Journal of Computer Technology and Applied Mathematics
Vol. 1, No. 1, 2024

Enhancing Small Object Detection in Remote Sensing Images Using Mixed Local Channel Attention with YOLOv8

WANG, Hao [1*] ABLAMEYKO, Sergey [1]

[1] Belarusian State University, Belarus

**** WANG, Hao is the corresponding author, E-mail: ahcenewang@gmail.com***

Abstract: Small object detection is very popular in computer vision, and the attention mechanism can automatically learn and selectively focus on important information in the input, improving the performance and generalization ability of the model. This paper proposes a new algorithm based on combination of YOLOv8 and Mixed Local Channel Attention (MLCA) to detect small objects. The results show that YOLOv8 using Mixed Local Channel Attention performs better than using other attention mechanisms and the original YOLOv8.

Keywords: YOLO, Small Object Detection, Mixed Local Channel Attention, MLCA.

DOI: https://doi.org/10.5281/zenodo.10986298

1 Introduction

Object detection is an important part of computer vision that helps us understand images better and enables various visual tasks like identifying objects, understanding scenes, tracking objects, describing images, and recognizing events. Detecting small objects has always been a tough challenge. The definition of small objects can differ based on their size compared to other objects. For instance, in the COCO [1] dataset, objects smaller than 32 × 32 pixels are considered small objects based on their absolute size criteria. When compared to regular-sized objects, small objects tend to occupy a minor portion of the image, featuring lower resolution and less pronounced visual characteristics.

The attention mechanisms enables networks to automatically learn and prioritize important information, thereby enhancing performance and generalization. By integrating attention mechanisms, neural networks dynamically adjust their focus, emphasizing significant features while disregarding noise. This adaptation improves the network's ability to capture complex patterns, leading to more effective model performance. In 2014, Volodymyr Mnih's article "Recurrent Models of Visual Attention" [2] applied the attention mechanism to the visual field. Later, with the proposal of the Transformer structure in Ashish Vaswani's "Attention is all you need" [3] in 2017, the attention mechanism was widely used in network design for computer vision-related issues. In computer vision, there are mainly these attention mechanisms, Spatial Attention, Channel Attention, Self-Attention, Multi-Head Attention, Cross-Modal Attention. As research develops, there are also different attention mechanisms and improved versions of the above attention mechanisms.

Lei Zhu et al. [4] proposed a streamlined yet effective method for implementing a two-layer routing attention system that leverages sparsity to conserve computational resources and memory usage, employing dense matrix multiplications compatible with GPUs. This approach forms the basis for a novel visual transformer called BiFormer, which capitalizes on the developed bi-layer routing attention architecture. Daliang Ouyang et al. [5] proposed a novel efficient multi-scale attention (EMA) module which can preserve information for each channel while reducing computational overhead. Qihang Fan et al. [6] introduced CloFormer, a lightweight visual transformer that leverages context-aware local enhancement, and proposed an effective and simple module to capture high-frequency local information. Yuxuan Li et al. [7] proposed Large Selective Kernel Network (LSKNet). It can dynamically adjust its large spatial receptive field to better simulate the ranging environment of various objects in remote sensing scenes. Dahang Wan et al. [8] proposed a lightweight Mixed Local Channel Attention (MLCA) module to improve the performance of the object detection network. It can integrate channel and spatial information simultaneously, incorporating both local and global details to enhance the network's expressive power.

In this paper, we propose an algorithm combining the Mixed Local Channel Attention with YOLOv8 [9], compare it with other methods and test it. The results show that YOLOv8 with Mixed Local Channel Attention has better result.

2 Methodology

The mixed local channel attention mechanism achieves a balance between detection effectiveness, speed, and model parameter count. It makes spatial information, channel information, local channel information, and global channel information simultaneously into the attention mechanism.

