Hindawi
Mobile Information Systems
Volume 2021, Article ID 6466051, 12 pages
https://doi.org/10.1155/2021/6466051

Research Article

Remote Sensing Data Detection Based on Multiscale Fusion and Attention Mechanism

Min Huang,[1] Cong Cheng,[1] and Gennaro De Luca[2]

[1]*South China University of Technology, Guangzhou, Guangdong 510000, China*
[2]*Department of Information Technology, Arizona State University, Tempe, AZ 85287-2180, USA*

Correspondence should be addressed to Cong Cheng; secongcheng@mail.scut.edu.cn

Received 13 August 2021; Accepted 4 October 2021; Published 26 November 2021

Academic Editor: Xingsi Xue

Remote sensing images are often of low quality due to the limitations of the equipment, resulting in poor image accuracy, and it is extremely difficult to identify the target object when it is blurred or small. The main challenge is that objects in sensing images have very few pixels. Traditional convolutional networks are complicated to extract enough information through local convolution and are easily disturbed by noise points, so they are usually not ideal for classifying and diagnosing small targets. The current solution is to process the feature map information at multiple scales, but this method does not consider the supplementary effect of the context information of the feature map on the semantics. In this work, in order to enable CNNs to make full use of context information and improve its representation ability, we propose a residual attention function fusion method, which improves the representation ability of feature maps by fusing contextual feature map information of different scales, and then propose a spatial attention mechanism for global pixel point convolution response. This method compresses global pixels through convolution, weights the original feature map pixels, reduces noise interference, and improves the network's ability to grasp global critical pixel information. In experiments, the remote sensing ship image recognition experiments on remote sensing image data sets show that the network structure can improve the performance of small-target detection. The results on cifar10 and cifar100 prove that the attention mechanism is universal and practical.

1. Introduction

At present, sensor information is a hot research target. The acquisition of sensor information and mobile computing are relatively mature [1]. There have been many outstanding achievements in the research of basic data types of sensor information, and many more successful algorithms have been proposed [2, 3]. For image sensors, information processing is faced with difficulties; because of the limitations of some devices, the pictures obtained by the sensor have the characteristics of large noise, small targets, and blurred targets.

With the development of deep learning, target detection technology based on deep learning has made significant progress [4–8]. However, due to the image sensors information's weakness, small target detection faces huge difficulties and challenges. First, it is easy to lose feature information during the convolution process, which affects the convergence and accuracy of the network. Second, traditional convolution mainly uses the accumulation of small convolution kernels to expand the receptive field. In the early stage of the network, the local convolution method of the small convolution kernel will treat the feature points and noise points equally, which is not conducive to the network's grasp of the feature points and affects the network convergence.

To improve the detection performance of small targets, researchers have carried out much research from network structure, training strategy, data processing, etc. However, compared with large- and medium-target detection, there is still a significant gap in the performance of small-target detection. The target scale is one of the critical factors that affect the performance of target detection. At present, the detection accuracy of small targets is far lower than that of

large targets and medium-sized targets in both open data sets and real-world images, and there are often missed and false detections. However, small-target detection has essential applications in many natural scenes.

In recent years, the deep convolutional neural network (CNN) has made significant progress in target detection [9–11]. CNNs realize the extraction of features, candidate regions, bounding boxes, and the discrimination of object categories. However, the CNN detector is not suitable for small-target detection due to the nature of the convolutional and pooling layers. These layers reduce the number of parameters in the network and the dimensionality of the image. The resolution of the feature image is therefore much lower than that of the original input image, which makes the classification and boundary box regression very difficult. Therefore, whether it is one-stage [12] Yolo [13] and SSD [14] or two-stage Faster R-CNN [15], the effect of small-target detection is not ideal. Since then, there have been some improved methods for small-target detection in the field of deep learning, such as multiscale fusion, scale invariance, and so on. Feature Pyramid Networks (FPNs) [16, 17] use low-level location information with high-level semantic information by propagating the high-level features down. The problem of small-target detection results in part from deep learning target detection algorithms only using top-level feature mapping for classification and prediction and ignoring the location information of low-level features. The pyramid structure of FPNs helps resolve this issue. Scale Normalization for Image Pyramids (SNIP) realizes multiscale image input, improves the precision of the preselector, and promotes the effect of small-target detection.

Since single-scale feature mapping is not good at representing objects of different sizes and shapes, extracting relevant information from different layers can naturally alleviate this contradiction. Multiscale deep convolutional neural network (MS-CNN) [18] extracts the proposal region from different-scale feature maps and uses the deconvolution replacement to sample the input image to improve the speed accuracy. Single-shot multibox detector (SSD) extends several additional convolution layers on the truncated Vgg-16 [19] as its backbone network and sets different default frame sizes according to different receptive fields, so it can better predict targets of various scales. This bottom-up pyramid hierarchy can detect objects of different sizes separately. However, although the use of low-level features is intentionally avoided, the shallow layer of a convolutional neural network cannot fully extract features, which still limits the performance of the detector in small-scale target detection.

Recently, a detection network was devised based on multiscale fusion features. HyperNet [20] is better than Fast R-CNN in processing small objects and generating a higher-quality proposal thanks to the interaction between multilayer feature fusion and different sampling strategies. FPNs alleviate this contradiction through an additional top-down architecture, enhancing semantic information through upsampling and adding details through horizontal connection to construct a high-level semantic feature map. Deconvolutional single-shot detector (DSSD) [21] uses a deconvolution module to build a feature pyramid on the SSD benchmark network. The detection network based on multiscale fusion features improves the detection accuracy by injecting large-scale context information. However, the convergence of the corresponding layers on the bottom-up and top-down architectures is not effective enough, and it depends on the quality of the top-level features.

To further alleviate the problem, spatial features of small areas can be lost in a deep network. The proposed method combines two adjacent layers to enrich the context information. Compared with other complex fusion methods, the representation ability of the feature graph fused by our method is not inferior because the two close-range features are highly complementary and related, and some fusion that seems to be beneficial to the feature representation is a kind of damage to both sides of the feature. Global pixel convolution (GPC) attention mechanism is introduced into the convolutional neural network so that the characteristics of different modules will change adaptively with the deepening of the network. Experimental results show that the proposed model improves the accuracy of small-target recognition in remote sensing images. In summary, our main contributions are threefold:

(i) We propose a multiscale feature fusion structure in this work, which can make full use of the attention mechanism with multiscale features to consider the supplementary effect of context information on semantics.

(ii) We propose a global pixel convolution attention mechanism, which helps to learn global pixel information, to a certain extent, overcomes the locality of traditional convolution, and better grasps the key features of the image.

(iii) In the benchmark data sets cifar10 and cifar100, our GPC attention mechanism is better than the current attention mechanisms such as CBAM [22] and SENet [23] on the accuracy and achieves the best SOTA results.

