RES-YOLO: A Real-Time Infrared Detection Framework for Intelligent Vehicle Traffic Monitoring

Junhao Dai and Kai Zhu *

Posted Date: 7 January 2026

doi: 10.20944/preprints202601.0503.v1

Keywords:
infrared thermal imaging; traffic object detection; deep learning; YOLOv8; receptive field adaptive convolution; efficient multi-scale attention

 doi:10.20944/preprints202601.0503.v1

Article

RES-YOLO: A Real-Time Infrared Detection Framework for Intelligent Vehicle Traffic Monitoring

Junhao Dai [1] and Kai Zhu [2,*]

[1] School of Mechanical Engineering, Jiangsu University of Technology, Changzhou 213001, China
[2] School of Automobile and Traffic Engineering, Jiangsu University of Technology, Changzhou 213001, China
* Correspondence: fatkyo@jsut.edu.cn

Abstract

Infrared traffic object detection faces challenges such as low resolution, weak thermal contrast, and inefficiency in detecting small objects. To address these issues, this paper proposes RES-YOLO, an enhanced YOLOv8n-based architecture. It incorporates Receptive Field Adaptive Convolution for improved multi-scale perception, Efficient Multi-scale Attention for better feature representation, and the Scylla-IoU loss for more accurate and faster bounding box regression. Additionally, a pseudo-color infrared dataset is constructed to enrich texture and contrast information beyond conventional white-hot images. Experiments on both the FLIR public dataset and a self-built dataset show RES-YOLO improves accuracy by 4.9% and 5.5% over the baseline while maintaining real-time performance. These results highlight the method's effectiveness in integrating lightweight deep learning and dataset enhancement for robust perception in intelligent vehicle systems, supporting AI-driven autonomous driving and driver assistance applications.

Keywords: infrared thermal imaging; traffic object detection; deep learning; YOLOv8; receptive field adaptive convolution; efficient multi-scale attention

1. Introduction

In road traffic systems, pedestrians and vehicles are the most critical subjects for monitoring, especially in the context of intelligent systems and autonomous vehicles, where reliable environment perception is crucial for AI-driven applications such as advanced driver assistance systems (ADAS) and autonomous driving functions. Under nighttime or low-visibility conditions, visible-light devices are often limited by poor illumination, making it difficult to capture accurate and reliable image information, thereby increasing the risk of traffic accidents. In this regard, infrared imaging technology, leveraging AI-based algorithms, offers a solution by capturing thermal radiation and providing stable visual input in no-light or low-light environments. This makes it particularly valuable for nighttime driving and intelligent systems in vehicles [1]. However, infrared images are often compromised by low resolution and weak contrast. Detecting small or distant objects in complex backgrounds with subtle thermal differences remains a significant challenge for AI models in real-world applications [2]. Addressing these challenges, this study explores novel AI-based solutions for improving infrared object detection, contributing to the advancement of intelligent vehicle technologies,

With the advancement of deep learning [3], methods based on neural network have largely supplanted traditional approaches in infrared object detection, offering superior feature extraction. Asad Ullah et al. employed Fast R-CNN for pedestrian detection [4], improving accuracy under low-contrast, cluttered backgrounds. Wang et al. proposed DNA-Net with a nested attention structure to enhance small-target retention [5]. Dai et al. proposed the Asymmetric Context Modulation module for multi-scale semantic fusion and boundary refinement [6]. Zhang et al. designed ISNet using Taylor differentials and bidirectional attention for shape modeling [7], while Liu et al. integrated image enhancement with Mask R-CNN to improve both accuracy and robustness [8].

 doi:10.20944/preprints202601.0503.v1

Despite these advances, most remain two-stage frameworks with high computational cost, limiting deployment on platforms with constrained resources. Contrastingly, one-stage detectors such as YOLO [9] provide higher efficiency. YOLOv3/v4 [10,11] maintained a compromise between accuracy and speed, while lightweight YOLOv8n [12] further reduces computational load and serves as a strong baseline for infrared detection. Lin et al. enhanced YOLOv4 [13] with multiscale fusion and attention for detecting dim and small targets. Li et al. introduced YOLO-FIRI [14], strengthening shallow feature extraction for traffic targets. Zhang et al. put forward EGISD-YOLO [15], an edge-guided dual-branch network preserving contours in maritime scenes, and Tang et al. developed IRSTD-YOLO [16], integrating edge-awareness and shallow feature enhancement to surpass YOLOv11-s on infrared UAV datasets.

Meanwhile, the issue of insufficient information dimensions in infrared images urgently needs to be addressed. Existing mainstream datasets, such as FLIR ADAS, are mostly presented in the form of white-hot images [17], where the contrast between targets and the background is limited, making it difficult to support high-precision detection. Therefore, this paper constructs a pseudo-color infrared image dataset under urban and rural road scenarios in China. By mapping grayscale values to the RGB channels, the image texture and edge information are enhanced, which not only improves the distinguishability and visibility of the targets but also promotes faster model convergence and better generalization capability.

Although incorporating pseudo-color images can bring certain performance improvements, existing detection models still face numerous challenges in infrared scenarios: such as insufficient detection accuracy for small or distant targets, poor discrimination ability in densely packed target scenes, and limited inference efficiency due to complex structures [18]. To approach these issues, this paper presents an upgraded infrared target detection algorithm which is based on the YOLOv8n baseline model, RES-YOLO, with optimizations in the following three aspects:

- Introduce Receptive Field Adaptive Convolution (RFAConv) module [19], which dynamically adjusts the receptive field to optimize the perception capability for targets of different scales;
- Integrate Efficient Multi-scale Attention mechanism (EMA) [20] into the backbone to improve the model's saliency modeling ability under complex backgrounds;
- Substitute the SIoU [21] loss function for the CIoU [22], introducing an angle penalty term to optimize bounding box localization, thus enhancing the regression accuracy on dense targets and accelerating training convergence.

