Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048755

ARTICLE

YOLO-MFD: Remote Sensing Image Object Detection with Multi-Scale Fusion Dynamic Head

Zhongyuan Zhang and Wenqiu Zhu*

School of Computer Science, Hunan University of Technology, Zhuzhou, 412007, China

*Corresponding Author: Wenqiu Zhu. Email: Zwq@hut.edu.cn

Received: 17 December 2023 Accepted: 25 March 2024 Published: 15 May 2024

ABSTRACT

Remote sensing imagery, due to its high altitude, presents inherent challenges characterized by multiple scales, limited target areas, and intricate backgrounds. These inherent traits often lead to increased miss and false detection rates when applying object recognition algorithms tailored for remote sensing imagery. Additionally, these complexities contribute to inaccuracies in target localization and hinder precise target categorization. This paper addresses these challenges by proposing a solution: The YOLO-MFD model (YOLO-MFD: Remote Sensing Image Object Detection with Multi-scale Fusion Dynamic Head). Before presenting our method, we delve into the prevalent issues faced in remote sensing imagery analysis. Specifically, we emphasize the struggles of existing object recognition algorithms in comprehensively capturing critical image features amidst varying scales and complex backgrounds. To resolve these issues, we introduce a novel approach. First, we propose the implementation of a lightweight multi-scale module called CEF. This module significantly improves the model's ability to comprehensively capture important image features by merging multi-scale feature information. It effectively addresses the issues of missed detection and mistaken alarms that are common in remote sensing imagery. Second, an additional layer of small target detection heads is added, and a residual link is established with the higher-level feature extraction module in the backbone section. This allows the model to incorporate shallower information, significantly improving the accuracy of target localization in remotely sensed images. Finally, a dynamic head attention mechanism is introduced. This allows the model to exhibit greater flexibility and accuracy in recognizing shapes and targets of different sizes. Consequently, the precision of object detection is significantly improved. The trial results show that the YOLO-MFD model shows improvements of 6.3%, 3.5%, and 2.5% over the original YOLOv8 model in Precision, map@0.5 and map@0.5:0.95, separately. These results illustrate the clear advantages of the method.

KEYWORDS

Object detection; YOLOv8; multi-scale; attention mechanism; dynamic detection head

1 Introduction

Remote sensing image processing represents an interdisciplinary convergence, bringing together insights and methods from diverse fields such as remote sensing, computer vision, geographic information systems, and machine learning. Its historical roots date back to the 20th century and

encompass image data captured by aircraft or satellites, providing a vast repository of information about the Earth's surface. Its central role extends to critical areas such as urban planning, agricultural management, and homeland security [1].

In the field of remote sensing image processing, Object detection and image processing are important problems in computer vision [2]. Conventional target detection techniques typically rely on the construction of artificial features but often have shortcomings in terms of both performance and efficiency. Initially, target detection involves manual feature extraction using Classic algorithms like Support Vector Machine (SVM) [3], Adaptive Boosting (AdaBoost) [4], Histogram of oriented gradient [5], and Deformable Part-based Model [6]. Girshick [7] and others pioneered convolutional neural networks for target detection by introducing a Region-Convolutional Neural Network (R-CNN) for this purpose. Firstly, two-stage algorithms typified by the R-CNN series, encompassing R-CNN [8], Spatial Pyramid Pooling Networks (SPP-Net) [9], Faster R-CNN [10], Mask R-CNN [11], and others, primarily employ the selective search method to extract candidate frames of interest. These frames undergo feature extraction within a CNN, followed by a secondary classification utilizing SVM. Finally, a regressor fine-tunes the localization coordinates. Secondly, a one-stage algorithm, exemplified by YOLO [12–15] and Single Shot MultiBox Detector (SSD) [16], differs from the two-stage approach by utilizing a single convolutional neural network to directly localize and classify all targets across the entire image. YOLO facilitates real-time target detection, outperforming two-stage detection algorithms by an order of magnitude in speed. Moreover, YOLO accounts for more background information, reducing the likelihood of misclassifying the background as foreground, making it more suitable for remote sensing image processing with diverse backgrounds. Presently, the latest research pertains to the YOLOv8 [17] algorithm, featuring a novel backbone network architecture and anchor-free detection head. This algorithm excels in feature extraction, target detection, and segmentation. However, it exhibits sensitivity to changes in target scale, limiting its efficacy in addressing the challenges posed by remote sensing images.

Challenges commonly encountered in remote sensing image processing encompass densely distributed targets, pronounced variations in target scales, and intricate backgrounds. Additionally, the presence of minuscule targets in distant images poses heightened difficulties in detection, elevating the likelihood of missed detections and false positives. The demanding performance standards for detection networks make it arduous for conventional networks to directly address remote sensing image detection. Considering these characteristics, this paper delves into the structure and technical enhancements of target detection algorithms tailored for remote sensing image scenes. Leveraging the YOLOv8 network structure as a foundation, a novel multi-scale and efficient remote sensing imaging target detection scheme, YOLO-MFD, is presented to increase the detection capability of the network. The main contributions of our work are as follows:

- Deep feature maps essentially encapsulate richer semantic information, while shallow feature maps focus mainly on providing precise location information. In the YOLOv8 model, we have increased the number of detection layers from three to four in the original backbone network. This increase allows the fusion of shallower layers, which ensures improved localization accuracy. Therefore, this improvement not only stabilizes the training process but also improves the accuracy of the model, making it more suitable for remote sensing image detection.
- To address miss and false detection rates in the YOLOv8 algorithm when applied to remote sensing images and to extract multi-scale feature information from complex backgrounds without incurring additional computational overhead, we propose the lightweight convolutional block Convolutional-Depthwise Convolution Block (CDW). CDW can extract more multi-scale feature information from complex backgrounds compared to the standard convolutional block,

without increasing computational complexity. Subsequently, the proposed Convolutional-Depthwise Convolution Block-Encoder Modulator Attention (CDW-EMA) is integrated with the CSP Bottleneck with 2 convolutions (C2f) modules in the backbone network of the YOLOv8 model to form a new CDW-EMA-C2f (CEF) module.

