mathematics

Article

Multibranch Attention Mechanism Based on Channel and Spatial Attention Fusion

Guojun Mao [1,2], Guanyi Liao [2], Hengliang Zhu [2] and Bo Sun [2,*]

[1] Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China
[2] School of Computer and Mathematics, Fujian University of Technology, Fuzhou 350118, China
[*] Correspondence: sunbo@fjut.edu.cn

check for updates

Citation: Mao, G.; Liao, G.; Zhu, H.; Sun, B. Multibranch Attention Mechanism Based on Channel and Spatial Attention Fusion. *Mathematics* **2022**, *10*, 4150. https://doi.org/10.3390/math10214150

Academic Editor: Radu Tudor Ionescu

Received: 17 September 2022
Accepted: 2 November 2022
Published: 6 November 2022

Abstract: Recently, it has been demonstrated that the performance of an object detection network can be improved by embedding an attention module into it. In this work, we propose a lightweight and effective attention mechanism named multibranch attention (M3Att). For the input feature map, our M3Att first uses the grouped convolutional layer with a pyramid structure for feature extraction, and then calculates channel attention and spatial attention simultaneously and fuses them to obtain more complementary features. It is a "plug and play" module that can be easily added to the object detection network and significantly improves the performance of the object detection network with a small increase in parameters. We demonstrate the effectiveness of M3Att on various challenging object detection tasks, including PASCAL VOC2007, PASCAL VOC2012, KITTI, and Zhanjiang Underwater Robot Competition. The experimental results show that this method dramatically improves the object detection effect, especially for the PASCAL VOC2007, and the mapping index of the original network increased by 4.93% when embedded in the YOLOV4 (You Only Look Once v4) network.

Keywords: object detection; multiscale module; spatial attention; channel attention; spatial attention; multibranch structure

MSC: 37M10

1. Introduction

Object detection is a fundamental task in the field of computer vision, where the main task is to locate all objects of interest in an image and determine their type and location [1]. An important challenge in this phase is to improve the detection accuracy of high-noise videos and images [2]. The noise environment is complicated, containing issues such as bad weather and blurred video obtained underwater; low-quality images obtained as a result of the image's size, shape, and position, as well as the lighting and shooting conditions; and interference factors such as occlusion and background. However, these noises are unavoidable and are the major cause of missed and false detections, so improving object detection performance in the presence of noise is an urgent problem.

Using attention mechanisms to improve the performance of object detection networks has been widely recognized [3]. The intuitive interpretation of the attention mechanism is to efficiently allocate limited computational resources to the analysis of salient regions of an object and, therefore, to improve the accuracy of object detection. This is consistent with the human visual system, which tends to focus on the useful information parts of an image and ignore the irrelevant information parts, although the overall framework of existing machine vision is still more focused on the holistic analysis of the image, and the detection accuracy of the large object is generally reasonable; however, the detection accuracy of the small and medium objects is poor. Therefore, introducing an attention mechanism can compensate for this disadvantage to a certain extent.

Attention mechanisms can be broadly classified into channel attention and spatial attention [4]. Channel is a feature detector, and channel attention is a mechanism to mine a set of representative features in a given image. The typical channel attention is the squeeze-and-excitation (SE) module [5], whose central idea is to learn the weights of different channels by compression and excitation operation and highlight the significant features. However, the disadvantages of SE are also apparent. It ignores the importance of spatial information [6]. Therefore, the bottleneck attention module (BAM) [7] and convolutional block attention module (CBAM) [8] can better combine channel attention and spatial attention to enrich feature maps. The above work on the attention mechanism is practical; however, there are still two fundamental problems to be solved. Firstly, determining how to mine and utilize the rich information in feature maps at different scales, and secondly, channel or spatial attention can only establish short-term channel dependence but cannot develop long-range channel dependence. Aiming at the above two problems, scholars have proposed Res2Net [9] from the multiscale aspect and nonlocal neural networks [10] from the long-range channel dependence aspect, respectively. Although the above two methods solve the problem to a certain extent, they bring a heavy computational burden to the network. Therefore, based on the above description, we believe it is necessary to develop attention that has low cost and combines multiscale feature extraction and long-range channel dependence. In this paper, we propose an effective and low-cost attention mechanism named the multibranch attention mechanism (M3Att). Our M3Att can process the input tensor at different scales. Specifically, our M3Att combines three parts: firstly, we use group convolutions with different sizes to build a pyramid structure and then enrich the feature map information after grouping convolutions through the channel shuffling mechanism. Then, the feature maps that pass the grouped convolutional pyramid are sent into the channel and spatial attention, respectively. Finally, we use the softmax function to realize the attention weights, thus establishing the long-range channel dependence. At the same time, introducing a skip connection can better compensate for the information loss problem after multiple convolutions.

In this paper, a multibranch attention (M3Att) mechanism that merges channel attention and spatial attention is proposed to address the two problems mentioned above, and the main contributions of this paper are summarized as follows.

(1) We propose a new multibranch attention mechanism (M3Att), which can be flexibly incorporated into existing object detection networks and improves the performance of the object detection network without a significant increase in the parameters of the network. Similarly, our M3Att can also be extended to other computer vision tasks.

(2) We propose a practical multiscale feature extraction module (MFE), which can learn richer multiscale feature representation. Most importantly, the output of each MFE is propagated to the next layer to generate the channel-wise attention vector by using hybrid attention.

(3) The attention mechanism in this paper is inserted into the object detection network of YoloV4 [11] and completes experiments on the detection accuracy. The experimental results show that the M3Att can significantly improve detection accuracy. M3Att achieved a 4.93% improvement in mAP over the YoloV4 for the PASCAL VOC2007 [12].

This paper is organized as follows: Section 1 presents relevant research works on attention mechanisms. Section 2 presents the multibranch attention model proposed in this paper. Section 3 compares the method of this paper with the existing mainstream dataset algorithm and gives experimental results. Section 4 gives a summary.

2. Related Work

The attention model was first applied to machine translation [13] and has become a central concept in convolutional neural networks. The attention model has two leading roles: first, to tell the computer which parts of the content to focus on; second, to allocate the limited computational resources to the important parts of the image. The attention mechanism is proven to be one of the most significant ways to improve the effectiveness and efficiency of neural network learning. Currently, the mainstream attention methods can be divided into three major categories, namely, channel attention, spatial attention, and hybrid attention [14].

In deep neural networks, different channels in different feature maps usually represent other objects [15]. Channel attention adaptively recalibrates each channel's weight and generates feature masks, a process of selecting objects to determine what to pay attention to. The squeeze-and-excitation (SE) can learn the consequences for each feature to obtain its importance and uses the critical metric to assign a weight value to each channel. GSop [16] builds on SENet and proposes a second-order pooling layer for aggregating richer features; however, the above attention has fully connected layers, leading to much computational redundancy. Therefore, ECANet [17] uses a one-dimensional convolutional layer to replace the fully connected layer, significantly reducing the model's number of parameters. Fcanet [18] reconsiders the influence of the pooling layer on the attention mechanism from the frequency domain perspective and proposes multispectral channel attention. Distinct from channel attention focused on essential features, some researchers have investigated where it is necessary to concentrate. Therefore, the notion of spatial attention was proposed by Zhu et al. [19], which converts the information in the image's spatial domain into the corresponding space to extract the critical data. GENet [20] uses feature context to aggregate features and then distributes them locally to tell the model which regions are essential.

