Large Kernel Disassembling Attention Mechanism for Remote Sensing Object Detection

Y.X. Geng, L. Wang and Y.G Wang

***Abstract*—In recent years, remote sensing object detection has become a research hotspot in computer vision tasks. However, previous approaches for remote sensing object detection often overlook the rich contextual information in images, which is crucial for accurately detecting occluded or interconnected objects using convolutional neural networks. To capture this contextual information, we propose a method called the Large Kernel Disassembling (LKD) Attention Mechanism. LKD breaks down large convolutional kernels to provide a larger receptive field to the convolutional neural networks, enabling them to capture rich contextual information in remote sensing images and enhance their performance. We employ an adaptive channel submodule and a deep convolutional spatial submodule. The adaptive channel submodule helps the network learn relationships between different channels, while the deep convolutional spatial submodule aids in extracting rich spatial features. We evaluate the proposed attention mechanism on the DIOR dataset and compare it with several recent attention mechanisms on the SSDD dataset. Experimental results demonstrate the superiority of LKD in terms of performance over other methods, validating the effectiveness of the Large Kernel Disassembling attention mechanism in remote sensing object detection tasks.**

***Index Terms*—Attention Mechanism, Remote Sensing Object Detection, Convolutional Neural Network, Yolov8.**

I. Introduction

REMOTE sensing object detection [1] is an application domain in computer vision that focuses on identifying and detecting various objects in remote sensing images, such as ships and airplanes. Currently, convolutional neural networks (CNNs) are widely applied in remote sensing object detection tasks. For example, R-CNN [2], Faster R-CNN [3], Transformer [4], and the YOLO [5], [6] series have shown good performance in various computer vision tasks. In recent years, the main research trend in remote sensing object detection has been to generate accurate bounding boxes that are suitable for the orientation of detected objects, rather than using simple horizontal boxes. Therefore, most research focuses on oriented bounding boxes for remote sensing object detection. These oriented bounding boxes are mainly implemented using specialized detection frameworks, such as RoI Transformer [7], Oriented R-CNN [8], and Oriented RepPoints [9].

Manuscript received April 14, 2024; revised July 26, 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 71472081.

Y. X. Geng is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: gen9yanx1n@ustl.edu.cn).

L. Wang is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: wangli9966@ustl.edu.cn).

Y. G. Wang is a postgraduate student at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: wyigeng@163.com).

With the development of deep learning, CNNs have been continuously improved and expanded. Researchers have proposed various innovative network structures and optimization techniques, such as residual connections, attention mechanisms [10], and depth-wise separable convolutions. The introduction of these techniques has further enhanced the performance of CNNs, achieving breakthrough results in various visual tasks. Attention mechanisms in neural network models guide the model to focus on important features, thereby improving feature extraction capabilities. By using attention mechanisms, neural network models can concentrate on the most informative and crucial features in an image while suppressing less important ones. Our research aims to leverage attention mechanisms to enhance the feature extraction capability of convolutional neural networks in remote sensing object detection tasks. To achieve this goal, we designed a novel attention mechanism module called "Large Kernel Disassembling (LKD) Attention Mechanism" that expands the receptive field of convolutional neural networks by disassembling large convolutional kernels, thereby capturing more contextual information in remote sensing images. The LKD attention mechanism sequentially applies spatial and channel submodules, allowing the model to better learn spatial and channel information and fuse them through convolutional operations to extract meaningful features.

II. Related Work

A. Remote Sensing Object Detection Framework

R-CNN [2] can be regarded as the pioneering work in utilizing deep learning for object detection. It first generates candidate regions in an image and extracts features using a deep network. Then, the features are fed into a classifier to determine the object class, and finally, regression is used to refine the positions of the candidate bounding boxes. High-performance frameworks for remote sensing object detection often rely on the R-CNN structure. Subsequently, several variants were proposed, including Fast R-CNN [11] and Faster R-CNN [3]. Fast R-CNN addressed the problem of not being able to output bounding boxes and labels simultaneously, while Faster R-CNN solved the issue of selective search inefficiency.