- Within the YOLOv8 model, the output of the backbone network generates a three-dimensional tensor structured as horizontal × spatial × channel. Consequently, the model substitutes its detection head with a Dynamic Head [18], allowing the simultaneous merging of scale-aware, spatial-aware, and task-aware attention mechanisms. This integration applies the attention mechanism to each distinct dimension of the feature tensor, enhancing the detection head's representational prowess. Consequently, this enhancement bolsters the algorithm's detection efficacy for smaller targets.

2 Structure of YOLOv8 and Algorithm Principles

The YOLOv8 model comprises several key components: The input layer, backbone network for feature extraction (Backbone), feature enhancement module (Neck), and detection output module (Detect). Fig. 1 illustrates the model's detailed structure and constituent parts.

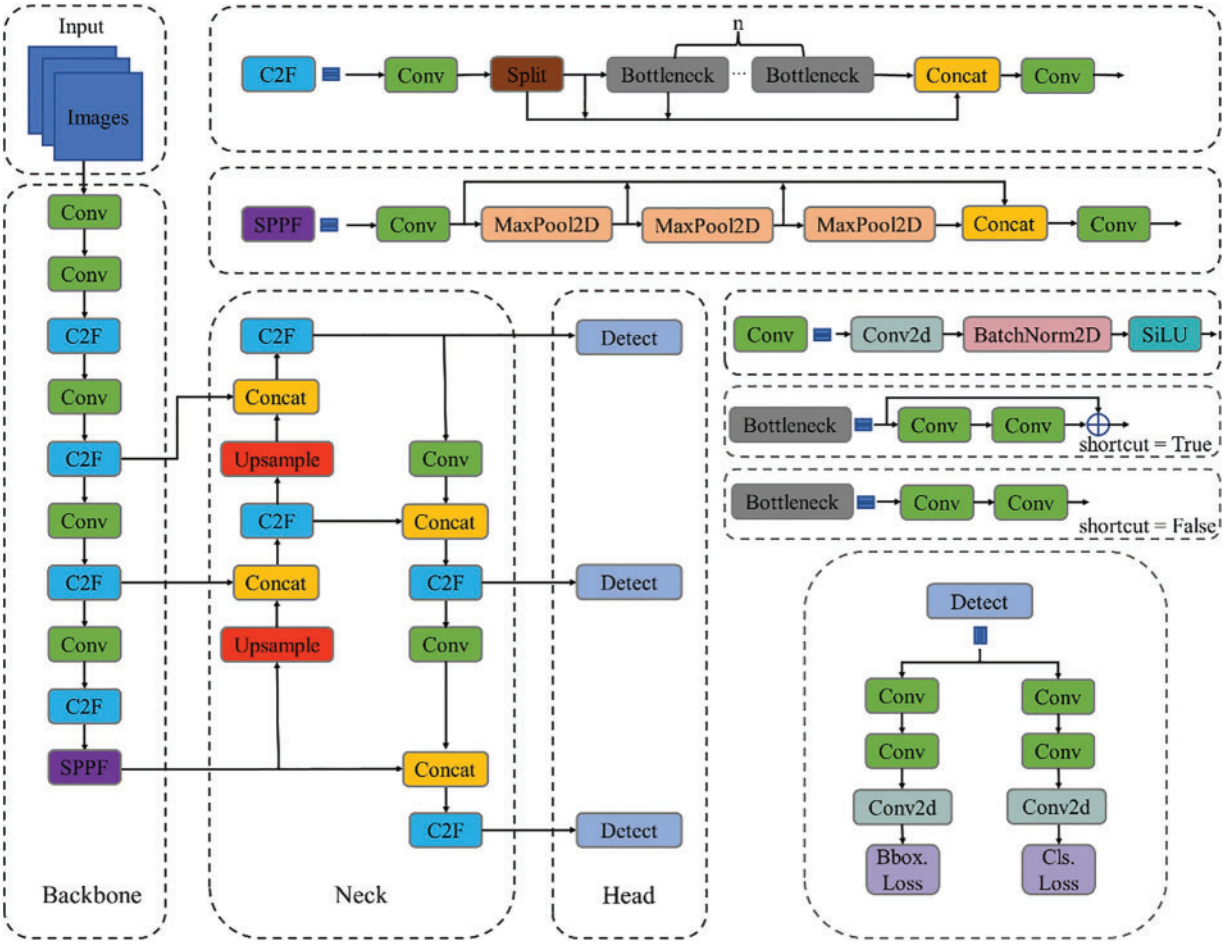

Figure 1: Structure and details of YOLOv8

The backbone network serves a primary function in extracting information from images and transmitting it to the neck and head modules. It comprises multiple convolution blocks, including the C2f and the terminal Spatial Pyramid Pooling-Fast (SPPF) modules. The convolution (Conv) module, featuring Conv2d, Batch Normalization 2D (BatchNorm2d), and activation functions, is pivotal for feature extraction and organizing the feature map. YOLOv8 incorporates elements from the residual structure of YOLOv5's CSP Bottleneck with 3 convolutions (C3) module [19] and the Efficient Layer Aggregation Network (ELAN) concept from YOLOv7 [20], merging them to create the C2f structure. This innovative design ensures model lightness while acquiring richer information, dynamically adjusting the channel as per the model scale.

The neck part mainly plays the role of feature fusion, leveraging backbone network features fully, using the Feature Pyramid Network + Path Aggregation Network (FPN+PAN) construction approach, which enhances its semantic representation and localization ability on multiple scales.

The head processes the last stage of the network. On the output side, it determines the category and location of the detected target based on the features from the first two processing parts, thus enabling detection. This has been replaced by today's prevalent decoupled head structure, which separates the classification and detection functions. This modification addresses the issue of different emphasis between classification and localization. In addition, it adopts the Anchor-Free [21] approach for target detection, which improves the detection speed. For the loss calculation, it adopts a dynamic allocation strategy for normal and negative samples and uses the varifocal loss (VFL loss) [22] as the classification loss.

3 Structure of YOLO-MFD

Our proposed YOLO-MFD architecture, illustrated in Fig. 2, works through several stages. First, the CEF module within the backbone captures comprehensive multi-scale feature data of the image. Next, the neck section fuses features from multiple layers to identify the target. Finally, the dynamic detection head outputs both target and category information.