2. Related Work

2.1. Network Structure. The research of network structure cannot be ignored in deep learning. At present, the primary way to improve the accuracy is to improve the network structure [24, 25]. Since the successful use of CNN, many studies have been presented in the improvement of network structure, and a variety of structures have been proposed. The VGGNet model shows that as the depth of the network increases, the accuracy of the network continues to improve, but as the gradient propagation increases with the depth of the network, the disappearance of the gradient becomes more and more obvious, which hinders the further optimization of the network. ResNet [26] proposed an identity-based jump connection to alleviate this problem. Based on the ResNet structure, the deepening of the network has become possible. In order to deal with the instability of gradient descent caused by the large difference in pixel convolution values, batch normalization improves the

stability of the learning process and makes the gradient descent smooth. At present, the network is usually optimized in three aspects: depth, width, and base. For example, WideResNet [27], which has more convolution filters and smaller depth, broadens its width. ResNeXt [28] uses block convolution, through grouped convolution operation and multibranch convolution, to avoid the parameter explosion problem caused by increasing depth and shows that cardinality can better improve classification accuracy. DenseNet [29] iteratively concatenates the input features with the output features so that each convolution can accept the original information and further smooth the vanishing gradient problem. However, we focus on another area: the attention mechanism [30, 31]. Attention is one of the curious aspects of the human visual system. We fuse the attention mechanism with multiscale features and get better results.

2.2. Attention Mechanism. For small target detection, whether the feature can be captured is particularly important. Helping the network learn features that are useful to the task while suppressing features that are not important to the task, the attention mechanism has gradually stepped onto the stage in recent years. The attention mechanism is different from the previous methods of enhancing the network. It does not increase the width, depth, or cardinality of the network but improves the performance of the network through selective and finer calibration and weighting of the existing feature maps.

Recently, there has been a relatively successful application of the attention mechanism in deep vision learning. Squeeze and excitation network (SENet) considers the relationship between feature channels and adds an attention mechanism to feature channels. SENet automatically obtains the importance of each feature channel through learning and uses the obtained importance to enhance features and suppress features that are not important to the current task. SENet focuses on channel attention, which significantly improves network performance but does not consider spatial attention. Collaborative block attention module (CBAM) combines the attention mechanism of feature channel and feature space. CBAM automatically obtains the importance of each feature channel by learning, similar to SENet. In addition, the importance of each feature space is automatically obtained through a similar learning method. Moreover, the importance degree is used to enhance the features and suppress the features that are not important to the current task. The CBAM method of extracting spatial feature attention is as follows: after channel attention, the feature graph selected by channel importance is finally sent to the feature spatial attention module. Similar to the channel attention module, the spatial attention is pooled by channel as the unit, and the results of the two are concatenated and then convoluted to $1 * w * h$ feature graph spatial weight. Then, the dot product of the weight and the input feature is used to realize the spatial attention mechanism. This kind of spatial attention mechanism improves the network performance to a certain extent, but it is still a pooled connection mode, which takes less consideration of the overall situation of space. Nonlocal neural networks are a self-attention model proposed by Wang Xiaolong in CVPR 2018. Nonlocal neural networks [32] and nonlocal means have a similar implementation. Ordinary filtering is a 3×3 convolution kernel and then moves to the whole picture. The processing is 3×3 local information. The nonlocal means operation combines a relatively large search range and carries out weighting. The proposed nonlocal operations directly capture remote dependencies by calculating the interaction between two locations instead of being limited to adjacent points. It is equivalent to constructing a convolution kernel with the same size as the feature map to maintain more information. Although this method has been proved to be effective, it involves the matrix multiplication of $B \times c \times w \times h$ characteristic graph, and the computational difficulty cannot be ignored. Later, GCNet [33] was optimized for the difficulty of calculation, but the optimization method still maintained the essential core of matrix multiplication and did not make any substantial changes. In our spatial attention mechanism of global pixel response, we use spatial attention based on an effective architecture while maintaining a small number of parameters and verifying this architecture's feasibility through experience. In addition, our module is proved to be effective in the identification task (cifar100 and cifar10) from experience. In particular, we can achieve the state-of-the-art performance just by placing modules on top of existing models in the cifar100 test set.

3. Method

3.1. Network Structure. To better capture the feature information of small target feature maps, our proposed network architecture not only uses the multiscale feature map information but also makes full use of the contextual information of the feature map to perform fusion processing of different sizes. In addition, in order to better process the fused feature maps, the global pixel point convolution attention mechanism is used to process the fused feature maps, which further improves the network performance.

It effectively detects small targets with low-level high-resolution feature maps because the receptive field of small targets in the feature map is relatively small, and the low-level accurate positioning feature is undoubtedly helpful for detection. In this paper, a feature extractor that is consistent with the original SSD is used to generate a multi-scale feature map by Vgg-16 at twice the scaling step (an additional convolution layer is extended at the end of the truncated Vgg-16). The iterative process can be expressed as follows.

Among them, C_i is the ith convolution block of the backbone network and $f_i(x)$ is the selected feature. With the deepening of the feature layer, the index i becomes more significant, and the feature layer becomes deeper. p_n is the prediction layer responsible for converting the feature graph into classification confidence and boundary box. In this paper, the multiscale detector is further optimized by two interconnected modules, and the two modules are organically combined into a novel network. The network structure is shown in Figure 1.

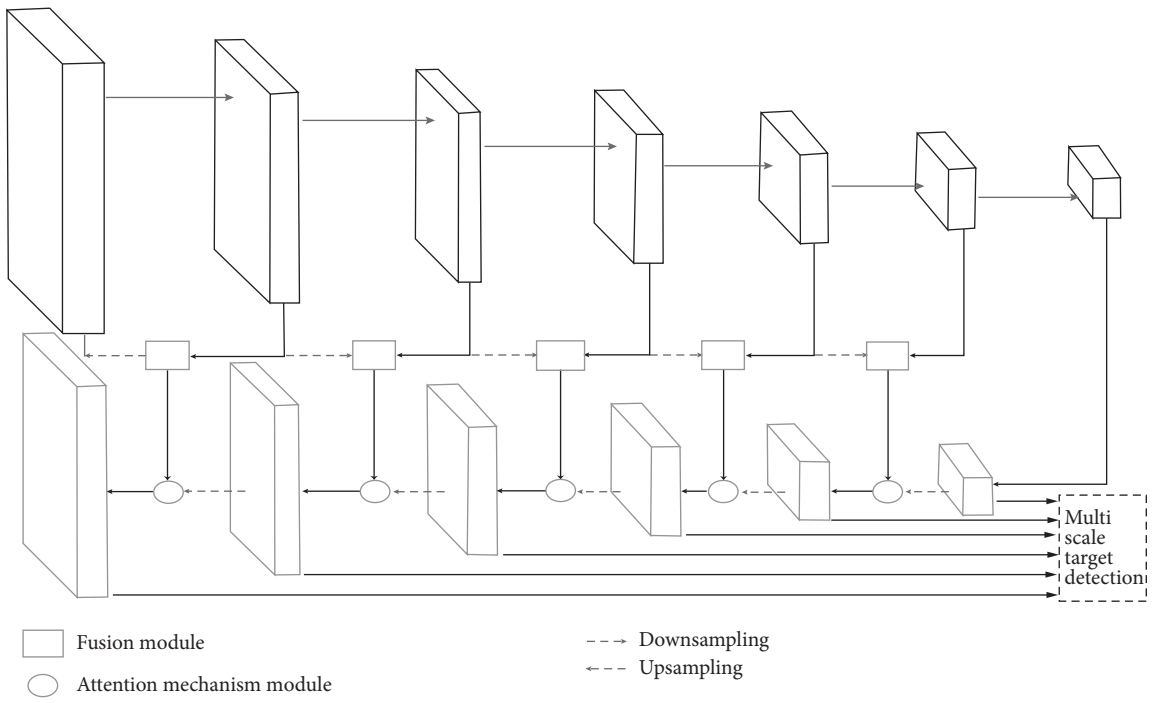

Figure 1: Network structure.

