

Article

Research on Fish Recognition in Complex Backgrounds Using ViT-Enhanced YOLOv11

Xiangshuo He ^{1,2,3,†}, Shenglong Yang ^{2,3,*,†} , Wei Wang ^{1,2}, Kai Zhu ⁴, Shengmao Zhang ², Yang Dai ², Keji Jiang ² and Fei Wang ^{2,*}

¹ College of Information, Shanghai Ocean University, Shanghai 201306, China

² Key Laboratory of Fisheries Remote Sensing, Ministry of Agriculture and Rural Affairs, East China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shanghai 200090, China

³ Key and Open Laboratory of Remote Sensing Information Technology in Fishing Resource, Chinese Academy of Fishery Sciences, Shanghai 200090, China

⁴ Zhejiang Marine Fisheries Research Institute, Zhoushan 316021, China

* Correspondence: ysl6782195@126.com (S.Y.); essiwf@163.com (F.W.);

Tel.: +86-137-6467-2421 (S.Y.); +86-157-0212-8229 (F.W.)

† The authors contributed equally to the work.

Abstract

To address the common challenges in fish recognition tasks under complex backgrounds, such as target overlap, occlusion, and chaotic spatial distribution, an improved YOLOv11 recognition model based on the Vision Transformer (ViT) is proposed. Traditional Convolutional Neural Networks (CNNs) and the YOLO series models are limited by their local receptive fields, making it difficult to capture global semantic correlations in dense and heavily occluded fish target detection, which often leads to feature confusion and false detections. By embedding ViT modules at the beginning of the Head and at the end of the Backbone of YOLOv11, the self-attention mechanism of ViT is leveraged to capture global dependencies in the image, re-integrate and enhance multi-scale features from the Backbone and Neck, thus constructing two improved ViT models. Comparative experiments are conducted on the FishRecognition-2025 dataset, which contains 955 high-resolution RGB images covering nine common coastal fish species across four categories: single fish species, multiple classes separated, slight overlap of multiple fish species, and severe overlap of multiple fish species. Under identical training strategies and evaluation metrics, the four models—original YOLOv11, traditional CNN, ViT-Head, and ViT-Backbone—are compared. The results show that the second improved ViT model (with ViT placed at the end of the Backbone) outperformed the first improved model (with ViT placed at the beginning of the Head) in terms of mAP50 and mAP50-95. Moreover, its overall accuracy across the four test data categories (single fish species, multiple classes separated, slight overlap of multiple fish species, and severe overlap of multiple fish species) surpassed that of YOLOv11, CNN, and the first ViT model. Although its accuracy in single fish species and multiple classes separated scenarios was slightly lower than that of the CNN model, it demonstrated significant advantages in scenarios with slight overlap of multiple fish species and severe overlap of multiple fish species. These findings validate the effectiveness of the ViT module in global feature modeling and adaptability to complex backgrounds, suggesting a promising technical direction for future real-time recognition in fishery field operations.



Academic Editor: Zhangying Ye

Received: 19 May 2026

Revised: 19 June 2026

Accepted: 25 June 2026

Published: 27 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: YOLOv11; fish recognition; Vision Transformer; complex backgrounds; object detection

Key Contribution: This study examined the impact of integrating a Vision Transformer (ViT) into the YOLOv11 framework for fish recognition under complex backgrounds with severe target overlap. The effects of different Transformer insertion positions within the one-stage detector were systematically analyzed, focusing on robustness and multi-scale feature consistency. Key findings were that backbone-level integration of global self-attention improved robustness while preserving multi-scale feature consistency. Direct comparison with baseline models on a complex-scene fish dataset demonstrated superior performance in densely cluttered scenarios. The study suggests that embedding Transformer modules at the backbone level effectively enhances one-stage detectors' ability to handle heavy occlusion and background complexity, which may offer valuable insights for future automated fish recognition in real-world aquaculture and ecological monitoring.

1. Introduction

With increasing global demands for stricter fisheries resource management and growing ecological conservation awareness, achieving accurate and automated identification and monitoring of fish species and quantities during fishing operations has become a critical technological challenge for ensuring the sustainable use of marine resources. For China's offshore fisheries, developing automated fish catch identification and quantitative monitoring technology suitable for the deck environment of fishing vessels is not only directly related to the scientific basis of catch quota management and the accuracy of fishery statistical data but also serves as a fundamental driver for the transition towards intelligent and standardized fisheries management. Traditional fish identification methods, such as manual observation based on morphological features or post hoc video analysis [1], are inefficient, subjective, and physical contact can easily induce stress responses in fish, affecting their growth [2].

In recent years, computer vision and deep learning technologies have provided new approaches for automated fish identification at fishing sites. Significant progress has been made by researchers in the identification and quantification of catch aboard fishing vessels, leading to the development of various deep learning-based automated identification models [3–6]. For instance, David C et al. [7] used specialized fishery remote electronic monitoring cameras to identify and quantify catches on five Peruvian small-scale gillnet vessels, showing effectiveness in detecting and quantifying target chondrichthyan catches and seal bycatch, although its performance was influenced by fishing gear. Geoff French et al. [8] conducted identification research based on closed-circuit television videos installed on trawlers, targeting discards, but it was susceptible to target movement, occlusion, and could not measure length. C Vilas et al. [9] analyzed fish on a conveyor belt during sorting by fishers, combining color, texture, and shape features for fish classification. These studies often rely on network architectures like Mask R-CNN, YOLO, or MobileNet, achieving high accuracy in identification and length measurement when individuals are well-separated. However, model performance significantly declines under conditions of severe fish overlap, complex lighting, or cluttered deck backgrounds [10]. Therefore, enhancing model robustness against multi-fish overlap and complex backgrounds under natural fishing conditions remains a technical challenge.

Since its introduction by Redmon et al. in 2016 [11], the YOLO algorithm has become a mainstream choice for real-time fish identification due to its balance between detection accu-

racy and speed. Numerous domestic and international studies have explored fish detection based on YOLO series and Convolutional Neural Network (CNN) architectures [12–16], achieving good results in underwater environments. However, when applied to real off-shore fishing scenarios, complex factors such as deck reflections, dense stacking of fish, and dramatic lighting variations lead to insufficient generalization capability of the models in actual fishing videos, making it difficult to meet practical identification needs.

The Vision Transformer (ViT), first proposed by Dosovitskiy et al. [17], effectively captures long-range feature dependencies by dividing images into patches and employing global self-attention, offering a new perspective for addressing issues like target overlap, occlusion, and complex spatial distribution in fish identification. For example, Gao H et al. [18] proposed a ViT-based point-level feature learning framework that improved accuracy in person re-identification under occlusion and pose changes. Rodrigo et al. [19] found that ViT demonstrated higher accuracy and robustness than CNNs in handling distance variations and occlusion in face recognition. Zhang Z et al. [20] proposed an occlusion suppression and restoration Transformer that achieved state-of-the-art performance in occluded person re-identification tasks. Furthermore, research on hybrid architectures, such as Latif S A et al. [21] combining CNNs with ViT, achieved 93.66% accuracy in iris recognition. Zheng X et al. [22] transferred CNN capabilities to ViT through cross-model knowledge distillation, achieving performance surpassing existing methods in semantic segmentation tasks. These cross-domain studies collectively validate the effectiveness of ViT and CNN hybrid architectures in tackling complex visual pattern recognition problems.

Nevertheless, in fish identification scenarios, while ViT possesses the advantage of global modeling, its standalone application faces issues such as poor real-time performance and low accuracy in small object detection. Moreover, no research has yet integrated it with YOLOv11 to specifically address the key challenges in fish identification. To overcome the performance degradation of convolution-based detectors in complex fish recognition scenarios, this study develops a ViT-enhanced YOLOv11 framework that integrates global self-attention into a one-stage detection architecture. Two hybrid models are constructed by embedding Vision Transformer modules at different stages of YOLOv11, namely at the beginning of the detection head and at the end of the backbone, enabling a systematic investigation of how Transformer placement affects multi-scale feature representation and detection robustness. Comprehensive experiments are conducted on the FishRecognition-2025 dataset, which is specifically designed to reflect varying degrees of target overlap and spatial disorder in real fishing environments. The experimental results demonstrate that incorporating ViT at the backbone level significantly improves global contextual modeling and target discrimination in highly cluttered scenes, while preserving the integrity of the feature pyramid. The main contributions of this work are summarized as follows: (1) A ViT-enhanced YOLOv11 detection framework is proposed to improve fish recognition performance under complex backgrounds characterized by dense target overlap and severe occlusion. (2) The influence of Transformer insertion positions within a one-stage detector is systematically analyzed, revealing that backbone-level global attention modeling yields superior robustness in complex scenes. (3) Extensive experimental evaluations on a complex-scene fish recognition dataset validate the effectiveness and practicality of the proposed approach compared with conventional CNN-based and baseline YOLO models.

2. Materials and Methods

2.1. Dataset Construction

Since on-site monitoring data from fishing vessels were not available for this study, a sample dataset was constructed through indoor photography. Given the wide variety of fish species caught by fishing vessels in the East China Sea, the main commercially permitted

species from the region were selected as the research subjects. Experimental samples were collected from the Shanghai Yangpu Aquatic Products Market, and high-definition images of the fish were captured using a high-resolution camera. The image collection covered nine fish species from the East China Sea with high economic value and ecological significance, including: *Octopus vulgaris*, *Cynoglossus semilaevis*, *Larimichthys polyactis*, *Sepiella japonica*, *Scomberomorus niphonius*, *Miichthys miiuy*, *Trichiurus lepturus*, *Pampus argenteus*, and *Platycephalus indicus*. The sample collection process accounted for variations in body length and posture (such as curved, stretched, or stacked positions), resulting in a total of 955 high-resolution RGB images with a uniform resolution of 4000×2256 pixels, ensuring comprehensive coverage of real-world fishing scenarios.

We adopted a multi-angle sampling design for data acquisition: fish targets were photographed from different orientations and pitch angles, while retaining the background and illumination variations in real underwater environments. The dataset explicitly provides a quantitative distribution of each category and experimental scenario: it includes 9 fish species, with 10 different individual samples collected for each species, resulting in a total sample size of 90 individuals. Each sample covers three size specifications: small, medium, and large. To simulate the complex distribution of fish catches on fishing vessels, the dataset was constructed according to the following four scenarios (Figure 1): (1) single fish species; (2) multiple classes separated (multiple fish species placed separately without overlap); (3) slight overlap of multiple fish species; (4) severe overlap of multiple fish species. The X-AnyLabeling annotation tool was used for manual labeling of the images, following the PASCAL VOC standard format. Each image was independently annotated by one professional and reviewed by another expert. The annotations included bounding boxes for the fish targets and their corresponding species category labels. The final constructed dataset was named FishRecognition-2025 and was randomly split into training (763 images), validation (95 images), and test (97 images) sets in an 8:1:1 ratio. During the split, a fixed random seed (e.g., seed = 42) is used, and the distribution across each category and the four scenarios (single fish species, multiple classes separated, slight overlap, and severe overlap) is balanced to avoid class bias in model training.