3.1 Four-Layer Detection Structure

The number of detection layers in the YOLOv8 model has been increased from three to four. The shallow layer of the convolutional neural network extracts features closer to the input, including some details and edge information of the image. Meanwhile, the deep network extracts more abstract semantic information after multiple convolutional pooling, and the features of small targets are easily masked or lost. However, the details and edge information of the target plays a crucial role in accurate classification and localization. Therefore, the small target detection layer adds a new output layer designed by us to the Neck network, which helps the network to better capture the shallow semantic information, thus improving the detection accuracy of small targets.

After two Upsample and Concat operations, the designed fourth output layer undergoes another Upsample operation to obtain a 160×160 feature map. Following Upsample and fusion of features at different scales, the feature map has stronger semantic information and finer spatial details, enabling better discrimination and detection of very small targets. The resulting feature map is then concatenated with the backbone network to further fuse feature information at different scales. A C2f module is introduced to process the fused feature map, keeping its size at 160×160. Finally, a very small target detection layer is obtained to help the network better capture detailed information about the target. Its improved overall structure is illustrated in Fig. 3.

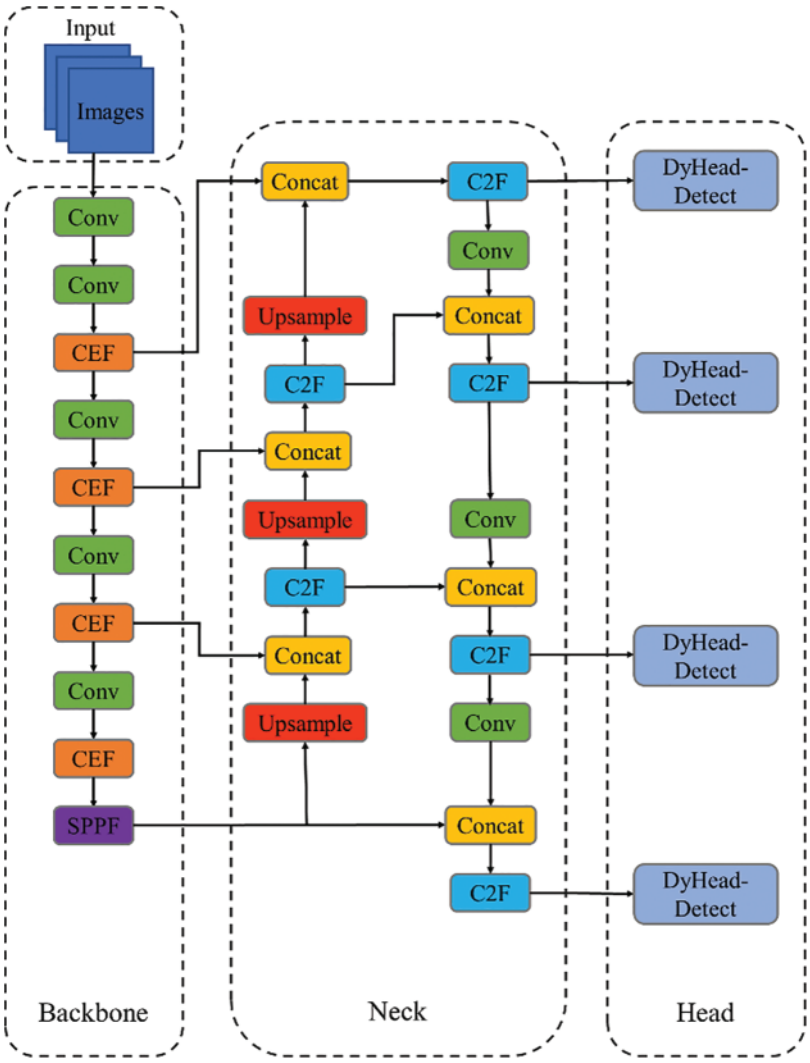

Figure 2: Structure of YOLO-MFD

3.2 The CEF Module

Attention mechanisms originated from neuroscience, describing the brain's mechanism when processing information. Later, this concept was integrated into machine learning to bolster model performance and accuracy. This mechanism works by automatically identifying critical features within the input data to increase the ability and precision of the model. Enhancing the model's ability to learn, attention mechanisms pave the way for further potential and prospects in image processing, computer vision, and their diverse applications.