The structure of MLCA is shown in the figure 1 and 2:

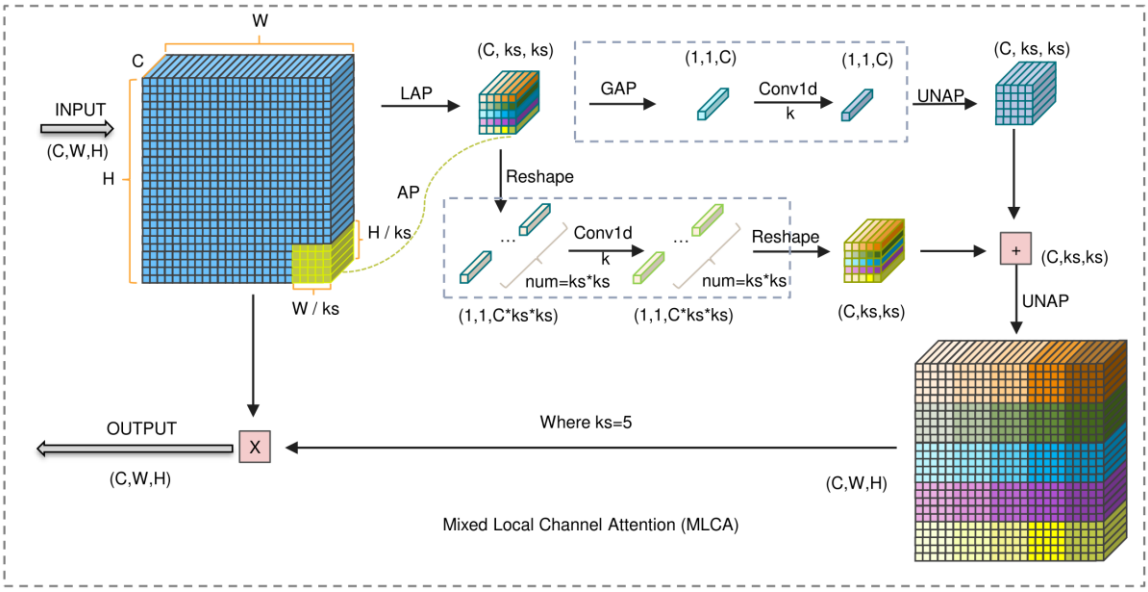

Figure 1. Structure of MLCA

From Figure 1, we can see there are several steps of MLCA.

1). The input feature map (C, W, H) is first processed by local average pooling (LAP) and global average pooling (GAP). Local pooling focuses on the features of local regions, while global pooling captures the statistical information of the entire feature map.

2). Both the local pooled features and the global pooled features undergo a 1D convolution (Conv1d) for feature conversion. The 1D convolution here is used to compress the feature channels while keeping the spatial dimensions unchanged.

3). After 1D convolution, the features are rearranged (Reshape) to adapt to subsequent operations.

4). The local pooled features are rearranged using 1D convolution, and then combined with the original input features through a "multiplication" operation. This process is equivalent to a kind of feature selection, which strengthens the focus on useful features.

5). After 1D convolution and rearrangement, the global pooled features are combined with the local pooled features through the "addition" operation. This step incorporates global contextual information in feature maps.

6). Finally, the feature maps processed by local and global attention are restored to the original spatial dimensions through the unpooling (UNAP) operation again.