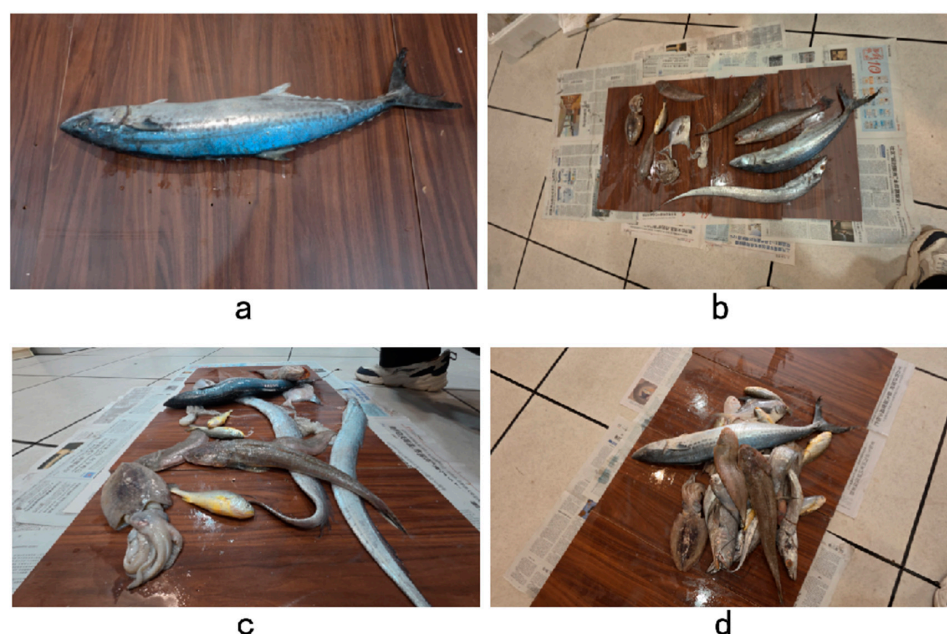


Figure 1. Photos taken under four different scene conditions. (a) Single fish species; (b) Multiple fish species placed separately without overlap; (c) Slight overlap of multiple fish species; (d) Severe overlap of multiple fish species.

2.2. Data Preprocessing

To enhance model performance and generalization capabilities, this study employed multiple data-preprocessing strategies. These primarily include image resizing, normalization, and data augmentation techniques. The data augmentation strategies consisted of random horizontal flipping, color space transformations, random cropping, and Mosaic composition. Color space transformations simulated natural lighting variations by adjusting the HSV channels: hue was randomly perturbed within a $\pm 1.5\%$ range, while saturation and value were dynamically adjusted at intensities of $\pm 70\%$ and $\pm 40\%$, respectively, improving the model's adaptability to diverse lighting and color conditions. Mosaic augmentation, applied with a probability of 0.7, combined four randomly selected images into a single composite image and merged their corresponding annotation boxes to simulate multi-object dense scenarios. By dynamically adjusting image arrangement and cropping regions, this technique increased the diversity of occlusion patterns among targets. These strategies effectively enhanced the model's adaptability to complex scenes while preserving data diversity.

2.3. YOLOv11 Model

YOLO series algorithms [23] have been widely adopted in recognition scenarios due to their high efficiency and accuracy in the field of object detection. The YOLOv11 model is an optimized detection architecture based on YOLOv8, comprising three main components: the Backbone, Neck, and Head (Figure 2) [24]. Compared to YOLOv8, YOLOv11 replaces the original C2f module with the C3K2 module, introduces a C2PSA module after the SPPF block [25], and incorporates design concepts from the YOLOv10 head into its own Head structure. These enhancements significantly improve the model's detection accuracy and inference speed, making it well-suited for real-time detection scenarios requiring rapid response, such as in situ fish identification in fisheries.

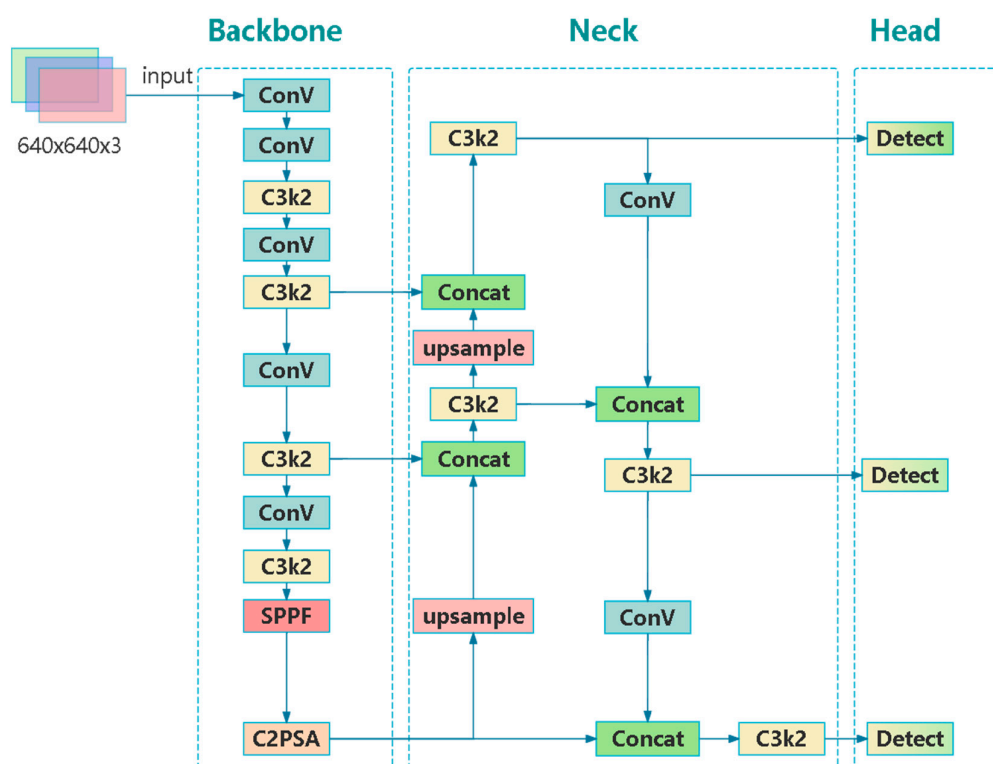


Figure 2. Structure of YOLOv11 Network.

2.4. ViT Technology

The Vision Transformer (ViT) model [17] represents a significant innovation by successfully adapting the Transformer architecture [26] from natural language processing to the field of computer vision. Unlike traditional convolutional neural networks, ViT divides an input image into fixed-size patches, embeds them into a sequence of high-dimensional vectors through linear projection, and incorporates positional encoding to retain spatial information (Figure 3). These vectors are then fed into a network composed of multiple Transformer encoder layers, where self-attention mechanisms capture global dependencies within the image. This approach enables ViT to model long-range feature relationships, making it particularly suitable for visual understanding tasks involving object occlusion, overlap, and complex backgrounds.

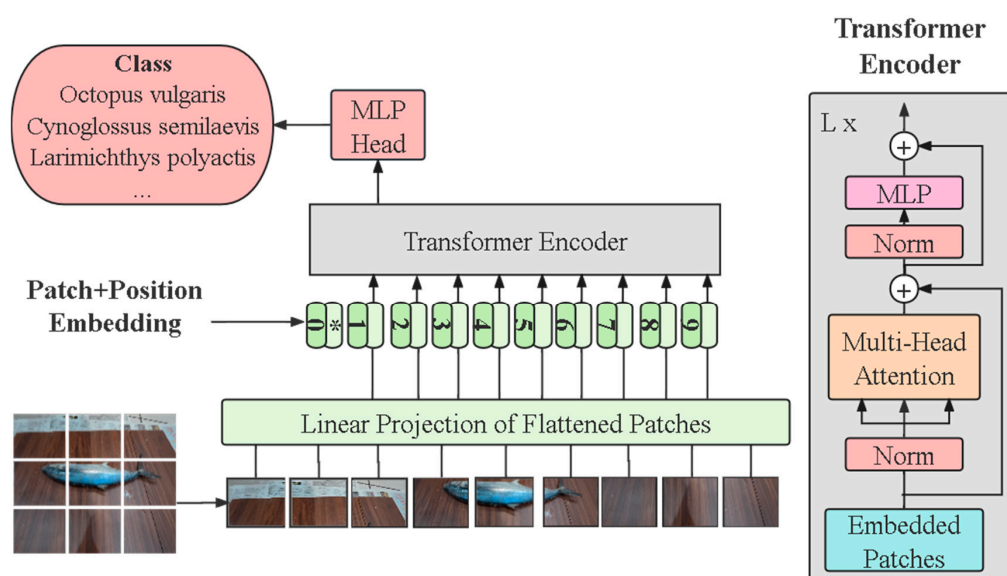


Figure 3. Structure of ViT Model.

2.5. ViT and YOLOv11 Fusion Model

While YOLOv11 demonstrates strong performance in fish identification tasks, it still exhibits limitations when dealing with complex scenarios characterized by severe target overlap and cluttered spatial distributions. To address this issue, this study proposes two hybrid models integrating ViT with YOLOv11, aiming to enhance the model's adaptability to complex backgrounds by incorporating global attention mechanisms.

- (1) **YOLOv11-ViT1 Model:** This model embeds a ViT module at the beginning of the YOLOv11 Head. The network architecture diagram is shown in Figure 4. The specific configuration is as follows: the feature map is divided into 20×20 patches, projected into a 256-dimensional vector space, and then fed into an encoder composed of 10 Transformer blocks. Each block employs 8 attention heads, with the MLP hidden layer dimension expanded to 2048, and a dropout rate of 0.1 is applied. This design enables the model to leverage ViT's global context modeling capability to reintegrate and enhance the multi-scale features from the Backbone and Neck before final detection, thereby improving target localization and classification accuracy.

