An Automatic Detection and Statistical Method for Underwater Fish Based on Foreground Region Convolution Network (FR-CNN)

Shenghong Li 1,2, Peiliang Li 1,2,*, Shuangyan He 1,2, Zhiyan Kuai 1,2, Yanzhen Gu 1, Haoyang Liu 1, Tao Liu 1, and Yuan Lin 1

1 Ocean College, Zhejiang University, Zhoushan 316021, China; 22234092@zju.edu.cn (S.L.); hesy@zju.edu.cn (S.H.); kuaizhiyan@zju.edu.cn (Z.K.); guyanzhen@zju.edu.cn (Y.G.); liuhaoyang@zju.edu.cn (H.L.); liutao08@zju.edu.cn (T.L.); samylin@zju.edu.cn (Y.L.)
2 Hainan Institute of Zhejiang University, Sanya 572025, China
* Correspondence: lipeiliang@zju.edu.cn; Tel.: +86-05-80-209235

Abstract: Computer vision in marine ranching enables real-time monitoring of underwater resources. Detecting fish presents challenges due to varying water turbidity and lighting, affecting color consistency. We propose a Foreground Region Convolutional Neural Network (FR-CNN) that combines unsupervised and supervised methods. It introduces an adaptive multiscale regression Gaussian background model to distinguish fish from noise at different scales. Probability density functions integrate spatiotemporal information for object detection, addressing illumination and water quality shifts. FR-CNN achieves 95% mAP with IoU of 0.5, reducing errors from open-source datasets. It updates anchor boxes automatically on local datasets, enhancing object detection accuracy in long-term monitoring. The results analyze fish species behaviors in relation to environmental conditions, validating the method’s practicality.

Keywords: underwater fish object detection; multiscale Gaussian background model; automated fish counting; deep learning

1. Introduction

The report by the Food and Agriculture Organization, titled “Outlook for Fisheries and Aquaculture 2030”, forecasts that aquaculture production is anticipated to reach 100 million tons by 2027, achieving this milestone for the first time, and is projected to rise further to 106 million tons by 2030 [1]. The increasing global population has resulted in a heightened demand for seafood, leading to the degradation of oceans and a continual decline in marine fisheries resources. To address this challenge, implementing modernized marine ranching systems is viewed as a promising approach to support the efficient restoration of marine fisheries resources and ecosystems [2]. Recently, the deployment of underwater observation networks has enabled the real-time monitoring of alterations in marine environments and the biodiversity of biological resources, presenting new opportunities for the scientific management of marine ranching. However, the underwater environment presents challenges for the statistical assessment of biological resources due to limited visibility caused by factors like low light conditions, scattering, and absorption resulting from suspended particles and algae in the water. Advancements in computer vision have led to the increasing use of deep learning-based object detection techniques in complex underwater settings, offering robust tools for detection and classification. This technology also assists in analyzing the population, behavior, and interactions of underwater organisms within their habitats, encompassing research on fish diseases and responses to hypoxic conditions [3].

The underwater fish resource assessment in the rapid development of marine ranching has primarily undergone three stages.
The first stage involves traditional counting methods, primarily relying on fishing techniques such as angling and trawling. Fishermen estimate the quantity of cultured fish through fishing, a process that incurs significant time and labor costs while also impacting the welfare and growth of the fish [4].

The second stage entails using counters, sonar, and other sensor-based counting methods. With advancements in information technology, counting devices based on sensor principles and acoustic technologies have been widely applied for fish identification and counting in this phase. A number of researchers utilize sensors based on different principles for counting, such as infrared optical counters [5] and resistivity fish counters [6]. The optics counter has limited transparency in turbid water, which depends on the channel structure to limit fish movement. As a result, the counting accuracy is affected by fish overlapping and swimming back and forth, resulting in estimation errors. Visible light is restricted by distance when imaging underwater, and the increased distance will lead to blurred images and fewer images. The advantages of sound waves, such as long-distance propagation, position them as the optimal method for remote detection and identification of underwater targets. Particularly in underwater environments where light conditions change over time, acoustic imaging has become the scholars’ preferred approach to obtaining information about underwater targets. Counting methods based on acoustic technology, unaffected by distance variations in underwater imaging, can be categorized into acoustics and acoustic imaging. These methods are employed for underwater measurements and counting. DIDSON is a high-definition imaging sonar that can provide high-quality acoustic images under dark and turbid underwater conditions [7–12]. However, studies based on DIDSON demonstrate poor performance in detecting and counting small fish. While optical and acoustic methods can achieve fish counting, they cannot classify fish into different categories.

The third stage involves using deep learning-based underwater video automatic resource counting methods for identification and enumeration. Traditional counting methods and sensor-dependent approaches face challenges such as time efficiency, cost-effectiveness, and classification recognition. Deep learning is crucial in underwater video automatic resource counting, leveraging its outstanding capabilities in adaptive feature extraction and non-linear mapping. With the rapid advancement of hardware devices and their widespread deployment in underwater observation, devices such as underwater cameras, sonar, and underwater drones enable real-time visualization of underwater biological resources, promoting scientific fisheries management and sustainable production [13–15]. Object detection, a computer vision technique for locating and classifying semantic objects in given images or videos, is integral to this stage. Presently, object detection algorithms have found widespread applications in various domains, including face recognition [16,17], text detection [18,19], pedestrian recognition [20,21], and vehicle detection [22,23], among others [24]. With the rapid development of deep learning-based target detection technology, these algorithms surpass human visual accuracy and find extensive applications in aquaculture, such as fish counting, body length measurement, and individual behavior analysis. In contrast to terrestrial environments, the lenses of underwater cameras are prone to influences from sediment and biological attachments like algae. Simultaneously, the turbidity of water exhibits significant variations over time, leading to pronounced changes in the illumination and color distribution of underwater images, as illustrated in Figure 1. However, the rapid development of attached sensors and advancements in underwater image enhancement makes it feasible to achieve automatic resource counting based on underwater videos.
Figure 1. The underwater image at different times in the same observation place. (a) The lens without attachments in a relatively clear water environment. (b) The lens polluted by silt and attached by algae. (c) The relatively turbid water environment. (d) The RGB color histogram of (a). (e) The RGB color histogram of (b). (f) The RGB color histogram of (c).

In recent years, with the advancements in computer vision, deep learning-based object detection technologies have played a crucial role in fish counting. Image segmentation struggles to accurately delineate targets in complex underwater environments, impacting counting precision. In contrast, object detection, by defining bounding boxes around object positions and categories, achieves object quantity by counting the number of bounding boxes, exhibiting higher generalization capability and accuracy. Anchor-based object detection frameworks can be categorized into two-stage and one-stage detection frameworks. The primary distinction between one-stage and two-stage object detection frameworks is that a one-stage detector network is an end-to-end model, directly regressing and outputting the positions and categories of detected objects.

In contrast, unlike its one-stage counterpart, a two-stage detector implements a two-step strategy involving anchor box filtering from coarse to fine and regression output from fine to coarse. This strategy sacrifices speed to enhance precision. Classic two-stage frameworks include Faster R-CNN [25] and Mask R-CNN [26], which have been applied in detecting and counting fish in underwater environments. Classic one-stage detection frameworks include the You Only Look Once (YOLO) series, YOLOv1–YOLOv10 series [27–36], and SSD [37] network structures.