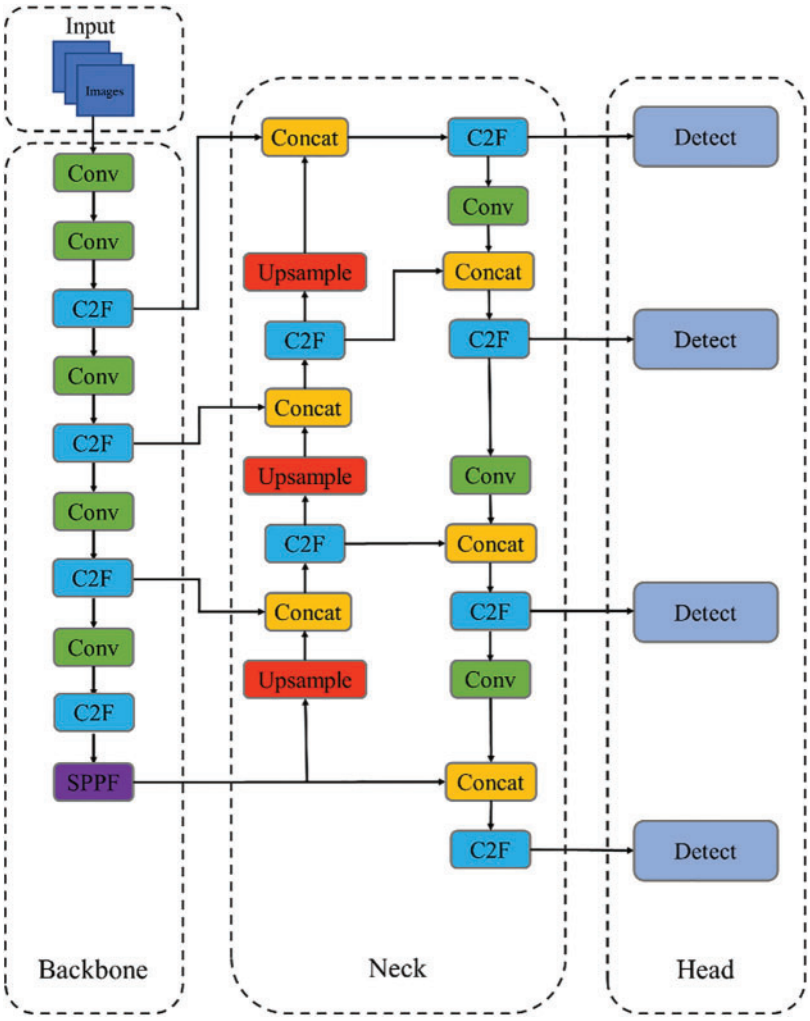

Figure 3: Add YOLOv8 detection layer

Commonly employed attention mechanisms in image processing include channel attention and spatial attention. Recent insights suggest the potential synergy of combining these mechanisms to address diverse challenges. For instance, the Squeeze-and-Excitation Networks (SE) [23] model, epitomized by the channel attention mechanism, facilitates the extraction of channel features by explicitly defining interactions across channel dimensions. Attention Module for Convolutional Blocks (CBAM) [24] is utilized to establish feature mappings spanning channel and spatial information, facilitating semantic dependencies in both spatial and channel dimensions. However, the conventional handling of inter-channel relations often involves dimensionality reduction, when extracting deep visual representations, this can lead to adverse effects. To address those limitations, researchers like Daliang Ouyang have adapted the Channel Attention (CA) [25] mechanism to retain crucial channel information while reducing computational costs. Their method includes reconfiguring specific channels into batch dimensions and organizing channel dimensions into separate sub-features [26,27]. This method encourages a more balanced distribution of spatial semantics across feature groups, offering novel solutions to reconcile the interplay between channel attention and spatial attention.

Additionally, it introduces a novel and efficient multiscale attention mechanism (EMA) [28] module, depicted in Fig. 4.

Figure 4: Structure of EMA

The EMA attention mechanism employs a distinctive modeling approach to handle cross-channel interaction information within a channel. Utilizing a global average pooling operation within the 1×1 convolutional branch, EMA encodes channels bidirectionally, emphasizing inter-channel information transfer. Conversely, the 3×3 convolutional branch excludes GN normalization and average pooling operations, focusing on extracting multi-scale feature representations. Moreover, EMA introduces a unique cross-space information aggregation method to enable more comprehensive feature aggregation. This method involves incorporating two tensors derived from the outputs of the 1×1 and 3×3 branches, respectively. Lastly, global dimensional information of the 1×1 branch output is encoded by global mean pooling. This transforms the smallest branch's output channel features into corresponding dimensional shapes, following the pooling operation described in Eq. (1).

$$Z_C = \frac{1}{H \times W} \sum_j^H \sum_i^W X_C(i,j) \tag{1}$$

A lightweight convolutional block CDW is introduced to capture diverse feature data in complex backgrounds without extensive computation. This serves as the basis for CDW-EMA, a lightweight multi-scale attention mechanism. Fig. 5 shows the detailed structures of CDW and CDW-EMA, correspondingly.