Hybrid attention has both advantages over the channel and spatial attention mentioned above and has attracted researchers' attention in recent years. In 2017, Fei et al. [21] pioneered the concept of hybrid attention. CBAM concatenates channel and spatial attention, and the generated feature vector has both channel and spatial attention advantages. The dual attention network [22] sums the outputs into two different attention branches and adaptively combines local and global features. Coordinate attention [23] embeds the location information into the channel attention so that the network can focus the computational cost on the sizeable important area. Unlike DANet and coordinate attention, relation-aware global attention (RGA) [24] is a new hybrid attention that emphasizes the importance of global structural information provided by pairwise relationships and uses it to generate attention maps. While some of the above attention mechanisms focus on designing more complex network structures, they will incur substantial computational costs. Some concentrate on creating lightweight network structures, but the improvement in model accuracy is not apparent. Thus, to further improve the model's detection accuracy and reduce the model's complexity, a novel attention mechanism, M3Att, is proposed; this mechanism aims to significantly improve the performance of the object detection network while reducing the complexity of the model, while channel attention and spatial attention are efficiently combined to generate complementary features.

In addition, some researchers have applied attention mechanisms to object detection networks in practical applications. Yang [25] integrated the CBAM attention mechanism into the YoloV4 object detection network, which significantly improved the ability and accuracy of wheat ears' extraction capability. Kim [26] proposed ECAP-Yolo, a modification of the feature extraction network of the core network by the channel attention module, which greatly optimizes the detection performance of small objects. These efforts provide helpful background and scenarios for this and similar studies.

This paper proposes an attention mechanism with a multibranch structure and successfully incorporates it into the YoloV4 object detection network. The multibranching attention mechanism first extracts rich multiscale features through a multilevel feature extraction module and then suppresses the interference of spatial and channel dimensions, respectively.

3. Materials and Methods

The multiband attention model proposed in this paper mainly consists of a multiscale feature extraction module, a spatial attention module, a channel attention module, and a skip connection technique.

3.1. Multiscale Feature Extraction Module

As shown in Figure 1, it is the multiscale feature extraction module MFE that implements feature extraction in M3Att. The input feature map $X \in R^{C \times H \times W}$ is divided into N parts, denoted by $[X_0, X_1, \cdots, X_{N-1}]$. For each partitioned feature map, the number of channels is C/S. Therefore, for the $i-th$ feature map, it can be represented as $X_i \in R^{\frac{C}{S} \times H \times W}, i = 1, 2, \cdots, n-1$. The model can process multiple scales of input tensor in parallel to obtain a topographic map with different scales. Accordingly, the input feature maps are processed using multiscale convolutional kernels, which can extract different spatial and depth information. As the size of the convolution kernel increases, the number of parameters will increase significantly. This paper introduces a clustered convolution approach to solve this issue so that the input tensor of different convolutional kernels can be processed without increasing the computational cost. The relationship between the convolution kernel size and the group size can be expressed as:

$$G = 2^{K-3} \tag{1}$$

where K is the size of the convolution kernel and G represents the group size. The kernel size of the convolution is 9×9, especially when k is equal to 9 and the default of G is 32. Equation refeq1 above will be demonstrated later in the paper by way of ablation experiments. In addition, the use of clustered convolution can significantly reduce the number of model parameters and increase the computation speed. However, communication between channels is weakened, which leads to a reduction in the feature extraction capability [27].

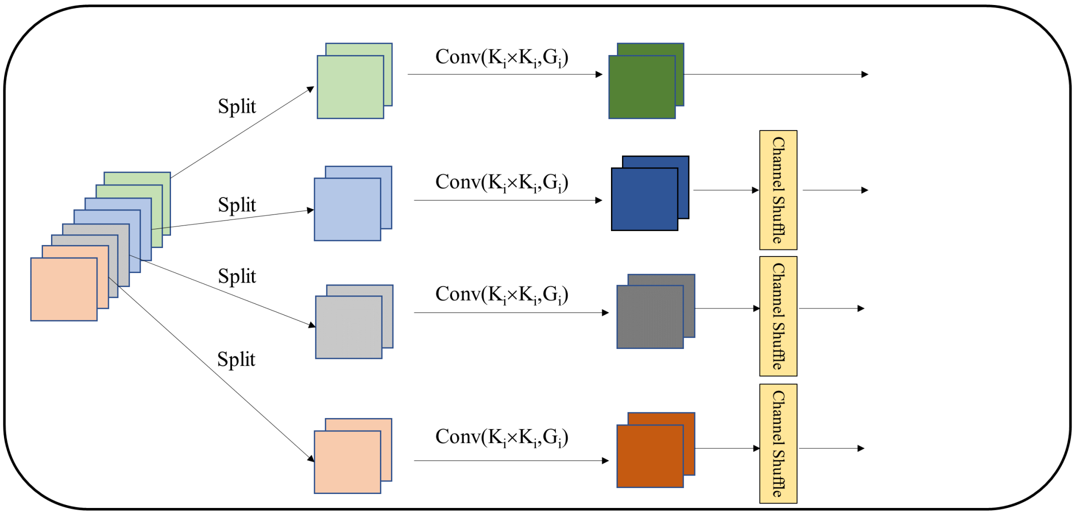


Figure 1. Multiscale feature extraction module.

Thus, this paper introduces channel shuffling to address the lack of feature communication between models. The essence of channel shuffling is to cluster each group of channels derived from the group convolution, randomly split them into new groups of channels, and then stitch them together to obtain the final feature map. The feature map after channel shuffling contains information from different channels, enriching the feature map information and avoiding the loss of feature information due to group convolution. Thus, the generate function of the multiscale feature map can be characterized as:

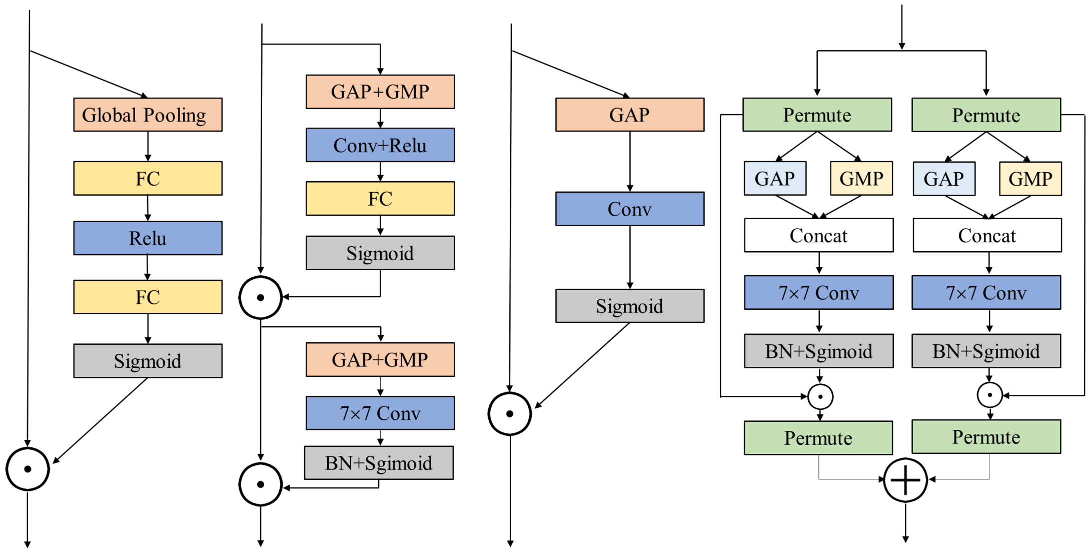
$$F_i = Shuffle(Conv(k_i \times K_i, G_i)), i = 0, 1, 2, \cdots, N-1 \tag{2}$$