Many researchers have applied object detection techniques to counting fish in aquaculture, utilizing both image- and video-based analysis approaches.

Based on images, researchers employ segmentation and detection techniques for fish counting. However, underwater images pose blurriness, noise, fish occlusion, and overlapping, making computer vision applications for counting inherently challenging. French et al. [38] proposed a method using the N4-field algorithm for scene segmenta-
tion and counting, achieving a calculation error for individual fish ranging between 2% and 16%. Nevertheless, the fish datasets they processed were obtained beneath the deck, exhibiting relatively high contrast, which differs from datasets acquired in nature underwater environments. To extend model applicability to complex underwater environments, Li et al. [39] introduced a fish detection system based on Fast R-CNN, which outperformed the Deformable Parts Model (DPM) in terms of mean Average Precision (mAP). The use of Selective Search for searching Regions of Interest (ROI) resulted in redundant computations of target features, impacting the training and inference speed of the object detection network. Faster R-CNN [40] incorporated a Region Proposal Network (RPN) to enhance detection speed, achieving end-to-end training. Li et al. [25] accelerated underwater fish detection using Faster R-CNN [40], obtaining an Average Precision (mAP) of 82.7%, with a detection time of one-third that of Fast R-CNN. However, calculating the number of fish based on frame-by-frame image analysis cannot track individual trajectories without incorporating an inter-frame individual matching algorithm. This omission may result in significant counting errors.

Based on video analysis for detection and counting, fish appearance and motion information can be fused for real-time fish detection. At the same time, the association of targets helps prevent repetitive counting in tracking. Ditria et al. [26] applied Mask R-CNN in aquatic ecology, demonstrating that deep learning exhibits higher accuracy and speed in fish abundance calculation. Arvind et al. [41] combined Mask R-CNN instance segmentation with the Generic Object Tracker for fish detection and tracking in large ponds and tanks. The study indicated that using a multi-region parallel detection method yielded the best results, achieving a fish detector with an F1 score of 0.91 at a rate of 16 frames per second. However, the limitation of this approach lies in the inability of drones to detect fish in turbid water environments. Mohamed et al. [42] employed a combination of multiple algorithms in fish farms, integrating the detection algorithm MSR-YOLO with optical flow algorithms for fish tracking. While the method performed well in fish farms, there was a significant gap compared to real underwater environments. Liu et al. [43] proposed a real-time fish detection and tracking method (RMCF) for multi-class fish counting in the real underwater environment of marine ranches, using YOLOv4 as the backbone network, achieving a recognition accuracy of 95.6%. However, the method’s recognition accuracy decreases when the experimental area changes. Additionally, the motion of fish fitted using a constant speed model may differ from the irregular activities of real underwater fish. In challenging underwater environments, Liu et al. [44] addressed the fish detection and segmentation difficulty by proposing an adaptive multiscale background modeling method applied to deep-sea cage videos. They utilized an online segmentation algorithm for fish detection and counting, maintaining robust results even without additional annotated videos for training and fine-tuning the framework. However, segmentation results from the background model pose the following issues: (1) The inference speed is relatively slow, rendering real-time computation unfeasible. (2) The segmentation of foreground moving objects can only achieve single-class semantic segmentation but not multi-class instance segmentation. (3) The segmentation result of a single moving object may correspond to several connected domains of different foreground objects, which leads to false statistics of the number of objects.

Similar to the multiscale regression Gaussian background model proposed by Liu et al. [44], our designed network simplifies the background model and the multiscale regression block Gaussian background model into a single-scale block Gaussian background model. Moreover, in the second stage of object detection, the network can determine the type of connected domains, eliminate all noisy regions, and refine the position of the target bounding boxes.

However, numerous challenges remain in underwater video detection and recognition algorithms. The primary reasons for these challenges include the following:

Variability in the distribution of video and image samples input to the network. In real underwater environments, especially with continuously sampled video data, the presence
of foreign objects and changes in water turbidity can significantly alter the illumination, color, and other characteristics of the images [45,46]. This leads to a decline in the model’s inferential capability over time [47,48]. One solution is to collect a sufficiently diverse training dataset that covers various periods and underwater conditions, simulating additional samples through data augmentation. However, due to the inability to encompass all possible scenarios, the model may exhibit limited adaptability to certain underwater environments. Regular analysis of misclassified samples and fine-tuning is essential to enhance the model’s adaptability to new samples.

To address these challenges, we propose a two-stage object detection network, named the Foreground Region Convolution Network (FR-CNN), designed to detect and enumerate fish in open sea areas. Meanwhile, our network eliminates the need for anchor frames. The FR-CNN utilizes foreground moving object segmentation results obtained through background modeling to establish a target region extraction network. Subsequently, a convolutional network is employed to extract features from candidate regions, facilitating target classification and refinement of the position coordinates for bounding boxes.

In our proposed two-stage object detection network FR-CNN, we use a combination of unsupervised and supervised approaches to identify and count fish species accurately. First, image pre-processing and enhancement techniques, including image clarity, contrast and brightness enhancement, and the creation of a dynamic background representation using a multiscale regression Gaussian background model, are used to overcome problems associated with low visibility, lighting variations, and turbidity. Foreground segmentation is then used to isolate potential fish targets from the background, and background subtraction techniques are used to identify regions of motion and serve as candidate regions for further analysis by the convolutional network. In the feature extraction and object detection stages, a convolutional neural network (CNN) is used to extract high-level features, and bounding boxes are generated using an anchor-free approach. A classification network is implemented for species identification to classify detected objects into specific fish species based on the extracted features. Finally, the number of fish is counted directly through the bounding boxes generated by the detection network, and the counts are aggregated across frames to provide a comprehensive assessment of the dynamics and behavior of the fish population. By combining these methods, FR-CNN effectively translates the signals of underwater images into fish species identification and counting, overcoming the challenges posed by the underwater environment and improving the accuracy and efficiency of marine resource assessment.

Overall, the contributions of this paper mainly include the following aspects:

1. We propose FR-CNN as an anchor-free two-stage object detection network. This network combines the advantages of both unsupervised and supervised methods, utilizing a supervised approach in the first stage of object detection for extracting target features from candidate regions. The unsupervised method effectively alleviates the prior errors introduced by publicly available datasets in the supervised stage, dynamically correcting for existing datasets and improving the precision of fine screening.

2. We introduce a multiscale regression Gaussian background model as an unsupervised method for dynamic calibration of existing datasets. In the context of two-stage object detection, addressing the issue of prior errors associated with anchor boxes on publicly available datasets, we employ an unsupervised approach that facilitates the automatic update of anchor boxes on local datasets, ensuring a closer alignment with the color variations resulting from underwater video imagery illumination changes.

