish objects in complex underwater environments can be challenging. On the one hand, underwater objects are often blurred due to water turbidity and insufficient brightness. On the other hand, there are camouflaged creatures in some underwater scenes, such as devil scorpion fish and flounder. To survive better, these creatures get evolved with low contrast with the background, making it difficult to quickly locate the object's position even in a clean underwater environment and more difficult to accurately segment the underwater camouflaged creatures. There are few studies for fish object segmentation in complex underwater videos,

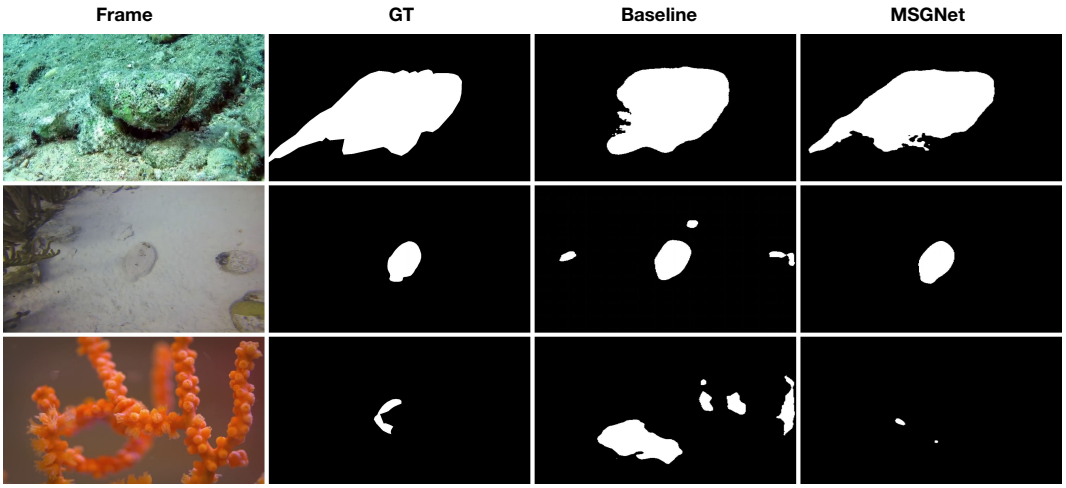

FIGURE 10
Tests with the MoCA-Mask dataset.

and most use only appearance information. Biological visual perception shows that in continuous scenes, people are easily attracted to moving objects, thus breaking the original blurred or camouflaged state. From this point, our work innovatively uses motion optical flow to help segment fish objects in underwater videos, and this is the most apparent difference between our proposed model and existing work (Haider et al., 2022; Shoffan, 2022). We first preprocess the underwater optical flow. After acquiring the underwater optical flow data, instead of simply summing the multimodal feature alignment, we selectively enhance and fuse the appearance and motion information to improve the accuracy and robustness of fish segmentation in underwater videos. The ablation tests show that through M-CAGM, the proposed model simultaneously considers the visual saliency of the fish itself and the correlation between appearance and motion, effectively achieving the interaction and enhancement of multi-source information in underwater scenes, highlighting the possible fish objects in the scenes. FAFM then further filters and fuses the features after the exchange to avoid the accumulation of errors, which may lead to segmentation failure. The results of comparison experiments on public datasets indicate that introducing motion information can effectively improve segmentation accuracy in underwater videos. Thus, our work can effectively segment blurred and camouflaged fish in underwater videos.

4.2 Application and future work

In 2021, the United Nations (UN) launched the Decade of the Ocean (OD) initiative to promote sustainable ocean development. Protecting marine ecosystems will be enhanced by increasing secure areas and severely protected areas for habitats and fish stocks salvage (European commission, 2020). Under such an initiative, our work focuses on pixel-level segmentation of fish in underwater videos. By accurately segmenting fish in underwater videos, information such as fish length and profile can be obtained, which may be helpful for visual verification or estimation of fish size and weight by human experts, facilitating habitat population monitoring. Hall et al. (2023) showed that the size of anadromous Baltic Sea perch females affects the quality of the offspring and the ability of the progeny to perform under different temperature conditions, and our proposed MSGNet allows for the accurate segmentation of underwater fish, which can be used to estimate fish size, thus enabling an assessment of the long-term sustainability of the population. Laradji et al. (2021) also pointed out that pixel-level segmentation masks are more helpful in evaluating the size and shape of fish to analyze fish habitat. The proposed segmentation methods can also be combined with counting and tracking and integrated into a system that automatically performs comprehensive monitoring to increase efficiency and reduce labor costs.

In addition, overfishing is a significant problem for the sustainable development of marine fisheries. The Food and Agriculture Organization of the United Nations (FAO) states that the proportion of unsustainably exploited fishery resources has increased from 10% to 35.4% since the 1970s (Food and Agriculture Organization of the United Nations, 2022). Employing pixel-level segmentation for obtaining fish sizes in underwater videos can reduce the risk of overfishing by avoiding catching fish that are not the right size. Another potential application is the lightweight deployment of the proposed model to underwater robots or other portable devices for automatically collecting vital information like the shape and size of fish in underwater videos, which can facilitate the survey and management of fishery resources. We hope our work will inspire relevant researchers and continue contributing to fish habitat monitoring and sustainable fisheries.