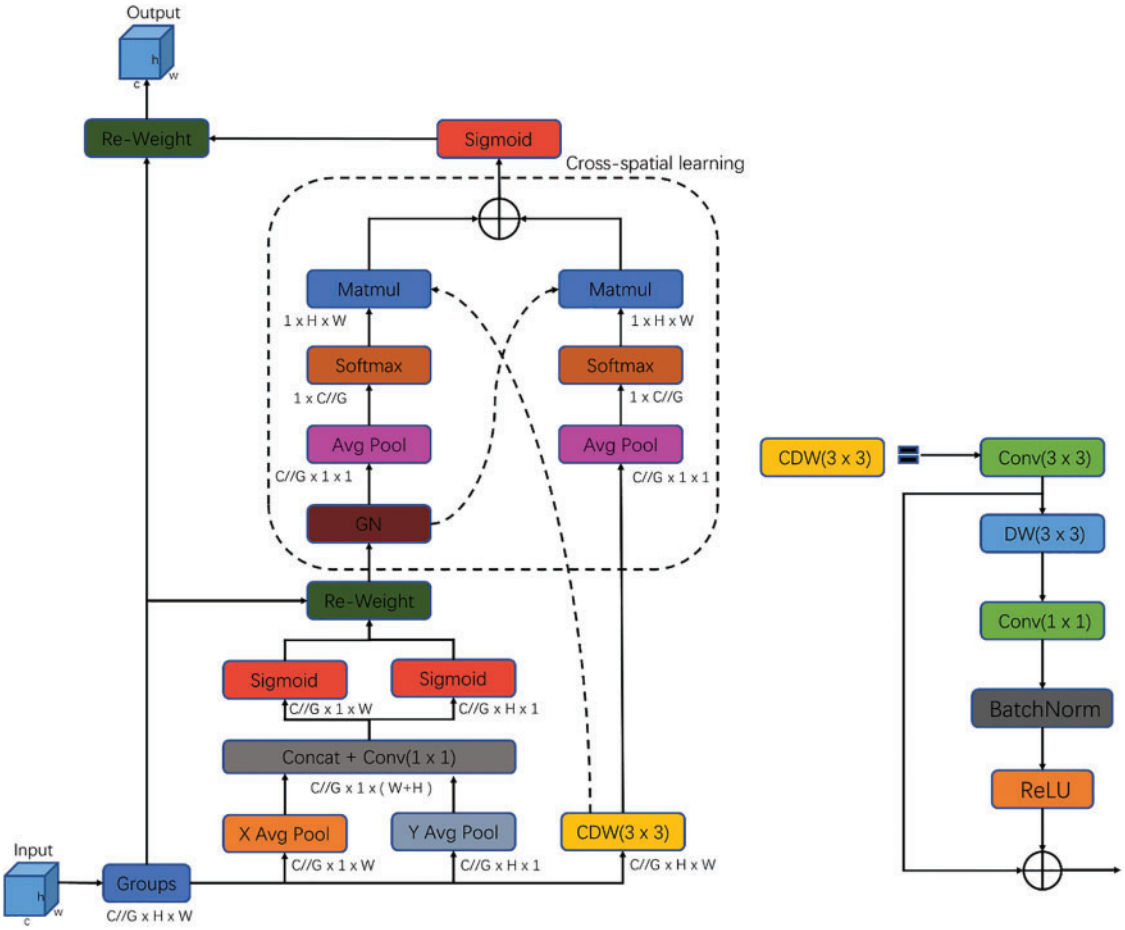

Figure 5: Structure of CDW and CDW-EMA

Where Depthwise Convolution (DW) [29] convolution is also called separable convolutional convolution, where the process of one convolution kernel corresponds to one channel, the output channel count matches the input channel count, which is M, and the dimensions are $D_F \times D_F$, the output channels are also M, with a convolution kernel size of $D_K \times D_K$, its calculated quantity is shown in Eq. (2).

$$C_{DW} = D_K \times D_K \times M \times D_F \times D_F \tag{2}$$

The number of its parameters is given by Eq. (3).

$$N_{DW} = M \times D_K \times D_K \tag{3}$$

The CDW convolution block consists of Conv 3 × 3, DW 3 × 3, Conv 1 × 1, BatchNorm, and ReLU activation function. Compared with ordinary convolution, DW convolution processes each channel of the input independently, which can be regarded as a spatial convolution operation on each channel, and has the advantages of fewer parameters and less computation, etc. Batch Normalization (BatchNorm) accelerates model convolution and enhances stability, while the Rectified Linear Unit (ReLU) [30] activation features amplify network nonlinearities and inhibit gradient disappearance. CDW convolution block consists of regular convolution and depth convolution. Regular convolution performs channel information adjustment, depth convolution performs efficient spatial feature extraction, and the two convolution layers are connected in the form of an inverse residual block, and the gradient disappearance problem of CDW is mitigated by jumping connection, so as to achieve better information extraction of small objects in remote sensing images and to minimize gradient disappearance problem.

CDW-EMA consists of four branches: Main Branch, Coordinate Branch, and CDW 3 × 3. We use CDW 3 × 3 instead of the original Conv 3 × 3. The improved CDW-EMA module reduces the amount of computation and the number of parameters while keeping the sensor field unchanged, which improves the speed and accuracy of the model. Then, we fuse the proposed CDW-EMA with the C2f module in the backbone part of the backbone network of the YOLOv8 model to form the new module CEF. The overall structure of CEF is shown in Fig. 6.

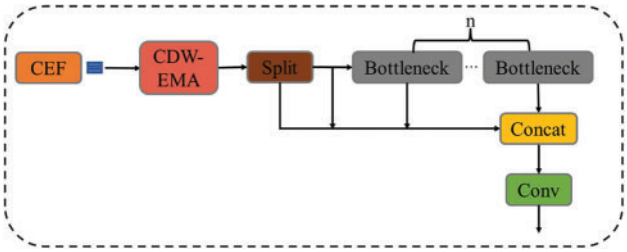

Figure 6: Structure of CEF

3.3 Dynamic Detection Head

In the YOLOv8 model, the output of the backbone network forms a three-dimensional tensor with dimensions of horizontal × spatial × channel. Consequently, the YOLOv8 model replaces its detection head with a dynamic detection head called DyHead (Dynamic Head). The DyHead unifies scale-aware, spatial-aware, and task-aware attention mechanisms, thereby integrating attention into each specific dimension of the feature tensor. Given the 3D feature tensor at the detection level $F \epsilon R^{L \times S \times C}$, Then applying attention to L, S, and C, respectively, yields three perceptual abilities. Given a feature layer, applying self-attention to it gives Eq. (4):

$$W(F) = \pi(F) \cdot F \tag{4}$$

If a subsequent process employs a fully connected layer, where $\pi(\cdot)$ represents an attention mechanism. If the next process uses a fully connected layer, this approach is too computationally intensive. So the attention is carried out in three dimensions, each attention over a certain perspective,

there is Eq. (5):

$$W(F) = \pi_C(\pi_S(\pi_L(F) \cdot F) \cdot F) \cdot F \tag{5}$$

Three distinct attention functions $\pi_C(\cdot)$, $\pi_S(\cdot)$, $\pi_L(\cdot)$ are used on dimensions C, S, and L, respectively. These attention sequences are imposed on the sensing head and can be stacked multiple times. In improving the design of the detection head model, four groups of modules are used to be stacked sequentially to give the detection head a stronger representational capability, this improves the algorithm's ability to find small targets. The DyHead structure is displayed in Fig. 7.

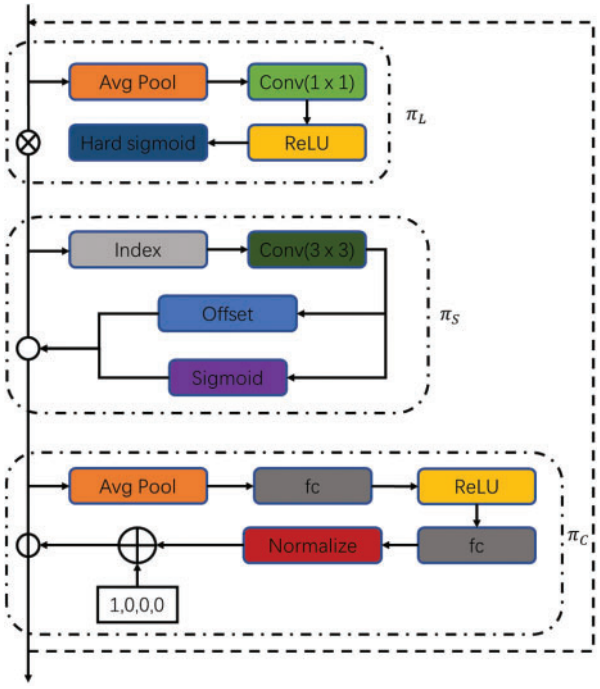

Figure 7: Structure of DyHead

4 Experimental Results and Analysis

Table 1 lists the environmental configuration used for the experiment.

