Infrared Road Object Detection Based on Improved YOLOv8

Zilong Luo, Ying Tian

Abstract—**In recent years, infrared target detection has played a crucial role in intelligent transportation and assisted driving. Addressing the current issues of low detection accuracy, poor robustness, and missed detections in infrared image detection, we propose an improved infrared road traffic detection algorithm, YOLOv8-EGP, based on YOLOv8s. Firstly, we replace the original C2f module with the SCConv convolution module to extract feature information of different sizes, thereby enhancing the perception of local features in infrared images and connecting spatial relationships between them. Then, in the head part, we use the dyhead detection head and combine three different dimensions with multi-attention, improving the expression ability of the detection head for infrared targets without increasing computational complexity. Finally, we add a small target detection layer (min) to reduce missed detections of small targets in infrared images and improve the final detection accuracy. The conducted ablation experiments show that on the FILIR public dataset, compared to YOLOv8s, the YOLOv8-EGP algorithm increases mAP50 by 6.1%, and precision and recall also increase by 5.8% and 1.6%, respectively, indicating that the improved model can better adapt to infrared target detection, validating the effectiveness of this method.**

Index Terms—**Deep Learning, Infrared Images, Object Detection, YOLOv8**

I. Introduction

With the rapid development of safe cities, smart transportation, and other fields, the application of infrared image object detection technology is becoming more and more widespread. This is especially true in the area of vehicle assisted driving, where it plays a crucial role. Due to the unique imaging method and the low resolution of IR images, they are also susceptible to environmental noise interference. This makes the task of object detection and recognition relatively challenging, leading it to become a hot topic of research.

Object detection is a crucial research area in computer vision, finding extensive applications across various domains [1]. Object detection algorithms have witnessed significant development in computer vision in recent years. Many classical representative algorithms have emerged, generally falling into two categories. One is the traditional object detection algorithm, in which the network only focuses on extracting representations from a single image and ignores the potential correlation between images [2]. Another type is deep learning-based object detection algorithms. In addition, object detection models can also be categorized into two types, one of which is the two-stage approach. For example, in 2014, Girshick et al [3]. A two-stage object detector model called R-CNN was introduced in, which was the first algorithm to apply deep learning to object detection. However, despite significant progress, it still faces many issues related to real-time performance and accuracy.

Subsequent models like Fast R-CNN [4] and Faster R-CNN [5] were developed to address the shortcomings of the R-CNN model. For instance, Fast R-CNN replaces the fully connected layers in R-CNN with ROI Pooling layers to improve the accuracy of object detection.

Faster R-CNN introduces the Region Proposal Network to generate candidate boxes, which reduces the computational complexity. However, it still faces challenges such as high computational requirements, long training times, and difficulties in handling multiple scales and small objects. Another class of models is the single-stage approach, which emerged in 2016. Representative algorithms include YOLO [6] and SSD [7]. These models successfully transform object detection into a regression problem, using deep neural networks to automatically learn and predict object positions and categories. This significantly improves the accuracy and efficiency of object detection, although at the cost of some accuracy. The computational speed is greatly enhanced.

The SSD algorithm is trained on both detection and classification tasks within the same network, simplifying the model structure. It uses different anchor boxes to accommodate various sizes, improving efficiency and detection accuracy. However, its speed was slower compared to the YOLO series [8-10] algorithms.

The YOLO series evolved rapidly, with continuous updates from YOLOv3 to YOLOv6 [11] in recent years. They addressed some of the shortcomings of the original version and improved both speed and accuracy. For example, the YOLOv7 [12] algorithm introduced in 2022 featured an efficient ELAN structure, dynamic label assignment strategy, and added auxiliary heads to the feature pyramid model. This not only improves the accuracy of object detection, but also increases the speed. However, YOLOv7 still faces challenges related to model complexity and computational requirements.

The latest YOLOv8 algorithm adopts a state-of-the-art (SOTA) model that draws inspiration from YOLOv7 ELAN's design principles. It replaces the C3 structure of YOLOv5 with the C2f structure, which provides a richer gradient flow. In addition, the head profile has been replaced with the current mainstream decoupled head structure, separating the

Manuscript received October 6, 2023; revised January 11, 2024.

Zilong Luo is a postgraduate student majoring in software engineering at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Liaoning 114051, China. (e-mail: 1753372894@qq.com).

Ying Tian is a Professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author to provide phone: +8613898015263; fax: 0412-5929818; e-mail: astianying@126.com).

classification and detection heads. Furthermore, YOLOv8 abandons the Anchor-Based approach in favor of Anchor-Free and employs the Task-Aligned Assigner [13] positive sample assignment strategy along with Distribution Focal Loss for loss calculation. In terms of data augmentation during training, YOLOv8 borrows the strategy from YOLOX [14] of disabling Mosaic augmentation in the final 10 epochs, effectively enhancing model accuracy.

Overall, YOLOv8 builds on the strengths of the YOLO family while introducing innovations that improve model flexibility and performance. However, when it comes to object detection in IR images, the complexity of IR images and the lack of distinctive object features, especially for small objects, can lead to poor model accuracy.