3.2. Fusion Module. Although most detectors use various multiscale structures to solve object diversity in images, the progress on small objects is not satisfactory. The reason why a large target can get reliable detection results is that its essential characteristics are not easy to be lost in propagating convolutional neural networks. Unfortunately, the detection of small objects is awkward because of the low-level features and the information defects of the high-level features.

This paper proposes a method to obtain complementary information by combining two close-range feature layers. (1) Because of the apparent difference between the two features, the reliability of deep prediction will be reduced due to the combination of shallow features. (2) The in-depth features with large receptive field usually introduce a lot of useless background noise to the shallow layer. (3) The feature layer of the close-range usually retains the most helpful information, and the convolution layer for small object detection is enhanced. Therefore, this paper focuses on the fusion between adjacent layers to capture their complementarity.

In the fusion module, as shown in Figure 2, a set of 1×1, 3×3, and 1×1 convolution kernels are used to process the shallow features with the size of $2W \times 2H \times D2$ (W, H, and $D2$ represent the width, height, and channel number of the feature graph, resp.). D in Figure 2 represents the number of convolution kernels, while $2W \times 2H \times D$ in another higher layer. In the process of feature graph processing, 1×1 convolution is added as buffers.

$$\begin{aligned} f_i(x) &= C_i(f_{i-1}(x)) = C_i(C_{i-1}(f_{i-2}(x))) \\ &= C_i(C_{i-1} \cdots C_1(f_1(x))), \end{aligned} \tag{1}$$

$$\text{detection result} = \{p_1(f_1),\ p_2(f_2), \ldots, p_n(f_n)\}. \tag{2}$$

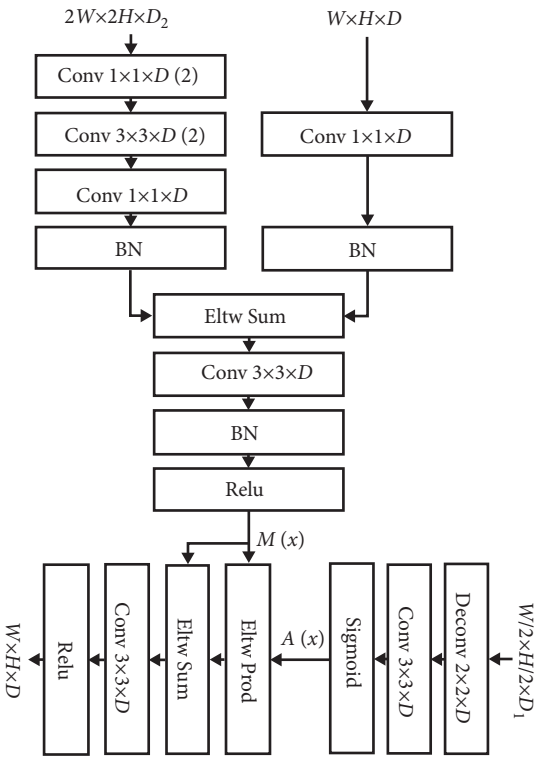

Figure 2: Internal details of the two modules.

3.3. Global Pixel Convolution Attention Mechanism. The traditional convolution operation often uses a smaller convolution kernel to perform local convolution on all feature map channels, and the receptive field is small. It relies on continuously repeating the convolution operation to

obtain a transitive receptive field. This method comprehensively processes the channel information of the feature map, but due to local convolution, it is possible to extract local noise points, which is not conducive to the extraction of key feature points by the network and affects the convergence and accuracy of the network. The global pixel point convolution operation proposed in this paper in Figure 3 convolves on the global pixel points under a specific channel, extracts useful information of the global pixel points, and can enhance the attention of the network, assist the traditional convolutional network to converge faster, and get better accuracy rate.

To make full use of the integrity of image pixels, we propose a spatial attention mechanism based on the convolution response of global pixels. It reduces the number of parameters and improves performance significantly. To avoid the problem of too many parameters caused by nonlocal, we propose the sequence change operation, using convolution to compress the pixels. Then, we realize three techniques of global pixel convolution through the weight multiplication of the original pixels: sequence arrangement, single-channel global pixel convolution, and sequence recovery in Figure 4.

Traditional convolution operation uses smaller-scale convolution kernel such as 3×3 for convolution operation. It is a kind of calculation method similar to the sliding window for local pixels in a multichannel feature map. By stacking convolution modules, the receptive field is improved.

However, the receptive field obtained by this method is still unstable. There is no way to help the network grasp the global receptive field key points at the first time. This paper proposes a spatial attention mechanism module, which aims at all the pixels in a specific channel of the feature map employing sequence change and convolution in Figure 5. This method can be expressed as follows:

$$F(x) = \text{recovery}\left(\sigma\left(f_{\text{ex}}^{3\times1}\left(\sigma f_{\text{co}}^{3\times1}\left(\text{arrangement}\left(x\right)\right)\right)\right)\right), \tag{3}$$

where arrangement and recovery are a pair of reverse operations whose main function is to adjust the dimension. Before and after the dimension of the feature image changes, the position of each pixel remains unchanged so that it can carry out the weighted multiplication operation with the original feature image pixel, X represents the input feature $x \in R^{C\times H\times W}$. For arrangement, $x \longrightarrow U \in R^{(H\times W)\times C\times 1}$. For recovery, recovery is a restore operation, $x \longrightarrow U \in R^{C\times H\times W}$.

σ refers to the sigmoid function and f^{3*1} represents a convolution operation with the filter size of 3×1. For $f_{co}^{3\times1}$, $x \longrightarrow U \in R^{(H\times W/r)\times C\times 1}$. For $f_{ex}^{3\times1}$, $x \longrightarrow U \in R^{(H\times W)\times C\times 1}$, where r is the reduction ratio and n is the filter size. In experiment, we set r to 16 and n to 3. It can be adjusted in Table 1.

The global pixel convolution attention mechanism proposed in this paper can be easily combined with almost any deep learning model, and the number of additional parameters is less, which will not affect the model itself. As a spatial attention mechanism, it can also cooperate with other attention models. In the later cifar100 ablation experiment, the global pixel convolution attention mechanism can be easily combined with almost any deep learning model. Our proposed attention mechanism can surpass the current mainstream SENet and CBAM when used alone. We take CBAM as the basic model. From the experimental results, we can see that our global pixel convolution attention mechanism, as a spatial attention mechanism, has much higher accuracy than the spatial attention mechanism in CBAM. When our GPC attention replaces CBAM's spatial attention mechanism, we even get the best result of cifar100. Taking ResNet-50 as an example, we give the way of GPC attention combination for reference in Figure 6.