We conduct experiments on the publicly available FLIR ADAS dataset as well as on the self-constructed pseudo-color infrared image dataset to validate the performance of RES-YOLO. Experimental analysis shows that the improved model delivers better outcomes than the original YOLOv8n across metrics including mAP, Precision, and Recall, while maintaining its lightweight characteristics in inference speed.

2. Materials and Methods

This chapter introduces the overall methodology adopted in this study. It begins with a brief overview of the YOLOv8n base network model to lay the foundation for subsequent improvements. Following that, a detailed description of the overall architecture and key module designs of the proposed RES-YOLO model for infrared image traffic target detection is presented.

2.1. YOLOv8n Network Model

YOLOv8n is the lightweight version within the YOLOv8 series released by Ultralytics. As a unified single-stage object detection framework, it significantly minimizes both model complexity and computational cost while maintaining high detection accuracy. Such characteristics make it ideal for applications in in-vehicle infrared detection, which demand real-time performance and have limited computational resources. Compared to earlier versions like YOLOv5s [23] and YOLOv6n [24], YOLOv8n has optimized its network architecture design, prediction strategies, and training

 doi:10.20944/preprints202601.0503.v1

mechanisms, further enhancing deployment flexibility and training stability without compromising detection performance.

Figure 1 demonstrates the YOLOv8n structure, consisting of four essential modules: input, backbone, feature fusion, and detection head. The input images are scaled to 640×640 pixels and normalized prior to being fed into the network. In the backbone, YOLOv8n employs a lightweight feature extraction module based on the C2f (Cross Stage Partial with concatenation) structure to extract semantic features at different scales. Compared to the CSPDarknet structure used in YOLOv5, the C2f module effectively reduces the number of model parameters while retaining feature representation capability, thereby improving training stability and convergence speed.

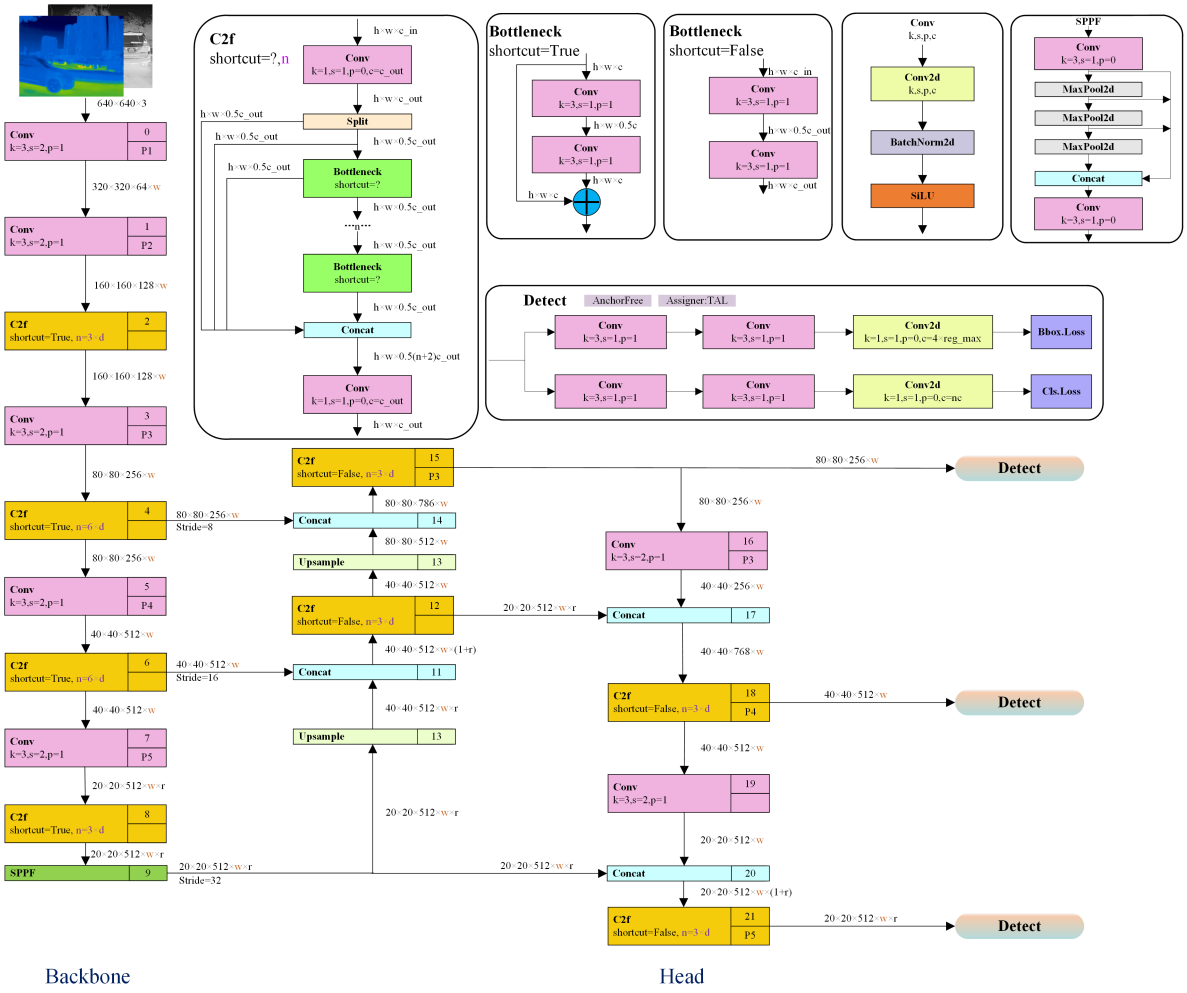


Figure 1. YOLOv8n network Structure.