In this paper, we introduce the YOLOv8-EGP algorithm for IR road object detection by proposing several improvements and optimizations based on YOLOv8s. The optimizations include:

1. Improving the C2f layer by replacing it with a more flexible and adaptive convolution module called SCConv [15], enhancing feature diversity in the output.

2. Adding a detection layer designed to better handle small targets, improving feature extraction for small objects, and addressing issues such as inaccurate detection caused by YOLOv8s' division of the image into large grids.

3. Enhancing detection capabilities using the attention-based detection head dyhead [16] to improve model generalization and detection accuracy.

II. Improved Model

This algorithm uses the YOLOv8s model as a reference and makes improvements to enhance object detection in IR images. First, it replaces the original C2f layer in YOLOv8s with SCConv. Second, an additional small object feature extraction layer is added to improve the feature extraction capability. In addition, the original Detect detection head of the model is replaced with an attention-based detection head called dyhead.

The entire model structure is divided into three levels. The YOLOv8-EGP model is structured as follows and is shown in Fig. 1.

1. The backbone main network, known as the "backbone," serves as the core feature extractor responsible for the initial feature extraction.

2. The neck part builds upon the backbone network to extract and fuse features to provide richer information for subsequent predictions.

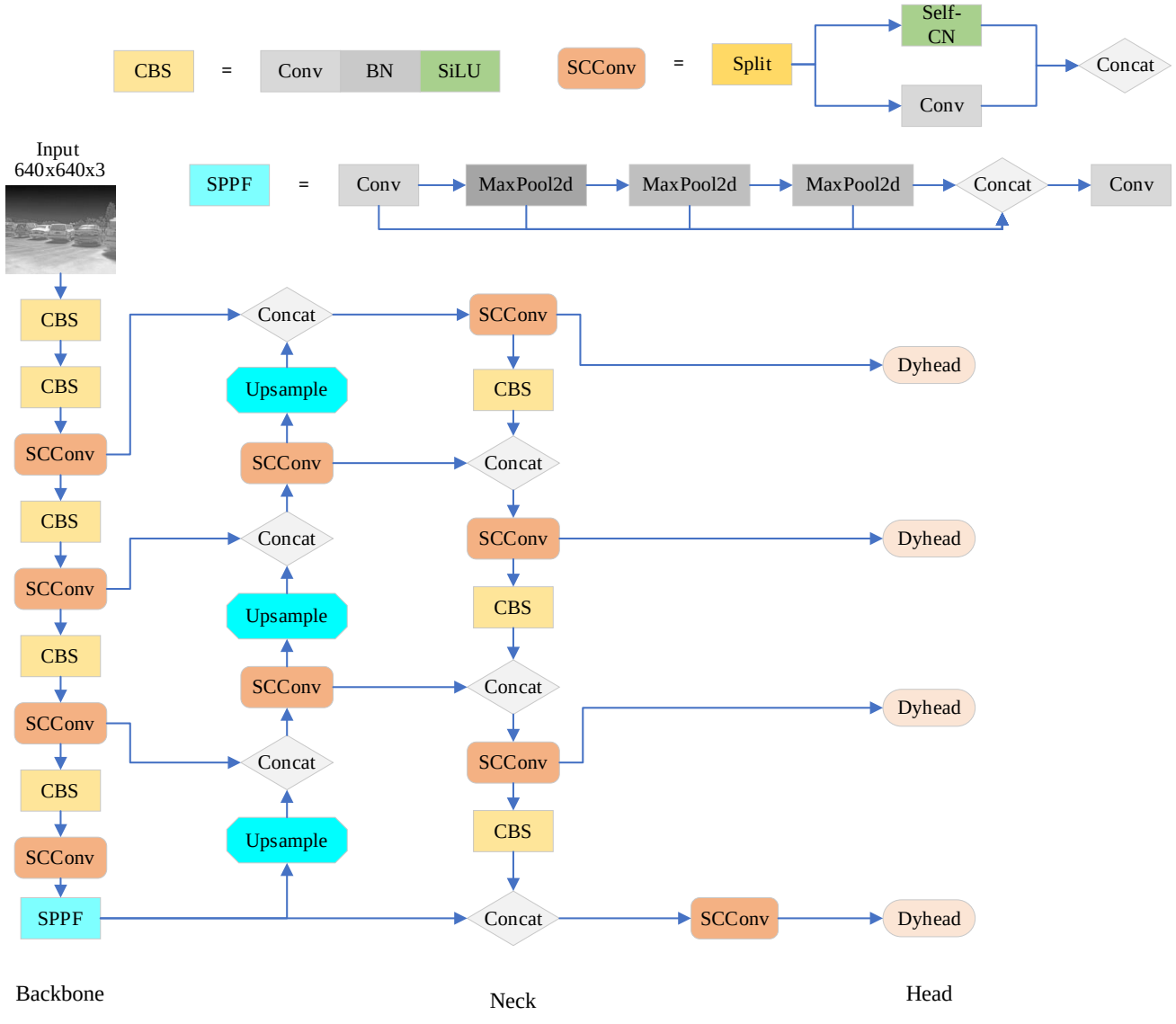


Fig. 1. YOLOv8-EGP model

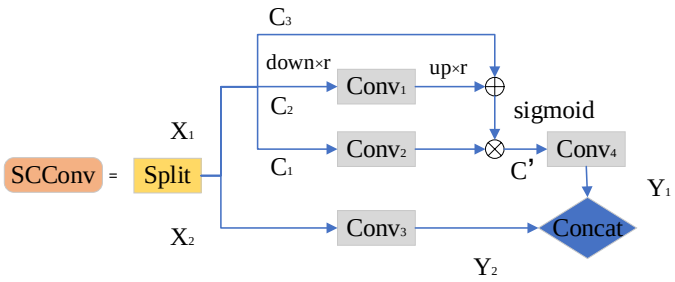


Fig. 2. SCConv module

3. The head part is the output layer for predicting bounding boxes and class information.

This architecture is designed to enhance the model's capability for infrared object detection by improving feature extraction and prediction.

A. *SCConv Module*

The SCConv module represents the Self-Calibrated Convolutions module.

The original YOLOv8s model uses the C2f module to maintain consistency in the number of convolution channels between cv1 and cv2. This is beneficial for preserving the consistency of the convolutional information. However, in the cv3 module, it has twice as many input channels as the previous layers. The reason for this is that the main gradient flow cv3 is obtained by a cat operation between the BottleNeck branch and the CBS branch of the secondary gradient flow cv2. This design is designed to allow the C2f module to capture richer gradient flow information.

The C2f module was designed with inspiration from the ELAN module, which aims to make the model more lightweight while capturing rich information. However, such a design may encounter issues when dealing with targets in infrared object detection. For instance, infrared images, which are influenced by temperature variations, may exhibit less prominent features when their grayscale values are low. The feature maps output by C2f are generated using the same formula repeatedly, and their spatial receptive field is primarily controlled by the predetermined kernel size. As a result, the extracted features may contain a significant amount of redundant information, leading to weak discriminative power.

In the context of infrared object detection, the challenge is to efficiently capture meaningful features from objects with low contrast or temperature sensitivity. Therefore, further improvements or adaptations of the C2f module may be necessary to address these specific challenges and enhance the performance of the model on infrared images.