4. Experiments and Results

4.1. Target Detection on Airbus Ship Data Set. In this paper, the experiment is carried out on a computer with two NVIDIA 2080ti GPUs using the MXNet framework. As with SSD, the pretrained Vgg-16 (on the ILSVRC CLS-LOC data set) is used to initialize the model for effective comparison. We use the remote sensing image ocean ship detection data set. The remote sensing image has been preprocessed by denoising, smoothing, and filtering. The data set includes a training set and test set. There are 1925526 images in the training set and 15606 images in the test set. The CSV file provides the stroke length code for the training image, which is used to locate the ship and generate the mask and bounding box of the image.

The coding information and statistical results of the training set images are shown in Table 2. The codes in Table 2 represent some rectangular boxes used to frame the vessels in the image. If the code is Nan, it means that there are no vessels in the image. The encoding string format is the start point, length, start point, length, . . ., and length of each pair (start point, length) representing a certain length of pixel line from the start point. The starting position is not the two-dimensional coordinates, but the index of a one-dimensional array so that the two-dimensional image is compressed into a one-dimensional pixel sequence. Read the run-length code. After decoding, 1 in the array represents the mask, and 0 represents the background.

The mask information is covered in the corresponding image and visualized with transparent color. The result is shown in Figure 7.

Figure 8 is a statistical chart of the images in the training set. It can be seen that 78% of the images have ships, and the number of ships in all images is 81723. Due to the class imbalance of the samples, the image without vessels is downsampled to prevent excessive noise during model training. In the images with ships, the images with 1-2 ships account for the vast majority. Too few tiny targets mean little information about small targets, which may cause the trained model to pay more attention to other information features. Therefore, to ensure the relative balance of sample types, this paper oversamples the samples containing a certain number of vessels.

Each model uses the same aerial remote sensing image as the detection data set. The processing process of different

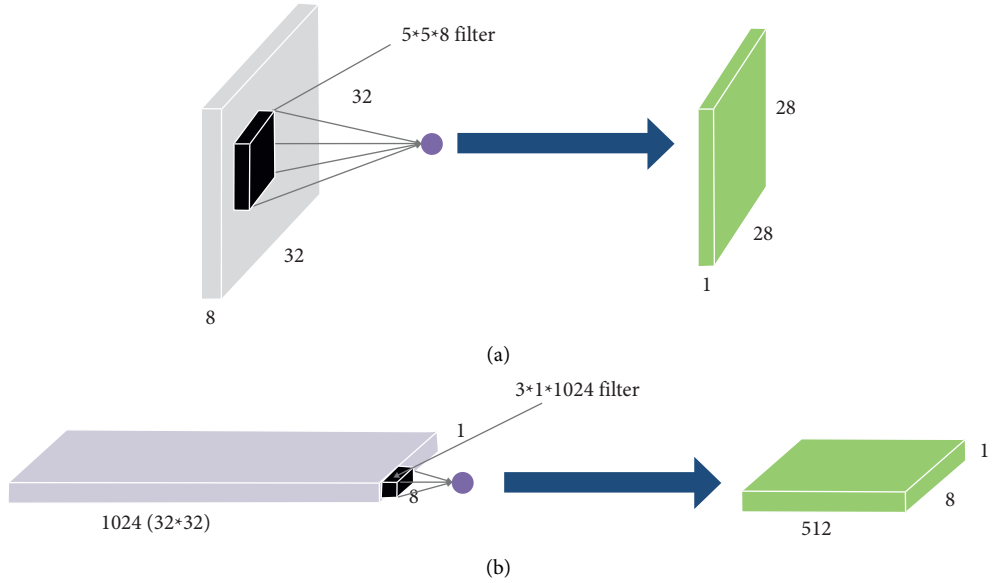

Figure 3: Comparison of GPC and traditional convolution. (a) Traditional convolution. (b) Global pixel convolution.

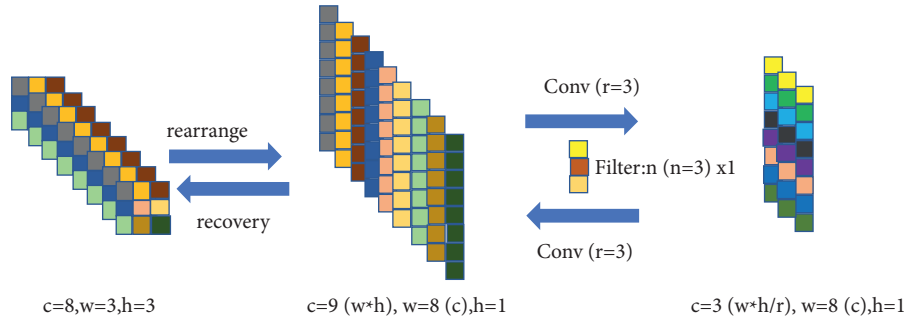

Figure 4: Global pixel convolution attention. To better illustrate, we take an $8 \times 3 \times 3$ feature map, in the left, with $r = 3$ as an example. Then, the process from left to middle is arrangement and recovery the convolutional image. The middle to right is applying global pixel convolution attention, which is different from traditional convolution.

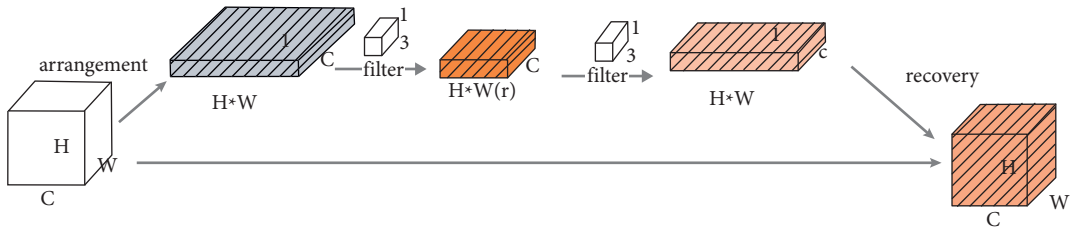

Figure 5: A global pixel convolution block.

models remains the same, training and testing in the case of taking different thresholds, recording its F2 score, and finally taking the average value of F2 score as the final F2 score of the model.

Table 3 shows the evaluation results of the seven models. The evaluation includes recall rate, accuracy rate, *F*2 score, and evaluation time. Mask R-CNN effectively utilizes mask information in the data set, so its indicators are better than the other three comparison models, but the time consumption is slightly higher. Our model adds a GPC attention mechanism in many stages, and the recommended region anchor frame is accurate, so the indicators are significantly better than other models, but the evaluation time is also slightly longer. Experimental results show that the combination of attention mechanism and SSD can effectively improve the small-target detection effect.

As can be seen from Figure 9, SSD has the phenomenon of missing detection. Our method is better than SSD. For small targets, the detection accuracy of our method is much higher than SSD with the GPC attention mechanism.

Table 1: These three columns refer to ResNet-50, SE-ResNet-50 based on the ResNet-50 backbone network, and the corresponding GPC-ResNet-50. Inside the brackets is the general shape of the residual block, including the filter size and feature size, and the optimal position for the insertion of the attention mechanism. The number of stacked blocks at each stage is shown outside the brackets. #P indicates the amount of network parameters.