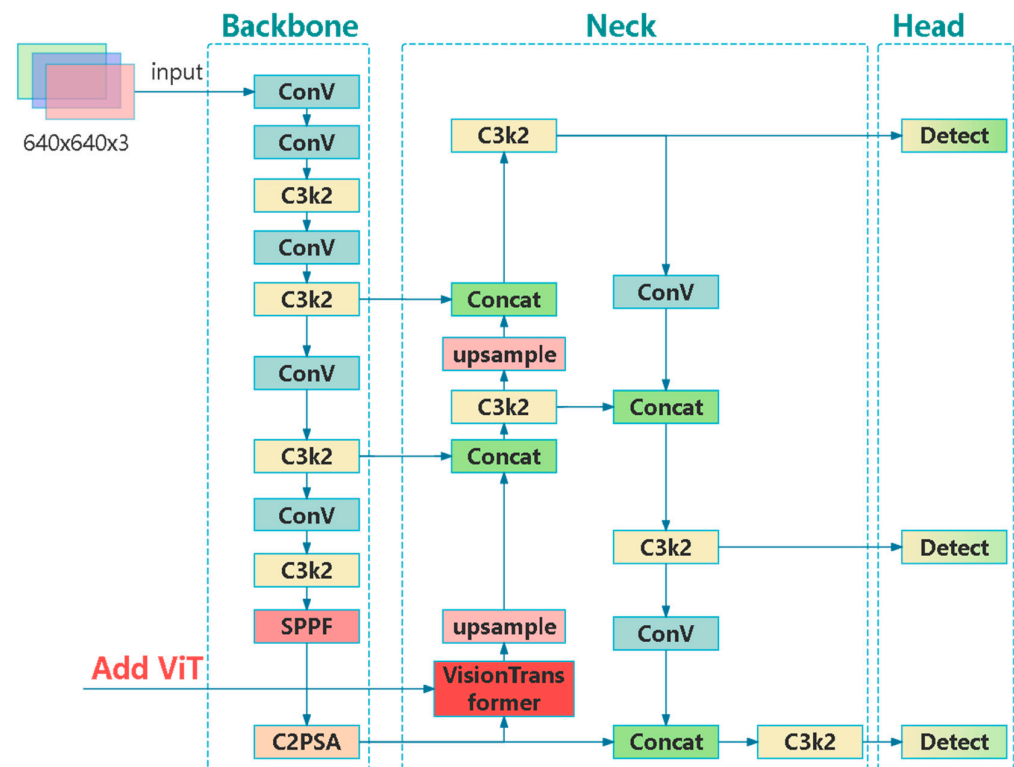


Figure 4. Network architecture diagram of the YOLOv11-ViT1 model.

- (2) YOLOv11-ViT2 Model: This model embeds an enhanced ViT module at the end of the YOLOv11 Backbone. The network architecture diagram is shown in Figure 5. Its ViT configuration is further optimized: the patch embedding dimension is increased to 512, the number of attention heads is raised to 4, the encoder layers are adjusted to 16, and a C2fAttn attention enhancement module is introduced into the Head. This module incorporates a multi-head self-attention mechanism into the standard C2f module. The specific structure is as follows: the input feature map is split and fed into two branches. One branch maintains convolutional feature extraction, while the other captures global dependencies through multi-head self-attention. Finally, the outputs from the two branches are fused and feature compression is performed via pointwise convolution. This design enables the model to balance local texture and long-range contextual information while remaining lightweight, thereby improving the recognition ability of key fish parts. This structure positions ViT as a powerful feature extractor, integrating global contextual information without disrupting the integrity of the feature pyramid, significantly boosting the model's feature representation and fusion capabilities.

2.6. Experimental Setting

To validate the effectiveness of the ViT modules, in this study, multiple baseline models and two ViT-improved models are set up for comparative experiments: Baseline models (control group): include VGG, YOLOv8, YOLOv11-CBAM (with a lightweight attention mechanism), original YOLOv11, and a traditional CNN (using ResNet50 as the backbone network). ViT-improved models (experimental group): YOLOv11-ViT1 and YOLOv11-ViT2 models. All models were evaluated based on the FishRecognition-2025 dataset, using identical training protocols (300 epochs with a 5-epoch learning rate warm-up phase) and the same test set (97 images) to ensure the comparability of the experimental results.

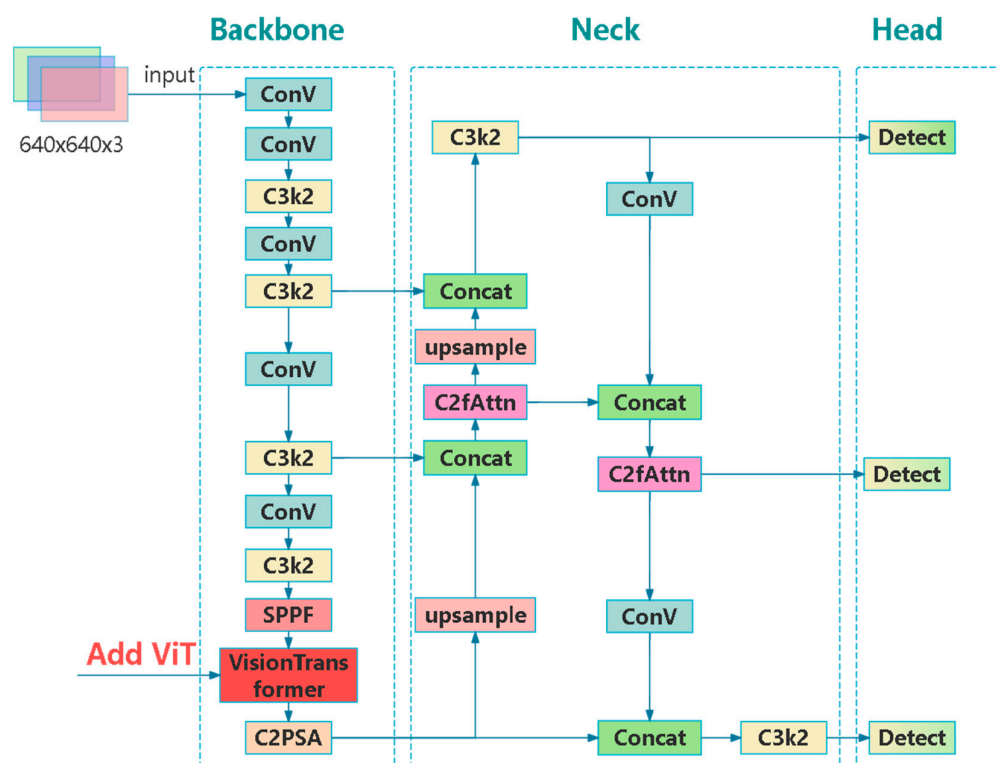


Figure 5. Network architecture diagram of the YOLOv11-ViT2 model.

2.7. Experimental Environment

The experimental environment was configured as follows: operating system Windows 11 Home Edition, processor Intel Core i5-11400KF (2.7 GHz), memory 32 GB, graphics card NVIDIA GeForce RTX 4090, deep learning framework PyTorch 2.5.1, and CUDA version 12.4. Some experiments were conducted on the AutoDL cloud platform to accelerate the training process. The key training hyperparameters are set as follows: the total number of epochs is 300, the batch size is 128, the input image size is uniformly resized to 640×640 pixels, and the number of data loading worker threads is set to 8. Regarding the optimizer configuration, the initial learning rate (lr0) is set to 0.1, the momentum is set to 0.937, and the weight decay coefficient is set to 0.0005.

2.8. Evaluation Index

This study employs two types of metrics to comprehensively evaluate model performance:

(1) Object detection metrics: Based on the IoU threshold, these metrics assess the model's ability to simultaneously localize and classify. The core metrics include mAP50 (mean average precision at IoU = 0.5) and mAP50–95 (mean average precision averaged over IoU thresholds from 0.5 to 0.95 with a step size of 0.05). These two metrics are standard evaluation measures for object detection tasks.

(2) Classification/confusion matrix metrics: After a detected box is successfully matched with a ground-truth box ($\text{IoU} \geq 0.5$), only the correctness of the class label is considered, ignoring localization deviations. Based on the confusion matrix, Precision, Recall, and F1-score are calculated for each class and overall, which are used to evaluate the purity of the model in class discrimination. The calculation formulas for these metrics are shown in Equations (1)–(3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

To comprehensively evaluate the real-time processing capability of the proposed method, we measured the runtime performance metrics of the model during inference, including the number of parameters (Params), floating point operations (FLOPs), GPU memory, average latency per image, and frames per second (FPS), as shown in Table S1.

In addition, for the object detection task, mAP (mean Average Precision) is introduced as a core evaluation metric. Here, mAP50 represents the average precision at an Intersection over Union (IoU) threshold of 0.5, while mAP50-95 denotes the average precision across IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05. These metrics collectively provide a comprehensive assessment of the model's overall performance under varying requirements for localization accuracy.

2.9. Ablation Experiment Setup

To quantify the individual contributions of the ViT module and the C2fAttn module, three ablation variants are constructed based on YOLOv11-ViT2 (ViT embedded at the end of the Backbone + C2fAttn added in the Head):

Model A (Baseline): Original YOLOv11 without any ViT or C2fAttn modules.

Model B (ViT only): ViT module embedded at the end of the Backbone (same ViT configuration as in YOLOv11-ViT2, but with C2fAttn removed from the Head).

Model C (C2fAttn only): C2fAttn module introduced only in the Head (without Backbone ViT).

Model D (Full version): YOLOv11-ViT2, containing both Backbone ViT and Head C2fAttn.

All models adopt the same training strategy (300 epochs, input size 640×640) and test set, and are evaluated on FishRecognition-2025 in terms of mAP50, mAP50-95, and F1-score.

3. Results

3.1. Comparison of Model Detection Accuracy

To comprehensively evaluate the performance of each model, mAP50 and mAP50-95 were calculated under the four scenarios, with the results presented in Tables 1 and 2. In the single fish scenario, all four models demonstrated high detection accuracy, with mAP50 values exceeding 0.95. Among them, the mAP50 of YOLOv11, YOLOv11-ViT2, CNN, YOLOv8, and YOLOv11-CBAM all reached 0.995, while YOLOv11-ViT1 achieved 0.9685, and VGG was the lowest (0.9546). In terms of the mAP50-95 metric, YOLOv11 was the highest (0.8908), followed by YOLOv8 (0.8782), whereas YOLOv11-ViT1 was the lowest (0.7825). These results indicate that in scenarios with simple single fish species, all models effectively accomplished the fish detection and identification task.

Table 1. mAP50 of the Seven Models.

Models	Yolov11	YOLOv11-ViT1	YOLOv11-ViT2	CNN	Yolov8	VGG	YOLOv11-CBAM
single fish species	0.995	0.9685	0.995	0.995	0.9950	0.9546	0.9950
multiple classes separated	0.9731	0.9741	0.9774	0.9807	0.9671	0.9428	0.9750
slight overlap of multiple fish species	0.8099	0.7850	0.8551	0.7976	0.8529	0.6978	0.7810
severe overlap of multiple fish species	0.6982	0.6662	0.7867	0.6749	0.7395	0.5626	0.7105

Note: Values in bold indicate the best performance among all compared models under the same metric.