In the feature fusion stage, YOLOv8n adopts a structure similar to PAN-FPN [25], which performs multi-scale fusion of feature maps from different hierarchical levels, thereby enhancing the detection capability for targets of varying sizes. Upsampling combined with lateral connections allows deeper semantic information to merge effectively with finer texture details. This design is particularly suitable for detecting targets in infrared images where edges are often blurred and thermal contrast is weak.

The detection head employs a decoupled structure, separating the classification and regression branches to make independent predictions. This approach avoids mutual interference between the two tasks and improves detection accuracy and robustness. The final outputs include the target class confidence, object presence probability, and bounding box location parameters. By default, YOLOv8n employs an integrated loss that simultaneously optimizes the losses for bounding box regression, object confidence, and class prediction.

 doi:10.20944/preprints202601.0503.v1

In terms of training strategies, YOLOv8n incorporates several advanced mechanisms, including an anchor-free prediction scheme and Mosaic data augmentation. Without using predefined anchors, the model's prediction process becomes simpler and is less prone to anchor-related inconsistencies. This approach helps enhance the model's adaptability in recognizing small and remote instances.

Although YOLOv8n already has relatively low computational complexity, it still faces challenges such as high missed detection rates and insufficient localization accuracy in scenarios with low-resolution infrared images, weak thermal contrast, and densely packed targets. Therefore, this paper selects YOLOv8n as the baseline model and further improves its architecture to better accommodate pseudo-color infrared image inputs.

2.2. RES-YOLO for Infrared Object Detection

RES-YOLO serves as the foundation for a pseudo-color infrared detection approach tailored to traffic scenarios with poor visibility, enabling robust recognition of multiple objects. It focuses on optimizations in three core aspects: receptive field adjustment, feature attention, and localization accuracy. Key modifications are applied to the convolutional components, attention modules, and the loss function used for bounding box regression, with the goal of improving both accuracy and robustness in detecting pseudo-color infrared images.

2.2.1. General Framework

RES-YOLO network architecture is illustrated in Figure 2. First, during the feature extraction stage, a receptive field adaptive convolution module is employed in place of the original convolutional layers, effectively strengthening the network's capacity to perceive targets at varying scales. Second, to improve feature focusing, a multi-scale cross-spatial attention mechanism is incorporated into the Neck and Head modules, enabling the model to more accurately capture target regions under low-contrast backgrounds. Finally, the Scylla-IoU (SIoU) is introduced to elevate the model's localization accuracy and convergence speed in dense target scenarios, thereby strengthening overall detection performance.

2.2.2. Receptive Field Adaptive Convolution Optimization

In traditional convolution operations, the convolution kernel shares the same parameters across each receptive field region. Although this parameter-sharing mechanism effectively reduces the model's computational complexity, it also limits the ability of convolution to express differences in features at different spatial locations, thereby constraining further improvements in model performance. To mitigate this problem, RFAConv is incorporated within the YOLOv8n backbone, substituting certain conventional convolutional layers and strengthening the network's sensitivity to local feature variations.

RFAConv achieves this by incorporating receptive-field spatial features, allowing each convolution sliding window region to obtain independent attention weights, which dynamically generate non-shared convolution kernel parameters for each sliding window. This mechanism fundamentally breaks the limitation of traditional convolution, where fixed kernel parameters are shared across all sliding windows. Unlike typical spatial attention mechanisms exemplified by CBAM [26] and CA [27], which assign uniform weights across the entire feature map, RFAConv generates independent attention weights at the sliding window level, effectively improving the model's ability to capture fine-grained target differences. This is especially suitable for infrared image target detection tasks with blurred object boundaries and large scale variations. The structural design of RFAConv module is shown in Figure 3. Initially, the input feature map undergoes Group Convolution to rapidly capture receptive field spatial information. Compared to Unfold operation, group convolution effectively reduces computational overhead while maintaining feature integrity. Next, the module obtains global contextual information through global average pooling and computes attention maps for each receptive field region via convolution. Then, the attention maps are multiplied element-wise with the extracted