Output size	ResNet-50	SE-ResNet-50 (M)	GPC-ResNet-50 (M)
32×32	Conv, $3 * 3$, 64, stride = 1		
32×32	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \\ \text{fc}, [16, 256] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 64 \\ \text{conv}, 3 \times 3, 64 \\ \text{conv}, 1 \times 1, 256 \\ \text{GPC}[n = 3, r = 16] \end{bmatrix} \times 3$
16×16	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \\ \text{fc}, [32, 512] \end{bmatrix} \times 4$	$\begin{bmatrix} \text{conv}, 1 \times 1, 128 \\ \text{conv}, 3 \times 3, 128 \\ \text{conv}, 1 \times 1, 512 \\ \text{GPC}[n = 3, r = 16] \end{bmatrix} \times 4$
8×8	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \\ \text{fc}[64, 1024] \end{bmatrix} \times 6$	$\begin{bmatrix} \text{conv}, 1 \times 1, 256 \\ \text{conv}, 3 \times 3, 256 \\ \text{conv}, 1 \times 1, 1024 \\ \text{GPC}[n = 3, r = 16] \end{bmatrix} \times 6$
4×4	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \\ \text{fc}[128, 2048] \end{bmatrix} \times 3$	$\begin{bmatrix} \text{conv}, 1 \times 1, 512 \\ \text{conv}, 3 \times 3, 512 \\ \text{conv}, 1 \times 1, 2048 \\ \text{GPC}[n = 3, r = 16] \end{bmatrix} \times 3$
1×1	Global average pool, 1000-d fc, softmax		
#P	23.7M	26.5	27.5

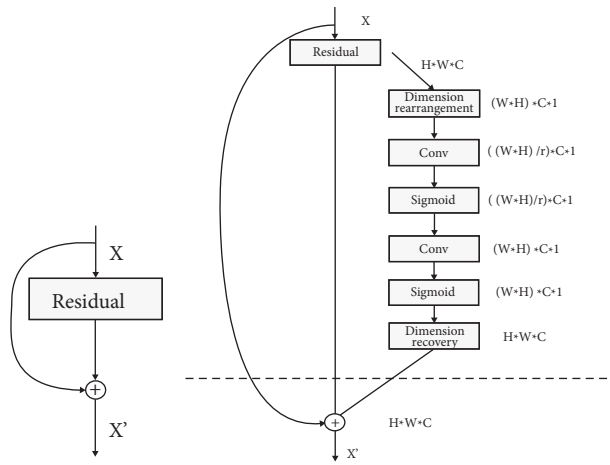

Figure 6: The schema of the original residual module (a) and the GPC-ResNet module (b).

Table 2: Encoded information of the training set images.

Image ID	File name	Encoded Pixels	Ship count
0	00003e153.jpg	NaN	0
1	0001124c7.jpg	NaN	0
2	000155de5jpg	264661 17 265429 33 266197 33 2669665 33 ...	1
3	000194a2d.jpg	360468 1 361252 4 362019 5 362785 8 ...	5
...	...	...	...
231719	fffedbb6b.jpg	NaN	0

4.2. Classification on cifar10. We combine the proposed GPC attention mechanism with the current CNN model and compare it with SOTAS based on CNN. We also compare it with the most widespread attention mechanism and ablation experiment. The experimental environment only uses a GPU based on customers: GeForce RTX 2080ti with 24 GB and a 12 Intel® Core™ i7-8700k CPU. The performance test of the GPC attention mechanism

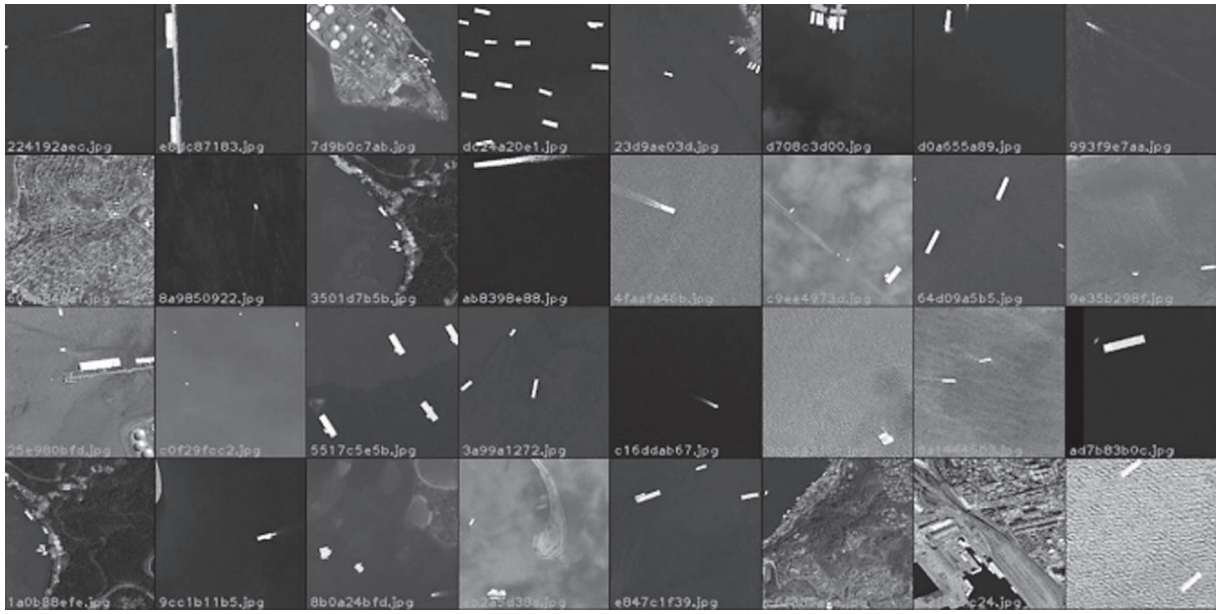

Figure 7: Parts of ships in training sets.

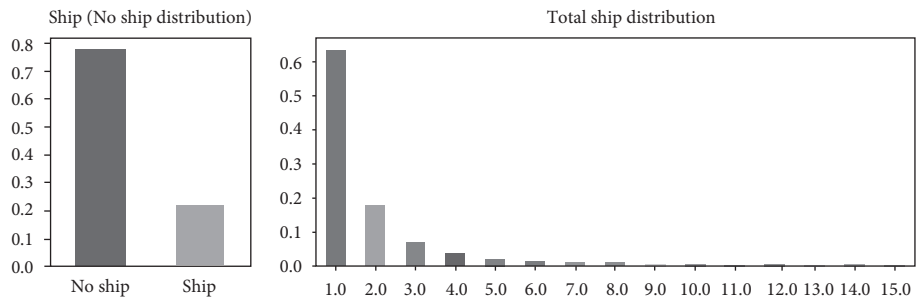

Figure 8: Statistics of ships in training set.