Table 2. mAP50-95 of the Seven Models.

Models	Yolov11	YOLOv11-ViT1	YOLOv11-ViT2	CNN	Yolov8	VGG	YOLOv11-CBAM
single fish species	0.8908	0.7825	0.82668	0.8349	0.8782	0.7321	0.8477
multiple classes separated	0.8037	0.8045	0.76125	0.8137	0.8032	0.6883	0.7746
slight overlap of multiple fish species	0.5312	0.5218	0.5568	0.5656	0.5979	0.3882	0.5024
severe overlap of multiple fish species	0.4287	0.3962	0.4502	0.4120	0.4962	0.2899	0.4098

Note: Values in bold indicate the best performance among all compared models under the same metric.

In the multiple classes separated scenario, the mAP50 of all seven models remained at a high level (all above 0.94). The CNN model achieved the highest value (0.9807), followed by YOLOv11-ViT2 (0.9774), while VGG was the lowest (0.9428). The YOLOv11-ViT2 model exhibited the lowest mAP50-95 (0.76125), suggesting room for improvement in localization accuracy.

In the scenario with slight overlap of multiple fish species, the detection accuracy of the models decreased significantly, with mAP50 values around 0.8 and mAP50-95 dropping to approximately 0.53 for all models. YOLOv11-ViT2 achieved the highest mAP50 (0.8551) and the second highest mAP50-95 (0.5568), while YOLOv8 achieved the highest mAP50-95 (0.5979) but its mAP50 (0.8529) was slightly lower than that of ViT2. VGG performed the worst (mAP50 0.6978, mAP50-95 0.3882). The mAP50 of the other models (YOLOv11, CNN, YOLOv11-CBAM, YOLOv11-ViT1) ranged between 0.78 and 0.81.

In the scenario with severe overlap of multiple fish species, accuracy further decreases. YOLOv11-ViT2 still achieves the highest mAP50 (0.7867), significantly outperforming YOLOv8 (0.7395), YOLOv11-CBAM (0.7105), YOLOv11 (0.6982), and other models. YOLOv8 achieves the highest mAP50-95 (0.4962), followed by YOLOv11-ViT2 (0.4502). VGG remains the worst (mAP50 0.5626, mAP50-95 0.2899).

Notably, the YOLOv11-ViT2 model consistently achieved the best performance in these two complex scenarios, with mAP50 values of 0.8551 and 0.7867 and mAP50-95 values of 0.5568 and 0.4502, respectively, significantly outperforming the other models. This demonstrates that the YOLOv11-ViT2 model retains strong recognition capability. Even in the complex background of severe overlap of multiple fish species, although its localization precision still requires further enhancement. Due to space limitations, the detailed AP50 and AP50-95 results of each model on different fish species are provided in Supplementary Materials Tables S2 and S3. Overall, YOLOv11-ViT2 achieves the highest AP50 and AP50-95 values for the vast majority of species. Although it is slightly lower than other models on a few individual species, its overall performance remains significantly better than that of the comparison models.

3.2. Comparison of Evaluation Metrics

To evaluate the model's class discrimination capability (ignoring localization accuracy), we calculated the Precision, Recall, and F1 score for each class based on the confusion matrix, and recorded real-time performance metrics. The results are shown in Table 3. From the classification performance metrics in Table 3, except for the VGG model, the Precision, Recall, and F1 scores of the remaining six models are mostly above 0.8. Specifically: VGG performs the worst (Precision 0.7713, Recall 0.6807, F1 0.7207); YOLOv8, YOLOv11-CBAM, YOLOv11, CNN, YOLOv11-ViT1, and YOLOv11-ViT2 all have Precision above 0.81. Regarding Recall, except for CNN (0.7815) and YOLOv11-ViT1 (0.7929), which are slightly below 0.8, all others exceed 0.8. The first improved model, YOLOv11-ViT1, achieves slightly lower classification performance than the original YOLOv11 and the CNN model. In contrast, YOLOv11-ViT2 shows improvements across all classification metrics, making

it the best-performing model among all compared models. Compared with the original YOLOv11 model, its Precision improves by 2.7% (0.8902 vs. 0.8671), Recall by 8.0% (0.8686 vs. 0.8044), and F1 score by 5.3% (0.8792 vs. 0.8346). Its F1 score reaches 0.8792, indicating that the model has well-balanced performance.

Table 3. Confusion matrix-based classification performance metrics and real-time metrics of each model on the test set.

Index	Precision	Recall	F1	Parameters (M)	FLOPs (G)	GPU Memory (MB)	Latency (ms)	FPS
VGG	0.7713	0.6807	0.7207	0.89	2.21	26.98	2.41	414.5
YOLOv8	0.8152	0.7766	0.8365	3.01	4.10	47.25	6.57	152.1
YOLOv11-CBAM	0.8540	0.7725	0.8092	2.59	3.22	81.77	9.06	110.4
Yolov11	0.8671	0.8044	0.8346	2.78	3.33	72.70	12.56	79.6
CNN	0.8587	0.7815	0.8183	3.53	3.64	134.81	9.60	104.2
Yolov11-ViT1	0.8232	0.7929	0.8078	3.15	3.75	102.5	13.2	75.8
Yolov11-ViT2	0.8902	0.8686	0.8792	2.59	3.22	119.27	8.69	115.0

Note: Values in bold indicate the best performance among all compared models under the same metric.

From the perspective of real-time metrics, the VGG model has the smallest number of parameters (0.89 M), the lowest FLOPs (2.21 G), the least GPU memory usage (26.98 MB), the shortest latency (2.41 ms), and the highest frame rate (414.5 FPS), but its detection accuracy (Precision 0.7713, Recall 0.6807) is relatively low. YOLOv8 shows moderate real-time performance (latency 6.57 ms, FPS 152.1). Compared with the original YOLOv11, YOLOv11-CBAM has slightly lower parameters and FLOPs, but higher GPU memory (81.77 MB vs. 72.70 MB) and higher latency (9.06 ms vs. 12.56 ms), and a lower frame rate (110.4 vs. 79.6), indicating that the introduction of the CBAM module increases computational overhead. The CNN model has the largest number of parameters (3.53 M), the highest GPU memory usage (134.81 MB), and poor real-time performance (latency 9.60 ms, FPS 104.2). YOLOv11-ViT1 exhibits unsatisfactory real-time performance, with a latency of 13.2 ms and a frame rate of only 75.8 FPS, making it the slowest among all models. In contrast, YOLOv11-ViT2, which achieves the best overall performance, also has excellent real-time metrics: 2.59 M parameters and 3.22 G FLOPs, both comparable to those of YOLOv11-CBAM; although its GPU memory (119.27 MB) is slightly higher than that of the original YOLOv11, its latency is only 8.69 ms and its frame rate reaches 115.0 FPS, achieving good real-time detection capability while maintaining high accuracy.

3.3. Ablation Experiment Results

Table 4 presents the detection performance and classification F1 scores of each ablation variant across four scenarios. From Model A to Model B, after only adding the ViT module at the end of the Backbone, mAP50 in the scenario with severe overlap of multiple fish species increases from 0.6982 to 0.7713 (an improvement of +10.5%), mAP50-95 increases from 0.4287 to 0.4412, and the F1 score increases from 0.8346 to 0.8512, indicating that global self-attention significantly helps in localizing and discriminating densely occluded targets. Model C (C2fAttn only) achieves an mAP50 of 0.7305 in the severe overlap scenario, with a smaller improvement than the ViT module, suggesting that C2fAttn, as a lightweight supplement for local-global interaction, contributes less than the full ViT module. Model D (full version) further improves mAP50 to 0.7867 and F1 to 0.8792, showing additive gains over both Model B and Model C, demonstrating that ViT and C2fAttn work synergistically: ViT provides backbone-level global context, while C2fAttn refines features in the Head, making their functions complementary rather than redundant.

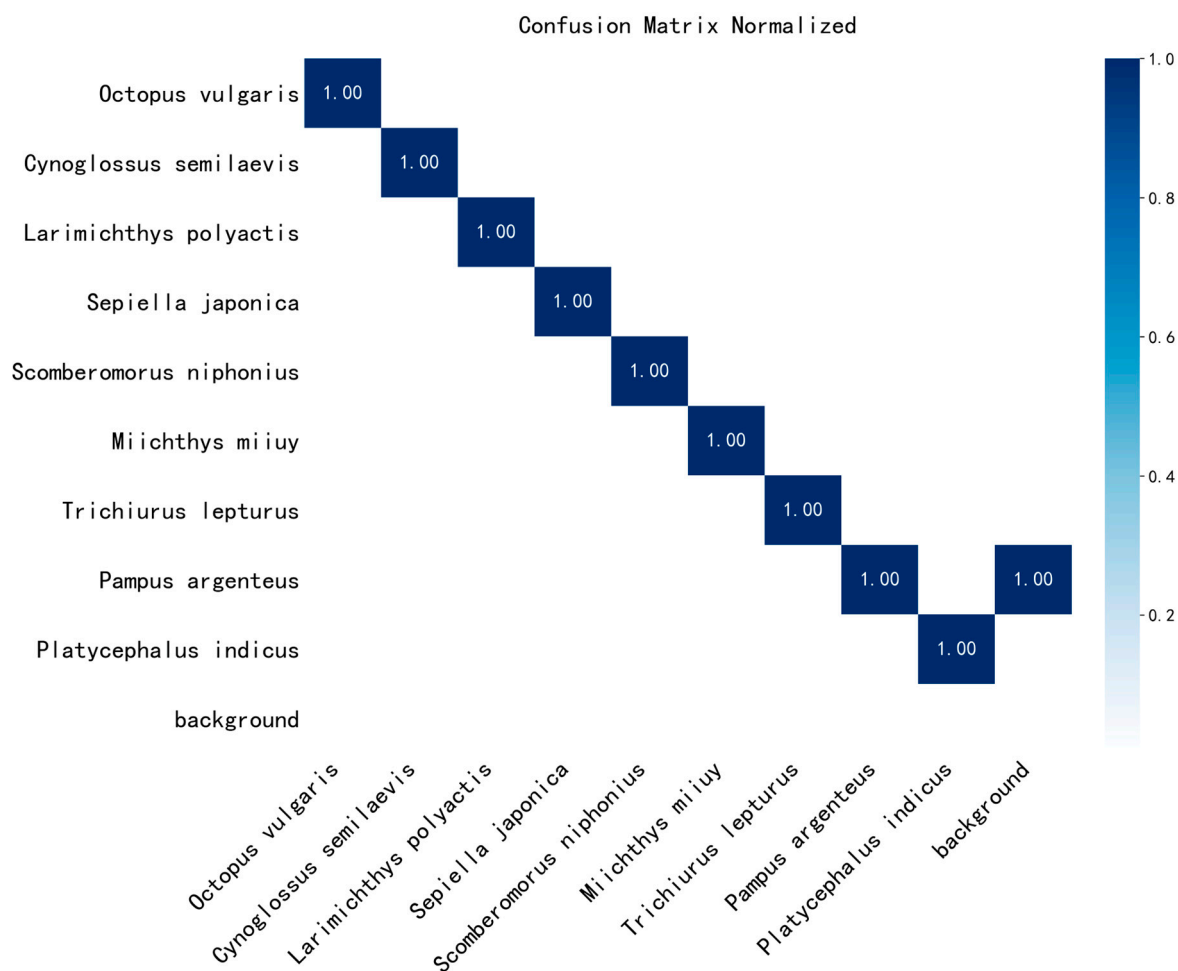
Table 4. The detection performance and classification F1 scores of each ablation variant across four scenarios.

