ENHANCING YOLOV8 FOR INFRARED OBJECT DETECTION VIA LEARNABLE MULTI-SCALE CONTEXT AND ATTENTION

Amankwah Consult[1]

[1]Research Department, Goethe Strasse 3, Kaufering, Germany

ABSTRACT

Infrared (IR) object detection poses unique challenges due to low texture, weak edges, sensor noise, and a strong dependence on global context. Modern convolutional object detectors, including YOLOv8, are primarily optimized for RGB imagery and often underperform when directly applied to large-scale infrared datasets. In this paper, we present a simple yet effective architectural enhancement to YOLOv8 by (1) replacing the standard Spatial Pyramid Pooling Fast (SPPF) module with a learnable multi-scale variant (SPPFPlus), and (2) integrating the Convolutional Block Attention Module (CBAM) into both the backbone and neck. The proposed approach introduces explicit parallel multi-scale context modeling and adaptive channel–spatial attention, addressing key representational limitations of infrared imagery. Extensive experiments on a large infrared dataset demonstrate an improvement of approximately 15% in detection performance over the YOLOv8 baseline, with notable gains in recall and small-object detection. The modifications are lightweight, modular, and can be seamlessly integrated into existing YOLOv8

KEYWORDS

YOLOv8, Attention, SPPFPlus

1. INTRODUCTION

Infrared object detection plays a critical role in applications such as night-time surveillance, autonomous driving, border monitoring, and search-and-rescue operations [1,2]. Compared to RGB imagery, infrared images exhibit fundamentally different characteristics, including reduced texture, blurred object boundaries, sensor noise, and significant dependence on global thermal context [3,4]. These properties make accurate localization and classification particularly challenging.

Recent advances in one-stage detectors, such as the YOLO family [5–7], have achieved impressive performance on RGB benchmarks. YOLOv8, in particular, incorporates efficient backbone design, feature pyramid networks, and fast inference pipelines [8]. However, its architectural components are largely inherited from RGB-optimized designs and are not explicitly tailored to the characteristics of infrared imagery. As a result, performance degradation is commonly observed when these models are directly applied to thermal datasets [9, 10].

Two major limitations arise in this context. First, standard multi-scale aggregation modules rely heavily on non-learnable pooling operations, which may suppress weak thermal signals. Second, the lack of adaptive feature selection makes it difficult to suppress background noise and emphasize salient thermal regions. To address these issues, we propose complementary

DOI: 10.5121/sipij.2026.17301

architectural modifications focusing on learnable multi-scale context modeling and attention-based feature refinement.

2. RELATED WORKS

2.1. Infrared Object Detection

Early infrared object detection methods relied on handcrafted features and threshold-based segmentation tailored to thermal contrast [11,12]. With the advent of deep learning, convolutional neural networks have been increasingly adopted for infrared tasks [13,14]. Several studies have explored adapting RGB-trained detectors to infrared domains through fine-tuning, data augmentation, or domain adaptation techniques [9,15]. Nevertheless, many approaches continue to employ architectures originally designed for RGB imagery, limiting their effectiveness in infrared scenarios.

2.2. Multi-Scale Feature Aggregation

Multi-scale feature aggregation is essential for detecting objects of varying sizes. Spatial Pyramid Pooling (SPP) [16] and Atrous Spatial Pyramid Pooling (ASPP) [17] are widely used to enlarge receptive fields and incorporate contextual information. Ultralytics introduced SPPF as an efficient approximation using sequential max-pooling operations [8]. While computationally efficient, SPPF relies exclusively on non-learnable pooling, which restricts its representational capacity, particularly for low-contrast infrared features.

2.3. Attention Mechanisms

Attention mechanisms have been extensively studied to improve feature representation by adaptively reweighting informative features. Channel attention approaches such as SENet [18] and ECA [19], as well as spatial attention mechanisms, have demonstrated effectiveness across vision tasks. CBAM [20] combines channel and spatial attention in a lightweight manner and has been successfully applied to object detection and remote sensing tasks [21,22]. However, its integration with modern YOLO architectures for infrared detection remains underexplored.