Table 3: Comparison of experimental results on remote sensing image test

Model	Recall (%)	Precision (%)	$F2$ (%)	Time cost (s)
UNet [34]	71.5	74.8	75.3	0.13
YOLOV3 [35]	69.4	73.1	71.6	0.15
RetinaNet [36]	68.0	71.5	70.1	0.15
Mask R-CNN [37]	74.7	75.3	76.1	0.18
SSD	76.2	78.6	76.5	0.15
SSD + CBAM	77.1	82.4	79.8	0.17
SSD + GPC (ours)	**78.1**	**83.4**	**80.2**	**0.17**

proposed in this paper is mainly carried out on benchmark data sets cifar10 and cifar100.

For cifar10 and cifar100, the same data enhancement combination is adopted on the data set: the size of the random crop is $32 * 32$, padding is 4, random horizontal flip, and normalization [38].

For cifar10, the setting of training parameters is as follows: every training process was implemented in a cosine annealing learning schedule with a half cycle. The initial learning rate is 0.1, momentum is 0.9, weight decay is 0.0005, and batch size is 100. The r (reduction ratio) is set to 16. The total epoch number is 200.

The main detection indicators are Top1 accuracy, and the parameters are compared to prove that our module will not cause a huge computational burden on the original model. The structure of reimplemented SOTAs followed the work in the link https://github.com/kuangliu/pytorch-cifar.

From Table 4, the addition of the GPC attention module improves the accuracy of the ResNet-50 baseline without causing significant influence on the parameters. Through the GPC spatial attention mechanism, the network can capture the critical features of the feature map faster, and the network can converge faster so that better results can be obtained without too deep networks. For example, the results of ResNet-50 + GPC in the table are even better than the ResNets101 and DenseNets201. After combining with channel attention in CBAM, the best result on the cifar10 data set is obtained.

4.3. Classification on cifar100. For cifar100, the setting of training parameters is as follows: the initial learning rate is 0.1, drops every 60 epochs, gamma is 0.2, momentum is 0.9,

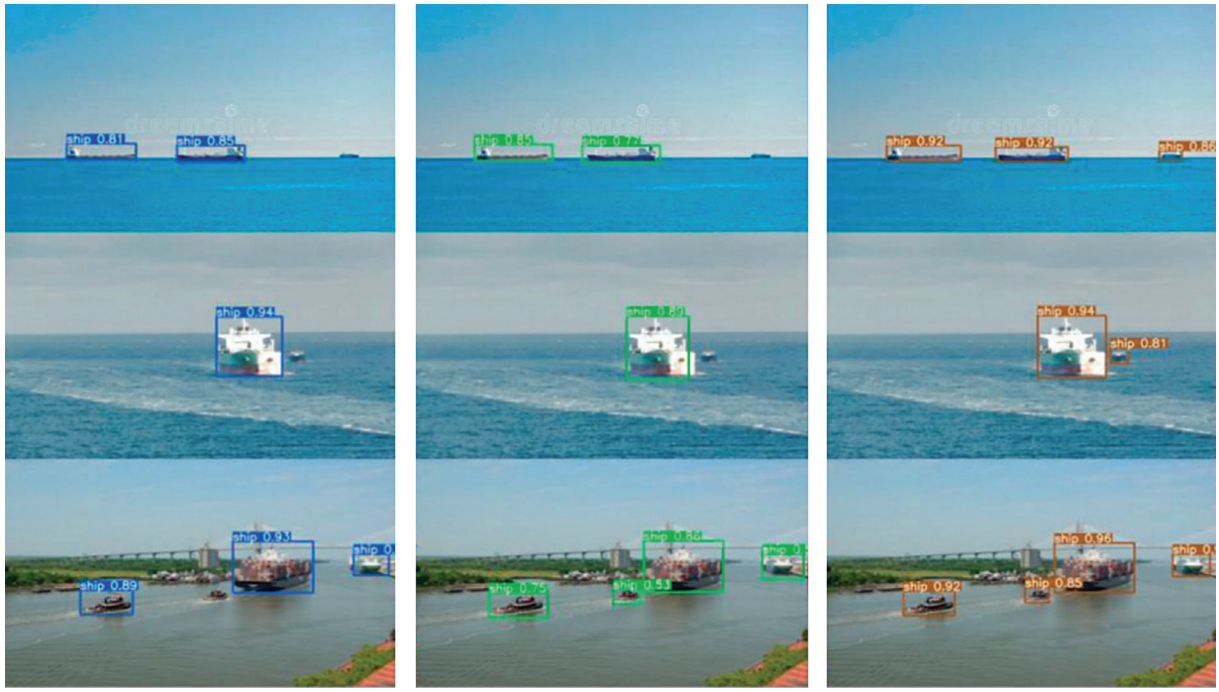

Figure 9: Comparison of detection visualization performance. (a) Original SSD. (b) SSD + CBAM. (c) Ours: SSD + GPC.

Table 4: GPCNets performance with SOTAs in cifar10.

Model	Size (M)	Test avg top1 accuracy↑
ShuffleNetsv2 [39]	1.26	91.26
EfficientNetsb0 [40]	3.60	91.35
MobileNetv2 [41]	2.30	91.71
SeNets18	1.34	93.71
VGG16	14.73	93.83
PreActResNets18	11.78	94.36
Resnets18	11.17	94.45
DenseNets121 16GR	1.76	94.81
SimpleDLA	15.14	94.99
Resnets50	**23.52**	**95.07**
DenseNets201	18.10	95.13
Resnets50 + GPC (ours)	**27.50**	**95.39**
RegNetY 400MF	5.71	95.46
DLA [42]	16.29	95.49
ResNets101	42.51	95.62
ResNeXt29 2 × 64 d	9.13	95.76
ResNeXt29 32 × 64 d	4.77	95.78
Resnets50 + channel (CBAM) + GPC	**30.00**	**95.82**

weight decay is 0.0005, and batch size is 128. The r (reduction ratio) is set to 16. The total epoch number is 200.

The primary detection indicators are Top1 error and Top5 error, and the parameters are compared to prove that our module will not cause a substantial computational burden on the original model. The structure of reimplemented SOTAs followed the work in the link https://github.com/weiaicunzai/pytorch-cifar100.

From Table 5, GPC spatial attention mechanism has better results on cifar100, which surpasses most models. From Figure 10, after adding the GPC spatial attention mechanism, ResNet-50 is better than the original model during the entire training process. Through the convolution response to the global space pixels, the network focuses on more valuable pixels to be able to achieve faster convergence and a higher accuracy rate. It improves the accuracy of the original model and has higher efficiency than the current popular SE attention mechanism and CBAM attention mechanism. Without using additional data sets and transfer learning, it gets the highest SOTAS results, and the increase of parameters is slight. In terms of accuracy, ResNet-50 + channel (CBAM) + GPC is the only model with an accuracy rate of over 80% without too many parameters.

5. Ablation Experiment on cifar100

In this section, we carry out ablation experiments to understand the impact of different parameters and configurations on the GPC module. All ablation experiments are performed on the benchmark data set cifar100, using ResNet-50 as the fundamental backbone architecture. All training parameters are the same as in Section 4.3. Only the GPC module is modified to study its impact on the results; the main comparison basis is top-1err, top-5 err5 and parameter amount.