 doi:10.20944/preprints202601.0503.v1

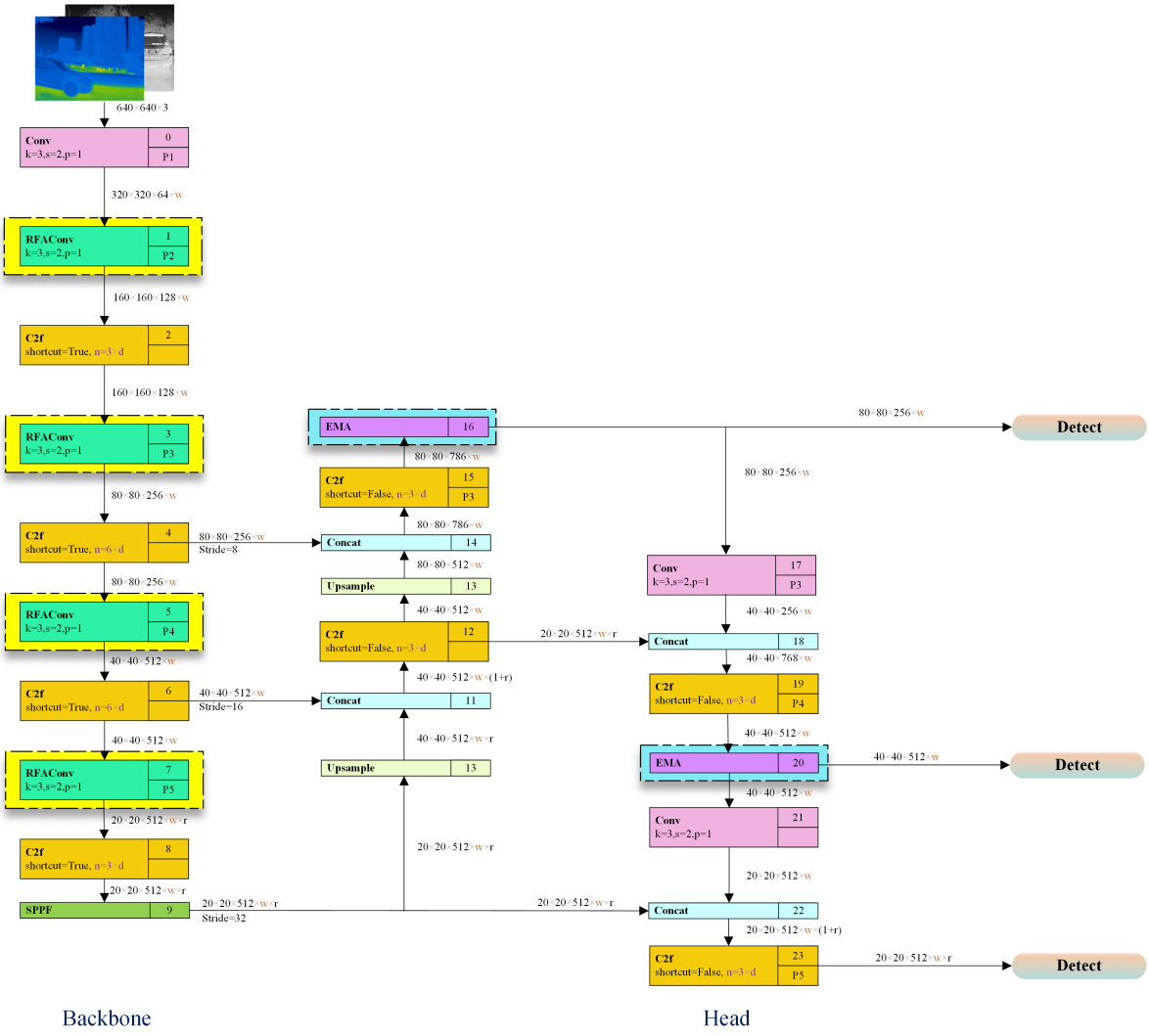


Figure 2. Res-YOLO network Structure.

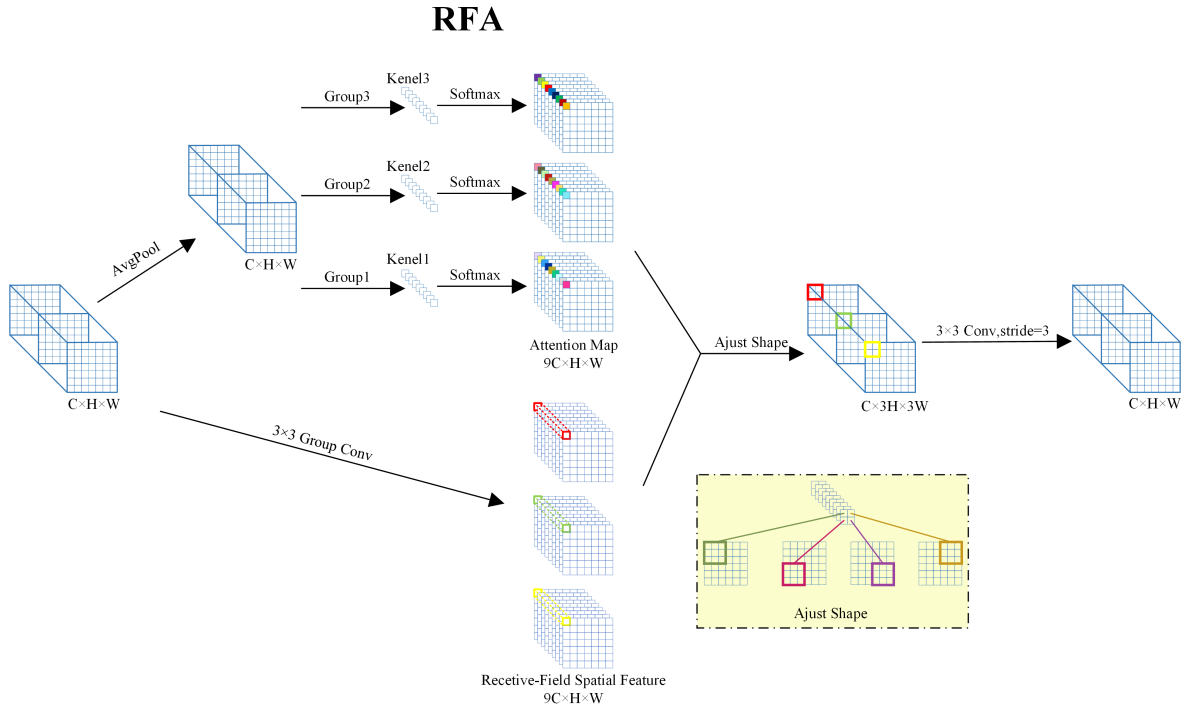


Figure 3. Network structure of the Receptive-Field Adaptive Convolution module.

 doi:10.20944/preprints202601.0503.v1

receptive field features to achieve dynamic adjustment of convolution weights. The entire procedure can be summarized in equation (1):

$$F = Softmax\left(g^{1\times1}(AvgPool(X))\right) \times ReLU\left(Norm\left(g^{k\times k}(X)\right)\right) \tag{1}$$

where $g^{1\times1}$ and $g^k \times k$ represent the 1×1 and $k \times k$ group convolution operations, respectively; X indicates the input features; F corresponds to the fused result; and denotes the normalization operation.

The introduction of RFAConv effectively alleviates the issue of indistinct features in infrared images caused by small temperature differences between targets and the background. By adaptively adjusting the convolution kernel's focus areas, the model gains stronger feature representation capability without significantly increasing computational overhead.