In the YOLOv8-EGP model, the SCConv module used is distinct from other convolutional modules. SCConv achieves this by employing convolutional kernels of various sizes, allowing it to capture spatial context relationships, have a stronger receptive field, and pay more attention to local information. The SCConv module, can encode multiscale information through adaptive operations, providing richer feature information for subsequent operations, thereby improving localization accuracy. The structure of the SCConv module is depicted in Fig. 2.

In Fig. 2, X represents the input feature map, Y represents the output feature map, $Conv_n$ denotes convolution layers with different kernels, C_n indicates results from different branches of X_1, $\oplus$ represents the XOR operation, and $\otimes$ signifies element-wise multiplication between matrices. The input feature map of dimensions C×H×W is divided into X_1 and X_2, with the channel count halved to C/2×H×W through Split. X_2 is processed by a regular convolutional layer to produce the output Y_2, aimed at preserving spatial context relationships. Meanwhile, X_1 undergoes an adaptive calibration operation to yield Y_1. For X_1, average pooling is applied:

$$C_2 = \text{avgpool}_r(X_1) \tag{1}$$

The down-sampling factor r for pooling is set to 2 in the model, and it is followed by up-sampling using bilinear interpolation.

The process involves upsampling the C_2 convolution and adding it to C_3, followed by a sigmoid operation to obtain weight values. Then, these weight values are applied as element-wise multiplication with the result of the C_1 convolution, yielding C'.

$$C' = \text{Conv}_2(X_1) \otimes \sigma(C_2 \oplus C_1) \tag{2}$$

After obtaining C', it undergoes a convolution operation using $Conv_4$ to produce the final output Y_1.

$$Y_1 = \text{Conv}_4(C') \tag{3}$$

X_1's role is to extract features in the downward direction, expand the receptive field, and capture attention mechanism weights. The structures of C_3 and C_2 serve a similar role in obtaining attention mechanism weights, while C_1 acts as the main branch module of the attention mechanism. X_2 and X_1 are concatenated without affecting each other, resulting in the acquisition of rich gradient information.

Overall, SCConv can significantly enhance the receptive field and improve the feature extraction capability for infrared targets with minimal changes in parameter count. This leads to more precise localization and higher accuracy, achieving better results.

B. *Dyhead Module*

The dyhead module represents the Dynamic Head module. Infrared images and visible-light images exhibit significant differences in visual characteristics, which can lead to various challenges when using object detection algorithms directly. Issues such as single-channel data and differences in thermal radiation need to be addressed for better performance. In the YOLOv8-EGP model structure, the dyhead detection head is utilized to combine multiple attention mechanisms at feature levels, spatial positions, and output channels. This enables scale awareness, spatial awareness, and task awareness, making it particularly effective for handling complex object shapes and scenes in infrared image targets. It enhances the robustness of object detection, and the design of dyhead allows high-performance object detection with relatively low computational resources. It can still deliver satisfactory results even when computational resources are limited. The block structure of this module is depicted in Fig. 3.