TABLE 5 Evaluation metrics with the MoCA-Mask datasets where the metrics with the highest rankings are shown in bold.

Method	*mPA*/%	*mIoU*/%
Baseline	85.38	77.23
MSGNet	**88.37**	**84.70**

However, some parts can still be improved at this stage of our work. As a result of our research, we found that in current ocean observation, the algorithms often need to be deployed on constrained hardware platforms (Novy et al., 2022) or portable underwater robots (Chatzievangelou et al., 2022), which require high computational power and running speed of the models. Although our proposed model has high segmentation accuracy for blurred or camouflaged fish in complex underwater videos, it contains many computational parameters and needs to be more lightweight. In applications requiring high portability, such as when divers carry equipment to explore marine resources and species, hardware limitations may make it impossible to balance inference speed with segmentation accuracy. Considering the limitations of the device in practical applications, in the future, we will investigate a way to make the model lighter and achieve more efficient ocean observation.

5 Conclusion

In this paper, a multi-source guidance network MSGNet was proposed to segment fish in underwater videos. It combined both video frames and motion optical flows, and can be used to facilitate ocean observation. To address the problems of low accuracy and poor robustness of the model caused by insufficient lighting and water turbidity in complex underwater environments, this paper proposes a method to segment fish in underwater videos combining both appearance and motion information. First, we apply a simple overlay to obtain high-quality underwater optical flow data. Then, we employ multiple co-attention mechanisms to facilitate the interaction and enhancement of the appearance and motion features of fish. Finally, we design a mutual gate to filter and adaptively fuse the different features to obtain the final segmentation results of fish in underwater videos through multiple iterations. The experimental results with several publicly available datasets validate the effectiveness and superiority of this study for fish segmentation in underwater videos. However, this study still needs to be improved, as the M-CAGM contains multiple global normalizations, resulting in a large computational overhead. Considering the equipment limitations in practical applications, we will research a method to make the model more lightweight in the future.

Data availability statement

This study was analyzed using publicly available datasets. These data can be found here: DeepFish, https://alzayats.github.io/DeepFish/; Seagrass, https://doi.pangaea.de/10.1594/PANGAEA.926930; MoCA-Mask, https://xueliancheng.github.io/SLT-Net-project/. Further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval were not required for the animal study because in this research, we used three open databases of DeepFish, Seagrass and MoCA-Mask, which can be freely used for academic purposes.

Author contributions

PZ: Conceptualization, Investigation, Writing – original draft. HY: Resources, Supervision, Writing – review & editing. HL: Methodology, Validation, Writing – original draft. XZ: Software, Writing – original draft. SW: Methodology, Writing – original draft. WT: Writing – original draft. ZY: Writing – original draft. JW: Funding acquisition, Writing – review & editing. YL: Funding acquisition, Writing – review & editing.

Funding

This work was supported by the Key Projects of Educational Department of Liaoning Province (LJKZ0729), and National Natural Science Foundation of China (31972846), Liaoning Province Natural Science Foundation (2020-KF-12-09), Foundation of Educational Department of Liaoning Province (LJKZ0730).

Acknowledgments

I received a lot of valuable advice and timely help in conducting my experiments and writing my thesis. I want to express my sincere gratitude to all those who have helped me. In particular, i would like to express my appreciation to my supervisor, Hong Yu, who gave me much professional guidance in experimental design and thesis writing. Secondly, I would like to thank Dr.Alzayat Saleh for his help and advice on optical flow, without which I would not have been able to carry out my subsequent work. Finally, I would like to thank my friends and family for your continuous support.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2023.1256594/full#supplementary-material

References

Beddow, T. A., Ross, L. G., and Marchant, J. A. (1996). Predicting salmon biomass remotely using a digital stereo-imaging technique. *Aquaculture* 146 (3-4), 189–203. doi: 10.1016/S0044-8486(96)01384-1

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 *Proc. IEEE int. conf. acoust. speech signal process* (Singapore: IEEE), 2629–2633. doi: 10.1109/ICASSP43922.2022.9746176

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, LA, USA: IEEE), 13854–13863. doi: 10.1109/CVPR52688.2022.01349

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 *IEEE International Conference on Image Processing* (Brussels, Belgium: IEEE), 3145–3148. doi: 10.1109/ICIP.2011.6116334

Contributors, M. (2020) *MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark*. Available at: https://github.com/open-mmlab/mmsegmentation (Accessed July 8, 2023).

Costa, C., Loy, A., Cataudella, S., Davis, D., and Scardi, M. (2006). Extracting fish size using dual underwater cameras. *Aquacultural Engineering* 35 (3), 218–227. doi: 10.1016/j.aquaeng.2006.02.003

De Boer, P. T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Ann. operations Res.* 134, 19–67. doi: 10.1007/s10479-005-5724-z

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL, USA: 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

Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazirbas, C., Golkov, V., et al. (2015). "FlowNet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision* (Santiago, Chile: IEEE), 2758–2766. doi: 10.1109/ICCV.2015.316

European commission (2020)Biodiversity strategy for 2030. In: *Energy, Climate change, Environment*. Available at: https://ec.europa.eu/environment/strategy/biodiversity-strategy-2030_en (Accessed July 8, 2023).