2.2.3. Efficient Multi-scale Attention Mechanism Optimization

In infrared images, due to the subtle differences in thermal radiation between the background and targets, the model is easily disturbed by redundant background information when extracting regional features. Therefore, this paper introduces EMA into the YOLOv8n network architecture. Through modeling features along channel and spatial dimensions, the model enhances responsiveness to key regions while maintaining computational efficiency.

Figure 4 illustrates the overall structure of the EMA module. The input feature map X first passes through the channel attention branch to generate a channel weight vector, then through the spatial attention branch to produce direction-sensitive spatial weight maps. Finally, these two are fused to attain the weighted output.

The final result generated by the EMA module can be expressed as:

$$Y = X \otimes C(X) \otimes S(X) \tag{2}$$

where $\otimes$ is defined as element-wise multiplication; $C(X)$ represents the channel attention weight map; and $S(X)$ denotes the two-dimensional spatial weight map that fuses direction-aware spatial attention.

The EMA module aims to simultaneously enhance the semantic representation and spatial localization capabilities of feature maps by jointly modeling channel attention and direction-aware spatial attention, thereby reinforcing both "what" the important features are and "where" they appear. Along the channel dimension, EMA uses global average pooling to compress semantic information and employs lightweight one-dimensional convolutions to efficiently capture inter-channel dependencies, reducing computational complexity while preserving nonlinear expressiveness. Along the spatial dimension, considering that targets in infrared images often distribute along horizontal or vertical directions, EMA performs separate pooling operations along these directions to extract direction-aware contextual structural information, which is then fused to generate refined spatial weight maps. This improves sensitivity to target structures, especially in complex scenarios with blurred contours and weak thermal contrasts.

 doi:10.20944/preprints202601.0503.v1

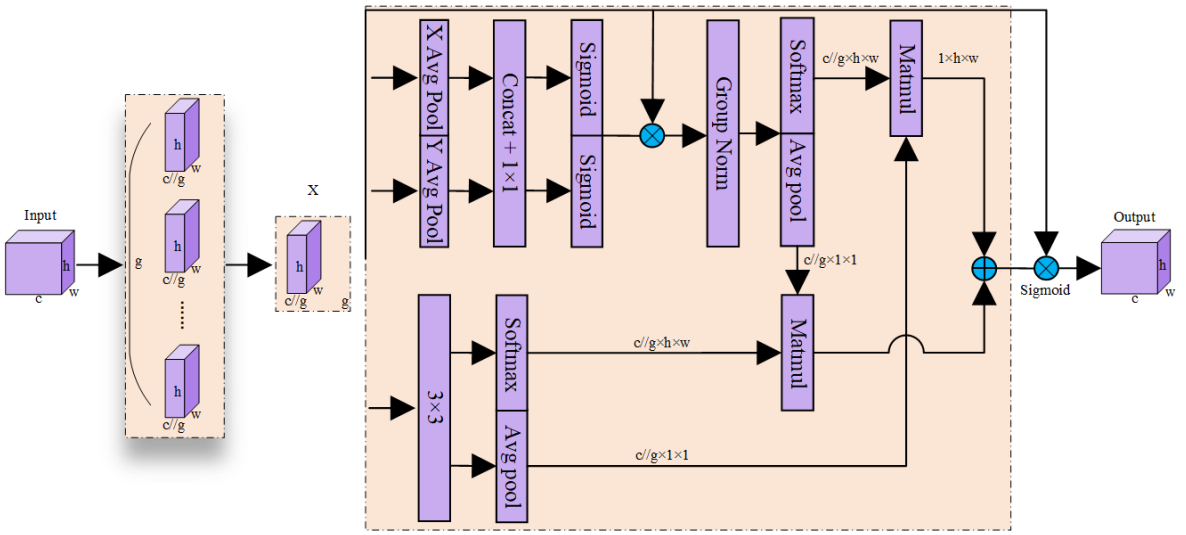


Figure 4. Structure of the EMA module. "g" denotes the number of groups, while "X Avg Pool" and "Y Avg Pool" represent one-dimensional global average pooling operations in the horizontal and vertical directions, respectively.

2.2.4. Loss Function Optimization

In pseudo-color infrared images, targets generally exhibit characteristics such as blurred edges, irregular sizes, and dense spatial distribution, which make bounding box localization prone to prediction offsets, drift, and slow training convergence during the regression process. Although the CIoU loss function used by default in YOLOv8n integrates center point distance, overlap area and aspect ratio information to some extent, it lacks finer modeling of the target's motion direction and spatial structure, resulting in suboptimal performance in scenarios involving complex target deformation or angle variations.

To address this issue, this paper introduces Scylla-IoU (SIoU) as the bounding box regression loss function, aiming to comprehensively describe the differences between predicted regions and ground-truth annotations from a spatial perspective. SIoU divides the bounding box regression process into four components: angle penalty, center distance, shape matching and area overlap, and constructs its total loss function as a weighted sum. The overall SIoU formula is shown in equation (3):

$$L_{SIoU} = L_{angle} + L_{distance} + L_{shape} + L_{IoU} \tag{3}$$

Figure 5 illustrates the contribution of calculation diagram of the angle cost to the loss function. It can be seen that SIoU no longer measures regression quality solely by area overlap but incorporates geometric direction, target shape, and positional relationships into the bounding box localization framework. This makes the optimization direction of the predicted box clearer and accelerates convergence speed.