The term π_L represents the Scale-aware Attention. It involves the following steps:

1. Global Pooling on H×W: The feature map's maximum value is obtained by performing global pooling over the H×W dimensions.

2. Channel Integration: A 1×1 convolutional kernel is used to process and integrate channels.

3. Activation Functions: The result goes through ReLU [17] and sigmoid activation functions.

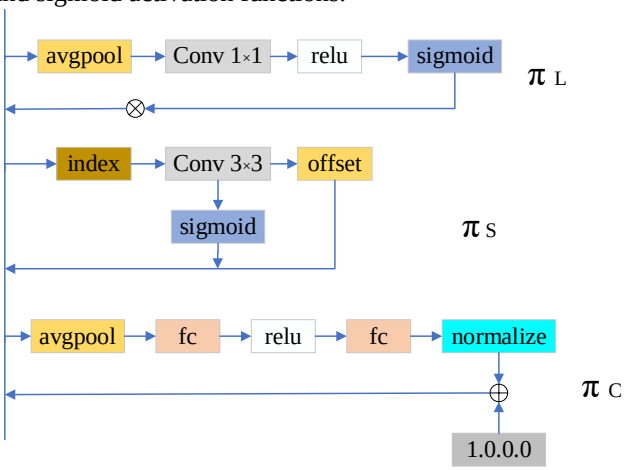


Fig. 3. Dyhead module

In the context of infrared images, which often have lower resolution and contrast, the ability of an object detector to perceive sizes is crucial. This structure helps in feature extraction by capturing the maximum value from the feature map. It is better suited to capture information about the size of the objects, making it effective for handling multi-scale targets. Therefore, it significantly enhances the scale-awareness capability for infrared object detection.

π_S represents the Spatial Attention, which is implemented using the deformable convolution v2 [18] structure. This structure enables a sparse treatment of the attention by adding an offset at each sampling point and modulating the feature amplitude. Conventional convolution operations are typically performed on a fixed grid, limiting their ability to handle irregular objects. However, deformable convolution v2 allows the convolutional positions to be deformed, ensuring diversity in the input data.

Deformable convolution v2 adds an offset at each sampling point, enabling convolution positions to adapt to the offset. When applied to infrared images, this adaptive convolution position adjustment allows the model to capture features of different sizes, shapes, and orientations. This, in turn, improves detection accuracy, performance, and the model's generalization ability for infrared object detection.

π_C represents the channel attention, which is constructed using a two-layer fully connected neural network. This channel attention mechanism facilitates joint learning and generalization of the target representation. In the context of infrared images, it can selectively weight different features extracted by various convolutional kernels to enhance or suppress specific features, thereby improving the effectiveness of infrared image processing.

For instance, high-frequency features in infrared images can capture information about texture structures. The channel attention mechanism enhances the perception of these features. On the other hand, low-frequency features reflect overall image characteristics, and the mechanism can enhance these features as well, thereby improving perception.

In summary, the channel attention mechanism selectively enhances different features based on their relevance, which contributes to better feature extraction and overall performance in processing infrared images.

C. Small-Object Detection Head (min)

One of the challenges in infrared image object detection is the difficulty in detecting small targets. When the target is distant, it has very few pixels, and the high background noise around the target can severely interfere with the detection process. Due to the low signal-to-noise ratio, it becomes challenging to distinguish the target from the noise, posing significant difficulties in detection. Additionally, distinguishing between infrared target features and background features can be challenging. In complex environments, targets can also be affected by various factors such as viewing angle, lighting conditions, and occlusion, making feature extraction for the target even more difficult. In the original YOLOv8s model, there are three detection heads designed for multi-scale object detection. The P3 head is used to detect targets larger than 8×8 pixels within a detection grid of 80×80.