3. METHODS

1. Overview

We enhance YOLOv8 by introducing learnable multi-scale context modeling and attention-based feature refinement. The proposed modifications consist of: 1. Replacing the standard SPPF module with the proposed SPPFPlus module at the end of the backbone. 2. Integrating CBAM blocks into both the backbone and neck to refine features along channel and spatial dimensions.

All other components of the YOLOv8 pipeline, including the detection head and training strategy, remain unchanged to ensure fair comparison.

2. SPPFPlus-Learnable Multi-Scale Context Module

The proposed SPPFPlus module serves as a drop-in replacement for the original SPPF. Unlike SPPF, which approximates multi-scale context via stacked max-pooling, SPPFPlus employs parallel pooling branches with fixed kernel sizes (5, 9, and 13), followed by learnable convolutional refinement.

Formally, the input feature map is first compressed using a 1×1 convolution to reduce channel dimensionality. The reduced feature map is then processed by multiple parallel branches, each consisting of a max-pooling operation with a distinct kernel size followed by a 3×3 convolution, batch normalization, and non-linear activation. The outputs of all branches, together with the unpooled feature map, are concatenated and fused via a final 1×1 convolution as shown in figure 1

This design explicitly captures multi-scale context while enabling the network to learn scale-specific feature refinements, which is particularly beneficial for infrared imagery where object discrimination often depends on surrounding thermal context.