Table 1: Configuration of the experimental environment

Project	Environment
OS	Windows11
CPU	i5-9400F
GPU	NVIDIA 2070 super
Language	Python3.10
Pytorch version	2.0.1
CUDA	11.1

4.1 Evaluation Metrics

Precision, Recall, mAP@0.5, and mAP@0.5:0.95 are taken as the experimental evaluation indexes, and multifaceted comparison data are used to ensure the model effect.

4.1.1 Precision

Precision, as defined in Eq. (6), measures the ratio of accurately predicted targets to the total number of targets predicted by the model:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Precision's denominator sums TP and FP, where TP signifies correctly predicted positive cases and FP represents incorrectly predicted positive cases when the actual cases were negative.

4.1.2 Recall

Recall, as shown in Eq. (7), measures the proportion of real targets correctly predicted by the model:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

Recall's denominator combines TP and FN, with TP representing cases predicted as positive where the actual value is also positive, and FN representing cases predicted as negative while being positive.

4.1.3 mAP

mAP represents the average accuracy of all labels divided by the total number of categories. A higher value of mAP indicates a higher average accuracy of the model. Eq. (8) illustrates mAP:

$$\text{mAP} = \frac{1}{N} \sum AP_i \tag{8}$$

4.2 Datasets

This experiment uses the RSOD dataset, a public dataset from Wuhan University for target detection in remote sensing images. It includes four categories: Airplanes, playgrounds, overpasses, and oil tanks, with 976 images and 6950 targets. The training parameters are detailed in Table 2.

Table 2: Experimental parameter setting

Parameters	Value
Division ratio	7:2:1 (train:Val:Test)
Optimizer	SGD
Batch size	8
Epochs	100

4.3 Analysis of Results

Firstly, we compared the CEF module with mainstream attention mechanisms (e.g., EMA, SE, CA, and Environment-Cognition-Action(ECA) [31]) individually combined with the YOLOv8 base model. This comparison helped determine the most effective attention mechanism for detection.

CEF is superior to other mainstream attention mechanisms in achieving optimal results across all metrics. Compared to the original Yolov8 algorithm, it improves precision by 2.2%, recall by 2%, mAP50 by 2.5%, and mAP95 by 1%, as detailed in Table 3.

Table 3: Comparison with other attention mechanisms

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8	0.862	0.912	0.916	0.600
+CA	0.822	0.892	0.885	0.571
+SE	0.805	0.879	0.842	0.582
+ECA	0.887	0.857	0.895	0.580
+EMA	0.865	0.888	0.895	0.601
+CEF	0.884	0.932	0.941	0.610

We performed ablation experiments on the YOLO-MFD algorithm using the validation set. Table 4 shows the detailed data from these experiments. Compared with the original YOLOv8, our proposed YOLO-MFD algorithm improves the precision, Recall and mAP@0.5:0.95 by 6.3%, 3.5%, and 2.5%, respectively, and these data completely demonstrate that the performance of the proposed YOLO-MFD algorithm achieves a huge improvement over the original YOLOv8.

Table 4: Ablation experiment

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv8	0.862	0.912	0.916	0.600
+detect	0.841	0.891	0.909	0.621
+CEF	0.884	0.932	0.941	0.610
+DyHead	0.891	0.907	0.921	0.597
YOLO-MFD	0.925	0.915	0.951	0.625

Finally, to confirm the detection capability of the YOLO-MFD algorithmic model, its target detection performance with the mainstream algorithms Faster R-CNN, SGD, YOLOv3, YOLOv4, YOLOv5 and YOLOv7 on the RSOD dataset was quantitatively analyzed, and the validation part of the dataset was examined using different models, and the results of comparing their metric values are shown in Table 5, which is not compared with the playground model in the dataset RSOD. The data of the playground model in the RSOD dataset are excellent and are not compared.

Table 5: Experimental comparison of mainstream algorithms

Model	mAP@0.5			
	Aircraft	Oil tank	Overpass	All
Faster R-CNN	0.631	0.841	0.769	0.805
SSD	0.521	0.668	0.567	0.764
YOLOv3	0.802	0.931	0.672	0.851
YOLOv4	0.813	0.965	0.688	0.873
YOLOv5	0.947	0.940	0.669	0.902
YOLOv7	0.935	0.981	0.841	0.924
YOLO-MFD	0.949	0.995	0.874	0.951

4.4 Comparison Test

Fig. 8 shows the loss and map@50 plots for each module after insertion, and it can be seen that YOLO-CEF outperforms the other algorithms in terms of both loss function convergence and map50 improvement.

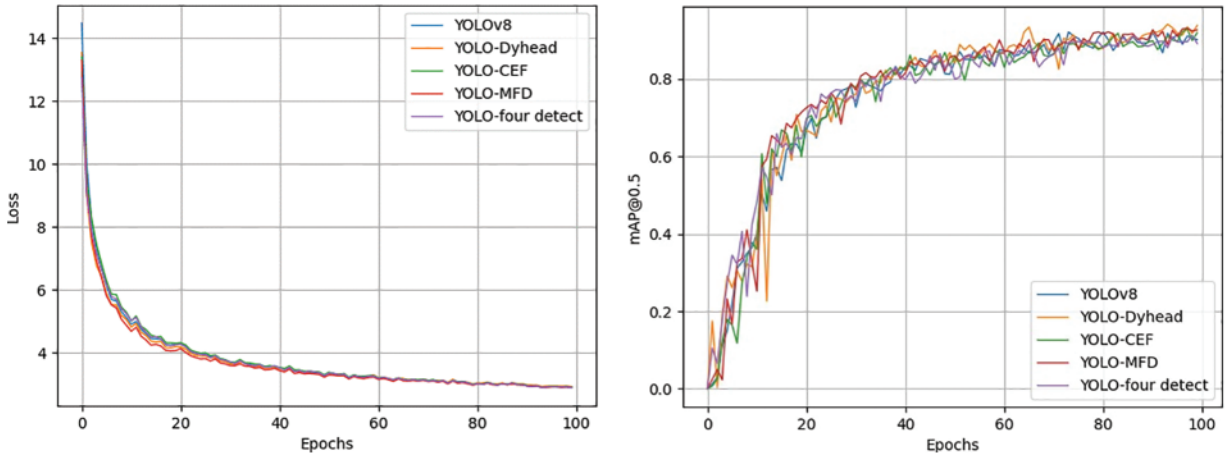

Figure 8: Loss and map50 comparison across models

To better showcase the algorithm's adaptation to various scenarios in solving the RSOD dataset detection issue, we performed visualization and comparison experiments. Fig. 9 shows the instances where small targets were missed, Fig. 10 illustrates false detections of small targets, and Fig. 11 presents a comparison of detection accuracy for large targets. This ensured a more comprehensive adaptation of the algorithm to diverse scenarios.