Models	ViT (Backbone)	C2fAttn (Head)	Single Fish Species mAP50	Multiple Classes Separated mAP50	Slight Overlap of Multiple Fish Species mAP50	Severe Overlap of Multiple Fish Species mAP50	Severe Overlap of Multiple Fish Species mAP50-95	F1 (Global)
A (baseline)	x	x	0.9950	0.9731	0.8099	0.6982	0.4287	0.8346
B (ViT only)	✓	x	0.9950	0.9762	0.8433	0.7713	0.4412	0.8512
C (C2fAttn only)	x	✓	0.9950	0.9745	0.8210	0.7305	0.4358	0.8428
D (full)	✓	✓	0.9950	0.9774	0.8551	0.7867	0.4502	0.8792

Note: 1. Values in bold indicate the best performance among all compared models under the same metric.
 2. “✓” indicates that the corresponding module is included in the model variant; “x” indicates that the module is removed (not used).

3.4. Classification Results for the Single Fish Species Scenario

In the single fish species scenario, all models demonstrate high recognition accuracy. As shown in the confusion matrices of Figures S1–S6, VGG, YOLOv8, YOLOv11-CBAM, the original YOLOv11, and YOLOv11-ViT2 achieve classification accuracy close to or reaching 100% for most fish species, with only a few models showing slight misclassifications for a small number of species (e.g., *Miichthys miiuy* and *Sepiella japonica*). The recognition effect diagrams (Figures S7–S13) show that each model can clearly localize and correctly annotate individual fish bodies, with well-fitting bounding boxes. Among them, YOLOv11-ViT2 (as shown in Figure 6) achieves 100% accuracy for all nine species, representing the best performance.

**Figure 6.** Confusion Matrix Diagram of the YOLOv11-ViT2 Model for Single Fish Species.

3.5. Classification Results for the Multiple Classes Separated Scenario

In the multiple classes separated scenario, the overall recognition accuracy of all models remains at a high level (mostly exceeding 95%). As shown in the confusion matrices of Figures S14–S19, VGG and YOLOv8 exhibit slight confusion for a few species (e.g., *Trichiurus lepturus* and *Pampus argenteus*); YOLOv11-CBAM and the original YOLOv11 perform stably, with only slightly lower accuracy (approximately 89%) for species such as *Sepiella japonica* and *Trichiurus lepturus*; the CNN model shows a decrease in accuracy for *Pampus argenteus*. In contrast, YOLOv11-ViT2 (as shown in Figure 7) achieves over 98% accuracy for all species, with its confusion matrix being nearly diagonal. The recognition effect diagrams (Figures S20–S26) show that the models can accurately distinguish and localize multiple non-overlapping fish individuals, with bounding boxes not interfering with each other.

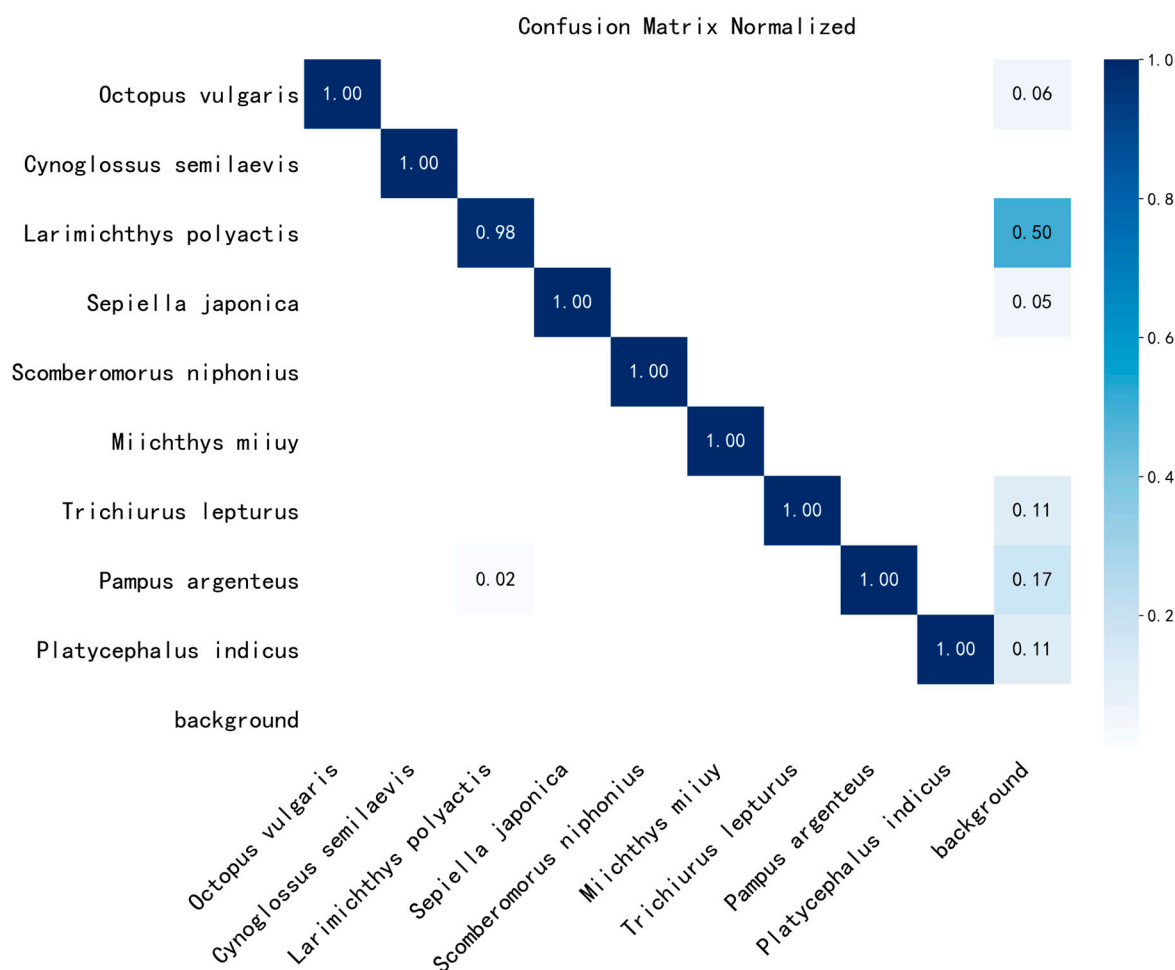


Figure 7. Confusion Matrix of the YOLOv11-ViT2 Model for Multiple Classes Separated.

3.6. Classification Results for the Slight Overlap of Multiple Fish Species Scenario

In the slight overlap scenario, the performance of the models begins to diverge. As shown in the confusion matrices of Figures S27–S32, the mAP50 values of models such as VGG, YOLOv8, and YOLOv11-CBAM drop to 0.70–0.85, and the recognition accuracy for some species (e.g., *Platycephalus indicus*, *Trichiurus lepturus*, and *Octopus vulgaris*) falls to around 70%, with increases in false positives and missed detections. YOLOv11-ViT1 performs slightly better than the baselines, but the improvement is limited. In contrast, YOLOv11-ViT2 (as shown in Figure 8) maintains over 98% accuracy across all categories, significantly outperforming the other models. The recognition effect diagrams

(Figures S33–S38) show that models such as YOLOv11-ViT1 tend to misidentify overlapping fish bodies as a single target or cause class confusion. In comparison, YOLOv11-ViT2 (as shown in Figure 9) can accurately delineate the bounding boxes of overlapping fish bodies and correctly assign category labels even under partial occlusion.

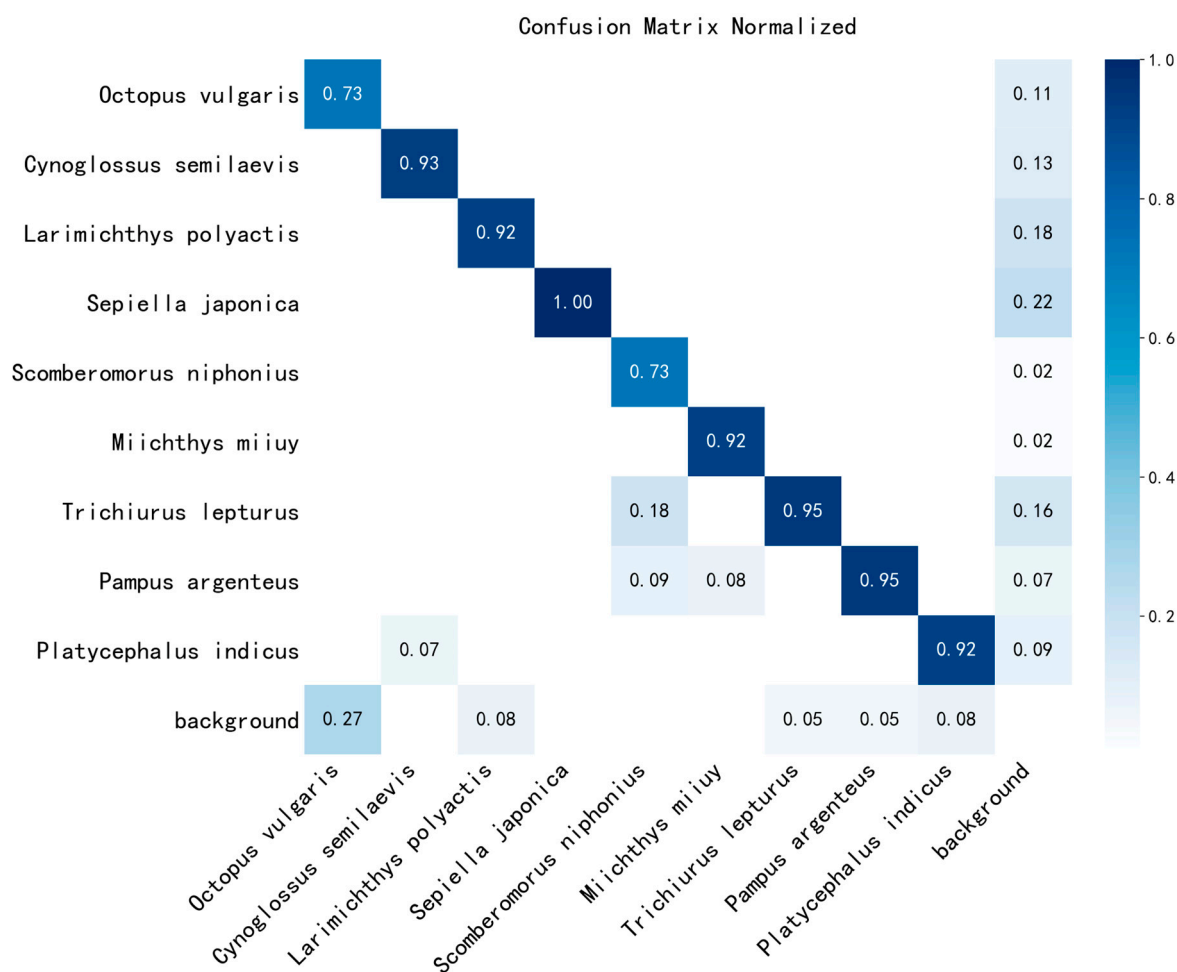


Figure 8. Confusion Matrix of the YOLOv11-ViT2 Model for The Slight Overlap of Multiple Fish Species Scenario.