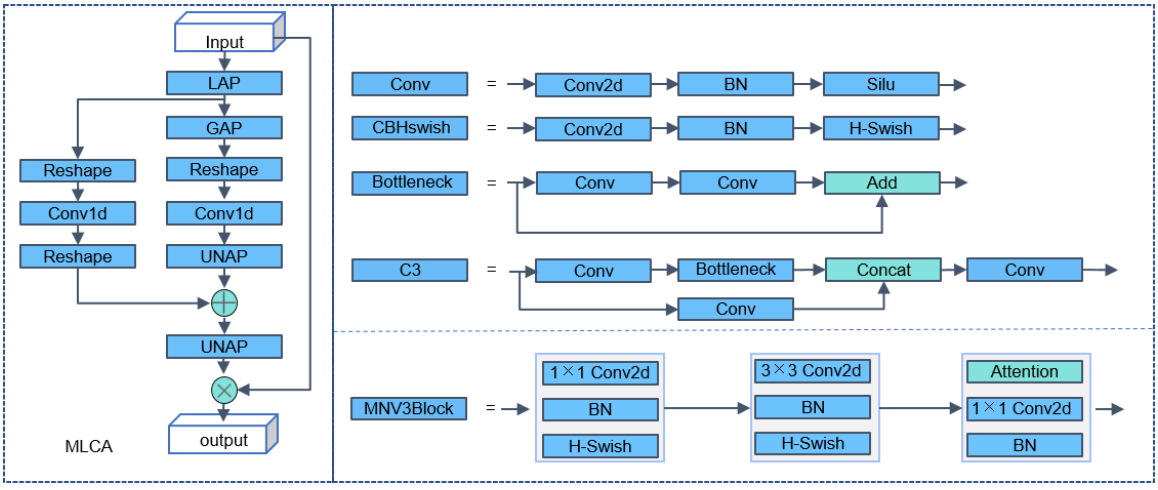

Figure 2. Structure of the flow of MLCA

Figure 2 provides a high-level flowchart of MLCA, showing the overall processing steps from input to output.

Generally speaking, the MLCA module is designed to enhance the network's ability to capture useful features while maintaining computational efficiency. By combining channel and spatial attention at local and global levels, MLCA effectively improves the accuracy of the algorithm.

3 Experiments and Results

In our experiments, we utilized the DOTA-v2.0 [10] dataset, renowned for its extensive coverage of aerial images, making it ideal for small object detection tasks. DOTA-v2.0 encompasses data from Google Earth, GF-2, JL-1 Satellites, and various aerial sources, offering a diverse range of scenarios for detector development and evaluation. It encompasses 16 common object categories such as planes, helicopters, ships, vehicles, and sports courts.

The experiments were conducted on the AutoDL platform, leveraging the computing power of the Nvidia GeForce RTX 3090 with 24268MiB memory capacity. Training was performed on images resized to 640 pixels, utilizing a batch size of 4, and training for 200 epochs. The experimental outcomes are presented in the table below.

Table 1. Comparison results.

Detector	precision	box(P)	recall	parameters	mAP50	mAP50-95
YOLOv8n	0.62357	0.623	0.302	3009768	0.332	0.198
YOLOv8-MLCA	0.69584	0.67	0.311	3008808	0.344	0.205
YOLOv8-CloFormer	0.67877	0.671	0.303	3144512	0.334	0.198
YOLOv8-EMA	0.63364	0.637	0.305	2312080	0.325	0.187
YOLOv8-Biformer	0.66012	0.664	0.307	3274496	0.339	0.205
YOLOv8-LSKNet	0.60915	0.607	0.259	5897134	0.283	0.165

As can be seen from Table 1, in several indicators, YOLOv8 used different attention mechanisms performs better than the original YOLOv8. And YOLOv8-MLCA is the best performer, it has fewer parameters and higher accuracy.