where the size of the $i - th$ convolution kernel k_i is $k_i = 2 \times (i+1) + 1$, and the size of the grouped is $G_i = 2^{K-3}$. To obtain the final output feature map, cascade stitching is performed.

$$F = Cat([F_0, F_1, \cdot\cdot\cdot, F_{N-1}]) \tag{3}$$

3.2. Channel Attention Module

Figure 2 compares channel attention in this paper with the current mainstream channel attention. The channel attention in this paper consists mainly of two parallel channel attentions, which are responsible for establishing cross-dimensional interactions between the channel dimension and the spatial dimension, respectively. Notably, instead of using a fully connected layer for dimensionality reduction, M3Att adopts a 7 × 7 convolutional kernel for dimensionality reduction. This reduces computational overhead and achieves greater efficiency when executing the forward propagation model [28].



(a) SENet (b) CBAM (c) ECANet (d) Ours

Figure 2. Comparison with different attention mechanisms.

For a given input feature map $F \in R^{C \times H \times W}$, it will be entered into the attention of two branches of the channel. In the first branch, the interaction between the channel information and the height information is constructed in this paper. Firstly, the input feature map is rotated 90° counterclockwise along the H axis via the permute function, which changes the shape of the input feature map to $F_H \in R^{W \times H \times C}$. Secondly, by arranging the GAP (global average pooling) and GMP (global max pooling) in parallel, the shape of the feature map is reduced $F'_H \in R^{2 \times H \times W}$. In addition, the tensor of the input feature map retains only two dimensions of information, which preserves the richness of the feature map and reduces the computational overhead at the same moment. The feature map information is then extracted via a standard convolutional layer with a convolution kernel size of 7 × 7. In turn, the final attention weights are generated by the BN layer and sigmoid activation function. Then, the generated attention weights are directly multiplied point by point with the original input feature map to obtain a feature map with cross-dimensional interaction information and then rotated 90° clockwise along the H axis to preserve the original shape of the feature map input for further operations.

Similarly, in the second branch, the interaction between width and height information is constructed in this paper. Firstly, by rotating the input feature map 90° counterclockwise along the W axis via the permute function, the shape of the input features map will be transformed to $F_W \in R^{H \times C \times W}$. The feature map is simplified to $F_W \in R^{H \times C \times W}$ by a parallel GAP and GMP layout, and then the feature map information is extracted via a standard convolutional layer with a convolution kernel size of 7 × 7. The final attention weights are generated via the BN layer and the sigmoid activation function. The generated attention weights are directly multiplied point by point with the original input feature map to obtain a feature map with cross-dimensional interaction information, and then rotated 90° clockwise along the W axis to keep the original shape of the input feature.

Ultimately, the outputs of the two different branches are then summed element by element and averaged uniformly to obtain the end output $Q_F \in R^{C' \times H' \times W'}$. In conclusion, for feature maps $F \in R^{C \times H \times W}$, the calculation of channel attention weight can be expressed mathematically as:

$$Q(F) = \frac{1}{2}(F_H \sigma(f^{7\times7}(F'_H)) + F_W \sigma(f^{7\times7}(F'_W))) \tag{4}$$

where the convolution operation and activation function are denoted as $f^{7\times7}$ and $\sigma(\cdot)$.

3.3. Spatial Attention Module

The spatial attention structure of this paper is shown in Figure 3. The spatial attention mechanism in this paper draws on the idea of the spatial attention mechanism from SAM and improves upon it.

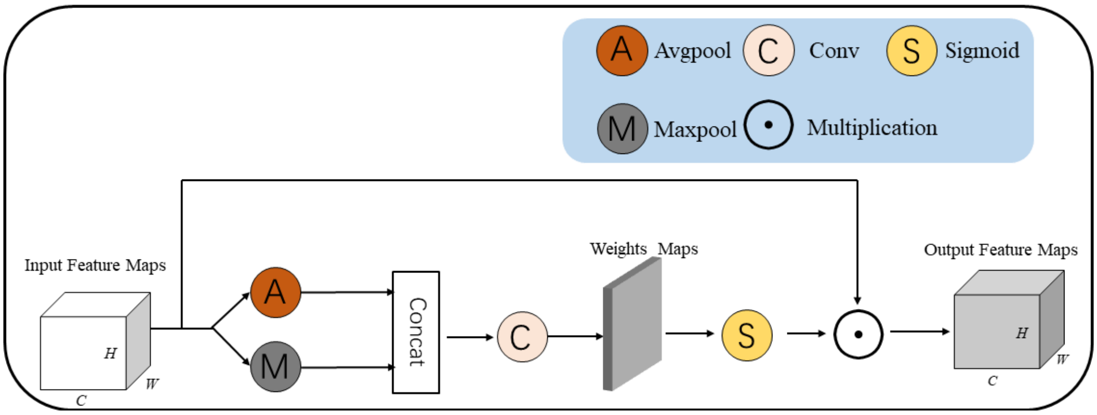


Figure 3. Diagram of our spatial attention.

The input feature map $F \in R^{C \times H \times W}$ is obtained by two feature maps, max pooling and average pooling, respectively, and the two feature maps are stitched together to obtain the feature map for the next level.

Next, the stitched feature map of the shape $F \in R^{2 \times H \times W}$ is fed into a standard convolutional layer with a convolution kernel size of 7 × 7 to generate a spatial attention map. In this way, the spatial features are further extracted. The feature map is also downscaled into a single-channel feature map, which completes the channel-matching process. Next, a sigmoid function is adopted to obtain a weight of spatial attention feature between (0,1) and apply it to the original feature map to obtain the final output feature map with spatial attention weights. Therefore, the calculation process of spatial attention weight can be presented as follows.

$$M(F) = \sigma((f^{7\times7}(C(\text{Maxpool}(F), \text{Avgpool}(F))))) \tag{5}$$

The max pooling layer, average pooling layers operation, convolution operation, splicing operation, and skip connection are used in $Maxpool(\cdot)$, $Avgpool(\cdot)$, $f^{7\times7}$, $C(\cdot)$.