The remaining sections of this paper are organized as follows: Section 2 offers a detailed introduction to the FR-CNN framework. Section 3 presents the experimental results, comparing them with various object detection networks on the validation set and providing discussions. Finally, Section 4 concludes the entire paper.
2. Proposed Methods

2.1. Observation Equipment and Datasets

The video data of fish in open sea area is obtained from the submarine cable online observation system (122°41′29.29″ E 30°11′34.77″ N) of Zhongjieshan Islands National Marine Ranch in Dongji Island, Zhejiang Province (Figure 2). The submarine cable online observation system comprises three subsystems: the control center human–machine inter-action information management subsystem, the power information transmission subsystem, and the submarine observation subsystem. The submarine observation subsystem, in particular, comprises a submarine observation platform and observation equipment sensors. The observation equipment sensors include an acoustic doppler current meter, multi-parameter water quality instrument, underwater high-definition camera, supplementary light lamp, water pump, and other equipment. These components enable the long-term, continuous, real-time monitoring of environmental parameters and underwater biological resource conditions. The integrated camera, specifically the OTWHC-500 underwater camera, supports 30× optical zoom, offers an approximate field of view of 84°, and operates within a pressure-resistant range of 1000 to 6000 m. It facilitates real-time 1080p video output. The accompanying supplementary lights ensure clarity and contrast in the underwater video, mitigating issues such as color distortion due to varying light wavelengths underwater and discontinuities in video brightness and clarity resulting from inconsistent lighting conditions during day and night. The water pump periodically cleans contaminants or attachments on the camera lens.

The underwater videos at the Dongji Island site utilize cameras directly positioned on the seabed, equipped with supplementary lights for illumination compensation. The shape and texture details of the fish in the images are notably clear. Object detection-based statistical methods are more optimal in open sea environments, where various fish species coexist with frequently changing sizes and postures. We extracted 8000 representative frames from the November 2020 underwater observation video at Dongji Island. We performed object detection annotation, encompassing labels for five categories: Thamnacoanus modestus, Microcanthus strigatus, Oplegnathus fasciatus, water plants, and unidentified fish species. For statistical analysis, we selected three typical fish species: Thamnacoanus modestus, Microcanthus strigatus, and Oplegnathus fasciatus (Figure 3). The rationale for choosing these three fish species for object detection and statistical analysis lies in the placement of the observation equipment at the seabed of the Dongji Island site, primarily capturing information about bottom-dwelling fish. Among the existing observation videos, these three fish species exhibit the highest occurrence frequency, rendering them of practical significance for subsequent statistical analysis work.

![Figure 2. The location of the submarine cable online observation system in the Dongji Island site in Zhejiang province and the submarine observation platform.](image-url)
2.2. Overall Architecture of Foreground Region Convolution Network

The FR-CNN proposed in our work employs a foreground-moving object segmentation framework and an object region extraction network constructed using convolutional networks (Figure 4). The FR-CNN network consists of three components: a multiscale feature fusion extraction network, a classification and position regression network, and a foreground target candidate box generation network. The multiscale feature fusion extraction network, represented by the blue and green regions of the Feature Pyramid Network (FPN) [49] and Region Proposal Network (RPN) [40], processes input video frames (the \( k \)-th frames) through five residual blocks in FPN [49]. This process extracts features at different scales, generating feature maps P2, P3, P4, and P5, corresponding to the smallest, second smallest, medium, and larger scales, respectively. P5 is max-pooled to obtain an even larger scale feature map, P6. Feature maps undergo a series of convolution and pooling operations from bottom to top, gradually reducing resolution, and are combined with the top-down generated feature maps, forming a multiscale feature pyramid. RPN integrates features from the five feature maps generated by FPN [49] and operates in parallel for two purposes. One feature map serves as candidate regions inputted into subsequent object detection networks for detection and classification.

Figure 3. (a) Thamnaconus modestus; (b) Microcanthus strigatus; (c) Oplegnathus fasciatus.

Figure 4. The architecture of foreground region convolutional neural network.
In contrast, the other serves as the input image for the foreground target candidate box generation network. The classification and position regression network, indicated by orange dashed lines, performs binary classification (background or foreground) on preset anchor boxes and regresses the coarse position of target bounding boxes. This process obtains all foreground target regions in the image and is used for multi-classification and precise regression of bounding box positions in the second stage. The foreground target candidate box generation network, shown by red dashed lines (Figure 4), takes foreground target regions generated by the background model as ROI regions to extract target features. It utilizes information from video frames ranging from the \((k - k_0)\) th to the \((k - 1)\) th frame for updating the background model. Simultaneously, the generated foreground target regions undergo morphological processing to merge adjacent small, connected regions, reducing the number of foreground targets. This can generate coarse-grained anchor boxes based on motion information, and after processing with a candidate box enhancement algorithm, serve as candidate boxes for the second-stage precise regression.

### 2.3. Multiscale Feature Fusion Extraction Network

During the extraction of target feature maps, we employ the ROI Align operation. The ROI Align operation refrains from rounding during the subdivision of sub-regions, as illustrated by the yellow dashed lines within the red ROI region (Figure 5). The red ROI region is directly partitioned into equidistant 3 × 3 sub-regions. Assuming a sampling point count of 2 × 2 for each sub-region, a further subdivision into a 2 × 2 grid is performed. Bilinear interpolation is then applied to determine the values corresponding to the centers of each grid within the sub-region. Subsequently, the average of the values at the centers of the four grids within the sub-region yields the aligned value. The resulting feature map for each augmented candidate box attains a final height and width of 7 (Figure 5), which illustrates the interpolation computation process for the ROI Align operation.

![Figure 5. The calculating flow of interpolation in ROI align.](image)

As shown in the right half of the blue dashed box in Figure 4, FPN [49] aims to enhance the semantic information of feature maps of different scales by fusing the output of the backbone network from the bottom up. Since shallower feature maps contain less semantic information, FPN [49] can propagate the rich semantic information from deeper to shallower feature maps through this bottom-up fusion process. Specifically, this is achieved by up sampling the feature map from the subsequent layer to match the feature map size from the preceding layer, typically using nearest neighbor interpolation.

In FPN, we use ResNet-50 as the backbone network to extract multiscale feature maps. The changes in feature map dimensions are detailed in Table 1. The input image size is 1920 × 1080 × 3, and the feature map sizes after processing through each layer of ResNet-50 are C2 \((480 \times 270 \times 256)\), C3 \((240 \times 135 \times 512)\), C4 \((120 \times 68 \times 1024)\) and C5 \((60 \times 34 \times 2048)\). The specific feature fusion steps are as follows:
Table 1. Feature map dimension changes.