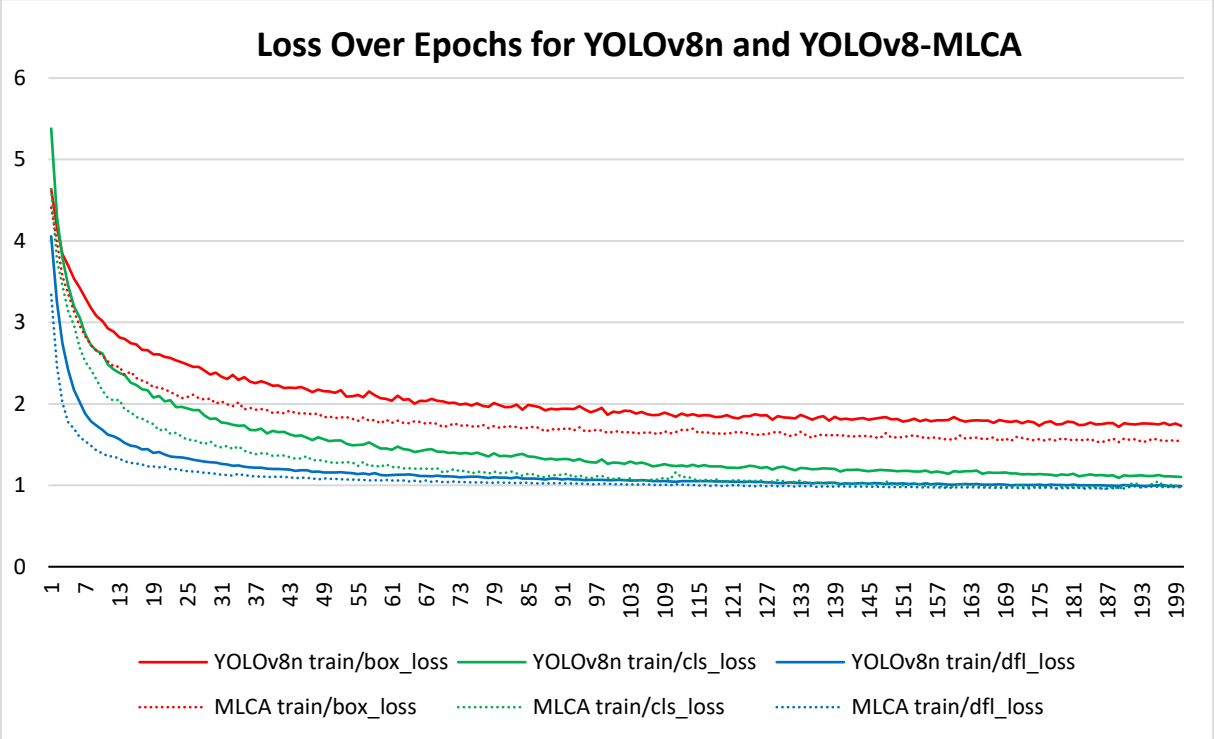

Figure 3. Loss over Epochs for YOLOv8n and MLCA

Figure 3 shows that the loss using the MLCA mechanism also has faster convergence speed and lower values.

From Figure 4 to Figure 7, the results of different curves of the original YOLOv8 and YOLOv8-MLCA are compared. We can see the performance of these two methods for different small objects in the dataset.

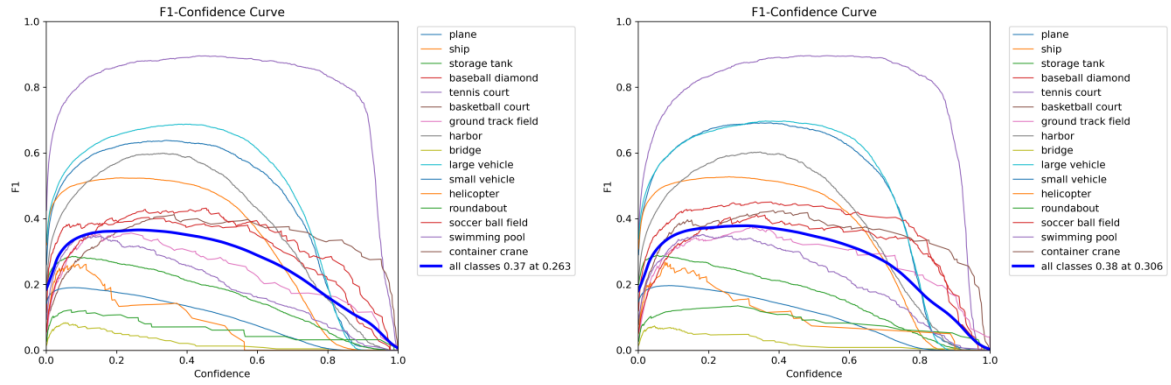

Figure 4. Comparison of F1 curves of YOLOv8n(left) and MCLA(right)