First, the angle loss term L_{angle} penalizes angular deviation according to the cosine of the vector connecting the centers of predicted and ground-truth boxes. Its formula is:

$$L_{\text{angle}} = 1 - \cos(\theta) \tag{4}$$

where θ is the angle formed by the line joining the centroids of the predicted and ground-truth boxes with respect to the horizontal axis. When the predicted direction is close to the true direction, this term approaches zero; otherwise, it increases rapidly, effectively preventing directional drift of the predicted box during early training.

 doi:10.20944/preprints202601.0503.v1

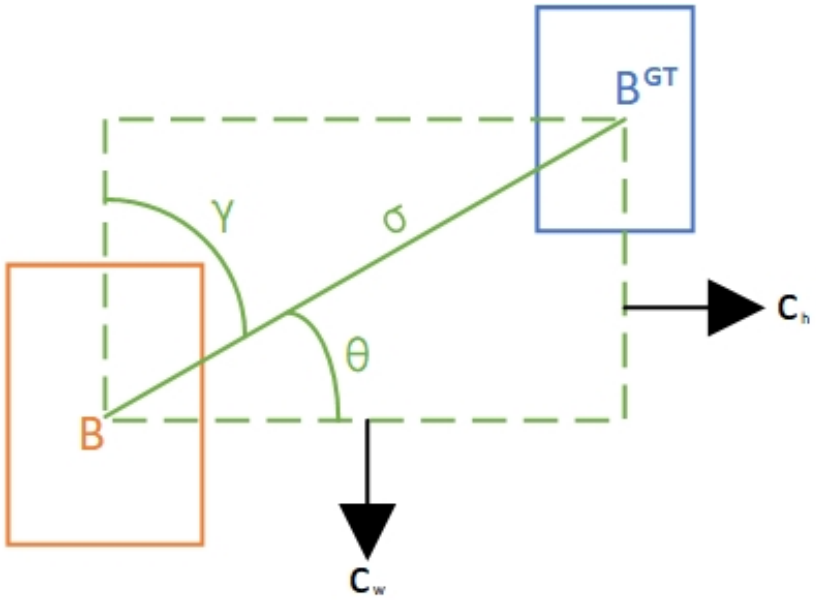


Figure 5. Schematic diagram of angle cost calculation in SIoU.

To further constrain the degree of center point offset between the predicted box and the target, SIoU employs an $L_{distance}$ loss term. This term normalizes the Euclidean distance between the centers, making it suitable for targets of different scales. The calculation is shown in equation (5):

$$L_{distance} = \frac{(\Delta x)^2 + (\Delta y)^2}{w_{gt}^2 + h_{gt}^2} \tag{5}$$

where Δx and Δy describe the positional shift along horizontal and vertical axes between the predicted and annotated box centers, while w_{gt} and h_{gt} specify the box's geometric dimensions.

In the shape loss component, SIoU considers how well the predicted box fits the true target shape along the width and height dimensions, and introduces an exponential decay function to enhance the model's adaptability to bounding box shapes. The formula is as follows:

$$L_{shape} = 1 - e^{\frac{|w_p - w_{gt}|}{w_{gt}}} + 1 - e^{\frac{|h_p - h_{gt}|}{h_{gt}}} \tag{6}$$

where w_p refers to the width and h_p to the height of the predicted bounding box. The introduction of the exponential term ensures that the smaller the shape difference between boxes, the faster the loss approaches zero, which helps the model converge quickly in terms of scale fitting.

SIoU retains the IoU loss term for evaluating the intersection area shared by the predicted box and the ground-truth box:

$$L_{IoU} = 1 - \frac{|B_p \cap B_{gt}|}{|B_p \cup B_{gt}|} \tag{7}$$

3. Experimental Dataset and Preprocessing

3.1. Evaluation Metrics

To thoroughly assess the detection capability of the model on traffic objects within vehicle-mounted infrared imagery, this study employs evaluation indicators including Precision (P), Recall (R), mean Average Precision (mAP), and Frames Per Second (FPS).

Precision represents the ratio of correctly predicted positives to the total predicted positives, whereas Recall reflects the proportion of true positives identified with respect to the total ground-truth positives. Their formal definitions are provided in equation (8) and (9):

 doi:10.20944/preprints202601.0503.v1

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

here, TP refers to correctly detected positive instances, FP indicates false alarms where negative cases are misclassified as positive, and FN corresponds to positives that fail to be recognized by the detector.

Average Precision (AP) measures the area under the Precision-Recall curve plotted at different confidence thresholds for each target category, reflecting the accuracy and completeness of detection for that category. The mean of AP values across multiple categories is called mean Average Precision, defined as:

$$mAP = \frac{1}{N} \sum_{j=1}^{N} AP_j \tag{10}$$

where N represents the class total, with AP_j denoting the mean precision for the j^{th} category.

FPS evaluates the inference efficiency of an object detector, reflecting how many image frames can be handled within one second. This metric reflects the model's real-time capability; a value above 50 is considered sufficient for real-time detection. Its formula is as follows:

$$FPS = \frac{M}{T} \tag{11}$$

where M denotes the test frame quantity, whereas T refers to the processing time consumed.

3.2. Experimental settings

This section outlines the setup of the experiments and summarizes the model configuration in the accompanying Table 1.

Table 1. Experimental environment and training parameter settings.