5.1. Reduction Ratio. The reduction ratio r introduced in the GPC attention module is a compression ratio of the global pixels. The setting value of r will significantly affect the number of parameters and computational complexity. We will use ResNet-50-GPC for experiments and set a series of different r values. Table 6 shows that $r = 16$ has achieved good results, but the practical effect does not change monotonously with r. It may be that $r = 16$ is just suitable for the cifar100 data set. The size of the feature map will change

Table 5: GPCNets performance with SOTAs in cifar100.

Model	Size (M)	Top-1 error	Top-5 error
MobileNet	3.3	34.02	10.56
ShuffleNetsv2	1.3	30.49	8.49
vgg11_bn	28.5	31.36	11.85
vgg13_bn	28.7	28.00	9.71
vgg16_bn	34.0	27.07	8.84
Resnet-18	11.2	24.39	6.95
ResNet-50	23.7	**21.**61	**5.42**
ResNet-101	42.7	21.53	5.30
DenseNet-121	7.0	22.99	6.45
DenseNet-161	26	21.56	6.04
ResNeXt-50	14.8	21.50	5.39
ResNeXt-101	25.3	21.48	5.99
SE-ResNet-18	11.4	23.56	6.68
SE-ResNet-50	26.5	21.22	5.58
SE-ResNet-101	47.7	20.98	5.41
ResNet-50 + CBAM	26.22	20.54	5.12
ResNet-50 + GPC	27.5	**20.33**	**5.04**
ResNets50 + channel (CBAM) + GPC	30.00	**19.76**	**4.75**

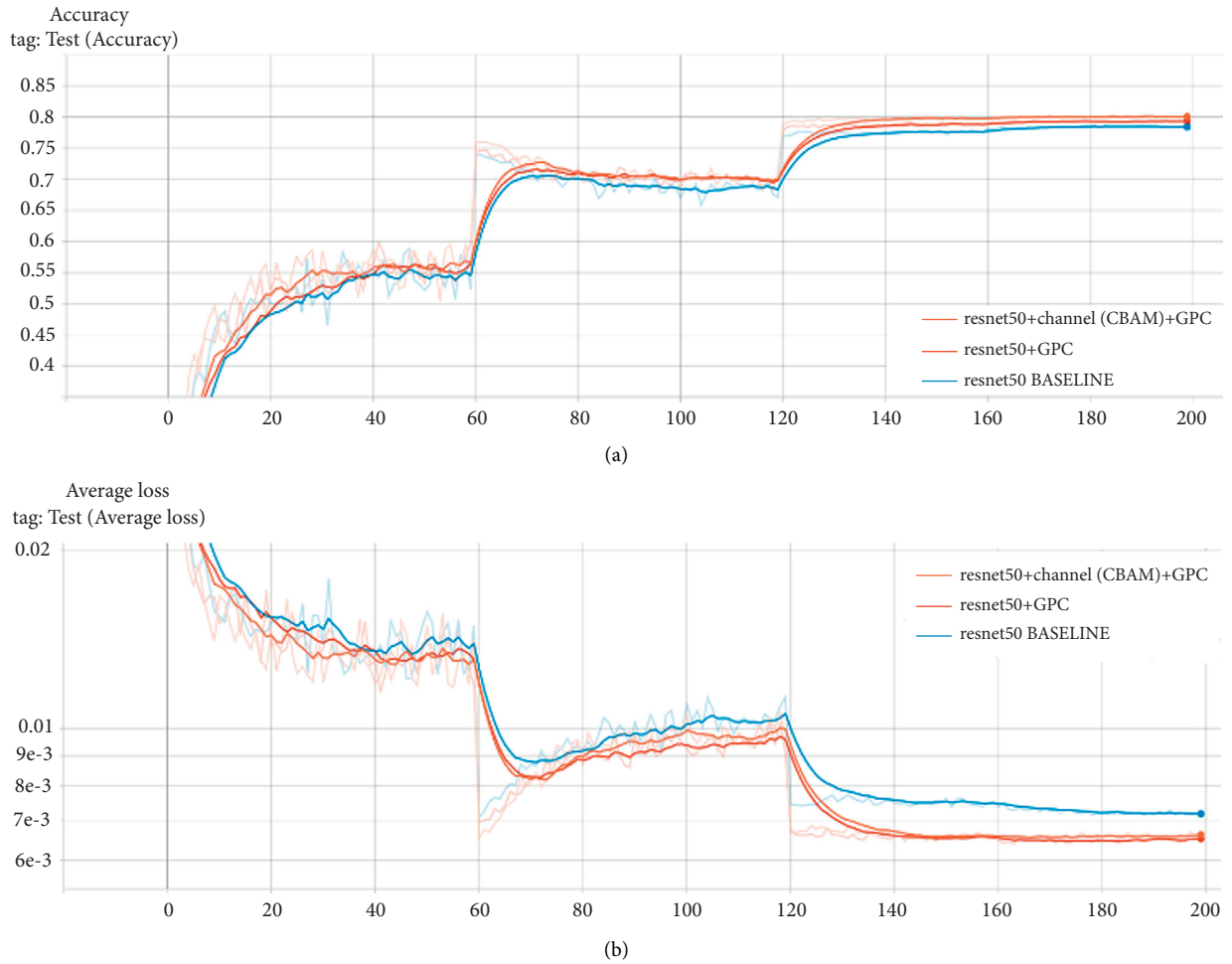

Figure 10: Acc curves and loss curves during cifar100 training. (a) Epoch-acc on cifar100. (b) Epoch-loss on cifar100.

Table 6: Single-crop error rates (%) on cifar100 and parameter sizes for GPC-ResNet-50 at different reduction ratios. Here, original refers to ResNet-50.

Ratio r	Top-1 err.	Top-5 err.	Params (M)
2	21.52	5.41	36.5
4	20.40	5.30	31.3
8	20.42	5.21	28.7
16	20.33	5.04	27.5
H ∗ W (output = 1)	20.62	5.02	26.2
Origin	21.61	5.42	23.7

Table 7: Single-crop error rates (%) on cifar100 and parameter sizes for GPC-ResNet-50 at different convolution kernels.

Design	Top-1 err.	Top-5 err.	Params (M)
1×1	20.39	5.06	24.13
3×1	**20.33**	**5.04**	27.5
5×1	20.62	5.21	28.3
7×1	20.71	5.33	29.24

Table 8: Single-crop error rates (%) on cifar100 and parameter sizes for different attention methods.

Description	Top-1 err.	Top-5 err.	Params (M)
ResNet-50 (baseline)	21.61	5.42	23.7
ResNet-50 + channel (SE)	21.22	5.58	26.5
ResNet-50 + channel (CBAM) + spatial (CBAM)	20.54	5.12	26.22
ResNet-50 + spatial (CBAM)	21.07	5.27	23.7
ResNet-50 + spatial (GPC)	**20.33**	**5.04**	**27.5**
ResNet-50 + channel (CBAM) + spatial (GPC)	19.76	4.75	30.00

when the number of layers of the network changes. With constant changes, such as in ResNet-50, the feature map will continue to decrease in half, so dynamic adjustment of the *r*-value may further improve the accuracy.