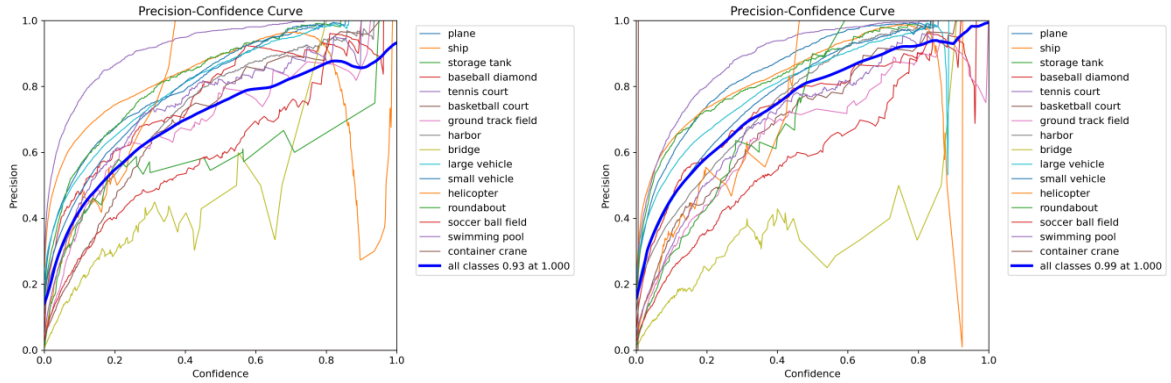

Figure 5. Comparison of P curves of YOLOv8n(left) and MCLA(right)

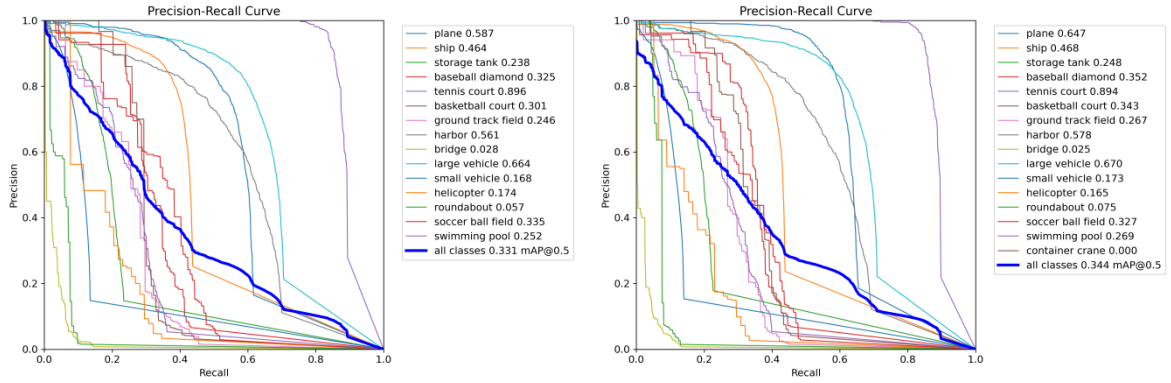

Figure 6. Comparison of PR curves of YOLOv8n(left) and MCLA(right)