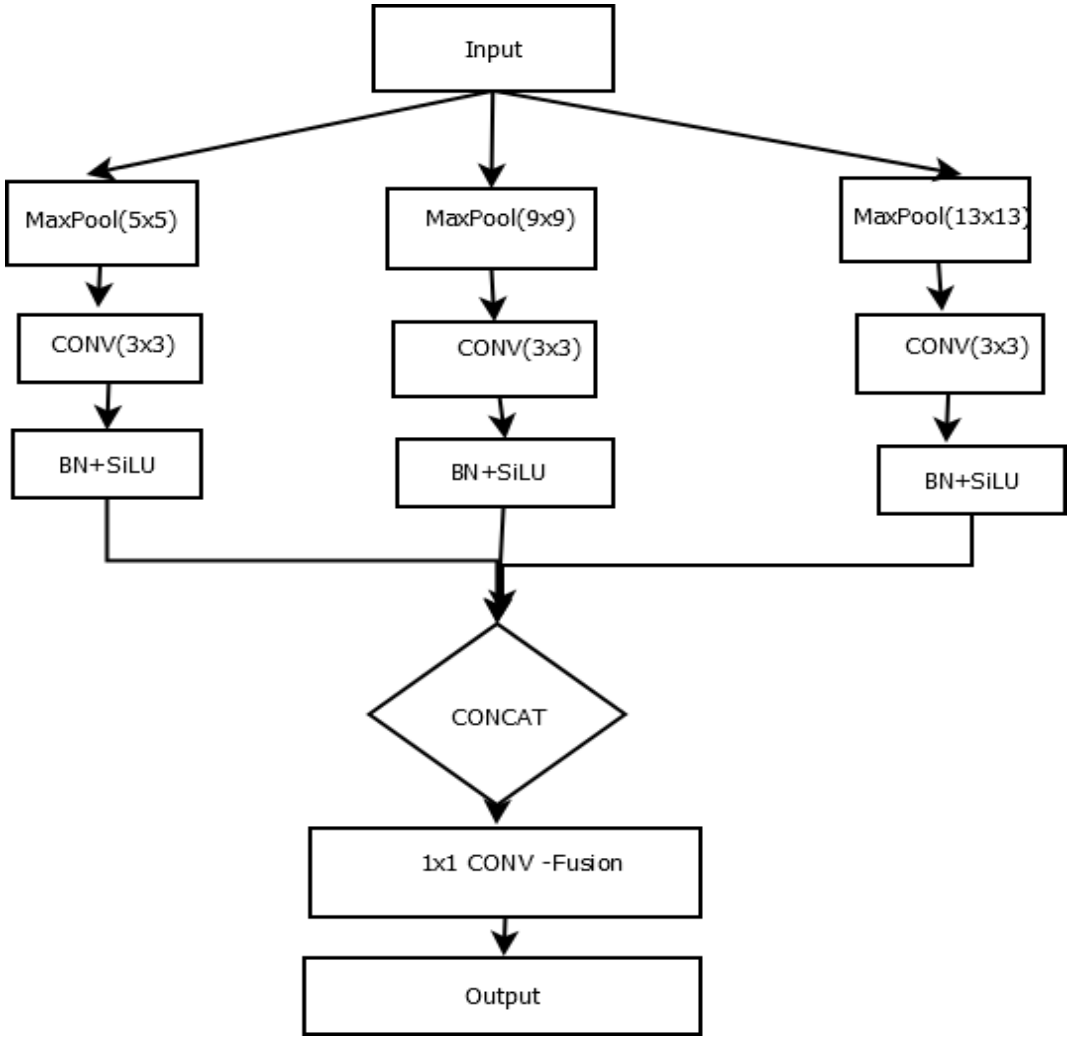

Figure 1 SPPFPlus Module

3. CBAM Integration in Backbone and Neck

CBAM is incorporated into both the backbone and neck of YOLOv8 to enhance feature representation. Channel attention allows the network to emphasize informative thermal responses, while spatial attention highlights relevant regions corresponding to object locations.

In the backbone, CBAM suppresses early-stage noise and enhances meaningful infrared patterns before multi-scale aggregation. In the neck, CBAM improves feature fusion across different resolutions in the FPN structure, facilitating better alignment of thermal cues across scales. Figure 2 show the proposed YOLOv8 method

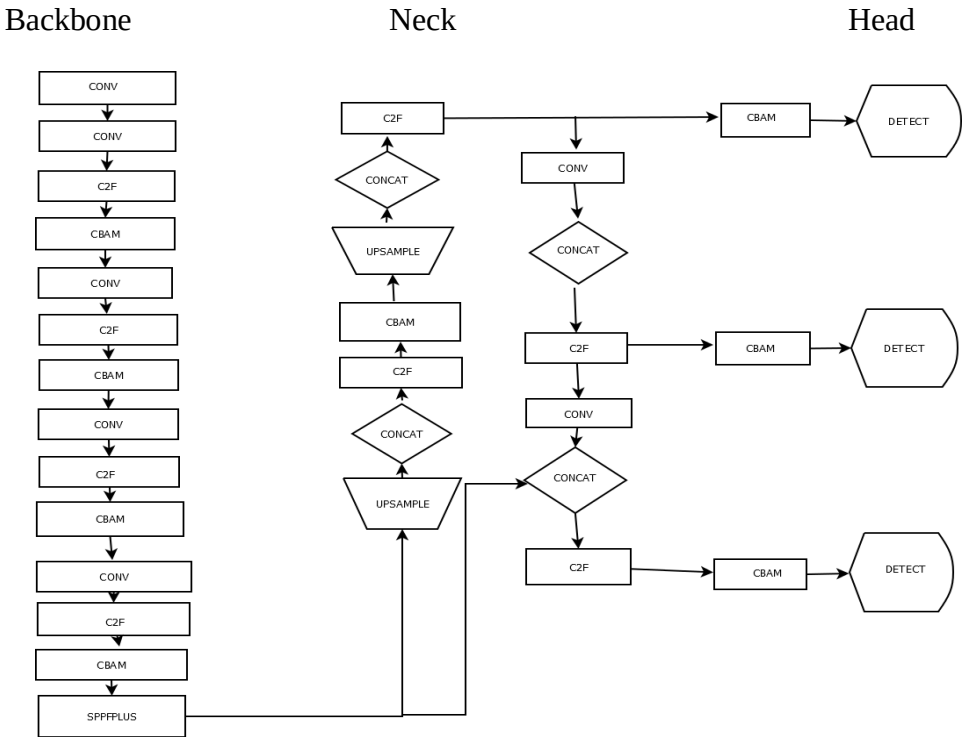

Figure 2 Proposed YOLOv8 Method

4. EXPERIMENTS AND RESULTS

Experiments were conducted on a large-scale infrared object detection dataset derived from FLIR thermal imagery [23]. The dataset contains approximately 23,000 training images and 2,500 validation images across three object categories: person, car, and dog. The data exhibit substantial variation in object scale, thermal contrast, and background clutter.