The YOLO series is a single-stage regression approach based on deep learning, while R-CNN, Fast R-CNN, and Faster R-CNN are two-stage classification methods. YOLO is a fast and accurate object detection algorithm widely used in computer vision. YOLO treats object detection as a regression problem, solving it with an end-to-end network that takes the original image as input and outputs the object positions and categories.

In recent years, several variants of the R-CNN framework have been proposed. The two-stage RoI Transformer [7] uses

fully connected layers in the first stage to generate rotated candidate horizontal anchor boxes. It then extracts features within these boxes for further regression and classification. Oriented RCNN [8] introduces a novel box encoding system to address the training loss instability caused by the periodicity of rotation angles. Oriented RepPoints [9] proposes an adaptive point learning approach to capture the geometric information of arbitrarily oriented objects.

B. Attention Mechanism

Currently, there have been several studies focusing on the performance of attention mechanisms in image detection tasks. Squeeze-and-Excitation Networks [12] (SENet) transform the feature maps of each channel into channel descriptors using global average pooling. Then, a fully connected layer applies a non-linear transformation to these channel descriptors, generating a weight vector as the excitation values for each channel. Finally, these excitation values are multiplied with the original feature maps to obtain weighted feature maps. SENet is computationally efficient and capable of effectively extracting global features, but it does not explicitly consider spatial correlations.

Convolutional Block Attention Module [13] (CBAM) sequentially places channel and spatial sub-modules. The channel attention module is used to compute the importance of each channel, enabling better differentiation between features in different channels. The spatial attention module calculates the importance of each pixel in the spatial domain, allowing better capture of the spatial structure in the image. However, CBAM can only capture local information and lacks the ability to capture long-range dependencies.

Coordinate Attention [14] (CA) incorporates positional information into channel attention, allowing mobile networks to access information from a broader area while minimizing computational overhead. To preserve positional details, CA decomposes channel attention into two parallel 1D feature encodings instead of using 2D global pooling. This approach efficiently integrates spatial coordinate information. The CA mechanism considers inter-channel relationships alongside positional details, enhancing the accuracy of localizing and recognizing target regions by capturing direction and position-sensitive information across channels.

Global Attention Mechanism [15] (GAM) is similar to CBAM as it also utilizes both channel attention and spatial attention mechanisms. In the channel attention, the input feature map undergoes max pooling and average pooling, followed by separate MLP processing, and finally passed through a sigmoid activation. In the spatial attention, the feature map is max pooled and average pooled, stacked together, convolved, and then passed through a sigmoid activation function.

Recent work has proposed the use of large convolutional kernels to capture feature maps. Large Kernel Attention [16] (LKA) incorporates local spatial convolution, spatial long-range convolution, and channel convolution. LKA combines the advantages of convolution and self-attention, including local structural information, long-range dependencies, and adaptability. It also avoids the drawbacks of ignoring adaptability in the channel dimension. Large Selective Kernel [17] (LSK) introduces a large selective kernel network that dynamically adjusts its large spatial receptive field, allowing for better distance estimation of various objects in remote sensing imagery.

TABLE I
THEORETICAL EFFICIENCY COMPARISON OF TWO REPRESENTATIVE EXAMPLES WITH 64 CHANNELS.K: CONVOLUTION KERNEL SIZE, D: DILATION RATE.

RF	(k, d)sequence	Params	FLOPs
23	(23, 1)	0.213M	8.9G
23	$(5,1) \rightarrow (7,3)$	0.143M	8.2G
29	(29, 1)	0.288M	9.8G
29	$(5,1) \rightarrow (7,4)$	0.143M	8.2G
29	$(3,1) \rightarrow (5,2) \rightarrow (7,3)$	0.179M	9.6G

III. Large Kernel Disassembling Attention Mechanism

Our work focuses on designing an attention mechanism module that enhances the receptive field of the neural network model and captures more contextual information in remote sensing imagery. We adopt a sequential spatial channel attention mechanism, which is the opposite of the order used in CBAM and GAM, and we have redesigned the sub-modules. The process is illustrated in Figure 1 and formalized in equations. Given the input feature map $\mathrm{F}_1 \in \mathrm{R}^{\mathrm{c}\times\mathrm{h}\times\mathrm{w}}$, the intermediate state F_2 and the output result F_3 are defined as follows:

$$F_2 = M_s\left(F_1\right) \otimes F_1 \tag{1}$$

$$F_3 = M_c\left(F_2\right) \otimes F_2 \tag{2}$$

Where M_s and M_c are spatial and channel attention maps, respectively, and $\otimes$ denotes element-wise multiplication.