3.4. M3Att Attention Module

The module M3Att in this paper consists of the following four parts, as shown in Figure 4. First, the feature map is divided by the MFE module to obtain different channel feature maps and rich multiscale data. Secondly, the channel feature maps are fed into the channel attention module and spatial attention module, respectively, for extracting attention information at different scales. Thirdly, the obtained channel and spatial attention weights are cascaded and spliced. By this, channel attention and spatial attention fusion can be achieved without destroying original attention, and more complementary attention weights can be obtained. Therefore, the entire hybrid attention vector at multiple scales can be described as follows:

$$V_{att} = C(Q_0(F), \cdots, M_{N-1}(F)) \tag{6}$$

Fourthly, the multiscale hybrid attention vector is fed into the softmax function for recalibration to enable better information interaction between channel attention and spatial attention, which can be represented as follows:

$$Z = softmax(V_{att}) = \frac{exp(V_{att})}{\sum_{i=0}^{N-1} exp(V_{att})} \tag{7}$$

The calibrated attention vector is element-wise multiplied with the resulting feature map after splitting and the original feature map. Lastly, a feature map with multiscale information is used as output. Thus, the whole process of attention training can be summed up:

$$Y = Z \otimes MFE(X) \otimes L(X) \tag{8}$$

$MFE(\cdot)$ is represented as a multiscale module; $l(\cdot)$ is skip connection.

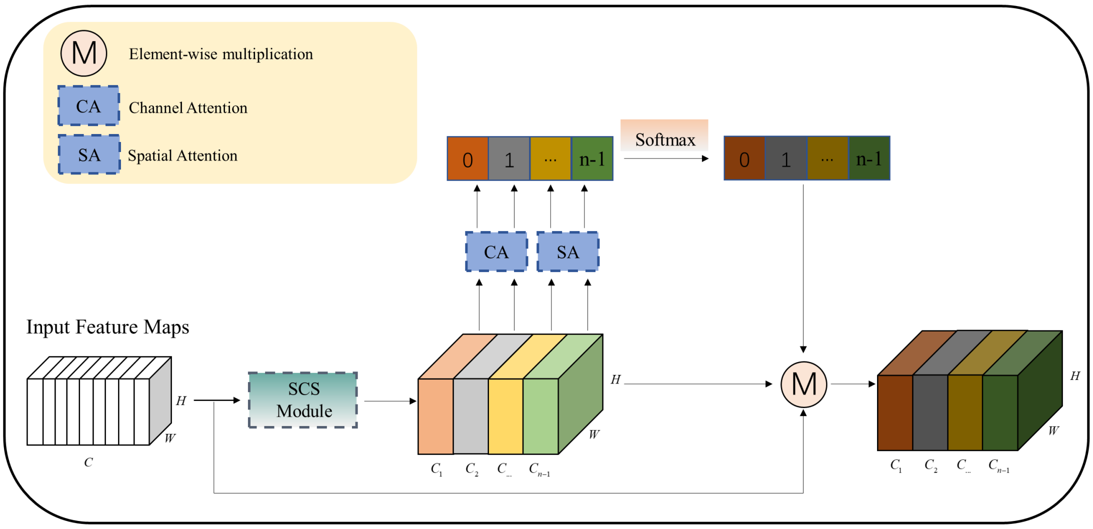


Figure 4. The structure of M3Att.

4. Experimental Results and Analysis

The effectiveness of M3Att was verified by comparing it with eight current top–down attention mechanisms on three public datasets, VOC2007, VOC2012, and KITTI [29], and one live—action photographed underwater critter dataset (contracted from the Zhanjiang Underwater Robot Competition 2020). The eight attention mechanisms were SENet, coordinate attention, CBAM, ECANet, DANet, EPSANet [30] and SPANet [31], and triplet attention [32]. In addition, to evaluate the effectiveness of our final model, ablation experiments were conducted to validate the model, which is the main focus of our study.

Six good object detection networks were then selected to validate the generalizability of our M3Att, thus demonstrating the "plug-and-play" nature of M3Att and its ability to improve object detection accuracy. The six object detection networks are YoloV3 [33], YoloV4, Yolov5, YoloX [34], SSD [35], and Faster R-CNN [36]. The integration in M3Att and YoloV4 is schematically represented as an example in Figure 5.

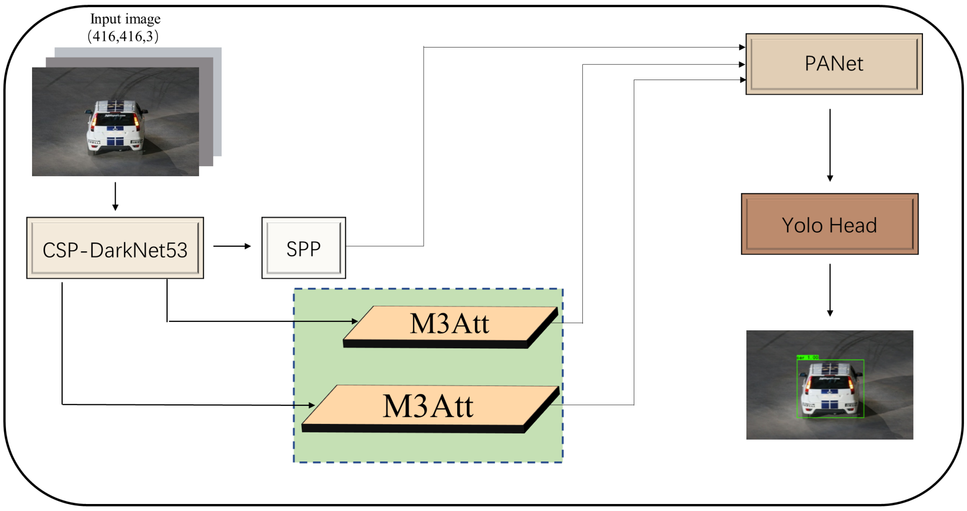


Figure 5. M3Att integrated in YoloV4.

4.1. Dataset

The public datasets used in this paper are shown in Table 1, and the main parameters of these three datasets are listed. Figure 6 shows some images of the dataset used in this paper.

Table 1. Key parameters of three public datasets.

Name	Train Image	Test Image	Class
VOC2007	5011	4952	20
VOC2007 + VOC2012	22,136	4952	20
KITTI	6883	765	4

(a) VOC2007 and VOC2012

(b) KITTI

(c) Underwater Dataset

Figure 6. Examples of some images from the datasets used in this paper.

Furthermore, to validate the model's performance in real-world scenarios, the marine life dataset from the 2020 National Underwater Robotics Competition (Zhanjiang) was included in this paper for empirical validation. Since the dataset consisted of 3824 images, the original image set was expanded using data augmentation techniques and after data preprocessing, which resulted in 7648 images after data preprocessing. The dataset was divided 9:1 between training and test sets, with 6883 training images and 765 test images, including Echinus, Starfish, Holothurian, and Scallop.