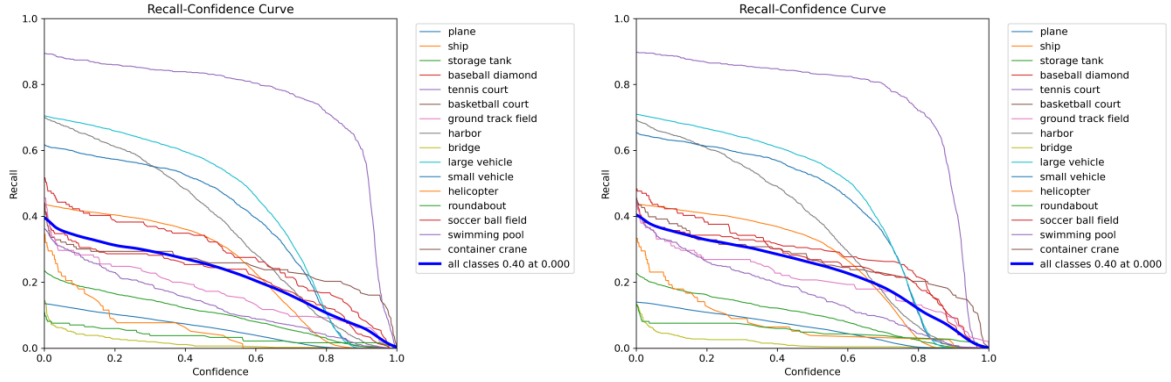

Figure 7. Comparison of R curves of YOLOv8n(left) and MCLA(right)

Experiments show that when using Mixed Local Channel Attention for small object detection in YOLOv8, the performance of this model is not only better than the original YOLOv8, but also better than other attention mechanism models tested in the experiment. On the DOTA-v2.0 dataset, this model has smaller loss, higher accuracy and fewer parameters.

4 Conclusion

This paper proposes a new algorithm based on a combination of YOLOv8 and Mixed Local Channel Attention (MLCA) to detect small objects, which is used to improve the performance of the YOLOv8 algorithm for small object detection. Our method aims to refine the YOLOv8 algorithm specifically for detecting small objects. We conducted comprehensive comparisons with some new attention mechanisms to validate the efficacy of our approach. The results demonstrate that our method not only surpasses the original YOLOv8 in small object detection in remote sensing images, but also outperforms other attention mechanisms (CloFormer, EMA, Biformer and LSKNet).

The integration of MLCA into YOLOv8 yielded significant enhancements in detection performance. Compared to the original YOLOv8, our method achieves superior small object detection performance in remote sensing images while slightly reducing the parameter count. For example, our method exhibits a 7.2% improvement in precision and a 1.2% increase in mAP50 over the original YOLOv8.

In future research, exploring alternative attention mechanisms is imperative. Our experiments revealed that most attention mechanisms exhibit promising enhancements in small object detection performance. Investigating these mechanisms further could lead to even more robust and effective algorithms for object detection in various domains.

Acknowledgments

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author Contributions

Not applicable.

About the Authors

WANG, Hao

Dept. of Mechanics and Mathematics Belarusian State University, Minsk, Belarus ahcenewang@gmail.com.

ABLAMEYKO, Sergey

Dept. of Mechanics and Mathematics Belarusian State University, Minsk, Belarus ablameyko@bsu.by.

References

[1] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context.” arXiv, Feb. 20, 2015. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/1405.0312

[2] L. Zhu, X. Wang, Z. Ke, W. Zhang, and R. Lau, “BiFormer: Vision Transformer with Bi-Level Routing Attention.” arXiv, Mar. 15, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2303.08810

[3] D. Ouyang et al., “Efficient Multi-Scale Attention Module with Cross-Spatial Learning,” in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096516.

[4] Q. Fan, H. Huang, J. Guan, and R. He, “Rethinking

Local Perception in Lightweight Vision Transformer." arXiv, Jun. 01, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2303.17803

[5] Y. Li, Q. Hou, Z. Zheng, M.-M. Cheng, J. Yang, and X. Li, "Large Selective Kernel Network for Remote Sensing Object Detection." arXiv, Mar. 19, 2023. Accessed: Apr. 15, 2024. [Online]. Available: http://arxiv.org/abs/2303.09030

[6] D. Wan, R. Lu, S. Shen, T. Xu, X. Lang, Z. Ren. (2023). Mixed local channel attention for object detection. Engineering Applications of Artificial Intelligence, 123, 106442, https://doi.org/10.1016/j.engappai.2023.106442.

[7] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8." 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[8] J. Ding et al., "Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges," IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3117983.