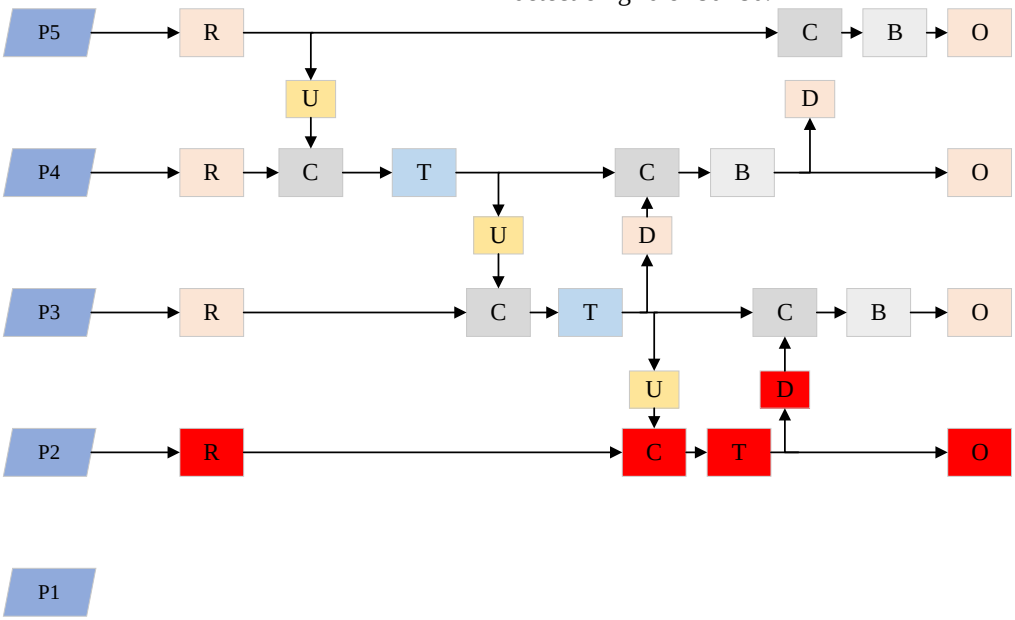


Fig. 4. Min module

The P4 head detects targets of 16×16 within a grid of 40×40, and the p5 head detects targets of size 32×32 within a grid of 20×20. However, since the model focuses on relatively large grids for detection, it can lead to missed detections or poor performance when dealing with small targets.

To address the issue of subpar detection of small objects, the proposed YOLOv8-EGP model adds a dedicated small object detection head on top of the existing structure. The structure diagram is shown in Fig. 4.

In the YOLOv8s model, there are typically three normal detection heads: P3, P4, and P5, each producing output at different scales. To enhance the model's ability to detect small targets, a fourth output layer, P2, has been added. This additional layer generates a feature map of size 80×80, allowing the model to detect targets as small as 4×4 in size. During experiments, this modification improved detection performance by 5% compared to YOLOv8s, specifically for detecting small targets.

III. Experimental Results And Discussion

A. Datasets

The dataset used for this experiment is the FLIR_ADAS_v2 [19] infrared dataset, released by FILIR Systems in 2022. Compared to previous versions, this dataset has expanded to include 15 different categories, added video data, and increased the total annotated frames to 26,442 frames, representing a 1% increase from the original version. Moreover, all images included in this dataset are annotated with labels for all 15 categories.

For this experiment, a subset of 10,467 infrared images was selected. These images were divided into training, testing, and validation sets using a 7:3:1 ratio, resulting in 7,326 images for training, 2,094 images for testing, and 1,047 images for validation. Due to the limited number of images available for some of the 15 categories, only six specific categories were chosen for experimentation. These categories include person, bike, car, bus, light, and sign.

B. Experimental Environment

The model was developed using the Python programming language and implemented with the PyTorch deep learning framework. The PyTorch version used was 1.8.1. The hardware specifications for the system included a GeForce GTX 1080ti GPU with 11,178MB of VRAM.

During training, the input images were resized to 640×640 pixels. The Stochastic Gradient Descent (SGB) function was employed as the optimizer. The training process consisted of 300 epochs, with a batch size of 8. The momentum and weight decay parameters were set to 0.937 and 0.0005, respectively. The initial learning rate was set to 0.01, and a cosine annealing learning rate schedule was applied. Mosaic augmentation was enabled for the last 10 epochs of training.

C. Experimental Evaluation Metrics

This article conducts research on three model evaluation metrics, namely precision, recall, and mean Average Precision (mAP) at a threshold of 0.5. Precision: Precision is an important evaluation metric in classification tasks. It measures the proportion of true positives (TP) in all samples predicted as positive. The calculation formula is as follows:

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

TP (True Positives) represents the number of true positive samples, which are the samples correctly classified as positive.

FP (False Positives) represents the number of false positive samples, which are the samples incorrectly classified as positive.

The calculation of precision helps assess the model's ability to recognize objects in infrared images. If the model has a high precision, it means that it can accurately identify the target, reducing false positives and false negatives, thereby improving the reliability and safety of the driving assistance system.

Recall is another important evaluation metric that represents the proportion of correctly predicted samples among all true positives.

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{5}$$

FN (False Negative) represents the number of false negatives, indicating samples that are incorrectly classified as negative.

The calculation of recall helps assess whether the model can effectively identify all true positive targets. If the model has a high recall rate, it means that it can correctly identify as many positive targets as possible, avoiding misses, and thus enhancing the system's reliability and safety.

mAP@0.5 stands for mean average precision (mAP) computed under the condition of an intersection over Union (IoU) equal to 0.5. It is commonly used to evaluate the performance of object detection tasks. mAP calculates the average precision for each class and then takes the mean of those values.