A. Spatial Attention Module

In the spatial attention module, in order to capture contextual information between objects of different scales in remote sensing images, we believe it is necessary to construct a series of multiple long-range contexts. Therefore, we propose to decompose the larger convolutional kernel into a series of depth-wise separable convolutions, where the size of the convolutional kernel and the dilation rate gradually increase. Specifically, the definition of the size k, dilation rate d, and receptive field RF of the i-th depth-wise separable convolution in the series of long-range contexts is as follows:

$$k_{i-1} \leq k_i; d_1 = 1, d_{i-1} < d_i \leq RF_{i-1} \tag{3}$$

$$RF_1 = k_1, RF_i = d_i(k_i - 1) + RF_{i-1} \tag{4}$$

Increasing the size of the convolutional kernel and the dilation rate allows for a rapid expansion of the receptive field. We set an upper limit on the dilation rate to ensure that dilated convolutions do not introduce gaps between feature maps. For example, in Table 1, we decompose a large convolutional kernel into 2 or 3 depth-wise separable convolutions, which theoretically have receptive field sizes of 23 and 29, respectively.

This design has two advantages. Firstly, disassemblinga large convolutional kernel allows us to generate multiple

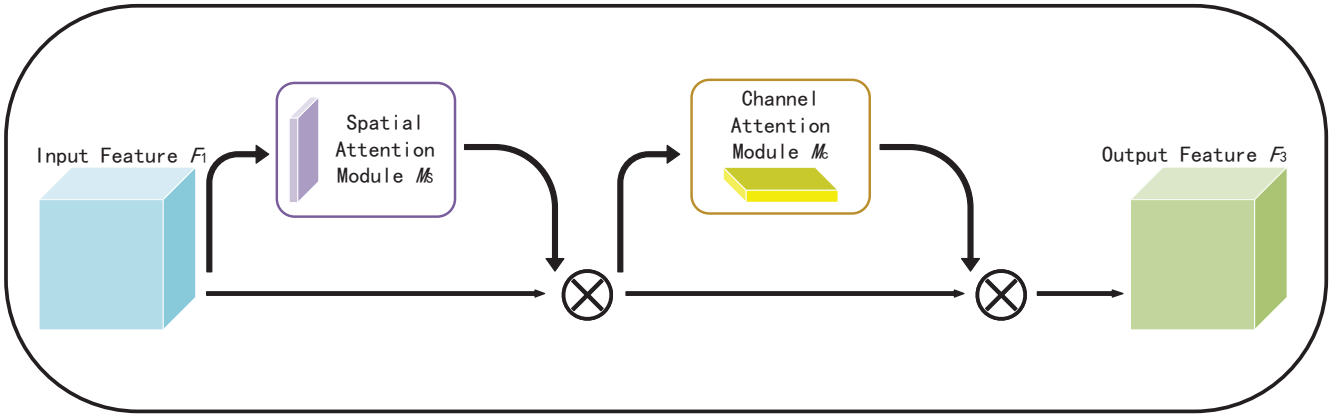

Fig. 1. The overview of LKD. The module has two sequential sub-modules: spatial and channel.

features with different receptive field sizes, enabling us to capture richer contextual information. Secondly, using multiple depth-wise separable convolutions is more efficient than using a single large convolutional kernel. The data in Table 1 shows that, under the same theoretical receptive field, our decomposition significantly reduces the number of parameters of the model, alleviating the training burden. In Figure 2, we illustrate the workflow of the spatial attention module, demonstrating intuitively how the disassemble large convolutional kernel sequence collects the receptive fields of different objects.