<table>
<thead>
<tr>
<th>Layer Name</th>
<th>Input Size</th>
<th>Operation Type</th>
<th>Kernel Size</th>
<th>Stride</th>
<th>Output Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1</td>
<td>1920 × 1080 × 3</td>
<td>Convolution</td>
<td>7 × 7</td>
<td>2</td>
<td>960 × 540 × 64</td>
</tr>
<tr>
<td>Max Pooling</td>
<td>960 × 540 × 64</td>
<td>Pooling</td>
<td>3 × 3</td>
<td>2</td>
<td>480 × 270 × 64</td>
</tr>
<tr>
<td>Conv2_x</td>
<td>480 × 270 × 64</td>
<td>Bottleneck Block</td>
<td>1 × 1, 3 × 3, 1 × 1</td>
<td>1</td>
<td>480 × 270 × 256</td>
</tr>
<tr>
<td>Conv3_x</td>
<td>480 × 270 × 256</td>
<td>Bottleneck Block</td>
<td>1 × 1, 3 × 3, 1 × 1</td>
<td>2</td>
<td>240 × 135 × 512</td>
</tr>
<tr>
<td>Conv4_x</td>
<td>240 × 135 × 512</td>
<td>Bottleneck Block</td>
<td>1 × 1, 3 × 3, 1 × 1</td>
<td>2</td>
<td>120 × 68 × 1024</td>
</tr>
<tr>
<td>Conv5_x</td>
<td>120 × 68 × 1024</td>
<td>Bottleneck Block</td>
<td>1 × 1, 3 × 3, 1 × 1</td>
<td>2</td>
<td>60 × 34 × 2048</td>
</tr>
</tbody>
</table>

In the top-down path, first set the number of channels of C5 to 256 by 1 × 1 convolution to obtain P5 (60 × 34 × 256). Next, P5 is up sampled to 120 × 68 and added, element by element, to the C4 feature map, whose channel count is adjusted to 256 by 1 × 1 convolution, to obtain P4 (120 × 68 × 256). Similarly, up sample P4 to 240 × 135 and add it element by element to the C3 feature map, whose channel count is adjusted to 256 by 1 × 1 convolution, to obtain P3 (240 × 135 × 256). Finally, P3 is up sampled to 480 × 270 and added, element by element, to the C2 feature map, whose channel count is adjusted to 256 by 1 × 1 convolution to obtain P2 (480 × 270 × 256).

Following the fusion of feature maps of different scales in the above top-down pathway, we further enhance the semantic information of the shallow feature maps through the bottom-up pathway. Specifically, the deeper feature maps (e.g., P3) are up sampled to the size of the shallower feature maps (e.g., P2) and fused. This bottom-up fusion process achieves the up sampling through nearest-neighbor interpolation, ensuring that the rich semantic information from the deep feature maps can be effectively propagated to the shallow feature maps.

In our study, the backbone network is ResNet-50. Within this network, each residual stage is preceded by a convolution operation with a stride of two, which down-samples the height and width of the feature map by half. Consequently, each subsequent layers’ size of the feature map output is one-fourth that of the previous layer. Therefore, the feature map from the subsequent layer can be up sampled using nearest-neighbor interpolation to double its height and width. However, direct fusion is not feasible due to the mismatch in the number of channels between different layers’ output feature maps. A lateral connection operation is necessary to fuse feature maps with different channel numbers.

In ResNet, the number of channels in the output feature map of each residual stage doubles compared to the preceding stage. Thus, a 1 × 1 convolution is applied to the feature map from the subsequent layer to reduce its number of channels to half, matching the number of channels of the feature map from the preceding layer. The two feature maps can then be fused by element-wise addition. Subsequently, a 3 × 3 convolution is applied to eliminate the aliasing effects caused by up sampling the feature map.

In the FPN [49], the authors also propose a method for combining FPN [49] with Faster R-CNN [40], where feature maps at different levels are used for extracting ROI features based on the size of the ROI. Specifically, smaller ROIs should utilize feature maps from shallower layers for feature extraction, while larger ROIs should use feature maps from deeper layers. This approach is based on the observation that deeper feature maps correspond to larger receptive fields in the original image. For smaller objects, deeper feature maps contain less relevant information about the target, making it difficult to extract sufficient information for effective learning and inference. The selection of feature map levels is determined by Formula (1), where \( k_0 \) represents the baseline value. At the same time, \( w \) and \( h \) correspond to the width and height of the ROI (Region of Interest) area, respectively. The default value for \( k_0 \) is set to 5, chosen due to ResNet’s pretraining on the ImageNet dataset [50]. Notably, ResNet utilizes the feature map from the fifth layer (depicted as P5 in the blue-shaded region of Figure 4) when processing standard input images sized at 224 × 224 for image classification. In this study, adhering to Formula (1),
the candidate boxes enhanced at multiple scales are assigned to different feature maps based on their respective area sizes to extract target features.

\[
k = \left( k_0 + \log_2 \left( \sqrt{w \cdot h} / 224 \right) \right)
\]

2.4. Classification and Location Regression Networks

The orange dashed-boxed region (Figure 4) represents the final output layer of the foreground region convolutional network. This layer is primarily employed for the multi-class classification of extracted target features and the precise adjustment of target candidate box positions to generate the ultimate predictions for target bounding boxes.

In the cited paper [51], employing a feature-sharing output layer was observed to degrade the accuracy of both object classification and position regression tasks. Consequently, this study adopts an output layer that separates classification and regression features to achieve the ultimate goals of classification and localization.

For the segregated feature output layer, the convolutional operations uses two distinct \(1 \times 1\) convolutional kernels to generate two new feature maps of equal size. Then, these maps are input into two branches, one for object classification and the other for regression of object bounding box positions. The feature maps from both branches subsequently undergo multiple convolutional operations, followed by flattening, and are input into different fully connected layers. This process yields two distinct \(d2\)-dimensional feature vectors for generating multi-class probability predictions for the corresponding targets and offset values for the target bounding boxes, respectively. Here, the softmax function is defined by Formula (2), where \(V\) is a column vector, and \(V_i\) denotes the vector element at index \(i\).

This function is a normalization function, ensuring that the sum of all elements in vector \(S\) after normalization equals 1. This guarantees that the sum of multi-class probabilities for the target equals 1, with vector \(S\) corresponding to the predicted probability values for different target categories. The index with the highest predicted probability represents the predicted class of the target.

The sigmoid function, as defined by Formula (3), accommodates \(x\) within the interval \((-\infty, +\infty)\). This function confines output values to the range \([0, 1]\), corresponding to the interval \([0, 1]\) for the position offset of bounding boxes.

\[
S_i = e^{v_i} / \sum e^{v_j}
\]

\[
S(x) = 1 / (1 + e^{-x})
\]

2.5. Foreground Object Candidate Box Generation Network

The Adaptive Multiscale Gaussian Background (AMGB) model [52] is established to differentiate foreground fish from background noise at different scales. Based on the probability density function, the background model incorporates both spatial and temporal distribution information of foreground and background. We have simplified the AMGB model in the following ways: (1) The foreground objects generated by the background model may include underwater noise. The object detection network, which we designed, classifies the connected components of the foreground generated by the background model in the second stage of the convolutional network, discriminating valid targets and removing underwater noise. (2) The background model’s foreground object output segmentation accuracy does not need to be excessively high, tolerating slight deviations. Our designed object detection network fine-tunes the position of object bounding boxes in the second stage, ensuring accurate final output of object positions.