TABLE I
Results of Ablation Experiments

YOLOv8s	SCConv	dyhead	min	P(%)	R(%)	mAP@0.5(%)
✓				84.0	68.2	76.8
✓	✓			83.2	69.3	77.2
✓		✓		82.7	69.4	77.1
✓			✓	84.6	73.3	81.8
✓	✓	✓		83.7	68.8	77.5
✓	✓	✓	✓	85.6	74.0	82.9

TABLE II
COMPARISON BETWEEN THE ORIGINAL YOLOV8S MODEL AND THE MODEL WITH ADDED SMALL OBJECT DETECTION HEAD

Approaches	alls(%)	person(%)	bike(%)	car(%)	bus(%)	light(%)	sign(%)
YOLOv8s(P)	84.0	85.9	85.2	0.85	82.1	87.4	78.5
YOLOv8s(R)	68.2	73.9	68.6	78.1	78.1	57.7	52.9
YOLOv8s(mAP@0.5)	76.8	82.8	76.9	85.0	84.1	70.2	61.6
min(P)	84.6	87.4	82.7	86.8	82.6	87.8	80.2
min(R)	73.3	78.1	73.9	81.0	77.0	68.3	61.6
min(mAP@0.5)	81.8	87.6	81.4	88.7	84.0	79.1	70.3

In object detection tasks, if a detection box has an overlap with a ground truth object box greater than 0.5 (i.e., IoU greater than 0.5), then that detection box is considered a True Positive. mAP@0.5 represents the average precision for all classes at IoU equals 0.5. The formula for calculating the average precision for n classes is as follows:

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1}\text{Precision(Recall)d(Recall)} \tag{6}$$

mAP calculates the average precision for each class and then computes the mean of these individual class average precision scores. It is used to evaluate the overall performance of a model across multiple classes.

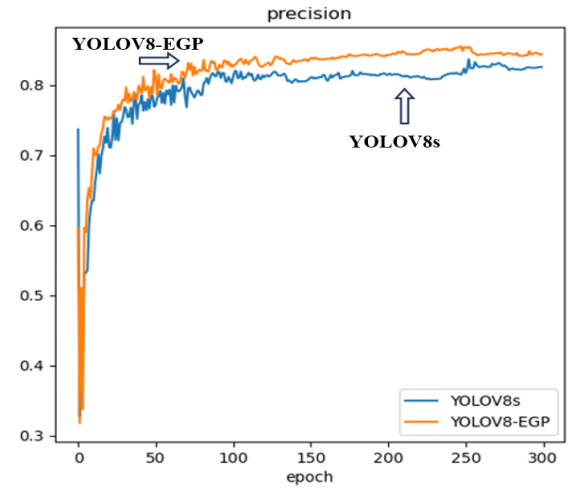
D. Ablation Experiment

To comprehensively evaluate the influence of various methods on the experimental outcomes, it is crucial to perform ablation experiments using the YOLOv8s model as a reference. This involves a step-by-step approach, starting with the replacement of the original C2f layer in YOLOv8s with SCConv, followed by the substitution of the native detection head (detect) with dyhead, and ultimately, the addition of an extra layer dedicated to small object detection. These sequential tests allow for a detailed depiction of parameter variations within each module.

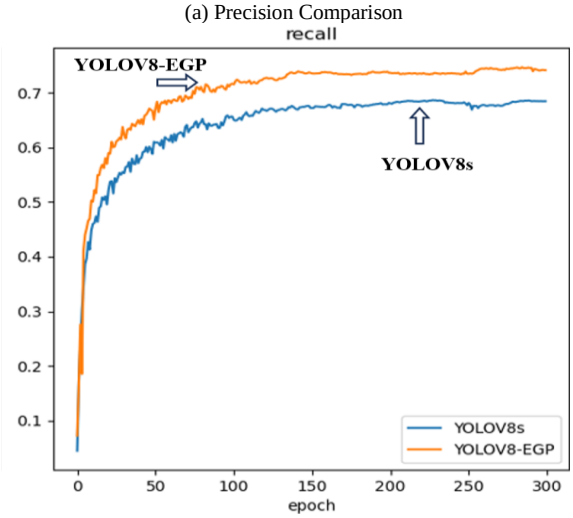
Table I demonstrates that when the original YOLOv8s model is enhanced with the SCConv and dyhead modules, there is a slight decrease in precision but an improvement in recall. Overall, the mAP values show a slight increase of 0.4% and 0.3%, respectively. This indicates that the application of SCConv and dyhead modules have a slightly positive impact on infrared object detection.

Moreover, in the original YOLOv8s model results, the precision, recall, and mAP for the "light" and "sign" classes exhibited significantly lower values compared to other classes. This discrepancy can be attributed to the adverse effects of low-resolution infrared images and substantial environmental noise. However, upon incorporating an additional small object detection layer, there is a conspicuous improvement in these metrics. This enhancement confirms the effectiveness of the added detection layer for improving the detection of small objects. The comparative experimental results are presented in Table II.