For the input feature F, we utilize a series of depth-wise separable convolutions with varying receptive field sizes, $\mathrm{Conv_i}^{dw}(\cdot)$ representing depth-wise separable convolutions with convolution kernels $\mathrm{k_i}$ and dilation rates $\mathrm{d_i}$:

$$U_0 = F, U_{i+1} = Conv_i^{dw}(U_i) \tag{5}$$

Assuming there are N disassemble convolutional kernels, each kernel is further processed through a 1×1 convolutional layer, denoted as $\mathrm{Conv}^{1\times1}(\cdot)$:

$$\widetilde{U}_\iota = Conv^{1\times1}(U_i), i \in [1, N] \tag{6}$$

To enhance the network's ability to focus on the most relevant contextual information, we concatenate the features obtained from different convolutional kernels:

$$\widetilde{U} = Cat(\widetilde{U_1}, \widetilde{U_2}) \tag{7}$$

For each spatial attention map $\widetilde{U}$, we apply the Sigmoid activation function to obtain the spatial mask of the disassemble large convolutional kernel:

$$SM = Sigmod(\widetilde{U}) \tag{8}$$

The features obtained from the disassemble large convolutional kernel sequence are weighted with the spatial mask and fused through the Convolutional layer ($\mathrm{Conv}(\cdot)$) to obtain the attention feature S:

$$S = Conv(\sum_{i=1}^{N}(SM \cdot \widetilde{U})) \tag{9}$$

The final output of the spatial attention module is the element-wise multiplication between the input feature F and the attention feature S, similar to some previous methods[18], [19]:

$$Y = F \otimes S \tag{10}$$

B. Channel Attention Module

Building upon the spatial attention module, we have incorporated a channel attention module that allows for capturing encoding information across channels. The entire workflow of the channel attention module is depicted in Figure 3.

In the channel attention module, we first utilize global average pooling to compress the features, which allows us to condense the global channel information into channel descriptors while addressing inter-channel dependencies:

$$X = GAP(Y) \tag{11}$$

To leverage the channel information obtained through the pooling operation, we perform the following operations to fully capture the dependencies among channels. In order to capture these dependencies effectively, we employ a simple gate mechanism with a Sigmoid function. We introduce a fully connected layer on top of the ReLU non-linear function to parameterize the gate mechanism:

$$\tilde{X} = Relu(Fc(X)) \tag{12}$$

Next is an additional fully connected layer that increases the dimensionality, which is used to transform the features back to the original channel dimension:

$$\check{X} = Fc(\tilde{X}) \tag{13}$$

The obtained features are weighted using the Sigmoid function to obtain the attention weight vector S on the channel dimension:

$$S = Sigmod(\check{X}) \tag{14}$$

The original input Y is element-wise multiplied with the obtained weight vector S to obtain the final weighted feature Z, which is then outputted:

$$Z = S \otimes Y \tag{15}$$

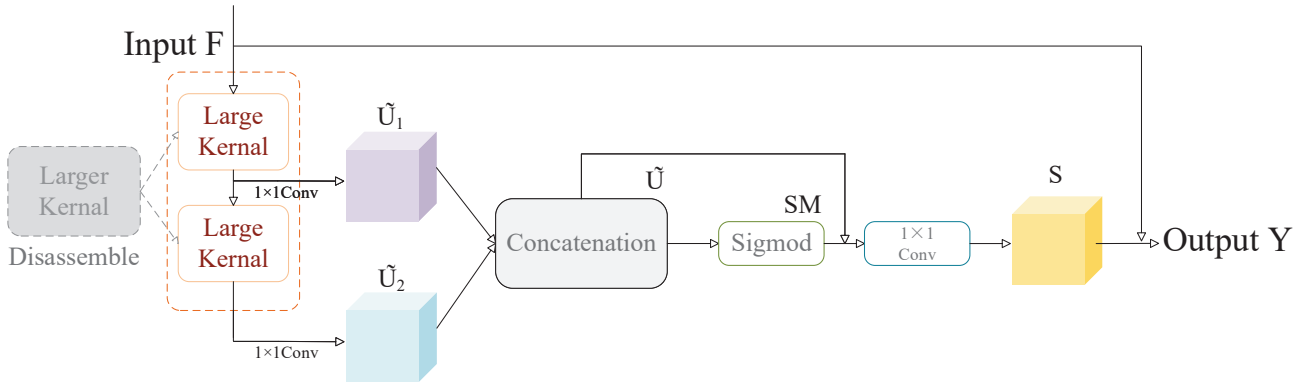

Fig. 2. The overview of Spatial Attention Module.