To reduce the computational burden of the background modeling process, the multiscale regression block Gaussian background model is simplified to a single-scale block Gaussian background model, utilizing only the smallest scale block for background modeling according to Formula (4) [44]. To further reduce computation, during background modeling, the height and width of the original video frame, initially \(1920 \times 1080 \times 3\), are
down sampled by 1/4. Thus, the output size of the background model is only 1/16 of the original image. This significantly reduces the computational workload of background modeling and, to some extent, mitigates the interference of underwater noise particles in background modeling. Figure 6 shows (a) the original input image and (b) the foreground target segmentation results generated by the single-scale block Gaussian background model. It can be observed in (b) that, apart from the three-object fish, numerous smaller foreground objects exist. Some smaller objects are floating particles in the water, while others are shrimp larvae. Due to their distinct volume compared to object fish, these noise objects can be easily filtered out using a simple area threshold.

\[
p^k_{i,j} = \begin{cases} 
1, & \sum_{c=1}^{3} \text{num}\left(\frac{|b^{k,c}_i - \mu_{k}^{c}|}{\sigma_k} < \sigma_k\right) > p \\
0, & \text{otherwise}
\end{cases}
\]  

\[
(4)
\]

**Figure 6.** (a) The origin input video frame of the size of 1920 × 1080 × 3. (b) The foreground object segmentation by single scale block-based Gaussian background model.

Following the paradigm of Faster R-CNN [40] two-stage object detection, we have improved the detection network, enabling the feature extraction network to utilize foreground object regions generated by the background model as Regions of Interest (ROI) for extracting object features. As illustrated in the red dashed box (Figure 4), with each new video frame input into the background model, the block-wise statistical information of the new frame is immediately utilized to update the background model. The current video frame’s foreground segmentation object regions are generated using Formula (4) [44]. Here, simple morphological operations are applied to the generated foreground object regions to merge neighboring small, connected components into larger ones, reducing the number of connected components in the foreground object regions. This approach also helps address the striped fish being split into multiple scattered connected components due to the dark stripes closely resembling the background.

The merging operation does not require extreme precision, meaning it is not necessary to ensure that all connected components corresponding to the same fish are merged into a single complete connected component. This is primarily because of the subsequent presence of the candidate box enhancement algorithm, as detailed in Section 2.6. It is crucial to note that, during the background modeling phase, the original video frames are down sampled to 1/16 of their original size to reduce computational load. Therefore, it is necessary to restore the bounding boxes corresponding to the merged foreground connected components back to the original image size. If the coordinates of the top-left and bottom-right corners of the bounding box for a foreground connected component generated by the single-scale block Gaussian background model are \((x_1, y_1)\) and \((x_2, y_2)\), respectively, then the coordinates after restoration to the original image size would be \((4x_1, 4y_1)\) and \((4x_2, 4y_2)\).

2.6. Candidate Box Enhancement Algorithm

We propose a further multiscale candidate box enhancement algorithm to enhance the integration of the foreground candidate box generation network and the FPN [49].
feature fusion network to effectively extract target features. This algorithm augments the foreground object bounding boxes obtained in Section 2.5 at different scales. The augmented candidate boxes are mapped onto feature maps fused through FPN [49] at different levels based on their area, facilitating the extraction of object features. This algorithm significantly improves the recall rate of object detection and effectively addresses the issue of overlapping multiple objects.

Due to the one-to-one correspondence between foreground object-connected regions generated by the background model and object-bounding boxes, there is a potential risk of missing detections, especially when connected regions encompass multiple objects or when a region contains multiple objects of different sizes. Single scale bounding boxes may struggle to handle such scenarios. Therefore, we introduce a candidate box augmentation algorithm to augment the foreground object bounding boxes generated by the background model at different scales. The red region (Figure 7) represents a candidate box extracted by the foreground candidate box extraction network, while two differently sized orange regions correspond to candidate boxes augmented at multiple scales. Assuming the coordinates of the original candidate box’s top-left and bottom-right points are \((x_1, y_1)\) and \((x_2, y_2)\), respectively, and the length scaling factor for multiscale augmentation is \(\alpha\), the coordinates of the corresponding augmented candidate box’s top-left and bottom-right points \((x_1', y_1')\) and \((x_2', y_2')\) satisfy Equations (5) and (6).

\[
\begin{align*}
\left\{ x_1' &= 0.5 \times [(1 + \alpha) \cdot x_1 + (1 - \alpha) \cdot x_2] \\
y_1' &= 0.5 \times [(1 + \alpha) \cdot y_1 + (1 - \alpha) \cdot y_2] \\
x_2' &= 0.5 \times [(1 - \alpha) \cdot x_1 + (1 + \alpha) \cdot x_2] \\
y_2' &= 0.5 \times [(1 - \alpha) \cdot y_1 + (1 + \alpha) \cdot y_2]
\right.
\end{align*}
\] (5)

2.7. Multitask Loss Function

Two tasks are involved for object detection: object classification and object localization, each with distinct evaluation metrics. The classification loss is measured using the multiclass cross-entropy loss function, as Formula (7) indicates. Here, \(L_i\) represents the loss for the \(i\) sample. When the class label for the \(i\) sample is \(c\), \(y_{i,c} = 1\) is equal to 1; otherwise,
\( y_{i,c} = 0 \) is 0. \( i \) denotes the probability of the \( i \) sample being assigned to class \( c \). Additionally, for the \( i \) sample, it holds that \( \sum_{c=1}^{M} p_{i,c} = 1 \).

\[
L_{\text{cls}} = \frac{1}{N} \sum_{i=1}^{N} L_i = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{M} y_{i,c} \log(p_{i,c})
\]

(7)

In this paper, the position loss of the candidate box and truth box is calculated based on CIoU, as shown in Formula (8). Here \( \alpha \) is the balancing coefficient used to adjust the proportion of the regularization center point difference and aspect ratio similarity in the loss function, where \( A_w, A_h, B_w, \) and \( B_h, \) respectively, denote the width and height of bounding boxes \( A \) and \( B \).

\[
\text{CIoU} = \text{DIoU} - \alpha \frac{4}{\pi^2} \left[ \arctan \left( \frac{A_w}{A_h} \right) - \arctan \left( \frac{B_w}{B_h} \right) \right]^2
\]

(8)

Integrating the classification loss for multi-class object detection and the bounding box position loss, the final loss function for training the foreground region convolutional network is represented by Formula (9), where \( L_{\text{cls}} \) corresponds to the classification loss function and \( L_{\text{reg}} \) corresponds to the bounding box position loss function.

\[
L = L_{\text{cls}} + L_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{M} y_{i,c} \log(p_{i,c}) + (1 - \text{CIoU})
\]

(9)

3. Results and Discussion

3.1. Experimental Results

**Dataset.** We extracted 8000 representative frames from the underwater observation video of Dongji Island in November 2020 and performed object detection annotation. The annotations encompass five categories: *Thamnaconus modestus*, *Microcanthus strigatus*, *Oplegnathus fasciatus*, aquatic plants, and an unspecified category for unidentified fish species. The annotated dataset is partitioned into training and validation sets in a 7:3 ratio. Specifically, 5600 annotated images are randomly selected for training the foreground region convolutional network, while the remaining 2400 annotated images are reserved for validating the performance of the object detection network. The annotated quantities for each fish category are detailed in Table 2.

<table>
<thead>
<tr>
<th>Type of Fish</th>
<th>Label the Number of Samples (Strips)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Thamnaconus modestus</em></td>
<td>2160</td>
</tr>
<tr>
<td><em>Microcanthus strigatus</em></td>
<td>1128</td>
</tr>
<tr>
<td><em>Oplegnathus fasciatus</em></td>
<td>1456</td>
</tr>
<tr>
<td>unknown</td>
<td>256</td>
</tr>
</tbody>
</table>

**Training.** The feature extraction network of our proposed FR-CNN uses ResNet-50 as the backbone network and incorporates the FPN [49] structure. During the training process, we can directly load the pre-trained weights of Faster R-CNN [40] on the COCO object detection dataset [11] as the initial weights for the feature extraction network, facilitating transfer learning and mitigating the issue of limited training samples [12].