Project	Environment	Parameters	Value
System	Windows 11	Epochs	300
CPU	Intel-core i7-13700KF	Learning rate	0.01
GPU	NVIDIA GeForce RTX 4070	Momentum	0.937
Memory	32G	Optimizer	SGD
Pytorch version	2.1.2	Batch size	16
CUDA	12.6	Image size	640

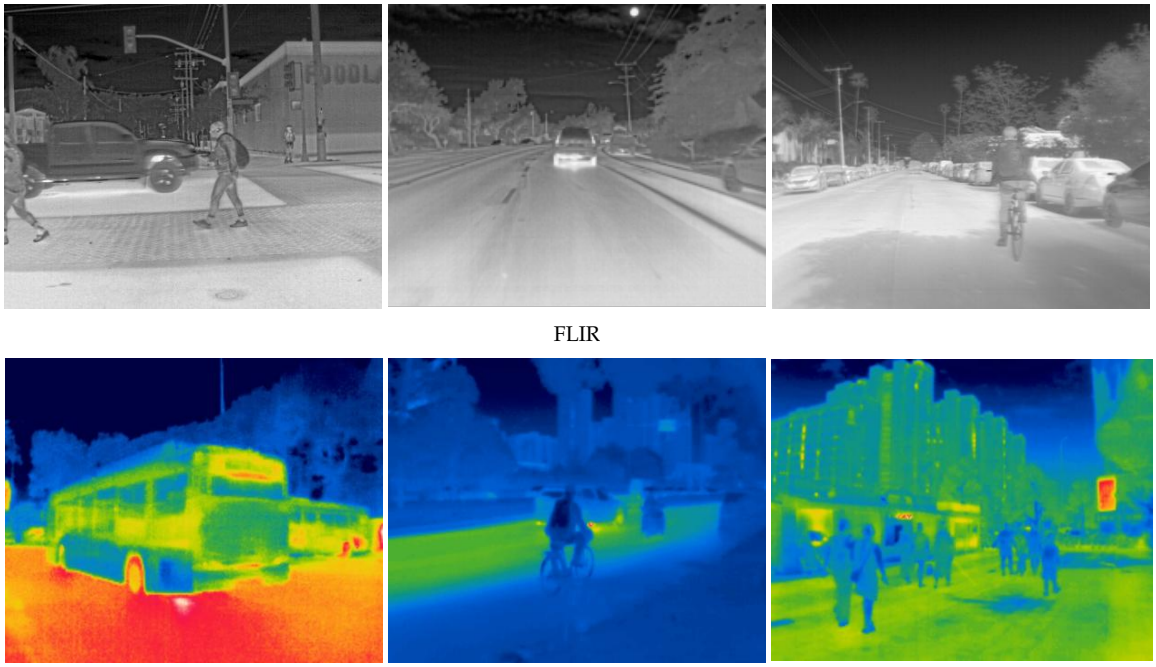
3.3. Datasets

This experiment utilizes two datasets. The first is the public FLIR ADAS dataset, consisting of 7,381 training and 3,167 validation grayscale infrared images covering Pedestrian, Bicycle, and Car categories across urban, rural, and highway scenes with varying weather conditions. The second is our self-collected pseudo-color infrared dataset from Chinese urban and rural roads, which includes more categories—Bicycle, Bus, Car, Pedestrian, and Truck—with 6,741 training and 2,892 validation images. To satisfy deep learning requirements for large-scale data and enhance model generalization, various data augmentation techniques such as random scaling, flipping, brightness adjustment, and Mosaic stitching were applied during training to improve image diversity and robustness. Sample images are shown in Figure 6.

In infrared traffic target detection, pseudo-color images offer clear advantages over grayscale images. By mapping grayscale values to colors, they enhance contrast between targets and background, improving the clarity of vehicle contours, road edges, and pedestrian shapes for easier identification and localization. In complex scenes with occlusion and background noise, grayscale images' limited

 doi:10.20944/preprints202601.0503.v1

information hinders accuracy, while pseudo-color images' rich color and texture features help models better detect obscured or blurred targets, boosting robustness and precision. Moreover, the additional feature dimensions improve generalization across varied conditions.



FLIR

Our dataset

Figure 6. Sample images from the FLIR dataset and our self-collected dataset.

4. Results and Analysis

To assess the performance of the RES-YOLO framework, comparative experiments were conducted against widely used object detection methods. Component-wise ablation analyses were carried out to examine the individual contributions of the convolutional module enhancements, attention mechanism refinements, and improvements in the loss function. Model training was performed using both the FLIR dataset and a self-constructed dataset. Throughout all tests, factors such as the hardware configuration, training parameters, and iteration counts were maintained consistently.

4.1. Experimental Results

4.1.1. Validation for Receptive Field Adaptive Convolution Optimization

Receptive Field Adaptive Convolution dynamically modifies the receptive field by applying location-specific convolutional kernel weights, which strengthens the network's capacity to capture indistinct boundaries and long-range targets. Keeping other parameters unchanged, evaluation trials were carried out using both the FLIR benchmark dataset and our self-collected pseudo-color infrared dataset to demonstrate the utility of this component.

The experimental results in Table 2 show that the introduction of the RFAConv module led to performance improvements on both datasets. On the FLIR dataset, the model's mAP increased from 83.89% to 85.56%, precision rose to 86.12%, and recall saw a slight improvement. On our self-collected dataset, mAP improved from 88.17% to 88.95%, precision increased to 89.19%, and recall also improved. Although the performance gains were modest, the module consistently delivered stable improvements across multiple scenarios, demonstrating its effectiveness in enhancing the model's perception of fine-grained targets.

 doi:10.20944/preprints202601.0503.v1

Table 2. Experimental Analysis of introducing Receptive Field Attention Convolution.

Dataset	Methods	mAP/%	P/%	R/%
FLIR	YOLOv8n	83.89	85.28	79.60
	YOLOv8n+RFAConv	**85.56**	**86.12**	**80.75**
Ours	YOLOv8n	88.17	88.34	83.66
	YOLOv8n+RFAConv	**88.95**	**89.19**	**84.79**

4.1.2. Validation for Efficient Multi-Scale Attention Mechanism Optimization

EMA module guides the network to focus on more discriminative regions. By applying directional spatial pooling operations, this module enhances the response of salient areas, effectively highlighting the differences between targets and background in feature representation.