4.2. Experimental Environment and Parameter Settings

The experimental equipment used in this paper was configured with an Intel i7-10700 CPU, an NVIDIA GeForce RTX 3090 GPU, video memory of 24 GB, and the Windows 10 operating system. The training model was constructed using the Pytorch deep learning framework based on the Windows 10 operating system, using the Python 3.7 programming language and Cuda 11.2. The main parameters of the experiment are shown in Table 2.

Table 2. Key parameters of the settings.

Experimental Parameters	Parameter Values
Image size	416 × 416
Learning rate	0.001
Batch size	16
Epochs	100
Optimizer	Adam

4.3. Evaluation of the Model Performance

To evaluate the model's performance more precisely, two metrics were chosen to measure the model: the average precision of AP50 (average precision) and mAP (mean average precision) for each type of object with an intersection ratio of 0.5. mAP is defined as shown in Equation (9).

$$mAP = \frac{\sum_{K}^{K=1} AP(P,R,K)}{K} \tag{9}$$

where P is precision, R represents recall, and K denotes the the number of class. In addition, the VOC dataset K = 20, the KITTI dataset K = 3, and the marine biology dataset K = 4. For a more intuitive evaluation of the model performance, the eight selected attention mechanisms were also validated using visual analytical plots and heat maps.

4.4. Analysis of the Generalizability of the Model

The PASCAL VOC 2007 dataset and the PASCAL VOC 07+12 dataset were selected for comparison in different networks to verify the generalizability of the model, and the results are shown in Table 3.

As shown in Table 3, after adding M3Att, the performance of the one-stage object detection network and the two-stage detection network improved compared to the original one. Specifically, the Yolo series object detection algorithms (YoloV3, Yolov4, YoloV5, and YoloX) improved by 1.41%, 4.93%, 2.05%, and 2.36%, respectively, over the original algorithms. This shows that the proposed M3Att is effective and can significantly improve object detection accuracy. This paper aims to improve the accuracy of the object detection algorithm, and the data indicate that the work in this paper is efficient. The model's generalization capability can be further tested for object detection tasks with more extensive datasets. Therefore, the VOC2007 dataset was fused with the VOC2012 dataset to test the network's performance further. When the M3Att module proposed in this paper was added, the mAP of all networks improved further, with YoloV3, for example, improving by 2.06%.

Table 3. Experiments with M3Att embedded in different object detection networks.

Model	Input Size	Dataset	mAP
Yolov3 [33]	416 × 416	07/07 + 12	79.77/81.76
Yolov3 [33] + M3Att	416 × 416	07/07 + 12	81.18/83.82
Yolov4 [11]	416 × 416	07/07 + 12	82.22/87.79
Yolov4 [11] + M3Att	416 × 416	07/07 + 12	87.15/87.41
Yolov5	416 × 416	07/07 + 12	85.01/87.62
Yolov5 + M3Att	416 × 416	07/07 + 12	87.06/88.48
YoloX [34]	416 × 416	07/07 + 12	82.33/85.91
YoloX [34] + M3Att	416 × 416	07/07 + 12	88.09/88.52
SSD [35]	300 × 300	07/07 + 12	68.0/74.3
SSD [35] + M3Att	300 × 300	07/07 + 12	76.41/77.50
Faster [36]	600 × 600	07/07 + 12	73.28/76.86
Faster [36] + M3Att	600 × 600	07/07 + 12	77.47/78.65

4.5. Experiment Comparing Different Attention Mechanisms

The experiments were conducted using the PASCAL VOC2007 dataset, and eight popular attentional mechanisms, such as SENet, coordinate attention (CA), CBAM, ECANet, SPANet, DANet, EPASNet, and triplet attention (abbreviated as triplet), were selected. Tables 4 and 5 show the accuracy comparison results along with the model complexity of the YoloV4-based object detection algorithm. With the addition of the M3Att module, the proposed M3Att–YoloV4 achieved 4.93% higher detection accuracy than the YoloV4 object detection network, and the number of parameters used was only 0.52 M higher, giving better results. In addition, compared to DANet, which is representative of the excellent hybrid attention mechanism in the last years, M3Att reduced the number of parameters and computational cost by 28.2% and 11.3% respectively, with a slight improvement in detection accuracy. In summary, based on the above results, our M3Att module achieved superior parametric number comparisons and detection accuracy comparisons at a lower computational cost.

For a more intuitive comparison of the performance of this method with other attention mechanisms for object detection, a visual comparison graph is introduced in this paper for evaluation purposes. The visual contrast diagram is shown in Figure 7. Figure 7a shows only one class of objects; however, there are occlusions and small objects, and the pose of each object class varies, so whether objects in the image can be fully detected can be a good test of the model's performance in dealing with various complex conditions.

As can be easily seen from the figure, it is easy to see that M3Att alone does not have any false or missed detection and detects all the objects completely. Figure 7b shows only one object class, a boat, but it is more difficult to detect due to the dark background of the image and the severe occlusion between objects. Only M3Att detects occluded objects, as shown in Figure 7.

To quantitatively and intuitively explain how M3Att makes full use of the object's salient features to enhance the network's performance, this paper introduces the Grad-CAM technique [37], which shows a comparative analysis of eight attentional mechanisms with the M3Att attentional map. The visualized class heat map allows one to identify the parts of the object detection task, in which the darker colors represent the parts that have the most significant impact on the results, that is, the most significant feature. A comparison diagram is presented in Figure 8. The visualization in Figure 8 demonstrates the inherent advantages of M3Att, which is better able to focus on salient features and cover salient regions than the rest of the attention mechanisms. In other words, M3Att can learn to use the information within the object region and cluster features from it very well. Therefore, this feature can significantly improve the performance of object detection networks.

Table 4. Experimental results of different attention mechanisms (based on YoloV4).

Method	mAP	Areo	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	Mbike	Person	Plant	Sheep	Sofa	Train	TV
YoloV4 [11]	82.22	91.78	89.98	86.21	73.60	74.48	90.65	93.33	91.35	68.92	92.03	76.22	89.48	93.61	92.96	90.34	54.78	86.67	91.41	91.99	86.49
YoloV4 [11] + SE [5]	85.54	93.72	92.33	84.48	78.70	77.82	91.50	93.62	89.74	72.93	89.15	80.21	86.44	91.32	93.31	89.93	61.42	85.55	79.61	91.70	87.37
YoloV4 [11] + CA [23]	84.31	90.15	91.90	83.78	75.05	74.32	89.86	93.36	90.57	68.90	91.39	78.50	85.14	91.73	90.60	89.41	60.21	86.36	76.22	92.08	86.71
YoloV4 [11] + CBAM [8]	85.43	93.26	92.59	84.45	75.67	76.32	89.66	93.30	89.82	72.48	92.15	80.52	87.76	92.29	91.77	89.21	62.12	87.73	80.59	90.31	86.57
YoloV4 [11] + ECA [17]	85.99	83.66	91.16	86.21	75.02	78.43	91.10	93.61	90.79	73.15	93.13	81.21	87.81	92.59	92.76	90.63	63.44	88.71	80.37	90.59	85.47
YoloV4 [11] + SPANet [31]	82.37	88.69	87.79	83.80	71.26	70.98	89.16	91.67	89.44	66.56	90.30	74.78	86.71	91.46	88.06	88.46	54.84	82.82	76.30	90.71	83.55
YoloV4 [11] + DANet [22]	87.09	93.24	90.63	87.62	79.90	80.03	92.03	91.34	90.82	74.23	93.35	82.96	91.01	92.84	92.98	90.94	63.41	91.84	79.77	95.06	87.92
YoloV4 [11] + ESPANet [30]	84.32	91.41	88.85	85.73	73.90	76.15	90.58	92.75	88.56	69.90	90.57	81.06	84.62	90.82	91.94	89.53	60.43	87.09	74.52	92.39	85.64
YoloV4 [11] + Triplet [32]	82.41	90.68	87.65	82.25	70.38	73.49	89.10	91.46	89.64	66.55	88.96	75.20	85.79	91.15	88.17	87.56	55.79	83.99	76.08	91.50	82.91
YoloV4 [11] + Residual [21]	85.81	93.93	91.78	84.96	78.95	77.61	89.25	92.84	90.37	71.57	92.14	81.91	88.09	93.14	92.43	89.69	60.39	88.02	80.72	92.18	86.29
YoloV4 [11] + M3Att	87.15	93.28	91.08	87.09	79.35	79.30	93.00	94.21	92.29	73.05	92.88	83.66	91.36	92.03	93.02	91.09	64.12	90.87	81.28	94.00	85.34