Furthermore, we introduce commonly used data augmentation techniques to augment the training dataset, including random rotations, brightness adjustments, contrast, saturation, hue, Gaussian blurring, and Cutout [10]. Initially, we employ a foreground candidate box generation network based on background modeling to generate candidate boxes for foreground targets in all images of the training and validation sets. Subsequently, we perform multiscale augmentation on the candidate boxes. The augmented foreground candidate boxes undergo feature extraction through the backbone network, enabling the
second stage of object classification and candidate box position regression. This training approach also contributes to reducing the required training time for the network, as it avoids redundant foreground candidate box generation and multiscale candidate box augmentation steps.

We use the same hyperparameters as in the Faster R-CNN [40] + FPN [49] detection network. For details, see the Detectron2 project published by Facebook [53]. The following are the detailed settings during the training process:

**Dataset:** In total, 8000 annotated frames extracted from the underwater observation video of Dongji Island are used, divided into 5600 training images and 2400 validation images in a ratio of 7:3.

**Pre-trained weights:** The Faster R-CNN [40] weights pre-trained on the COCO dataset are loaded to facilitate transfer learning.

**Model structure:** ResNet-50 is the backbone network, combined with the FPN [49] structure.

**Optimizer:** The Adam optimizer is used, with the initial learning rate set to 0.001 and the learning rate decaying during training.

**Learning rate scheduler:** A learning rate scheduler is built to manage learning rate changes during training.

**Batch size:** The batch size is set to 16 during training.

**Number of training epochs:** The training process is set to 300 epochs.

**Hardware:** Training and inference are performed on a server with two Intel(R) Xeon(R) Gold 6342 CPUs and four NVIDIA A100 40 GB GPUs.

**Loss function:** Combining the classification loss and the bounding box position loss of multi-class object detection, the final loss function is used to train the weights of the foreground convolutional network, as shown in Formula (9) in Section 2.7.

**Inference.** During the inference phase, the foreground candidate box generation network dynamically generates foreground target candidate boxes for each video frame based on background modeling. After augmentation with multiple scales, these candidate boxes are input into the deep convolutional network, enabling end-to-end object detection.

**Results on the test set.** The multi-class average precision at the IoU threshold of 0.5 (mAPIoU = 0.5) of our proposed FR-CNN detection network is the highest, reaching 95%, as shown in Table 3. The average precision for detecting *Thamnacorus modestus, Microcanthus strigatus*, and *Oplegnathus fasciatus* is 98%, 91%, and 96%, respectively, as shown in Table 4. YOLOv5 achieves the second-highest multi-class average precision at 93%, while Faster R-CNN [40] + FPN [49] exhibits the lowest at 90%, respectively, as shown in Table 3. Regarding inference speed, YOLOv5 demonstrates the highest speed, averaging 52 FPS, followed by Faster R-CNN [40] + FPN [49] at 40 FPS, and FR-CNN with the lowest speed at 32 FPS. This is primarily attributed to the fact that the background modeling computation is performed on the CPU, which cannot leverage GPU parallel acceleration, leading to a certain performance bottleneck.

| Table 3. The comparison of detection results of different object detection models. |
|---------------------------------|----------------|----------------|
| **Object Detection Network**    | **mAPIoU = 0.5** | **Inference Speed (FPS)** |
| Faster R-CNN [40] + FPN [49]    | 0.90           | 40             |
| YOLOv5                          | 0.93           | 52             |
| FR-CNN                          | 0.95           | 32             |

| Table 4. APIoU = 0.5 comparison of different fish categories in object detection. |
|---------------------------------|----------------|
| **Type of Fish**                | **APIoU = 0.5** |
| *Thamnacorus modestus*          | 0.98           |
| *Microcanthus strigatus*        | 0.91           |
| *Oplegnathus fasciatus*         | 0.96           |
Device. The network training and inference in the device experiment are carried out on the server configured with Intel (R) Xeon (R) Gold 6342 CPU × 2 and NVIDIA A100 40 GB × 4.

The fish object detection results for the underwater monitoring video of Dongji Island are illustrated in Figure 8. The detection results show that the FR-CNN detection network performs well in the face of variations in object pose, changes in brightness, partial occlusion, and size variations caused by varying distances from the camera. In most cases, the confidence scores for the detected objects are consistently above 0.85, indicating robust detection performance under diverse conditions.

![Fish Detection Results](image)

Figure 8. The fish object detection results in Dongji Island. The top left label of each bounding box is its category and classification confidence. The Thamnaconus modestus is class 0, the Microcanthus strigatus is class 1, and the Oplegnathus fasciatus is class 4. (a) The detection results for class 0 (Thamnaconus modestus). (b) The detection results for class 1 (Microcanthus strigatus). (c) The detection results for both class 0 (Thamnaconus modestus) and class 1 (Microcanthus strigatus). (d) The detection result for class 4 (Oplegnathus fasciatus).

The primary purpose of designing the FR-CNN detection network is to address the challenge posed by the shifting illumination and color distribution in underwater monitoring video images over time, which can impact object detection accuracy. To validate the effectiveness of FR-CNN in addressing this issue, 1000 frames containing three fish species were randomly selected from the underwater monitoring videos of Dongji Island in December 2020 and January 2021. These frames are manually annotated as ground truth samples. Subsequently, three detection networks, namely Faster R-CNN [40] + FPN [49], YOLOv5, and FR-CNN, perform object detection on the additional 2000 frames. The detection metrics are calculated and compared for different months, specifically the multi-class average precision at IoU threshold of 0.5 (mAPIoU = 0.5). Due to the lack of manual maintenance of the observation equipment from November 2020 to January 2021, sediment accumulation on the underwater camera lens became more pronounced over time.

Additionally, algae and other organisms’ attachment becomes more severe, as depicted in the upper half of Figure 9. Therefore, evaluating the performance of the three networks on frames from different months during this period effectively demonstrates the generalization and robustness of the detection networks when facing changes in illumination and color distribution in video images. As shown in the lower half of Figure 9, over time, both YOLOv5 and Faster R-CNN [40] + FPN [49] exhibit a noticeable decrease...
in object detection performance, with a maximum decrease of 5% in mAPIoU = 0.5. In contrast, the detection performance of the proposed FR-CNN network remains relatively stable, with a maximum decrease of only 1% in mAPIoU = 0.5. This experiment indicates that the FR-CNN object detection network proposed in this study can effectively be used for long-term, continuous, and real-time multi-class fish object detection and statistics in underwater monitoring videos.