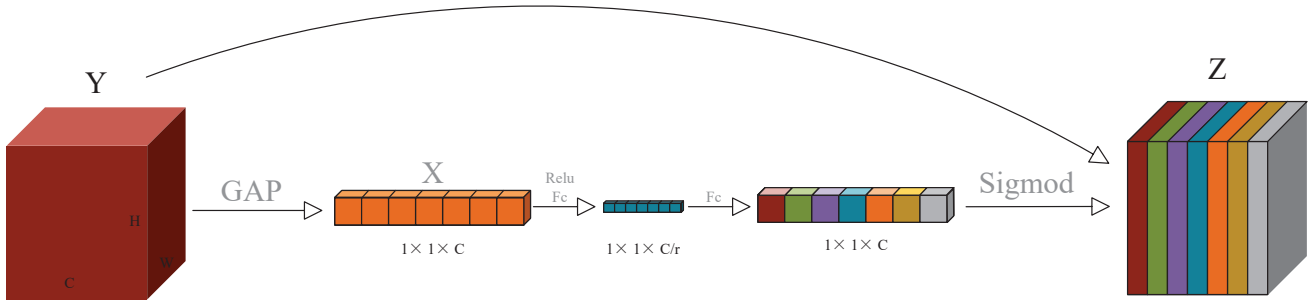

Fig. 3. The overview of Channel Attention Module.

IV. Experiments

A. Datesets

DIOR [20] is a large-scale benchmark dataset for optical remote sensing image object detection, consisting of 23,463 remote sensing images. It comprises 190,288 instances of 20 object categories: airplane, airport, baseball field, basketball court, bridge, chimney, dam, highway service area, highway toll station, harbor, golf course, ground track field, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. The images in the dataset have a size of 800×800 pixels and spatial resolutions ranging from 0.5m to 30m.

SSDD [21] is a synthetic aperture radar (SAR) dataset specifically designed for ship detection. It consists of 1,160 images and 2,456 ship instances, with an average of 2.12 ships per image. The dataset was created by the Department of Electronic and Information Engineering at the Naval Aeronautical University. SSDD dataset is the first publicly available dataset dedicated to ship object detection based on SAR images.

DOTA [25] is a large-scale dataset designed for object detection in aerial images. It serves to develop and evaluate detectors for objects captured by various sensors and platforms. The images range in size from 800×800 to 20,000×20,000 pixels, depicting objects of diverse scales, orientations, and shapes. Currently, DOTA has three versions. DOTA-v1.0 includes 15 common categories with 2,806 images and 188,282 instances.

TABLE II
Ablation of Spatial and Channel Attention Modules

Architecture	Parameters	FLOPs	mAP50	mAP50-95
Yolov8	3.01M	8.1G	84.76%	60.84%
Yolov8+sp	3.15M	8.2G	85.67%	61.44%
Yolov8+ch	3.02M	8.1G	84.83%	60.72%
Yolov8(sp+ch)	3.16M	8.2G	85.84%	61.54%
Yolov8(ch+sp)	3.16M	8.2G	85.37%	61.25%

B. Ablation Study

We conducted ablation experiments using YOLOv8 on the DIOR dataset to evaluate the contributions of the spatial attention module and the channel attention module separately. Additionally, we experimented with different orders of placement between these modules.

To better understand the contributions of the spatial and channel attention modules, we conducted ablation experiments by separately adding each module. For instance, 'sp' denotes the addition of only the spatial attention module, while 'ch' denotes the addition of only the channel attention module. The results are presented in Table 2. We observed performance improvements in these experiments, indicating that both the spatial and channel attention modules contribute to enhancing performance. Moreover, the order of placement also influences the experimental outcomes, and we found that the best results were achieved when the spatial and channel attention modules were placed in a specific sequence.