All models were trained using the Ultralytics YOLOv8 framework with identical hyperparameters. The input resolution was fixed at 640×640, and training was performed for 100 epochs using stochastic gradient descent with momentum, cosine learning rate scheduling, and Automatic Mixed Precision (AMP). Experiments were conducted on a single NVIDIA RTX 2050 GPU. We evaluate performance using mean Average Precision at an IoU threshold of 0.50 (mAP@0.50) and recall. mAP@0.50 measures the area under the precision–recall curve using a fixed IoU of 0.50, while recall represents the proportion of correctly detected objects among all ground-truth instances [24].

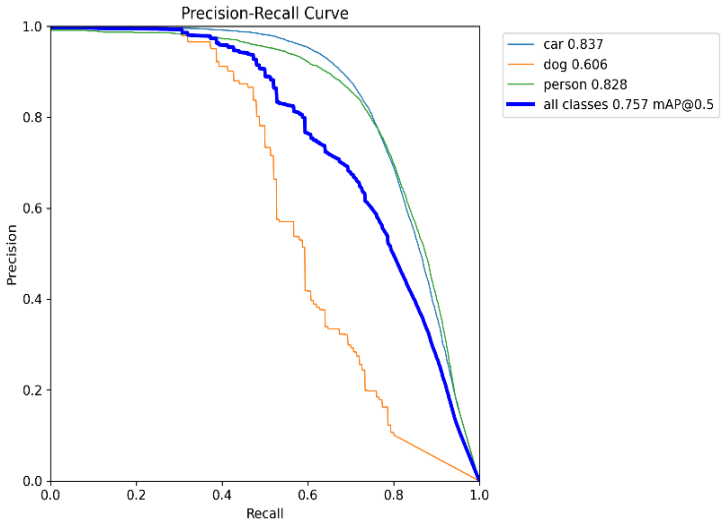

Figure 3 Precision-Recall Curve for standard YOLOv8

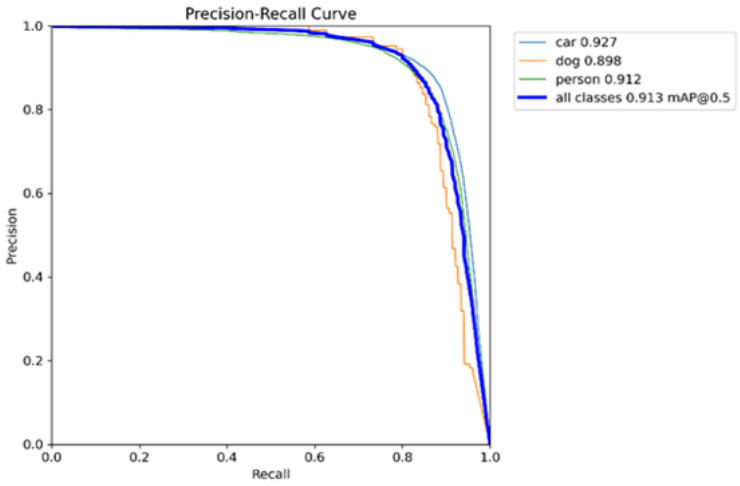

Figure 4 Precision-Recall Curve for proposed YOLOv8+ CBAM + SPPFPlus

Figures 3 and 4 show precision-recall curves for standard YOLOv8 and the proposed YOLOv8+ CBAM + SPPFPlus respectively. The proposed YOLOv8 architecture delivers a substantial performance improvement over the baseline YOLOv8 model employing the standard SPPF module. The baseline detector achieves approximately 0.76 mAP@0.50, whereas the proposed model reaches 0.913 mAP@0.50 and 0.559–0.560 mAP@0.50–0.95. Furthermore, recall increases by approximately 0.20, indicating a significant reduction in missed detections. The precision–recall curves demonstrate that the proposed model maintains higher precision across a wider recall range, resulting in a substantially larger area under the curve. These results confirm that the introduction of learnable multi-scale context modeling via SPPFPlus, together with channel–spatial attention from CBAM, enhances feature discriminability and confidence calibration, particularly for low-contrast infrared targets. Figure 5 shows the visualisation of object detection using proposed YOLOv8 on test images from FLIR

Figure 5. Object detection using proposed YOLOv8 on FLIR dataset

Despite the additional modules, the proposed model remains computationally efficient, with approximately 14M parameters and an average inference time of 26 ms per image at 640×640 resolution. This demonstrates a favorable accuracy–efficiency trade-off suitable for real-world deployment.