5.2. Convolution Kernel. The default convolution kernel set in the GPC attention module is 3×1 because the feature map has been flattened in the GPC module to facilitate convolution operations, so the convolution kernel can only be set to $n \times 1$. We set a series experiment with the value of n. Table 7 shows that when the value of n is slight, such as 1, 3, better results can be obtained. Multichannel improves the robustness of the network, but when there are too many channels, it will interfere with the results of global pixel convolution under different channels, which will lead to the slow network convergence.

5.3. Comparison of Different Spatial Attention Methods. We have compared the GPC spatial attention mechanism with the current mainstream attention mechanism. Table 8 shows that using GPC alone as a spatial attention module is effective, and when GPC is combined with other channel attention modules, it can produce better results, which fully illustrates the high efficiency and practicality of GPC as a spatial convolution attention mechanism.

6. Conclusion

This paper presented a global pixel convolution spatial attention mechanism that could compress the global pixel response under the same channel. Thus, it improved the overall grasp of space, integrated it with the multiscale feature fusion module proposed in this paper. The paper also developed a feature pyramid framework with an adaptive attention mechanism. The framework considered the positive influence of rich context information on classification and location, as well as a guidance of advanced semantic features on global features. Tasks on multiple data sets achieved the best performance. In addition, the GPC attention mechanism proposed in this paper was a new convolution method, which avoided the problem of too many parameters caused by the nonlocal matrix operation. The weight assignment of pixel points was completed by convolution, and good results were obtained. We believe that this convolution method offers an alternative way to consider the network architecture in the future.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

All the authors declare no conflicts of interest regarding the publication of this paper.

Acknowledgments

This project was partly sponsored by Guangdong Natural Science Foundation Project (Grant no. 2021A1515011496) and Guangdong Province Science and Technology Projects (Grant no. 2020A1010020041).

References

[1] X. Xue, X. Wu, C. Jiang, G. Mao, and H. Zhu, "Integrating sensor ontologies with global and local alignment extractions," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 6625184, 10 pages, 2021.

[2] X. Xue and J. Zhang, "Matching large-scale biomedical ontologies with central concept based partitioning algorithm and adaptive compact evolutionary algorithm," *Applied Soft Computing*, vol. 106, pp. 1–11, 2021.

[3] X. Xue, C. Yang, C. Jiang, P.-W. Tsai, G. Mao, and H. Zhu, "Optimizing ontology alignment through linkage learning on entity correspondences, complexity," vol. 2021, Article ID 5574732, 12 pages, 2021.

[4] Y. Xu and T. T. Qiu, "Human activity recognition and embedded application based on convolutional neural network," *Journal of Artificial Intelligence Technology*, vol. 1, p. 10.

[5] L. K. Sára and L. Ágnes, "Tracking statistical learning online: word segmentation in a target detection task," *Acta Psychologica*, vol. 215, 2021.

[6] C. Gao, Y. Wu, and H. Xiaohui, "Hierarchical suppression based matched filter for hyperspertral imagery target detection," *Sensors*, vol. 21, no. 1, 2020.

[7] Y. Chen and D. L. Gennaro, "Technologies supporting artificial intelligence and robotics application development," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 1, pp. 1–8.

[8] A. P. Kaur, T. Sattar, R. Anvo, and M. O. Tokhi, "Development of a robot for in-service radiography inspection of subsea flexible risers," *Journal of Artificial Intelligence and Technology*, vol. 1, no. 3, pp. 180–187.

[9] H. Wang, H. Wang, X. Tong, and F. Lu, "Deep learning based target detection algorithm for motion capture applications," *Journal of Physics: Conference Series*, vol. 1682, no. 1, 2020.

[10] H. S. Basavegowda and G. Dagnew, "Deep learning approach for microarray cancer data classification," *Journal of Intelligent Technology*, vol. 5, no. 1, pp. 22–33, 2020.

[11] Y. Xing and J. Zhu, "Deep learning-based action recognition with 3D skeleton: a survey," *Journal of Intelligent Technology*, vol. 6, no. 1, p. 13, 2016.

[12] S. DeliaGeorgiana, C. RaduIoan, and D. Ciprian, "Vehicle detection in overhead satellite images using a one-stage object detection model," *Sensors*, vol. 20, no. 22, 2020.

[13] X. Hou, J. Ma, and S. Zang, "Airborne infrared aircraft target detection algorithm based on YOLOv4-tiny," *Journal of Physics: Conference Series*, vol. 1865, no. 4, 2021.

[14] W. Liu, D. Anguelov, D. Erhan et al., "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, October 2016.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Proceedings of the Conference on Neural Information Processing Systems*, Montreal, Canada, December 2015.

[16] T. Y. Lin, P. Dollar, and R. Girshick, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, Honolulu, HI, USA, July 2017.

[17] F. Li, W. Jin, C. Fan, and Z. Lian, "PSANet: pyramid splitting and aggregation network for 3D object detection in point cloud," *Sensors (Basel, Switzerland)*, vol. 21, no. 1, 2020.

[18] Z. Cai, Q. Fan, and R. S. Fe Ris, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proceedings of the European Conference on Computer Vision*, Springer International Publishing, Amsterdam, Netherlands, October 2016.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[20] T. Kong, A. Yao, and Y. Chen, "HyperNet: towards accurate region proposal generation and joint object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las vegas, NV, USA, June 2016.

[21] C. Y. Fu, W. Liu, and A. Ranga, "DSSD: Deconvolutional Single Shot Detector," ArXiv, 2017.

[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, Munich, Germany, September 2018.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, Salt Lake City, UT, USA, June 2018.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[25] C. Szegedy, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[26] X. He and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[27] S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016. arXiv preprint arXiv:1605.07146.

[28] S. Xie, R. Girshick, and P. Dollár, "Aggregated Residual Transformations for Deep Neural Networks," IEEE, ArXiv, 2016.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, Honolulu, HI, USA, July 2017.

[30] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *Proceedings of the ICLR*, Toulan, France, April 2017.

[31] A. Vaswani, "Attention is all you need," arXiv preprint arXiv: 1706.03762, 2017.

[32] X. Wang, R. Girshick, and A. Gupta, "Non-local neural networks," ArXiv, 2017.

[33] Y. Cao, J. Xu, and S. Lin, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," arXiv, 2019.

[34] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation*, Springer, Cham, Switzerland, 2015.

[35] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," arXiv e-prints, 2018.

[36] T. Y. Lin, P. Goyal, and R. Girshick, "Focal loss for dense object detection," in *Proceedings of the IEEE Transactions on Pattern Analysis & Machine Intelligence*, pp. 2999–3007, IEEE, Venice, Italy, August 2017.

[37] Y. Gan, Y. Gan, S. You et al., "Object detection in remote sensing images with mask R-CNN," *Journal of Physics: Conference Series*, vol. 1673, no. 1, 2020.

[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," arXiv preprint arXiv: 2012.12877, 2020.

[39] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, Munich, Germany, September 2018.

[40] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," ArXiv, 2019.

[41] M. Sandler, A. Howard, and M. Zhu, "MobileNetV2: inverted residuals and linear bottlenecks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Salt Lake, UT, USA, June 2018.

[42] F. Yu, D. Wang, and E. Shelhamer, "Deep layer aggregation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Salt Lake, UT, USA, June 2018.