![Figure 9. Object detection results of mAPIoU = 0.5 of three detection networks in different months.](image)

3.2. Discussion

Due to the observational equipment at the Dongji Island site being specifically designed for monitoring various fish species in the open sea and constrained by the camera’s visual range, it can only achieve localized visualization monitoring of the seabed conditions. In this scenario, fish frequently enter or leave the shooting area, which hinders the statistical analysis of fish resources. Even with continuous tracking of detected fish objects through methods such as object tracking, there is no guarantee that fish leaving and re-entering the shooting area can be correctly re-matched to associated targets. Therefore, directly using quantity metrics based on multi-category fish target detection in monitoring videos for statistical analysis is unreliable. To address this issue, we convert the total quantity of different fish species into the time-averaged frequency of being captured by the camera in that area. This frequency represents the average number of occurrences per second for a particular fish species in the visual area, providing a basis for statistical analysis. This metric effectively indicates the local area’s activity level or relative abundance of fish species. Specifically, for the same fish species, the average value of this statistical indicator over a period reflects the activity level of that species in the corresponding area during that time interval.

After manually reviewing and analyzing the underwater monitoring videos from the Dongji Island site spanning from November 2020 to January 2021, it was observed that the video quality was optimal in November 2020, with the highest clarity and contrast. Moreover, the frequency of various fish species was relatively high during this period. Therefore, utilizing the FR-CNN object detection network proposed in this study, we conduct object detection on all monitoring videos for November’s three specified fish categories. The goal is to calculate the time-averaged frequency of their occurrences in the monitoring videos. A compromise was made because the object detection model was not performing real-time online inference, and the computational load for continuous monthly
video detection was substantial. Despite the FR-CNN model’s ability to meet real-time multi-category fish object detection and quantity statistics at an average inference speed of 32 FPS, completing the fish object detection and statistical task for a month of video would require nearly a month. To address this, the volume of the video data to be analyzed is reduced. Frames from the monitoring videos are sampled at a frame rate of 1 FPS, and object detection is then performed on the sampled video frames. This approach reduces the detection time to approximately 1/30 of the original time. Additionally, the continuous monitoring videos are divided into multiple consecutive time segments for parallelized object detection, reducing the time required for detection.

As shown in Figures 10 and 11, the daily and hourly average statistical results of the occurrence frequency of three fish species in the underwater monitoring videos at the Dongji Island site in November 2020 are presented. *Thamnaconus modestus* is a warm-water fish species that predominantly inhabits near-bottom layers, exhibiting strong thermal preferences. It has the highest yield in the East China Sea, with its fishing season typically commencing in mid to late November. This may explain why *Thamnaconus modestus* has the highest average occurrence frequency among the three fish species in the November underwater monitoring videos at Dongji Island, with daily and hourly average values of 0.66 fish/second and 0.64 fish/second, respectively. These values significantly surpass those of the other two statistically analyzed fish species. *Thamnaconus modestus* exhibits nocturnal bottom-dwelling behavior and diurnal surface swimming, aligning with the hourly average statistical results of the occurrence frequency shown in Figure 11. In the underwater areas of Dongji Island in November 2020, *Thamnaconus modestus* exhibits higher hourly average occurrence frequencies during the periods of 0:00–10:00 and 19:00–20:00.

*Oplegnathus fasciatus* is a nearshore, warm-temperate, rocky reef-dwelling fish species with a wide distribution range, primarily found in the Yellow Sea, East China Sea, and Taiwan Strait. However, it does not exhibit a distinct fishing season. In the November underwater monitoring videos at Dongji Island, the average occurrence frequency of *Oplegnathus fasciatus* is second only to that of *Thamnaconus modestus*, with daily and hourly average values of 0.23 fish/second and 0.25 fish/second, respectively. The statistical results indicate that the frequency of occurrence of *Oplegnathus fasciatus*, which is mainly higher during the night, is offset from that of *Thamnaconus modestus*. The high occurrence frequency periods for *Oplegnathus fasciatus* in the underwater area are mainly concentrated between 11:00–18:00 and 21:00–23:00.

*Microcanthus strigatus* is a warm-water fish species living in rocky reefs near the shore, mainly distributed in the South China Sea and less commonly found in the East China Sea and Yellow Sea. In the November underwater monitoring videos at Dongji Island, *Microcanthus strigatus* has the lowest average occurrence frequency, with daily and hourly average values of 0.18 fish/second and 0.19 fish/second, respectively. Higher hourly average occurrence frequencies in the underwater area occur during the periods of 0:00–3:00 and 13:00–14:00. Compared to the other two fish species, *Microcanthus strigatus* has a shorter span of high occurrence frequency periods. This may be attributed to the lower abundance of *Microcanthus strigatus* in the Zhoushan waters, resulting in a lower probability of being recorded in the monitoring videos. *Microcanthus strigatus* adults tend to appear near artificial structures. As a fish species with occurrence frequencies second only to those of *Thamnaconus modestus* and *Oplegnathus fasciatus* in the underwater monitoring videos, this finding also suggests that artificial reefs placed in the nationally designated marine ranch of the Zhoushan Archipelago have a certain attractiveness for rocky reef-dwelling fish species, providing them with suitable shelter and foraging grounds.
Constrained by the natural conditions of real water bodies, it is not feasible to artificially manipulate various environmental factors in different aquatic environments and analyze their impact on different fish species’ movement characteristics and behavioral patterns. Temperature, salinity, and dissolved oxygen are the primary environmental factors influencing the composition and distribution of fish species in local marine areas [23]. Therefore, this study utilizes the fish activity level metric proposed earlier to analyze the correlation between the activity levels of three benthic fish species at the Dongji Island site in November 2020 and environmental factors. As described in Section 2.1 of this paper, the underwater cabled observation system at the Dongji Island site not only facilitates long-term, continuous, real-time visualization monitoring of underwater biological resources but
also concurrently measures various environmental parameters in the local water, including temperature, salinity, pH, chlorophyll, and dissolved oxygen. Table 5 presents the hourly average range and mean values of various environmental factors in the seawater at the Dongji Island site in November 2020. Figure 11 illustrates the hourly average variation curves of these environmental factors. It can be observed that the underwater seawater environmental factors at the Dongji Island site exhibit relatively small variations within a day during this month. Table 6 displays the correlation analysis results between the activity levels of three different fish species and underwater seawater’s hourly average environmental factors. The activity level of *Thamnaconus modestus* shows a higher correlation with the underwater seawater environmental factors compared to the other two fish species. The activity level of the *Thamnaconus modestus* exhibits a significant correlation with changes in seawater salinity and dissolved oxygen and a significant correlation with seawater temperature. Notably, it shows a positive correlation with dissolved oxygen changes and a negative correlation with temperature and salinity changes. Considering the relatively small range of changes in hourly average environmental factors for this month and the fish’s ability to adapt to minor variations in environmental factors, the higher correlation between *Thamnaconus modestus*’s activity level and changes in seawater salinity, dissolved oxygen, and temperature is likely attributed to its ecological habits. Given that *Thamnaconus modestus* has a nocturnal bottom- dwelling and diurnal rising ecological behavior, exhibiting higher activity levels in the underwater area during the night, and considering the lower temperature, lower salinity, and higher dissolved oxygen concentrations in the underwater seawater at night, as shown in Figure 12, the significant correlation between *Thamnaconus modestus*’s activity level and changes in seawater salinity, dissolved oxygen, and temperature is likely more influenced by its ecological habits. Compared to *Thamnaconus modestus*, the correlation between the activity levels of *Microcanthus strigatus* and *Oplegnathus fasciatus* and environmental factors is slightly lower. This may be attributed to two reasons: (1) The sensitivity of *Microcanthus strigatus* and *Oplegnathus fasciatus* to changes in environmental factors with a small range of hourly average variations within a day is relatively low. (2) *Microcanthus strigatus* and *Oplegnathus fasciatus* have lower frequencies of occurrence in the monitoring videos compared to *Thamnaconus modestus*, which may have a certain impact on the accuracy of the correlation analysis results. Specifically, the activity level of *Microcanthus strigatus* exhibits a highly significant positive correlation with changes in chlorophyll concentration, while the activity level of *Oplegnathus fasciatus* shows a significant positive correlation with changes in salinity and a significant negative correlation with changes in dissolved oxygen.