Table 5. Comparison of mechanisms parameters.

Model	Input Size	mAP(%)	Parameters (M)	GFlops (G)
Yolov4 [11]	416 × 416	82.22	64.040	29.948
Yolov4 [11] + SE [5]	416 × 416	85.54	64.083	59.898
Yolov4 [11] + CA [23]	416 × 416	84.31	64.075	59.659
Yolov4 [11] + CBAM [8]	416 × 416	85.43	64.731	59.900
Yolov4 [11] + ECA [17]	416 × 416	85.99	64.040	59.898
Yolov4 [11] + SPANet [31]	416 × 416	85.54	64.513	59.901
YoloV4 [11] + DANet [22]	416 × 416	87.09	89.806	68.606
YoloV4 [11] + EPSANet [30]	416 × 416	84.32	65.755	60.482
YoloV4 [11] + Triplet [32]	416 × 416	82.41	64.041	59.902
YoloV4 [11] + Residual [21]	416 × 416	85.81	64.999	61.204
YoloV4 [11] + M3Att	416 × 416	87.15	64.467	60.822

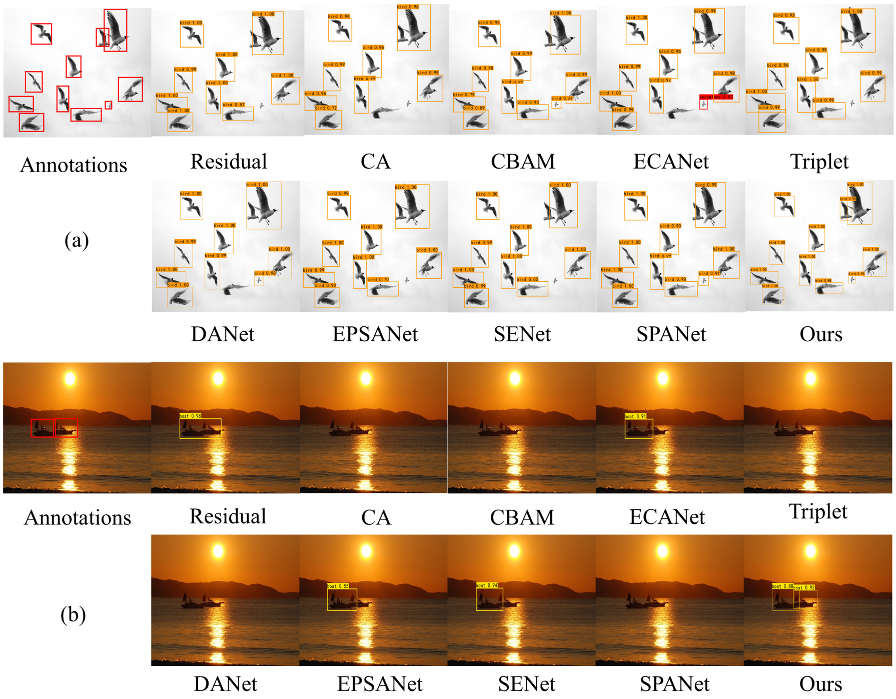


Figure 7. Visual comparison of VOC datasets.

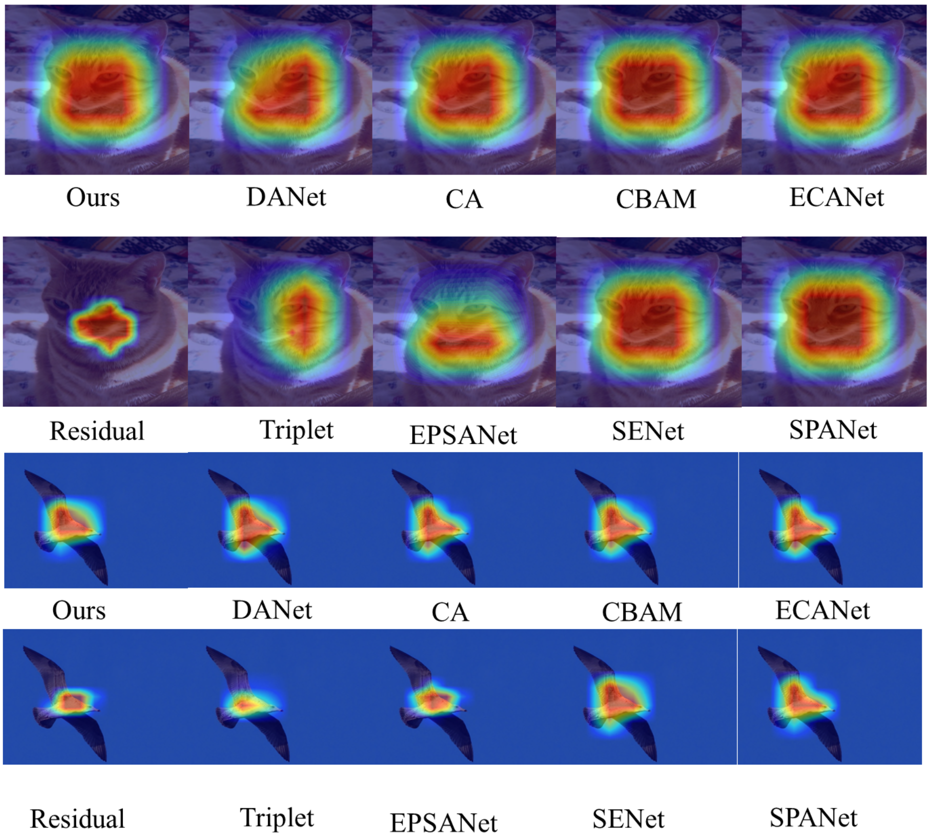


Figure 8. Visualization of Grad-CAM.