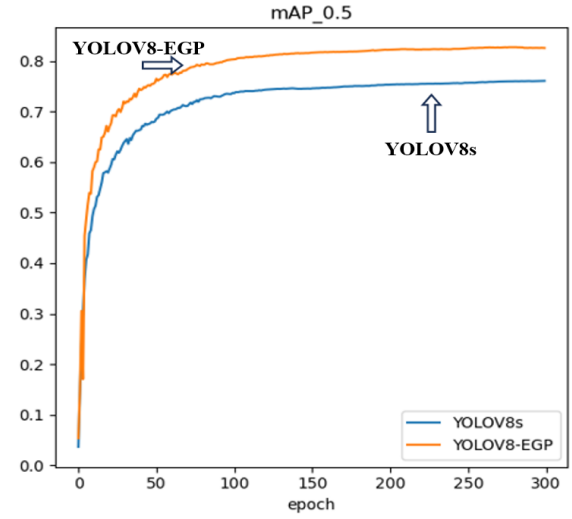
In the final comparison between YOLOv8s and YOLOv8-EGP, we observed consistent improvements across all three measures of evaluation: Precision increased by 1.6%, Recall showed a notable improvement of 5.8%, and mAP exhibited a significant boost of 6.1%. These findings affirm that, within the domain of infrared target detection, the YOLOv8-EGP model surpasses the original model, underscoring its capacity for substantially enhancing the extraction of target features. To offer a more comprehensive depiction of these advancements, we have included the actual model progression curves in Fig. 5.



(a) Precision Comparison



(b) Recall Comparison



(c) mAP Comparison

Fig. 5. YOLOv8s and YOLOv8-EGP

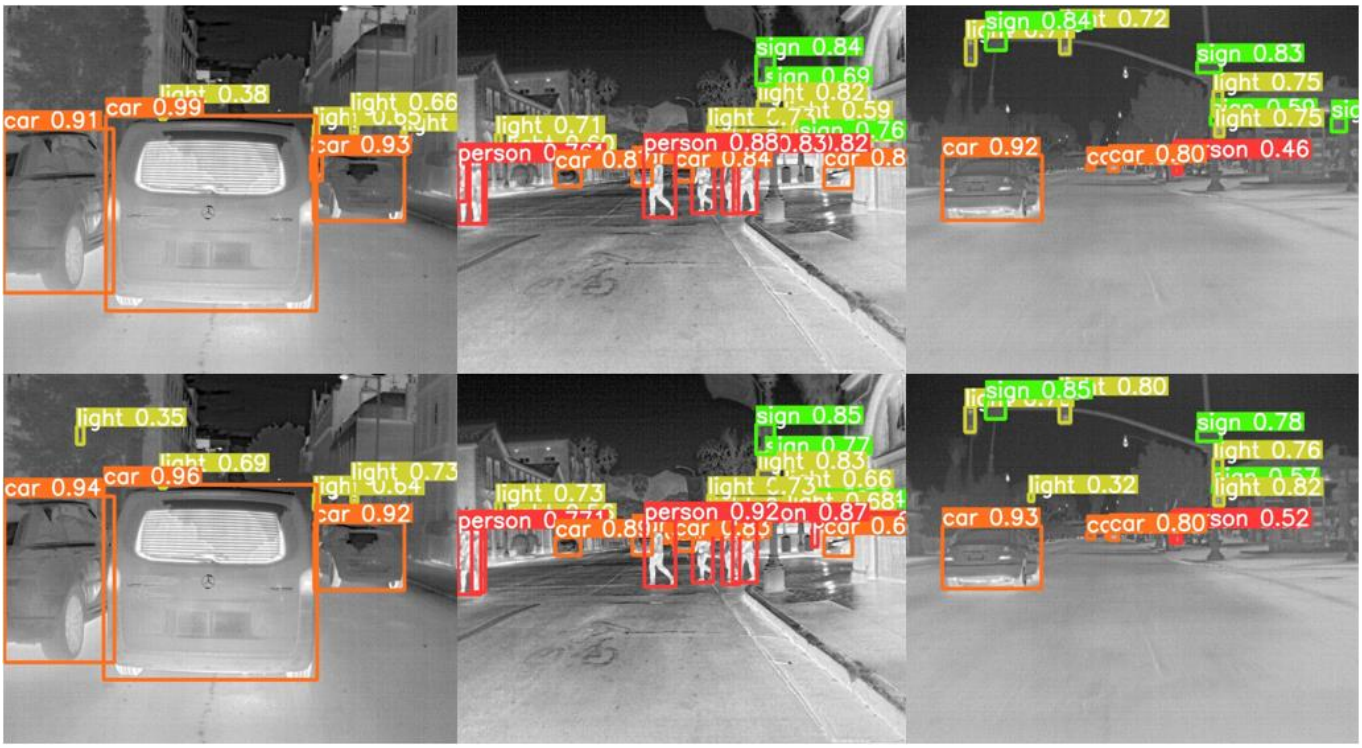


Fig. 6. YOLOv8s (top) and YOLOv8-EGP (bottom)

After 300 rounds of training, the Precision, Recall, and mAP curves for both YOLOv8 and YOLOv8-EGP models are shown in the figure. It is evident that both models essentially converge within the first 100 rounds of training, and their trends remain relatively stable, demonstrating excellent model robustness. By analyzing the curves, one can estimate that the actual difference in Recall and mAP between the two models is approximately around 6%. This further confirms that the improved model is better suited for infrared image object detection tasks. Therefore, applying the enhanced model to practical production and daily life is expected to deliver superior detection performance. To provide a clearer illustration of the model's real-world effectiveness, we have included some actual result images in Fig. 6.