3.7. Classification Results for the Severe Overlap of Multiple Fish Species Scenario

In the most challenging severe overlap scenario, the performance of most models drops sharply. As shown in the confusion matrices of Figures S40–S45, the accuracy of most categories for VGG, YOLOv8, YOLOv11-CBAM, the original YOLOv11, and CNN falls below 60%, with severe inter-class confusion. YOLOv11-ViT1 shows only limited improvement. In contrast, YOLOv11-ViT2 (as shown in Figure 10) demonstrates remarkable robustness, with accuracy exceeding 70% for most categories. Although the accuracy for *Trichiurus lepturus* and *Platycephalus indicus* is relatively lower, it still surpasses or matches the best levels achieved by other models. The recognition effect diagrams (Figures S46–S51) show that other models frequently suffer from missed detections and false positives (e.g., misidentifying *Pampus argenteus* as *Platycephalus indicus*). In comparison, YOLOv11-ViT2 (as shown in Figure 11) can accurately identify target species even when fish bodies are heavily overlapped and only local features (e.g., fins, tails) are exposed, with small bounding box localization deviations and confidence scores mostly remaining above 0.5. This fully demonstrates the effectiveness of embedding the ViT module at the end of the backbone for improving recognition capability in densely occluded scenarios.

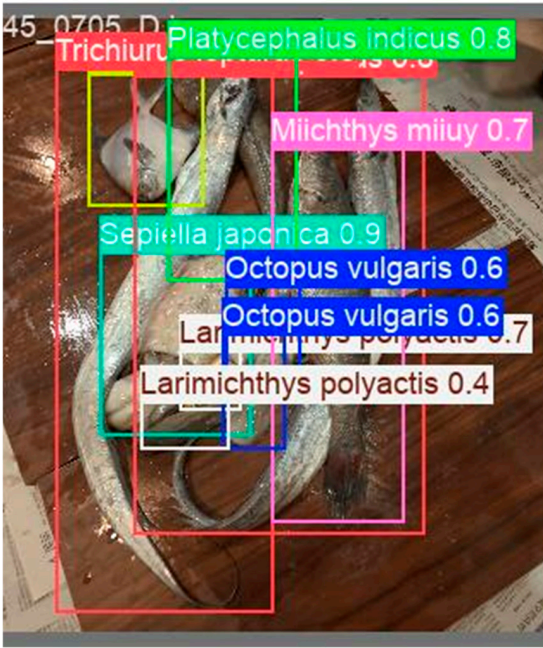


Figure 9. Recognition Effect Diagram of the YOLOv11-ViT2 Model for The Slight Overlap of Multiple Fish Species Scenario. Note: Different colors represent different fish species; the same color indicates the same species.

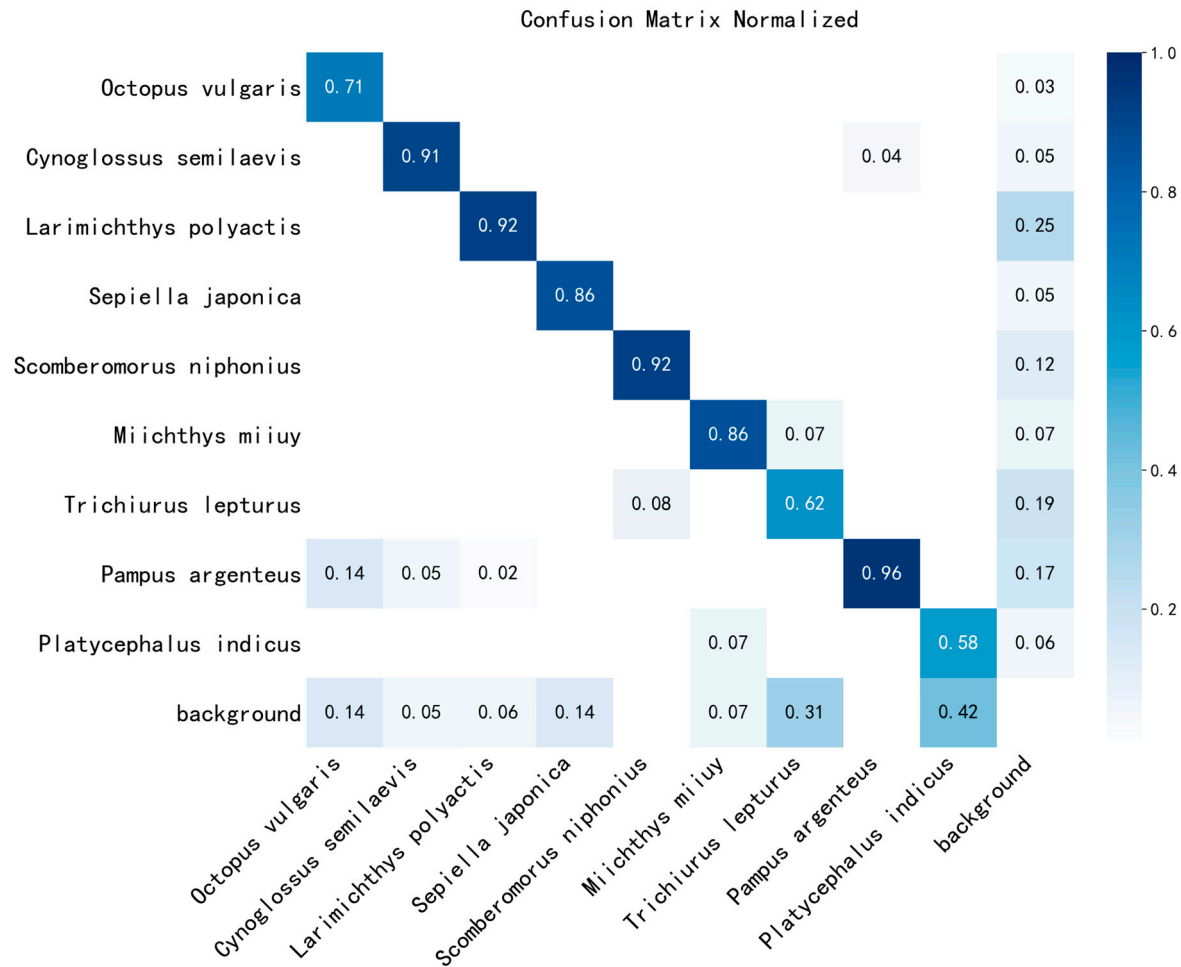


Figure 10. Confusion Matrix of the YOLOv11-ViT2 Model for the Severe Overlap of Multiple Fish Species.



Figure 11. Recognition Effect Diagram of the YOLOv11-ViT2 Model for the Severe Overlap of Multiple Fish Species. Note: Different colors represent different fish species; the same color indicates the same species.

4. Discussion

4.1. The Model's Recognition Performance in Complex Background

The YOLOv11-ViT model proposed in this study demonstrates high robustness and stability in complex backgrounds. Compared with existing research, this model exhibits more outstanding recognition capabilities in complex scenarios involving multi-target overlap, lighting variations, and deck glare. For example, studies by French (2020) [7] and Vilas (2020) [8] also indicate that complex backgrounds and target occlusion remain key factors limiting recognition accuracy. By introducing the ViT module, this study enhances the model's holistic feature perception ability in complex situations such as blurred fish edges, uneven lighting, and stacking occlusions. This enables the model not only to maintain high accuracy in simple scenarios but also to demonstrate more comprehensive detection capabilities and stability in complex scenarios, reflecting its strong adaptability in real-world fishing environments.

Ovalle et al. (2022) [9] conducted research on identifying caught fish in scenes with varying degrees of overlap based on the Mask R-CNN model. Their results show that when target overlap is low, the model achieves high recognition accuracy (average precision approximately 98%, recall approximately 95%). However, when individuals are severely overlapped, the system performance declines significantly, especially in the complex operational environments of commercial fishing vessels, where the Weighted Absolute Percentage Error (WAPE) reaches as high as 289%. The authors noted, "When individual overlap is significant, the model struggles to effectively detect and identify targets, and even well-trained human observers find it difficult to accurately identify targets from images."

In contrast, the YOLOv11-ViT2 model proposed in this paper demonstrates clear advantages in scenarios with substantial multi-target overlap. Experimental results show that under the condition of severe overlap of multiple fish species, YOLOv11-ViT2 achieves an mAP50 of 0.7867. Furthermore, in confusion matrix analysis, YOLOv11-ViT2 maintains recognition accuracy above 70% for most categories even in the scenario with severe overlap of multiple fish species, whereas the model in Ovalle et al.'s study [9] largely fails under similar conditions. This indicates that by embedding the ViT module at the end of YOLOv11's backbone, the model can better model global feature dependencies, enhancing its ability to discern occluded and overlapping targets, thereby mitigating to some extent the performance degradation of traditional detection models in dense target scenarios. In summary, while maintaining high accuracy in low-overlap scenarios, the

proposed model exhibits stronger adaptability and stability in medium- to high-overlap complex scenes, demonstrating potential for real-time recognition of densely caught fish in shipboard environments.

4.2. Model Improvement Effect and Performance Improvement Analysis

The performance improvement of the proposed YOLOv11-ViT framework is fundamentally driven by the complementary interaction between convolutional local feature extraction and Transformer-based global contextual modeling. In dense fish recognition scenarios, convolutional operations alone tend to focus on local texture and edge information, which becomes ambiguous when targets are heavily overlapped or partially occluded. The introduction of self-attention enables the model to establish long-range dependencies across spatially distant regions, thereby enhancing semantic coherence and reducing local feature confusion.