4.6. KITTI Dataset Experiment

To test the performance of our algorithm in complex scenarios, the KITTI dataset was selected for the experiments. The dataset contained 7482 images with eight object categories, including 6733 images in the training set and 749 in the test set. For statistical and analysis purposes, the eight categories in the KITTI dataset were combined into three categories, namely, Car, Cyclist, and Pedestrian. The images in this dataset were captured from real street scenes and therefore involved many small and medium objects in complex environments. Experiments on the KITTI dataset can further demonstrate the model's performance. Table 6 shows the comparison results of the proposed module M3Att with the other eight attention mechanisms on the KITTI dataset. Compared with the other eight mainstream attention mechanisms on mAP, M3Att achieved the best results. Compared with other methods, this algorithm achieved the best improvement over YoloV4 with a 5.38% improvement in mAP. For these eight classical attentional mechanisms, the simple global mean pool ignores the local information in the channel, so the algorithm does not perform exceptionally well in datasets with large numbers of small and medium objects.

Experimental results of different attention mechanisms on the KITTI dataset are shown in Figure 8. The P-R graph of the three objects types on the KITTI dataset are shown in Figure 9.

Experimental results of different attention mechanisms and the P–R graph of the three object types on the KITTI dataset are shown in Figure 9. An irregular curve enclosed by the vertical axis of accuracy and the horizontal axis of completeness is called a P–R curve. P and R values should be as high as possible for better experimental results, but precision and recall are contradictory. A higher precision rate tends to be associated with a lower recall rate. Thus, the P–R curve plotting can better study the model performance. From Figure 9, the P–R graph of this algorithm is significantly better than the other algorithms. The proposed M3Att module in this paper can successfully capture object information. It improves object detection accuracy significantly compared to YoloV4, SENet, CA, CBAM, ECANet, SPANet, and DANet, as well as EPSANet networks, triplet networks, and other object detection and attention mechanisms. The final mAP achieved 86.94% with good detection results.

Table 6. Experimental comparison results on the KITTI dataset.

Model	Input Size	mAP
Yolov4 [11]	416 × 416	81.56
Yolov4 [11] + SE [5]	416 × 416	84.55
Yolov4 [11] + CA [23]	416 × 416	85.11
Yolov4 [11] + CBAM [8]	416 × 416	83.39
Yolov4 [11] + ECANet [17]	416 × 416	85.92
Yolov4 [11] + SPANet [31]	416 × 416	85.21
Yolov4 [11] + DANet [22]	416 × 416	86.09
Yolov4 [11] + EPSANet [30]	416 × 416	82.72
Yolov4 [11] + Triplet [32]	416 × 416	82.99
YoloV4 [11] + Residual [21]	416 × 416	83.35
Yolov4 [11] + M3Att	416 × 416	86.94

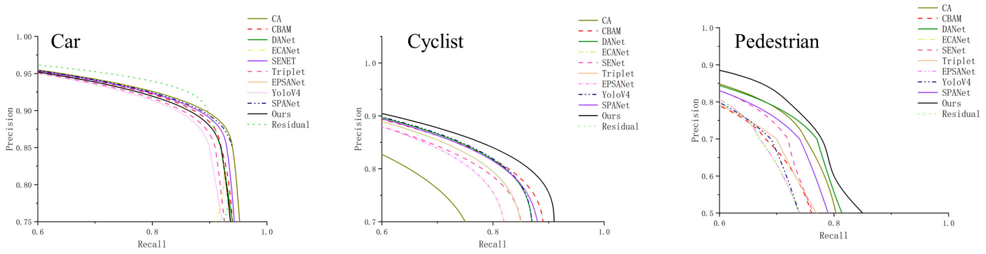


Figure 9. P–R graph for objects in the KITTI dataset.

4.7. Ablation Studies

As indicated in Table 7, this set of experiments verified the effectiveness of M3Att on the PASCAL VOC2007 dataset by adjusting the cluster size of the clustered convolution. Since parallel computing will significantly increase the number of model parameters, clustered convolution is introduced to deal with the increasing number of parameters.

Table 7. Results of group size ablation experiments.

Kernel Size	Group Size	Input Size	mAP (%)	Parameters (M)
(3,5,7,9)	(4,8,16,16)	416 × 416	86.28	64.409
(3,5,7,9)	(4,4,4,4)	416 × 416	87.29	64.972
(3,5,7,9)	(16,16,16,16)	416 × 416	87.08	64.343
(3,5,7,9)	(1,4,8,16)	416 × 416	87.01	64.674
(3,5,7,9)	(1,4,16,32)	416 × 416	87.15	64.467
(3,5,7,9)	(1,8,16,32)	416 × 416	86.97	64.496
(3,5,7,9)	(8,8,8,8)	416 × 416	87.15	64.553
(3,3,3,3)	/	416 × 416	86.55	64.870
(3,2 × 3,3 × 3,3 × 3)	(1,4,16,32)	416 × 416	86.22	65.008

Using parallel computing significantly increases the number of model parameters, so we introduced clustered convolution to deal with the increasing number of parameters.

As can be seen from the results in the table, grouping size directly affects the performance and complexity of the model. Hence, this paper determined the convolutional kernel size and adjusted the group size to balance model performance and complexities. Lastly, this paper used a convolutional kernel size of (3,5,7,9) and a convolved clustering of (1,4,16,32).

Notes: 2 × 3, 3 × 3 means two cascades connected to 3 × 3 convolutional kernels, and three cascades connected to 3 × 3 convolutional kernels.

To test the impact of the three modules proposed in this paper on object detection results, we performed ablation experiments on the PASCAL VOC2007 dataset. The results can be seen in Table 8. It is easy to see from the experimental results that adding the MFE module allows the network to extract more detailed features, and mAP increases from 82.22% to 85.41% compared to the original object detection network Yolov4. On this basis, mAP was increased from 82.22% to 85.41% by adding the channel attention mechanism, and the model focused on each key characteristic. Secondly, by introducing the spatial attention mechanism, the model could focus on key regions, and the mAP value increased to 86.97%. Lastly, a skip connection mechanism was introduced to transfer the bottom features from the shallow layer to the deep layer, compensating for a large amount of detail lost in the deep one. The final model mAP value reached 87.15%, which achieved a relatively good result.

Table 8. Experiments with M3Att embedded in different object detection networks.

mAP	YoloV4	MFE	Channel Attention	Spatial Attention	Skip Connection
82.22	√				
85.41	√	√			
85.99	√	√	√		
86.97	√	√	√	√	
87.15	√	√	√	√	√

4.8. Practical Scenario Experiments

To test the robustness of the algorithm in real-world environments, we experimented with the 2020 Zhanjiang Underwater Robotics Competition dataset, visualizing the comparison as shown in Figure 10. The proposed M3Att performed better than the current mainstream attention mechanisms in object detection tasks in a real-world environment. In particular, M3Att performed well in the small objects, with all objects detected with high confidence. In some specific cases, M3Att detected unlabeled objects, as shown in Figure 10a, demonstrating the robustness of our M3Att. To verify the model's performance in complex underwater environments, Figure 10b tested the performance of the M3Att in fuzzy and occluded conditions, in which only M3Att detected the entire contents of all annotated images and the unannotated scallops in the upper right corner of the image with high confidence. This is because M3Att uses a multibranching structure that combines the advantages of channel attention and spatial attention to achieve additional complementary properties that effectively highlight object features and suppress irrelevant information, thus increasing the confidence of each object and further validating the model.