In the comparative visualization of actual results, it is evident that the YOLOv8s model performs less effectively in detecting small objects compared to YOLOv8-EGP and exhibits occasional instances of missed detections. Moreover, YOLOv8-EGP demonstrates a substantial enhancement in overall accuracy, while there is a slight reduction in the detection of a few infrared targets, this decrease is significantly outweighed by the substantial overall improvement in performance.

E. Comparative Experiments

To further validate the algorithm's performance, this study conducted a comparative analysis between YOLOv8-EGP and other mainstream algorithms using the FILIR dataset. The evaluation was based on three key metrics: precision, recall, and mAP@0.5. This approach allowed for a comprehensive assessment of the algorithm's capabilities in Table III.

To highlight the improvements brought by the YOLOv8-EGP model, it was rigorously compared with the original YOLOv8s and the larger, more accurate, albeit computationally intensive YOLOv8m models, all under comparable configurations. The results were striking: YOLOv8-EGP not only effortlessly outperformed YOLOv8s but also surpassed YOLOv8m with a remarkable 3.6% higher mAP, all while maintaining lower parameter and computational complexities. When compared to the YOLOv7-tiny algorithm in the same category, YOLOv8-EGP achieved an outstanding 12.3% higher mAP. Furthermore, in comparison to YOLOv5m, which shares a similar computational complexity, YOLOv8-EGP showcased a 5.5% higher mAP. These findings unequivocally establish that, among models in its category on infrared image datasets, YOLOv8-EGP stands out as a superior choice, making it exceptionally well-suited for real-world applications.

TABLE III
COMPARATIVE EXPERIMENT RESULTS OF DIFFERENT MAINSTREAM MODELS

Approaches	P(%)	R(%)	mAP@0.5(%)
YOLOv8s	84.0	68.2	76.8
YOLOv8m	84.7	71.5	79.3
YOLOv7-tiny	78.5	62.9	70.6
YOLOv5m	86.3	68.4	77.4
YOLOv8-EGP	85.6	74.0	82.9

IV. CONCLUSION

This paper presents a YOLOv8-EGP model algorithm for infrared target detection, which addresses various issues in the YOLOv8s model, such as low recognition rates, low accuracy, and missed detection of small targets in infrared images. The SCConv module enhances the receptive field of infrared targets, leading to improved feature extraction. The dyhead module achieves high-performance target detection with low computational cost, and experimental results show some improvements compared to the original model. Additionally, the addition of the additional small target

detection layer efficiently enhances the detection capability of small targets in infrared images. The original model suffers from low accuracy due to low pixel resolution and significant noise interference in infrared images, while the improved model shows a significant increase in accuracy in this regard. When compared to other models, it demonstrates similar or superior performance.

The experimental results also indicate that the YOLOv8-EGP model outperforms the original model by 6.1%. This enhancement has practical applications in fields such as vehicle-assisted driving, nighttime road recognition, and intelligent transportation.

REFERENCES

[1] Yuhang Bai, Zhengpeng Li, Jiansheng Wu, and Xinmiao Yu, "DUCAF-Net: An Object Detection Method for UAV Imagery," *Engineering Letters*, vol. 31, no.4, pp. 1374-1382, 2023.

[2] X. P. Chen, and Y. Xu, "A Multi-Dimensional Attention Feature Fusion Method for Pedestrian Re-identification," *Engineering Letters*, vol. 31, no.4, pp. 1365-1373, 2023.

[3] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580-587.

[4] R. Girshick, "Fast R-CNN," *IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec. 2015. pp. 1440**-**1448.

[5] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779-788.

[7] W. Liu et al., "SSD: Single Shot MultiBox Detector," in Computer Vision - ECCV 2016, Lecture Notes in Computer Science, 2016, pp. 21-37.

[8] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017.

[9] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*, Apr. 2018.

[10] Bochkovskiy A, Wang C Y, Liao H Y M. "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv preprint arXiv*:2004.10934, 2020.

[11] Li, Chuyi, et al. "YOLOv6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv*:2209.02976, 2022.

[12] C.-Y. Wang, A. Bochkovskiy, and H.-Y. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors." *arXiv preprint arXiv*, 2022.

[13] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned One-stage Object Detection," *IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021.

[14] Ge, Zheng, et al. "YOLOX: Exceeding YOLO Series in 2021," *arXiv preprint arXiv*:2107.08430, 2021.

[15] J.-J. Liu, Q. Hou, M.-M. Cheng, C. Wang, and J. Feng, "Improving Convolutional Networks With Self-Calibrated Convolutions," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020.

[16] X. Dai et al., "Dynamic Head: Unifying Object Detection Heads with Attentions," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 7373-7382.

[17] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *Journal of Machine Learning Research, Journal of Machine Learning Research*, Jan. 201.

[18] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets v2: More Deformable, Better Results." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 9308-9316.

[19] V. Venkataraman, G. Fan, and X. Fan, "Target Tracking with Online Feature Selection in FLIR Imagery," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, Jun. 2007, pp. 1-8.