5. CONCLUSIONS

This paper proposes an enhanced YOLOv8 architecture for infrared object detection by integrating CBAM into the backbone and neck together with an improved SPPFPlus module. The modifications strengthen feature selection and multi-scale aggregation, which is especially beneficial for low-contrast thermal imagery. Experiments on a large infrared dataset demonstrate an approximate 15% absolute improvement over the baseline YOLOv8 model and recall rises by about 0.20. Precision–Recall curves further confirm that the proposed model maintains higher precision across all recall levels. These results demonstrate that combining attention mechanisms with enhanced spatial pyramid pooling is an effective strategy for improving infrared target detection. The proposed modifications are lightweight, modular, and broadly applicable, making them a strong baseline for future infrared detection research.

REFERENCES

[1] M. Vollmer and K.-P. Möllmann, Infrared Thermal Imaging: Fundamentals, Research and Applications. Weinheim, Germany: Wiley-VCH, 2018.

[2] J. W. Davis and M. A. Keck, “A two-stage template approach to person detection in thermal imagery,” in Proc. IEEE Workshop on Applications of Computer Vision (WACV), 2005.

[3] A. Torabi, G. Massé, and G.-A. Bilodeau, “Thermal–visible image registration using mutual information,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012.

[4] A. Berg, J. Ahlberg, and M. Felsberg, “Challenges in infrared image analysis,” Pattern Recognit., vol. 47, no. 3, 2016.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
[6] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018.
[7] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934, 2020.
[8] Ultralytics, "YOLOv8 documentation," 2023. [Online]. Available: https://docs.ultralytics.com
[9] C. Li, Y. Zhao, and Y. Wang, "Object detection in thermal infrared images based on deep learning," Sensors, vol. 19, no. 16, 2019.
[10] L. Zhang, L. Peng, T. Zhang, and Z. Cao, "Infrared small target detection based on multiscale local contrast learning," IEEE Trans. Geosci. Remote Sens., vol. 59, no. 3, 2021.
[11] S. D. Deshpande, M. H. Er, R. Venkateswarlu, and P. Chan, "Max-mean and max-median filters for detection of small targets," in Proc. SPIE, vol. 3809, 1999.
[12] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 1998.
[13] S. Liu, D. Huang, and Y. Wang, "Deep learning for infrared imagery: A review," Neurocomputing, vol. 275, 2018.
[14] J. Zhao, Y. Zhang, and J. Huang, "Thermal pedestrian detection using deep neural networks," Pattern Recognit., vol. 102, 2020.
[15] D. König, M. Adam, C. Jarvers, G. Layher, and M. Teutsch, "Domain adaptation for thermal person detection," in Proc. ECCV Workshops, 2018.
[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014.
[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv:1706.05587, 2017.
[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.
[19] Q. Wang et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020.
[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018.
[21] Y. Zhang, J. Xiao, and Z. Zhang, "Attention-based convolutional neural networks for remote sensing image classification," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 10, 2020.
[22] X. Li, M. Zhang, and Y. Zhang, "Attention-based deep learning for object detection," Neurocomputing, vol. 452, 2021.
[23] FLIR Systems, "FLIR thermal dataset for algorithm training," 2018. [Online]. Available: https://www.flir.com/oem/adas/adas-dataset-
[24] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common objects in context,"in Proc. Eur. Conf. Comput. Vis. (ECCV), 2014.

AUTHORS

Anthony Amankwah earned his B.Sc. degree in Metallurgical Engineering from Kwame Nkrumah University of Science and Technology in 1996. He subsequently pursued further studies in Germany, obtaining both B.Sc. and M.Sc. degrees in Electrical Engineering and Computer Science from University of Duisburg-Essen in 2003. He later completed his Ph.D. in Electrical and Computer Science at University of Siegen. Dr. Amankwah currently works in the Machine Vision industry in Germany, where he applies his multidisciplinary expertise in engineering, computer science, and intelligent vision technologies.