TABLE III
EVALUATION STUDY OF LARGE CONVOLUTION KERNEL DISASSEMBLY

(k1, d1)	(k2, d2)	(k3, d3)	Parameters	mAP50
(23, 1)	-	-	0.213M	85.2%
(5, 1)	(7, 3)	-	0.143M	85.84%
(29, 1)	-	-	0.288M	85.13%
(5, 1)	(7, 4)	-	0.143M	85.60%
(3, 1)	(5, 2)	(7, 3)	0.179M	85.30%

TABLE IV
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DIOR DATASET.

Architecture	Parameters	FLOPs	mAP50	mAP50-95
Yolov8	3.01M	8.1G	84.76%	60.84%
Yolov8+SE	3.02M	8.1G	84.83%	60.72%
Yolov8+CA	3.01M	8.1G	84.72%	60.51%
Yolov8+CBAM	3.08M	8.1G	85%	60.58%
Yolov8+GAM	3.85M	8.8G	84.99%	60.92%
Yolov8+LKA	3.09M	8.2G	85.71%	61.20%
Yolov8+LSK	3.13M	8.2G	85.60%	61.22%
Yolov8+LKD	3.16M	8.2G	85.84%	61.25%

C. Evaluation Study

In the spatial attention module, under the same theoretical receptive field, we demonstrated the effectiveness of disassemblinglarge convolutional kernels and showed that the performance and parameter efficiency are better compared to using a single large convolutional kernel. This validates the effectiveness of our decomposition approach. The determination of the number of disassembled kernels in the LKD module is crucial, and we follow Formula 4 to configure the disassembled kernels. We experimented with different numbers of disassembled kernels when the theoretical receptive fields were 23 and 29, and the results are shown in Table 3. Under the theoretical receptive field of 23, disassemblingthe large convolutional kernel into two depth-wise separable convolutions achieved the best performance and had fewer parameters compared to the single large convolutional kernel.

D. Main Results

1) Results on DIOR: We evaluated the performance of LKD, along with six other attention mechanisms, using the YOLOv8 model on the DIOR remote sensing image dataset. The results in Table 4 demonstrate that our LKD outperforms the other methods in terms of performance.

To ensure fairness, our model training was conducted in the same manner. The experiments were performed on two NVIDIA GeForce RTX 3090 GPUs, with a total of 100 epochs. The batch size was set to 32, and we utilized the SGD optimizer with an initial learning rate of 0.01, a final learning rate of 0.00001, a momentum of 0.937, and a weight decay of 5e-4. We employed a warm-up phase of 3 epochs with a warm-up momentum of 0.8.

We performed feature visualization analysis using Eigen CAM [22] on several attention modules, and the results are shown in Figure 4. In the figure, we can observe that disassembling the large convolutional kernel allows the model's attention to focus more on the objects of interest.