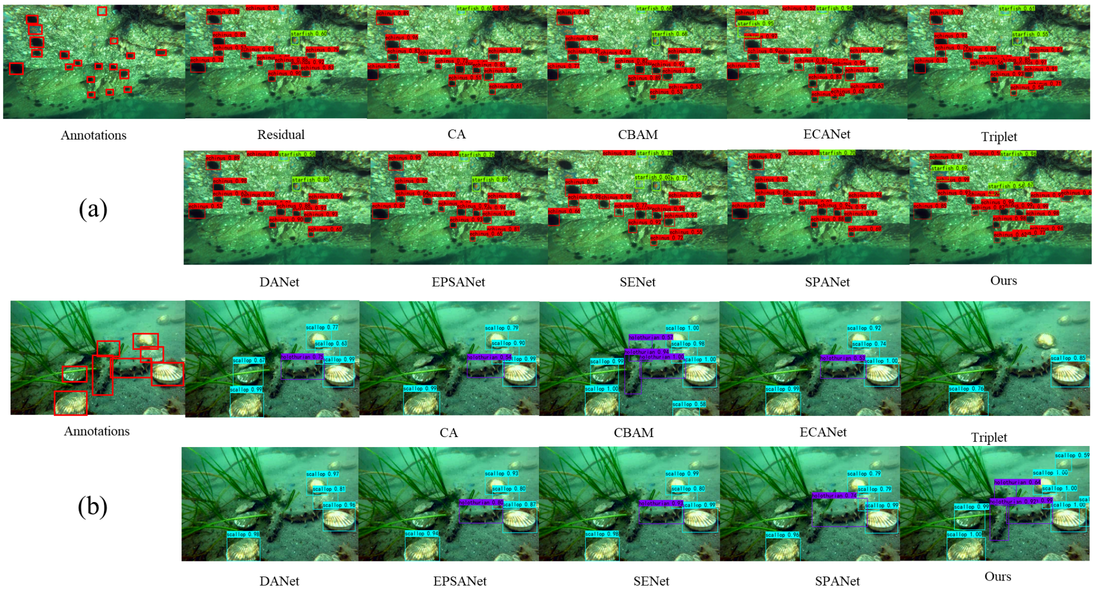


Figure 10. Comparison of the actual detection effect of different attention mechanisms.

5. Conclusions

In this paper, an effective and lightweight attention mechanism named M3Att was proposed. It is plug-and-play and can be easily added to any object detection network to improve network performance. The multiscale feature extraction module (MFE) can fully extract multiscale information and enrich feature information by channel shuffling. At the same time, a hybrid form of the channel and spatial attention is used to construct long-range channel dependence, which is more helpful in obtaining complementary features. Finally, we introduced a skip connection mechanism to connect the shallow information with the profound information after multiple convolutions, avoiding the information loss problem after numerous convolutions. Experimental results demonstrate that our M3Att is successful. It was verified that our M3Att can achieve the best performance in object detection compared with other attention mechanisms. In future work, we will continue to explore the role of M3Att in the remaining computer vision tasks.

Author Contributions: Methodology, G.M.; writing—original draft preparation, G.L.; data curation, G.L.; review and editing, H.Z. and B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (61773415), the National Key Research and Development Program of China (2019YFD0900805) and Science and Technology Project of Fujian University of Technology (GY-Z220205).

Informed Consent Statement: Not applicable.

Data Availability Statement: PASCAL VOC2007 dataset (http://host.robots.ox.ac.uk:8080/pascal/VOC/voc2007/, accessed on 1 September 2022.), PASCAL VOC2012 dataset (http://host.robots.ox.ac.uk/pascal/VOC/voc2012/, accessed on 1 September 2022.), KITTI dataset (https://www.cvlibs.net/datasets/kitti/, accessed on 1 September 2022.), 2020 Zhanjiang Underwater Robotics Competition dataset. Restrictions apply to the availability of these data. The data were obtained from the Fujian University of Technology and are available from the authors with the permission of Institute for Machine Learning and Intelligence Sciences.

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

References

1. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. *Neurocomputing* **2020**, *396*, 39–64. [CrossRef]
2. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
3. Huang, Z.; Li, W.; Xia, X.G.; Wu, X.; Cai, Z.; Tao, R. A novel nonlocal-aware pyramid and multiscale multitask refinement detector for object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–20. [CrossRef]
4. Guo, M.; Xu, T.; Liu, J.; Liu, Z.; Jiang, P.; Mu, T.; Zhang, S.; Martin, R.; Cheng, M.; Hu, S. Attention mechanisms in computer vision: A survey. *arXiv* **2021**, arXiv:2111.07624.
5. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
6. Chaudhari, S.; Mithal, V.; Polatkan, G.; Ramanath, R. An attentive survey of attention models. *ACM Trans. Intell. Syst. Technol. (TIST)* **2021**, *12*, 1–32. [CrossRef]
7. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
8. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
9. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef]
10. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
12. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
13. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
14. Wang, F.; Tax, D.M. Survey on the attention based RNN model and its applications in computer vision. *arXiv* **2016**, arXiv:1601.06823.
15. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.
16. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–21 June 2019; pp. 3024–3033.
17. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
18. Qin, Z.; Zhang, P.; Wu, F.; Li, X. Fcanet: Frequency channel attention networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 783–792.
19. Zhu, X.; Cheng, D.; Zhang, Z.; Lin, S.; Dai, J. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6688–6697.
20. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31* .
21. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
22. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–21 June 2019; pp. 3146–3154.
23. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13713–13722.
24. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3186–3195.
25. Yang, B.; Gao, Z.; Gao, Y.; Zhu, Y. Rapid detection and counting of wheat ears in the field using YOLOv4 with attention module. *Agronomy* **2021**, *11*, 1202. [CrossRef]
26. Kim, M.; Jeong, J.; Kim, S. ECAP-YOLO: Efficient Channel Attention Pyramid YOLO for Small Object Detection in Aerial Image. *Remote Sens.* **2021**, *13*, 4851. [CrossRef]
27. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
28. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

29. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
30. Zhang, H.; Zu, K.; Lu, J.; Zou, Y.; Meng, D. Epsanet: An efficient pyramid split attention block on convolutional neural network. *arXiv* **2021**, arXiv:2105.14447.
31. Guo, J.; Ma, X.; Sansom, A.; McGuire, M.; Kalaani, A.; Chen, Q.; Tang, S.; Yang, Q.; Fu, S. Spanet: Spatial pyramid attention network for enhanced image recognition. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2019; pp. 1–6.
32. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision(WACV), Waikoloa Beach, HI, USA, 5–9 January 2021; pp. 3139–3148.
33. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
34. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
35. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
36. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
37. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.