The detection results of the YOLO-MFD (left) and the YOLOv8 (right) are shown below.

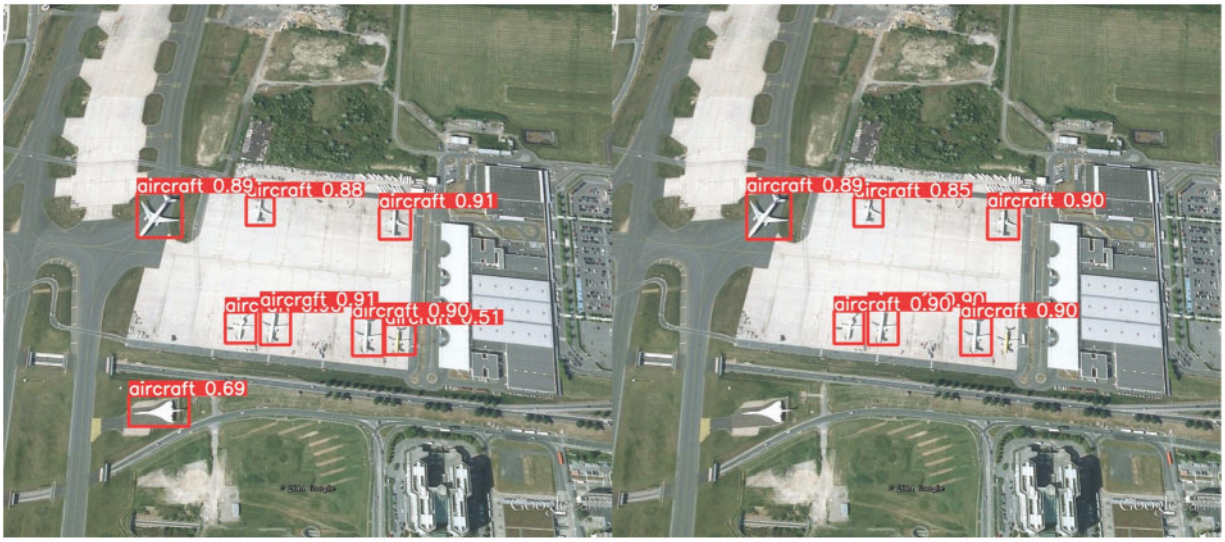

Figure 9: Missed detection of small targets in more complex backgrounds

Figure 10: False detection of small targets in more complex contexts

Figure 11: Target detection accuracy comparison

5 Conclusions

This paper introduces a novel algorithm designed for remote sensing image object detection. YOLO-MFD is an algorithm that improves the YOLOv8 model. First, we developed a lightweight multiscale module called CEF. Second, an additional detection layer was added to improve the accuracy of target localization in remote sensing images. Finally, the integration of a dynamic attention mechanism further refined the accuracy of target detection.

In the evaluation conducted on the RSOD dataset, the YOLO-MFD algorithm exhibited a detection accuracy of 92.5%, a map@0.5 of 95.1%, and a map@0.5:0.95 of 62.5%, all of which surpass the existing mainstream methods. Compared to the original YOLOv8 model, the YOLO-MFD algorithm demonstrated a 6.3% increase in accuracy, with increases of 3.5% and 2.5% in map@0.5 and map@0.5:0.95, respectively. Collectively, these improvements effectively address challenges in remote sensing image detection, including low average detection accuracy, false alarms, and missed detections due to scale diversity, small target areas, and complex backgrounds.

In future research, we will focus on exploring model pruning and lightweight strategies. The aim is to optimize the algorithm while maintaining detection accuracy, facilitating more efficient deployment in practical applications.

Acknowledgement: Thanks are extended to the editors and reviewers.

Funding Statement: A project supported by the Scientific Research Fund of Hunan Provincial Education Department (23A0423).

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design: Zhongyuan Zhang and Wenqiu Zhu; data collection: Zhongyuan Zhang; analysis and interpretation of results: Zhongyuan Zhang and Wenqiu Zhu; draft manuscript preparation: Zhongyuan Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The publicly RSOD dataset can be found at: https://github.com/RSIA-LIESMARS-WHU/RSOD-Dataset-. The data and materials used to support the findings of this study are available from the author upon request after acceptance.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] Y. Zhang, G. Ning, S. Chen, and Y. Yang, "Impact of rapid urban sprawl on the local meteorological observational environment based on remote sensing images and GIS technology," *Remote Sens.*, vol. 13, no. , pp. 2624, 2021. doi: 10.3390/rs13132624.

[2] M. F. Alotaibi, M. Omri, S. Abdel-Khalek, E. Khalil, and R. F. Mansour, "Computational intelligence-based harmony search algorithm for real-time object detection and tracking in video surveillance systems," *Math.*, vol. 10, no. 5, pp. 733, 2022. doi: 10.3390/math10050733.

[3] W. Huang, H. Liu, Y. Zhang, R. Mi, and B. Shuai, "Railway dangerous goods transportation system risk identification: Comparisons among SVM, PSO-SVM, GA-SVM, and GS-SVM," *Appl. Soft Comput.*, vol. 109, no. 5, pp. 107541, 2021. doi: 10.1016/j.asoc.2021.107541.

[4] A. Shahraki, M. Abbasi, and Y. Haugen, "Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost, and Modest AdaBoost," *Eng. Appl. Artif. Intell.*, vol. 94, pp. 103770, 2020. doi: 10.1016/j.engappai.2020.103770.