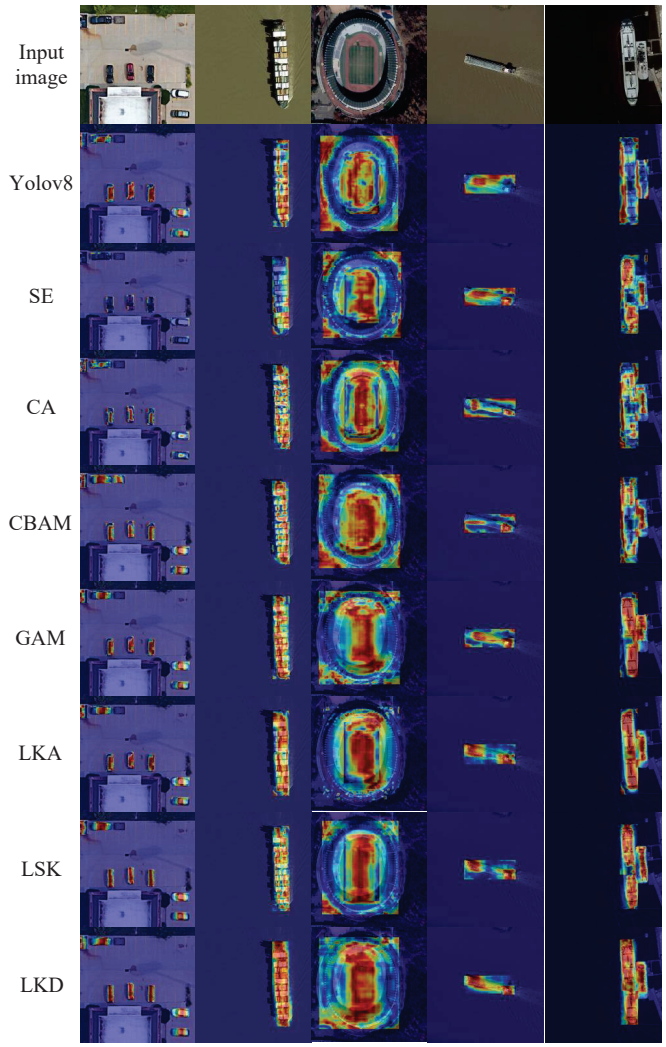

Fig. 4. Feature visualization with Eigen CAM.

This demonstrates the effectiveness of our LKD attention module.

2) Results on SSDD: With the assistance of the MMRotate [23] open-source toolkit, we conducted detection experiments on the SSDD [21] dataset with rotated remote sensing images. MMRotate decouples the task of rotated bounding box detection into various modular components. By combining different modular components, we can easily construct customized algorithms for rotated bounding box detection.

We utilized ResNet50 [24] as the backbone network for our model. The rotation method was defined as "le135". We selected Oriented RepPoints as the detector, with the FPN serving as the neck. The batch size was set to 2, and we conducted experiments for a total of 40 epochs. The main results of the experiments are presented in Table 5. In the results, we can observe that our LKD attention module improves the accuracy of remote sensing image detection tasks, even when using different models, datasets, and detection methods. We conducted tests on the experimental results, and the visualization of the effects is shown in Figure 5.

3) Results on DOTAv1.0: We compare our LKD with the other methods on the DOTAv1.0 dataset, as reported in Table 6. Our LKD achieve state-of-the-art performance with

TABLE V
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE SSDD DATASET.("INSHORE" REPRESENTS THE AREA CLOSE TO THE COASTLINE, "OFFSHORE" REFERS TO THE AREA FAR FROM THE COASTLINE, AND "ALL" REPRESENTS A COMBINATION OF BOTH.)

Architecture	Parameters	FLOPs	mAP(Inshore)	mAP(Offshore))	mAP(All))
ResNet50	36.6M	66.86G	45.77%	75.80%	66.88%
ResNet50+SE	39.11M	66.9G	49.71%	76.26%	68.04%
ResNet50+CBAM	56.54M	66.93G	49.78%	76.68%	71.41%
ResNet50+LKA	57.63M	94.22G	46.64%	76.95%	67.82%
ResNet50+LKD	60.34M	96.8G	50.76%	78.69%	72.89%