Food and Agriculture Organization of the United Nations (2021) *Empowering women in small-scale fisheries in the United Republic of Tanzania. EAF-Nansen Programme, FAO*. Available at: https://www.fao.org/in-action/eaf-nansen/news-events/detail-events/en/c/1413988/ (Accessed July 8, 2023).

Food and Agriculture Organization of the United Nations (2022)The State of World Fisheries and Aquaculture 2022. In: *Towards Blue Transformation* (Rome: FAO). doi: 10.4060/cc0461en (Accessed July 8, 2023).

Garcia, R., Prados, R., Quintana, J., Tempelaar, A., Gracias, N., Rosen, S., et al. (2020). Automatic segmentation of fish using deep learning with application to fish size measurement. *ICES J. Mar. Sci.* 77 (4), 1354–1366. doi: 10.1093/icesjms/fsz186

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. 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

He., K., Zhang., X., Ren., S., and Sun, J. (2016). "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV, USA: IEEE). 770–778. doi: 10.1109/CVPR.2016.90

Huang, P. X., Boom, B. J., and Fisher, R. B. (2015). Hierarchical classification with reject option for live fish recognition. *Mach. Vision Applications* 26 (1), 89–102. doi: 10.1007/s00138-014-0641-2

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 *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Las Vegas, NV, USA: 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 *IEEE/CVF International Conference on Computer Vision* (Montreal, QC, Canada: 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. doi: 10.3389/fmars.2022.1003568

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* (Berlin, Heidelberg: Springer-Verlag), 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, L., Dong, B., Rigall, E., Zhou, T., Dong, J., and Chen, G. (2021a). Marine animal segmentation. *IEEE Trans. Circuits Syst. Video Technol.* 32, 2303–2314. doi: 10.1109/TCSVT.2021.3093890

Li, L., Rigall, E., Dong, J., and Chen, G. (2021b). MAS3K: An open dataset for marine animal segmentation. *Proc. Symp. Benchmarking Meas. Optim* 12614, 194–212. doi: 10.1007/978-3-030-71058-3_12

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Boston, MA, USA: IEEE), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., and Porikli, F. (2019). "See more, know more: unsupervised video object segmentation with co-attention Siamese networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA, USA: IEEE).

Muñoz-Benavent, P., Martínez-Peiró, J., Andreu-García, G., Puig-Pons, V., Espinosa, V., Pérez-Arjona, I., et al. (2022). Impact evaluation of deep learning on image segmentation for automatic bluefin tuna sizing. *Aquacultural Engineering* 99, 102299. doi: 10.1016/j.aquaeng.2022.102299

Novy, D., Kawasumi, L., Ferguson, J., Sullivan, M., Bell, P., Chow, J. S., et al. (2022). Maka Niu: A low-cost, modular imaging and sensor platform to increase observation capabilities of the deep ocean. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2022.986237

Pei, G., Shen, F., Yao, Y., Xie, G. S., Tang, Z., and Tang, J. (2022). "Hierarchical feature alignment network for unsupervised video object segmentation," In *European Conference on Computer Vision* (Switzerland: Springer Nature), 596–613. doi: 10.1007/978-3-031-19830-4_34

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 Engineering* 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: Comput. Vision Pattern Recognition*. 2208.10662, 1–16. doi: 10.48550/arXiv.2208.10662

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 (4), 1295–1307. doi: 10.1093/icesjms/fsz025

Shoffan, S. (2022). "K-means and morphological approach on image segmentation for fish detection," in *International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology* (Prachuap Khiri Khan, Thailand: IEEE), 1–4. doi: 10.1109/ECTI-CON54298.2022.9795404

Teed, Z., and Deng, J. (2020). "Raft: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision* (Glasgow, UK: Springer, Cham), 2 (16), 402–419. doi: 10.1007/978-3-030-58536-5_24

Woo, S., Park, J., Lee, J. Y., and Kweon, I. S. (2018). "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision* (Cham: Springer), 3–19. doi: 10.1007/978-3-030-01234-2_1

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 *IEEE/CVF International Conference on Computer Vision* (Montreal, QC, Canada: IEEE), 1544–1553. doi: 10.1109/ICCV48922.2021.00159

Zhang, W., Wu, C., and Bao, Z. (2022). DPANet: Dual Pooling-aggregated Attention Network for fish segmentation. *IET Comput. Vision* 16 (1), 67–82. doi: 10.1049/cvi2.12065

Zhao, Y.-p., 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 Engineering* 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 Processing* 29, 8326–8338. doi: 10.1109/TIP.2020.3013162

Zhuang, P., Wang, Y., and Qiao, Y. (2020). Wildfish++: A comprehensive fish benchmark for multimedia research. *IEEE Trans. Multimedia* 23, 3603–3617. doi: 10.1109/TMM.2020.3028482

Zivkovic, Z., and van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition Lett.* 27 (7), 773–780. doi: 10.1016/j.patrec.2005.11.005