The comparison between YOLOv11-ViT1 and YOLOv11-ViT2 further indicates that the effectiveness of global attention is highly dependent on its integration stage within the detection architecture. Applying the ViT module at the detection head introduces global interactions after multi-scale feature fusion, which weakens the contribution of low-level spatial details and disrupts feature pyramid consistency. Consequently, this configuration yields limited gains and may even degrade performance in simpler scenarios.

By contrast, embedding the ViT module at the end of the backbone allows global contextual modeling to be incorporated during feature extraction, before multi-scale aggregation. This design preserves the hierarchical structure of the feature pyramid while enriching backbone features with global semantic information. As a result, the downstream neck and head modules operate on more discriminative and context-aware representations, leading to improved robustness in highly cluttered and overlapping scenes.

It is worth noting that YOLOv11-ViT2 not only significantly outperforms other models in detection metrics (mAP50, mAP50-95) but also maintains a leading position in classification metrics (Precision, Recall, F1), indicating that its performance improvement stems from effective modeling of global context, which both improves object localization (detection metrics) and reduces class confusion (classification metrics).

From the perspective of network architecture and feature propagation, the performance difference between YOLOv11-ViT1 and YOLOv11-ViT2 arises from the different timing of ViT module insertion in the detection pipeline, which directly affects the consistency and discriminability of multi-scale features.

YOLOv11-ViT1 places the ViT at the beginning of the Head, i.e., after the Backbone and Neck have completed multi-scale feature extraction and fusion, before introducing global self-attention. At this stage, the feature maps have already undergone multiple downsampling and convolution operations, resulting in relatively low spatial resolution (typically around 20×20), with severe loss of detailed information. Although the ViT can establish global dependencies on such low-resolution feature maps, its self-attention operation treats all patches equally, weakening the local localization cues (e.g., edges, textures) originally preserved by the Neck. More critically, the multi-scale features (P3, P4, P5) output by the Neck originally have different receptive fields and semantic levels, and the global modeling of the ViT disrupts this hierarchical structure, causing small-target information in shallow features to be overwhelmed by deep semantics. Consequently, YOLOv11-ViT1 shows no improvement or even performance degradation in simple scenarios (single fish species, multiple classes separated), and fails to fully leverage the advantages of global attention in complex scenarios.

YOLOv11-ViT2, in contrast, embeds the ViT at the end of the Backbone, before the Neck. At this stage, the Backbone has already extracted rich multi-scale features (still

maintaining relatively high spatial resolution and hierarchical structure). The ViT performs global self-attention on these features, simultaneously capturing long-range semantic correlations and local details without disrupting the subsequent feature pyramid aggregation process of the Neck. Specifically, after the feature maps output by the Backbone are enhanced by the ViT, each token contains global contextual information, yet the spatial dimensions do not change drastically. These globally contextualized features are then fed into the Neck for top-down and bottom-up multi-scale fusion, enabling small targets (e.g., fish fins, tails) to receive semantic support from larger target regions. In addition, YOLOv11-ViT2 introduces the C2fAttn module, which further strengthens local-global interaction while remaining lightweight. Therefore, YOLOv11-ViT2 can more accurately distinguish different individuals in scenarios with severe overlap and occlusion, reducing feature confusion and achieving optimal performance in both detection and classification metrics.

In summary, the ViT module should be placed before the construction of the feature pyramid (i.e., at the end of the Backbone) rather than after it (at the beginning of the Head). The former preserves the hierarchy of multi-scale features, allowing global context to “infuse” into each level of features; the latter disrupts the structure of the feature pyramid, causing interference between global attention and local localization information. This mechanistic analysis provides clear guidance for the design of hybrid CNN-Transformer detectors.

These findings suggest that, for one-stage object detectors deployed in complex environments, backbone-level integration of Transformer-based attention provides a more effective balance between global perception and multi-scale feature consistency. This insight offers a practical guideline for designing hybrid CNN-Transformer detection models targeting dense object distributions and severe occlusion.

Ablation experiments (Table 4) further quantify the contribution of each module. Introducing only the Backbone ViT (Model B) achieves a 10.5% mAP50 improvement in the scenario with severe overlap of multiple fish species, indicating that global self-attention is the core factor in overcoming occlusion. The C2fAttn module alone provides limited improvement (4.6% mAP50 gain), but when combined with the ViT, it yields an additional 2.0% gain, demonstrating its auxiliary role in local-global feature refinement. More importantly, the synergistic effect of ViT and C2fAttn is not a simple superposition: the F1 score of the full model (0.8792) far exceeds the sum of their individual effects ($0.8512 + 0.8428 - 0.8346 = 0.8594$), indicating that C2fAttn can more effectively aggregate discriminative features based on the global context provided by the ViT. This finding supports the design principle of this paper: “global context as the mainstay, local attention as the supplement.”

4.3. Limitations Regarding ViT Hyperparameter Selection

In the YOLOv11-ViT2 model proposed in this paper, the hyperparameters of the Vision Transformer (ViT) module (including embedding position, patch size, number of attention heads, number of encoder layers, and embedding dimension) were selected based on preliminary experiments and computational resource constraints, without conducting systematic ablation studies on all hyperparameters. It should be noted that the primary goal of this study is to explore the feasibility of integrating ViT with YOLOv11 and to reveal the optimal insertion position for global self-attention in a single-stage detector, rather than exhaustively searching for the optimal hyperparameter combination. Given limited computational resources, we prioritized ensuring the adequacy of core comparative experiments (different insertion positions, with/without ViT module, C2fAttn ablation, and systematic comparison with multiple baselines).

Specifically, the most critical design dimension—embedding position—has been validated through a direct comparison between YOLOv11-ViT1 (at the beginning of the Head) and YOLOv11-ViT2 (at the end of the Backbone). Experimental results show that YOLOv11-ViT2 significantly outperforms YOLOv11-ViT1 in the scenario with severe overlap of multiple fish species, and this conclusion is robust to the specific values of other hyperparameters. The remaining hyperparameters (e.g., 16 encoder layers, embedding dimension of 512, 4 attention heads, etc.) are consistent with mainstream ViT designs, ensuring the reproducibility of the model. Although the current configuration may not be globally optimal, the existing experimental evidence (Tables 1–4) sufficiently supports the core conclusion: “integrating ViT at the end of the Backbone can effectively improve fish recognition performance under complex backgrounds.”

It must be acknowledged that the lack of systematic ablation studies on ViT depth, patch size, embedding dimension, etc., is a limitation of this study. This limitation does not affect the validity of the current conclusions, but it suggests that when generalizing the model to other datasets or application scenarios, hyperparameter tuning for the specific task may be necessary. In the future, we will conduct systematic hyperparameter ablation experiments based on this work to explore superior ViT configurations.

4.4. The Actual Application of the Model Is Insufficient and Needs to Be Improved

While the YOLOv11-ViT model demonstrates strong recognition accuracy and stability in complex backgrounds, it still exhibits certain limitations. Firstly, the incorporation of the ViT module increases model complexity and computational load, leading to a slight reduction in inference speed, which makes it difficult to fully meet real-time recognition requirements for onboard terminals. Secondly, the dataset constructed in this study is based on static images and does not adequately account for dynamic factors in fishing videos, such as motion blur, water droplet glare, and continuous target movement. The model's stability in video sequences requires further validation.

Future research could focus on the following three areas:

- (1) Integrating lightweight Transformer architectures (such as Swin Transformer or MobileViT) to reduce model parameters and computational complexity while maintaining accuracy, thereby improving real-time inference performance;
- (2) Introducing temporal modeling mechanisms (e.g., 3D convolutions or temporal Transformers) to incorporate inter-frame correlation features during training, enabling dynamic recognition of fish movement behaviors and posture changes;
- (3) Fusing multi-modal information (such as infrared images, depth images, and environmental illumination data) to build a multi-source perception framework for fish recognition, enhancing model robustness in low-light, strong-glare, and severe occlusion scenarios.

Furthermore, by leveraging technologies such as 5G communication, the Internet of Things (IoT), and edge computing, the optimized YOLOv11-ViT model could potentially be deployed in intelligent fishing vessel terminal systems in the future. This could help enable real-time recognition, classification, and data reporting of catches, potentially contributing to the advancement of fisheries resource monitoring and management in China towards intelligent and automated operations, thereby offering a potential technical reference for sustainable fisheries utilization.

5. Conclusions

To address the issue of declining fish recognition accuracy under complex backgrounds, this study proposes an improved YOLOv11 recognition model based on the Vision Transformer. The main conclusions are as follows:

- (1) The improved YOLOv11-ViT model achieves superior comprehensive performance on the FishRecognition-2025 dataset. The recall rate is increased to 81.99%, mAP50-95 is improved to 0.6460, and the F1 score reaches 0.8731. The model demonstrates more stable recognition results in complex backgrounds.
- (2) The ViT module enhances the model's ability to distinguish overlapping and occluded targets through its global self-attention mechanism. This compensates for the limitations of YOLO series models in terms of local feature receptive fields, significantly reducing false detection and missed detection rates.
- (3) The improved model outperforms traditional convolutional models in simulated complex deck environments and stacked fish bodies, suggesting a promising technical direction for intelligent fisheries monitoring and automated identification of fishing data.