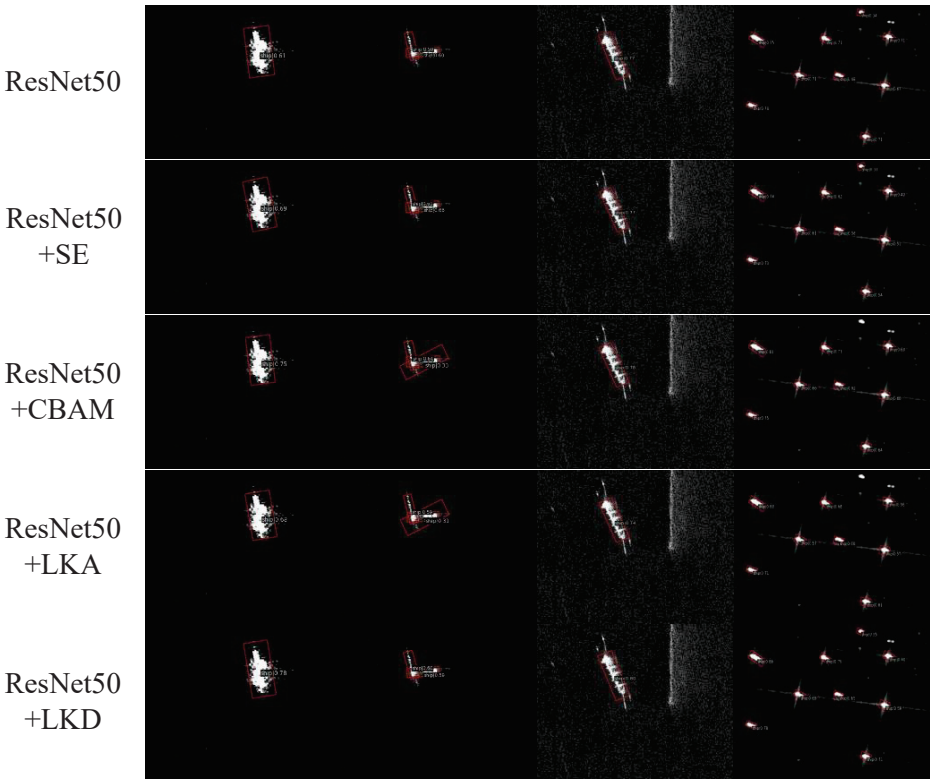

Fig. 5. The visualization of the results on the SSDD dataset.

TABLE VI
COMPARISON WITH STATE-OF-THE-ART METHODS ON THE DOTAV1.0 DATASET.

Architecture	Parameters	FLOPs	mAP50	mAP50-95
Yolov8	3.08M	8.3G	74.4%	56.8%
Yolov8+SE	3.09M	8.4G	76.9%	58.3%
Yolov8+CBAM	3.15M	8.4G	77.3%	59.3%
Yolov8+LKA	3.17M	8.4G	77.39%	59.33%
Yolov8+LKD	3.23M	8.5G	77.74%	59.69%

mAP50 of 77.74% respectively.

V. CONCLUSION

In this work, we proposed the Large Kernel Disassembling (LKD) Attention Mechanism and applied it to remote sensing image detection tasks. The experimental results demonstrate that LKD consistently improves the performance of neural network models. We evaluated LKD on multiple remote sensing image datasets, considering both horizontal and rotated detection scenarios. The evaluation results show that the LKD module outperforms other attention modules. In future research, we plan to extend the application of the LKD module to lightweight models such as EfficientNet and VGG. Additionally, we intend to explore the potential of applying the LKD module to other computer vision tasks. Attention mechanisms have wide-ranging applications across various tasks, and we believe the LKD module could play an important role in other tasks as well.

REFERENCES

[1] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, p. 103514, 2022.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A review of yolo algorithm developments," *Procedia Computer Science*, vol. 199, pp. 1066–1073, 2022.

[6] T. Diwan, G. Anirudh, and J. V. Tembhurne, "Object detection using yolo: Challenges, architectural successors, datasets and applications," *Multimedia Tools and Applications*, vol. 82, no. 6, pp. 9243–9275, 2023.

[7] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning roi transformer for oriented object detection in aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2849–2858.

[8] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-cnn for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3520–3529.

[9] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1829–1838.

[10] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[11] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

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

[13] 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)*, 2018, pp. 3–19.

[14] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.

[15] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," *arXiv preprint arXiv:2112.05561*, 2021.

[16] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Computational Visual Media*, vol. 9, no. 4, pp. 733–752, 2023.

[17] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large selective kernel network for remote sensing object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 794–16 805.

[18] M.-H. Guo, C.-Z. Lu, Q. Hou, Z. Liu, M.-M. Cheng, and S.-M. Hu, "Segnext: Rethinking convolutional attention design for semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1140–1156, 2022.

[19] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style convnet for visual recognition," *arXiv preprint arXiv:2211.11943*, 2022.

[20] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 296–307, 2020.

[21] J. Li, C. Qu, and J. Shao, "Ship detection in sar images based on an improved faster r-cnn," in *2017 SAR in Big Data Era: Models, Methods and Applications (BIGSARDATA)*. IEEE, 2017, pp. 1–6.

[22] M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[23] Y. Zhou, X. Yang, G. Zhang, J. Wang, Y. Liu, L. Hou, X. Jiang, X. Liu, J. Yan, C. Lyu *et al.*, "Mmrotate: A rotated object detection benchmark using pytorch," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7331–7334.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[25] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3974–3983.