Table 5. Hourly average minimum, maximum, and average value of seawater environmental factors in the Dongji Island site in November 2020.

<table>
<thead>
<tr>
<th>Environmental Parameters</th>
<th>Minimum Value</th>
<th>Maximum Value</th>
<th>Average Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water depth (m)</td>
<td>11.762</td>
<td>13.367</td>
<td>12.835</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>20.448</td>
<td>21.133</td>
<td>20.844</td>
</tr>
<tr>
<td>Salinity (psu)</td>
<td>28.645</td>
<td>30.517</td>
<td>29.561</td>
</tr>
<tr>
<td>pH</td>
<td>7.900</td>
<td>7.910</td>
<td>7.906</td>
</tr>
<tr>
<td>Chlorophyll (mg/m³)</td>
<td>0.913</td>
<td>1.051</td>
<td>0.989</td>
</tr>
<tr>
<td>Dissolved oxygen (mg/L)</td>
<td>7.988</td>
<td>8.286</td>
<td>8.129</td>
</tr>
</tbody>
</table>

Table 6. Correlation coefficients between the activity of three kinds of fish and the hourly average of seawater environmental factors in the Dongji Island site in November 2020.

<table>
<thead>
<tr>
<th>Type of Fish</th>
<th>Temperature</th>
<th>Salinity</th>
<th>pH</th>
<th>Chlorophyll</th>
<th>Dissolved Oxygen</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Thamnaconus modestus</em></td>
<td>−0.685 *</td>
<td>−0.654 **</td>
<td>0.112</td>
<td>−0.342</td>
<td>0.708 **</td>
</tr>
<tr>
<td><em>Microcanthus strigatus</em></td>
<td>0.346</td>
<td>0.182</td>
<td>0.228</td>
<td>0.502 **</td>
<td>−0.310</td>
</tr>
<tr>
<td><em>Oplegnathus fasciatus</em></td>
<td>0.351</td>
<td>0.487 *</td>
<td>0.003</td>
<td>−0.048</td>
<td>−0.402 *</td>
</tr>
</tbody>
</table>

Note: ** denotes a statistically significant correlation at the 0.01 level (two-tailed); * denotes a statistically significant correlation at the 0.05 level (two-tailed).
Type of Fish | Temperature | Salinity | pH | Chlorophyll | Dissolved Oxygen
--- | --- | --- | --- | --- | ---
Thamnaconus modestus | $-0.688 *$ | $-0.654 **$ | $0.112$ | $0.708 **$ | $-0.342$
Microcanthus strigatus | $0.346$ | $0.182$ | $0.228$ | $0.502 **$ | $-0.310$
Oplegnathus fasciatus | $0.351$ | $0.487 *$ | $0.003$ | $-0.048$ | $-0.402 *$

Note: ** denotes a statistically significant correlation at the 0.01 level (two-tailed); * denotes a statistically significant correlation at the 0.05 level (two-tailed).

Figure 12. The changing curves of multiple hourly average environmental factors in the Dongji Island site in November 2020.

4. Conclusions

With the continuous advancement of computer vision technology, object detection techniques have been widely applied in aquaculture, particularly in fish counting. This study proposes a video-based automatic fish detection and analysis method for open-sea scenarios. Based on the statistical results of fish species in open-sea marine ranches, we analyze three benthic fish species’ ecological habits and activity variations. Additionally, we attempt to explore the correlation between the activity levels of these fish species and underwater environmental factors, as well as the sensitivity of their activity levels to changes in underwater environmental factors. The main aspects covered are as follows:

(1) We introduce an improved two-stage object detection network, FR-CNN, designed to detect multiple fish species. We classify and analyze the activity changes of three benthic fish species based on the detection results. Experimental results demonstrate that FR-CNN effectively alleviates the problem of image feature shift caused by long-term underwater video monitoring under changing lighting and water quality conditions, thereby improving object detection accuracy. Furthermore, the multiscale candidate box enhancement algorithm exhibits improved matching of foreground candidate boxes at different scales and fusion with feature maps, alleviating training difficulties and accuracy issues associated with too many predefined anchor boxes.

(2) We propose a multi-category fish statistical indicator for open-sea bottom monitoring videos, namely the time-averaged occurrence frequency of fish, to measure the activity levels of different fish species. By analyzing the activity levels of three typical benthic fish species at the Dongji Island site in November 2020, we found variations in their activity levels during different periods, which are correlated with their ecological habits. Additionally, we analyze the hourly average variations in underwater environmental factors, revealing correlations between the activity levels of different fish species and environmental factors. However, further monitoring data from more months and water quality conditions are needed to verify the accuracy of these conclusions.

We have preliminarily explored the correlation between three benthic fish species and underwater environmental factors. However, due to the limited duration of monitoring data and sparse prior research, an in-depth analysis of the reasons for these correlations...
was not conducted. Therefore, obtaining monitoring data spanning a longer period is necessary to validate and further analyze the results. This study primarily aims to address the influence of complex aquatic environments on the effectiveness of detection statistical methods. Future research can delve deeper into the impact of oceanic physical processes on fish statistics and environmental factor variations.

**Author Contributions:** Conceptualization, S.L., S.H. and Y.L.; methodology, S.L., S.H., Y.G. and T.L.; software, S.L. and Y.G.; writing—original draft, S.L. and H.L.; supervision, P.L. and Y.L.; data curation, Z.K.; investigation, Z.K.; experiment, H.L.; validation, H.L. and T.L.; writing—review & editing, P.L. and T.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (Grant No. 2021JLH0053, supported by Mingjun He), and the scientific and technological projects of Zhoushan (Grant Nos. 2021CXLH0020 and 2022C01004, supported by Yanzhen Gu).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Acknowledgments:** We would like to express our sincere gratitude to the following funding projects for their support for this study: the Hainan Provincial Joint Project of Sanya Yazhou Bay Science and Technology City (Grant No. 2021JLH0053 and 2021CXLH0020), and the scientific and technological projects of Zhoushan (2022C01004). We also acknowledge any additional support provided that is not covered in the Funding or Author Contribution sections, including administrative and technical assistance, and donations in kind. We would also like to thank the Hainan Yazhou Bay Ecological Environment and Fishery Resources Observation and Research Station for its support. All these supports played a vital role in the smooth progress of this study.

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

**References**

1. FAO. *The State of World Fisheries and Aquaculture (SOFIA)*; FAO: Roma, Italy, 2022. [CrossRef]

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.