Future research can further optimize the model in areas such as model lightweighting, temporal feature fusion, and multimodal data application, which could eventually facilitate the practical implementation of integrated systems for real-time onboard recognition and remote monitoring.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/fishes11070385/s1>, Table S1: Explanation of real-time performance metrics; Table S2: AP50 per class for each model; Table S3: AP50-95 per class for each model; Figure S1: Confusion Matrix of VGG on Single Fish Species; Figure S2: Confusion Matrix of YOLOv8 on Single Fish Species; Figure S3: Confusion Matrix of YOLOv11-CBAM on Single Fish Species; Figure S4: Confusion Matrix of the YOLOv11 Model on Single Fish Species; Figure S5: Confusion Matrix of the CNN Model on Single Fish Species; Figure S6: Confusion Matrix of the YOLOv11-ViT1 Model on Single Fish Species; Figure S7: Recognition Effect Diagram of the VGG Model on Single Fish Species; Figure S8: Recognition Effect Diagram of the YOLOv8 Model on Single Fish Species; Figure S9: Recognition Effect Diagram of the YOLOv11-CBAM Model on Single Fish Species; Figure S10: Recognition Effect Diagram of the YOLOv11 Model on Single Fish Species; Figure S11: Recognition Effect Diagram of the CNN Model on Single Fish Species; Figure S12: Recognition Effect Diagram of the YOLOv11-ViT1 Model on Single Fish Species; Figure S13: Recognition Effect Diagram of the YOLOv11-ViT2 Model on Single Fish Species; Figure S14: Confusion Matrix of VGG on Multiple Classes Separated; Figure S15: Confusion Matrix of YOLOv8 on Multiple Classes Separated; Figure S16: Confusion Matrix of YOLOv11-CBAM on Multiple Classes Separated; Figure S17: Confusion Matrix of the YOLOv11 Model on Multiple Classes Separated; Figure S18: Confusion Matrix of the CNN Model on Multiple Classes Separated; Figure S19: Confusion Matrix of the YOLOv11-ViT1 Model on Multiple Classes Separated; Figure S20: Recognition Effect Diagram of the VGG Model on Multiple Classes Separated; Figure S21: Recognition Effect Diagram of the YOLOv8 Model on Multiple Classes Separated; Figure S22: Recognition Effect Diagram of the YOLOv11-CBAM Model on Multiple Classes Separated; Figure S23: Recognition Effect Diagram of the YOLOv11 Model on Multiple Classes Separated; Figure S24: Recognition Effect Diagram of the CNN Model on Multiple Classes Separated; Figure S25: Recognition Effect Diagram of the YOLOv11-ViT1 Model on Multiple Classes Separated; Figure S26: Recognition Effect Diagram of the YOLOv11-ViT2 Model on Multiple Classes Separated; Figure S27: Confusion Matrix of VGG on Slight Overlap of Multiple Fish Species; Figure S28: Confusion Matrix of YOLOv8 on Slight Overlap of Multiple Fish Species; Figure S29: Confusion Matrix of YOLOv11-CBAM on Slight Overlap of Multiple Fish Species; Figure S30: Confusion Matrix of the YOLOv11 Model on Slight Overlap of Multiple Fish Species; Figure S31: Confusion Matrix of the CNN Model on Slight Overlap of Multiple Fish Species; Figure S32: Confusion Matrix of the YOLOv11-ViT1 Model on Slight Overlap of Multiple Fish Species; Figure S33: Recognition Effect Diagram of the VGG Model on Slight Overlap of Multiple Fish Species; Figure S34: Recognition Effect Diagram of the YOLOv8 Model on Slight Overlap of Multiple Fish Species; Figure S35: Recognition Effect Diagram of the YOLOv11-CBAM Model on Slight Overlap of Multiple Fish Species; Figure S36: Recognition Effect Diagram of the YOLOv11 Model on Slight Overlap of Multiple Fish

Species; Figure S37: Recognition Effect Diagram of the CNN Model on Slight Overlap of Multiple Fish Species; Figure S38: Recognition Effect Diagram of the YOLOv11-ViT1 Model on Slight Overlap of Multiple Fish Species; Figure S39: Recognition Effect Diagram of the YOLOv11-ViT2 Model on Slight Overlap of Multiple Fish Species; Figure S40: Confusion Matrix of VGG on Severe Overlap of Multiple Fish Species; Figure S41: Confusion Matrix of YOLOv8 on Severe Overlap of Multiple Fish Species; Figure S42: Confusion Matrix of YOLOv11-CBAM on Severe Overlap of Multiple Fish Species; Figure S43: Confusion Matrix of the YOLOv11 Model on Severe Overlap of Multiple Fish Species; Figure S44: Confusion Matrix of the CNN Model on Severe Overlap of Multiple Fish Species; Figure S45: Confusion Matrix of the YOLOv11-ViT1 Model on Severe Overlap of Multiple Fish Species; Figure S46: Recognition Effect Diagram of the VGG Model on Severe Overlap of Multiple Fish Species; Figure S47: Recognition Effect Diagram of the YOLOv8 Model on Severe Overlap of Multiple Fish Species; Figure S48: Recognition Effect Diagram of the YOLOv11-CBAM Model on Severe Overlap of Multiple Fish Species; Figure S49: Recognition Effect Diagram of the YOLOv11 Model on Severe Overlap of Multiple Fish Species; Figure S50: Recognition Effect Diagram of the CNN Model on Severe Overlap of Multiple Fish Species; Figure S51: Recognition Effect Diagram of the YOLOv11-ViT1 Model on Severe Overlap of Multiple Fish Species; Figure S52: Recognition Effect Diagram of the YOLOv11-ViT2 Model on Severe Overlap of Multiple Fish Species..

Author Contributions: Conceptualization, X.H., S.Y. and W.W.; methodology, X.H., S.Y. and W.W.; software, X.H. and S.Y.; formal analysis, X.H. and S.Y.; data curation, X.H. and S.Y.; writing—original draft preparation, X.H.; writing—review and editing, X.H. and S.Y.; validation, S.Y., K.Z., F.W. and S.Z.; visualization, Y.D., F.W. and W.W.; supervision, K.J., X.H. and S.Y.; project administration, X.H. and S.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2024YFD2400801).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: If data are needed, interested parties may contact the corresponding author.

Acknowledgments: The authors would like to thank Wei Wang for his valuable suggestions.

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

References

1. Zhang, B.; Zhang, F.; Sun, Y.; Li, X.; Liu, P.; Liu, L.; Miao, Z. Underwater Target Recognition via Cayley-Klein Measure and Shape Prior Information in Hyperspectral Imaging. *Appl. Sci.* **2023**, *13*, 7854. [\[CrossRef\]](#)
2. Zhao, Y.Y.; Yang, P. Review of research on fish body length measurement based on machine vision. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 207–218.
3. Lu, Y.C.; Tung, C.; Kuo, Y.F. Identifying the species of harvested tuna and billfish using deep convolutional neural networks. *ICES J. Mar. Sci.* **2020**, *77*, 1318–1329.
4. Yusup, I.M.; Iqbal, M.; Jaya, I. Real-time reef fishes identification using deep learning. *IOP Conf. Ser. Earth Environ. Sci.* **2020**, *429*, 012046. [\[CrossRef\]](#)
5. Villon, S.; Mouillot, D.; Chaumont, M.; Darling, E.S.; Subsol, G.; Claverie, T.; Villéger, S. A deep learning method for accurate and fast identification of coral reef fishes in underwater images. *Ecol. Inform.* **2018**, *48*, 238–244. [\[CrossRef\]](#)
6. Knausgard, K.M.; Wiklund, A.; Sørvald, T.K.; Halvorsen, K.T.; Kleiven, A.R.; Jiao, L.; Goodwin, M. Temperate fish detection and classification: A deep learning based approach. *Appl. Intell.* **2022**, *52*, 6988–7001.
7. Bartholomew, D.C.; Mangel, J.C.; Alfaro-Shigueto, J.; Pingo, S.; Jimenez, A.; Godley, B.J. Remote electronic monitoring as a potential alternative to on-board observers in small-scale fisheries. *Biol. Conserv.* **2018**, *219*, 35–45. [\[CrossRef\]](#)
8. French, G.; Mackiewicz, M.; Fisher, M.; Holah, H.; Kilburn, R.; Campbell, N.; Needle, C. Deep neural networks for analysis of fisheries surveillance video and automated monitoring of fish discards. *ICES J. Mar. Sci.* **2020**, *77*, 1340–1353.
9. Vilas, C.; Antelo, L.T.; Martín-Rodríguez, F.; Morales, X.; Pérez-Martin, R.; Alonso, A.; Valeiras, J.; Abad, E.; Quinzan, M.; Barral-Martínez, M. Use of computer vision onboard fishing vessels to quantify catches: The iObserver. *Mar. Policy* **2020**, *116*, 103714. [\[CrossRef\]](#)

10. Ovalle, J.C.; Vilas, C.; Antelo, L.T. On the use of deep learning for fish species recognition and quantification on board fishing vessels. *Mar. Policy* **2022**, *139*, 105015. [[CrossRef](#)]
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 779–788.
12. Zhang, Z.; Qu, Y.; Wang, T.; Rao, Y.; Jiang, D.; Li, S.; Wang, Y. An improved YOLOv8n used for fish detection in natural water environments. *Animals* **2024**, *14*, 2022. [[CrossRef](#)] [[PubMed](#)]
13. Zouin, B.; Zahir, J.; Baletaud, F.; Vigliola, L.; Villon, S. Improving CNN fish detection and classification with tracking. *Appl. Sci.* **2024**, *14*, 10122. [[CrossRef](#)]
14. Yuan, H.C.; Tao, L. Detection and identification of fish in electronic monitoring data of commercial fishing vessels based on improved Yolov8. *J. Dalian Ocean Univ.* **2023**, *38*, 533–542.
15. Robillard, A.J.; Trizna, M.G.; Ruiz-Tafur, M.; Panduro, E.L.D.; de Santana, C.D.; White, A.E.; Dikow, R.B.; Deichmann, J.L. Application of a deep learning image classifier for identification of Amazonian fishes. *Ecol. Evol.* **2023**, *13*, e9987. [[CrossRef](#)] [[PubMed](#)]
16. Tseng, C.H.; Kuo, Y.F. Detecting and counting harvested fish and identifying fish types in electronic monitoring system videos using deep convolutional neural networks. *ICES J. Mar. Sci.* **2020**, *77*, 1367–1378. [[CrossRef](#)]
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
18. Gao, H.; Hu, C.; Han, G.; Mao, J.; Huang, W.; Guan, Q. Point-level feature learning based on vision transformer for occluded person re-identification. *Image Vis. Comput.* **2024**, *143*, 104929. [[CrossRef](#)]
19. Rodrigo, M.; Cuevas, C.; García, N. Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks. *Sci. Rep.* **2024**, *14*, 21392. [[CrossRef](#)] [[PubMed](#)]
20. Zhang, Z.; Han, S.; Liu, D.; Ming, D. Focus and imagine: Occlusion suppression and repairing transformer for occluded person re-identification. *Neurocomputing* **2024**, *578*, 127442. [[CrossRef](#)]
21. Latif, S.A.; Sidek, K.A.; Hashim, A.H.A. An efficient iris recognition technique using cnn and vision transformer. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2024**, *34*, 235–245.
22. Zheng, X.; Luo, Y.; Zhou, P.; Wang, L. Distilling efficient vision transformers from cnns for semantic segmentation. *Pattern Recognit.* **2025**, *158*, 111029. [[CrossRef](#)]
23. Xu, Y.; Li, J.; Dong, Y.; Zhang, X. Survey of development of YOLO object detection algorithms. *J. Front. Comput. Sci. Technol.* **2024**, *18*, 2221–2238.
24. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv* **2024**, arXiv:2410.17725.
25. Hussain, M. Yolov1 to v8: Unveiling each variant—a comprehensive review of yolo. *IEEE Access* **2024**, *12*, 42816–42833. [[CrossRef](#)]
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems 2017*; NeurIPS: La Jolla, CA, USA, 2017; Volume 30.

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