[5] L. Kalake, Y. Dong, W. Wan, and L. Hou, "Enhancing detection quality rate with a combined HOG and CNN for real-time multiple object tracking across non-overlapping multiple cameras," *Sensors*, vol. 22, no. 6, pp. 2123, 2022. doi: 10.3390/s22062123.

[6] E. J. Cheng, M. Prasad, J. Yang, P. Khanna, and C. T. Lin, "A fast fused part-based model with new deep feature for pedestrian detection and security monitoring," *Measurement*, vol. 151, no. 4, pp. 107081, 2020. doi: 10.1016/j.measurement.2019.107081.

[7] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comp. Vision*, Santiago, Chile, Dec. 7–13, 2015, pp. 1440–1448.

[8] N. Kesav and M. G. Jibukumar, "Efficient and low complex architecture for detection and classification of brain tumor using RCNN with two channel CNN," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6229–6242, Aug. 2022. doi: 10.1016/j.jksuci.2021.05.008.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, 2015. doi: 10.1109/TPAMI.2015.2389824.

[10] M. M. Mijwil, K. Aggarwal, R. Doshi, K. K. Hiran, and M. Gök, "The distinction between R-CNN and fast RCNN in image analysis: A performance comparison," *Asian J. Appl. Sci.*, vol. 10, no. 5, pp. 429–437, 2022. doi: 10.24203/ajas.v10i5.7064.

[11] H. Qian, Y. Ma, W. Chen, T. Li, Y. Zhuo and W. Xiang, "Traffic signs detection and segmentation based on the improved mask R-CNN," in *2021 40th Chinese Control Conf. (CCC)*, Shanghai, China, Jul. 26–28, 2021, pp. 8241–8246.

[12] D. Pestana *et al.*, "A full featured configurable accelerator for object detection with YOLO," *IEEE Access*, vol. 9, pp. 75864–75877, 2021. doi: 10.1109/ACCESS.2021.3081818.

[13] C. Kim *et al.*, "Implementation of YOLO-v2 image recognition and other testbenches for a CNN accelerator," in *2019 IEEE 9th Int. Conf. Consum. Electron. (ICCE-Berlin)*, Berlin, Germany, IEEE, Sep. 08–11, 2019, pp. 242–247.

[14] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Conf. Comp. Vision Pattern Recogn. (CVPR)*, Salt Lake City, USA, 2018.

[15] C. Wang, X. Tong, R. Gao, and L. Yan, "Mobile-YOLO: A lightweight and efficient implementation of object detector based on YOLOv4," in *Advances in Computer Science for Engineering and Education*, Z. Hu, I. Dychka, S. Petoukhov, and M. He, Eds. Cham: Springer, 2022, pp. 221–234. doi: 10.1007/978-3-031-04812-8_19.
[16] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Comp. Vision–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer International Publishing, 2016, pp. 21–37.
[17] H. Lou *et al.*, "DC-YOLOv8: Small-size object detection algorithm based on camera sensor," *Electron.*, vol. 12, no. 10, pp. 2323, 2023. doi: 10.3390/electronics12102323.
[18] X. Dai *et al.*, "Dynamic head: Unifying object detection heads with attentions," presented at the IEEE/CVF Conf. on Comp. Vision and Pattern Recogn., Jun. 19–25, 2021, pp. 7373–7382.
[19] G. Jocher *et al.*, "Ultralytics/yolov5: V5.0-YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Zenodo, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244964519
[20] V. Pham, D. Nguyen, and C. Donan, "Road damage detection and classification with YOLOv7," in *Proc. 2022 IEEE Int. Conf. Big Data*, Osaka, Japan, Dec. 17–20, 2022, pp. 6416–6423.
[21] Y. Yan, J. Li, J. Qin, P. Zheng, S. Liao and X. Yang, "Efficient person search: An anchor-free approach," *Int. J. Comput. Vis.*, vol. 131, no. 7, pp. 1–20, 2023. doi: 10.1007/s11263-023-01772-3.
[22] A. Fu, X. Zhang, N. Xiong, Y. Gao, H. Wang and J. Zhang, "VFL: A verifiable federated learning with privacy-preserving for big data in industrial IoT," *IEEE Trans. Ind. Inform.*, vol. 18, no. 5, pp. 3316–3326, 2020. doi: 10.1109/TII.2020.3036166.
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comp. Vis. Pattern Recogn.*, State Utah, USA, Jun. 18–22, 2018, pp. 7132–7141.
[24] Fu, X. Hui, G. Song, and Y. Wang, "Improved YOLOv4 marine target detection combined with CBAM," *Symmetry*, vol. 13, no. 4, pp. 623, 2021. doi: 10.3390/sym13040623.
[25] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recogn.*, Nashville, TN, USA, Jun. 20–25, pp. 13713–13722.
[26] H. Liao and W. Zhu, "YOLO-GCRS: A remote sensing image object detection algorithm incorporating a global contextual attention mechanism," *Electronics*, vol. 12, no. 20, pp. 4272, 2023. doi: 10.3390/electronics12204272.
[27] H. Liao and W. Zhu, "YOLO-DRS: A bioinspired object detection algorithm for remote sensing images incorporating a multi-scale efficient lightweight attention mechanism," *Biomimetics*, vol. 8, no. 6, pp. 458, 2023. doi: 10.3390/biomimetics8060458.
[28] D. Ouyang *et al.*, "Efficient multi-scale attention module with cross-spatial learning," in *Proc. ICASSP 2023–2023 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Rhodes Island, Greece, Jun. 4–10, 2023, pp. 1–5.
[29] Y. Guo, Y. Li, L. Wang, and T. Rosing, "Depthwise convolution is all you need for learning multiple visual domains," presented at the Proc. AAAI Conf. Artif. Intell., Hawaii State, USA, Jan. 27–Feb. 01, 2019, vol. 33, no. 1, pp. 8368–8375.
[30] T. Jiang and J. Cheng, "Target recognition based on CNN with LeakyReLU and PReLU activation functions," presented at the Proc. 2019 Int. Conf. Sens., Diagn. Progn. Control (SDPC), Beijing, China, Aug. 15–17, 2019, pp. 718–722.
[31] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," presented at the Proc. IEEE/CVF Conf. Comp. Vis. Pattern Recogn., Jun. 14–19, 2020, pp. 11534–11542.