As shown in Table 3, the EMA module brought consistent performance improvements across both datasets. Notably, precision surpassed 90% on our self-collected pseudo-color dataset. Precision and recall on the FLIR dataset also improved compared to the baseline model, further confirming the module's positive impact on target modeling in complex backgrounds. The EMA module significantly enhanced the model's responsiveness to key regions in infrared images, strengthening robustness in detecting edge and boundary targets.

Table 3. Experimental Analysis of introducing Efficient Multi-scale Attention.

Dataset	Methods	mAP/%	P/%	R/%
FLIR	YOLOv8n	83.89	85.28	79.60
	YOLOv8n+EMA	**84.72**	**87.19**	**80.67**
Ours	YOLOv8n	88.17	88.34	83.66
	YOLOv8n+EMA	**89.13**	**90.26**	**84.52**

4.1.3. Validation for Loss Function Optimization

Bounding box regression is a fundamental component in target detection, influencing both the precision of object localization and the convergence rate during training. To enhance localization capabilities in scenarios with densely packed targets, the conventional CIoU loss used in YOLOv8n has been substituted with SIoU, which offers a more thorough representation of the geometric relationships among bounding boxes.

From the data in Table 4, after incorporating SIoU, mAP on the FLIR dataset increased from 83.89% to 87.45%, and recall improved from 79.60% to 85.18%. On our self-collected pseudo-color dataset, mAP rose to 91.83%, with recall reaching 89.82%. Although precision saw little change, the significant increase in recall indicates that SIoU enhanced the model's control over bounding box regression quality, effectively mitigating overlapping regression errors in dense target distributions.

Table 4. Experimental Analysis of introducing SIoU loss function.

Dataset	Methods	mAP/%	P/%	R/%
FLIR	YOLOv8n	83.89	85.28	79.60
	YOLOv8n+SIoU	**87.45**	**88.27**	**85.18**
Ours	YOLOv8n	88.17	88.34	83.66
	YOLOv8n+SIoU	**91.83**	**92.76**	**89.82**

 doi:10.20944/preprints202601.0503.v1

4.1.4. Ablation Experiment

The individual contributions of each module to detection performance were experimentally verified. Systematic ablation studies integrating these modules into YOLOv8n evaluated their combined effects. As shown in Table 5, the three modules significantly improved mAP across datasets, proving their effectiveness in infrared object detection. RFAConv enhances low-level perception, EMA boosts mid-level feature representation, and SIoU refines high-level decision-making, collectively addressing shortcomings in feature extraction, target focus, and bounding box regression. While keeping the model lightweight, their combination improves detection accuracy and localization precision. The final RES-YOLO model achieves a 5.5% Precision improvement over the YOLOv8n baseline on our dataset while maintaining an inference speed of 305 FPS, demonstrating its capability to meet the real-time perception requirements of intelligent electric vehicle systems in infrared traffic scenarios.

Table 5. Results of ablation experiment.

Dataset	RFAConv	EMA	SIoU	mAP/%	FPS
FLIR				83.89	311
	✓	✓		86.38	308
	✓		✓	87.55	306
		✓	✓	86.91	308
	✓	✓	✓	**87.98**	305
Ours				88.17	311
	✓	✓		89.42	308
	✓		✓	92.06	306
		✓	✓	91.51	308
	✓	✓	✓	**93.04**	305

Figure 7 shows RES-YOLO achieves high main diagonal values across categories, indicating accurate classification. Misclassifications mostly occur between “Bicycle” and “Pedestrian” due to similar shapes and thermal features, a common real-world challenge. Overall, errors are limited with no severe confusion, demonstrating strong generalization and discrimination.

Figure 8 displays smooth PR curves, showing effective reduction of false positives and redundant detections. For recall above 0.85, some categories maintain precision over 0.85, highlighting the positive impact of SIoU and EMA on dense target detection accuracy.

 doi:10.20944/preprints202601.0503.v1

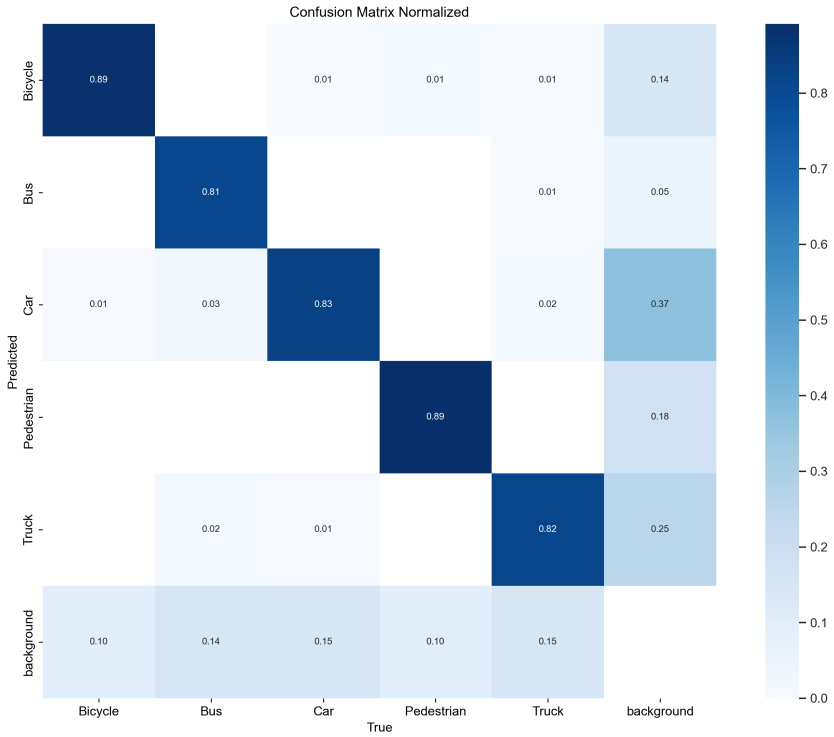


Figure 7. Confusion matrix of RES-YOLO on our dataset.

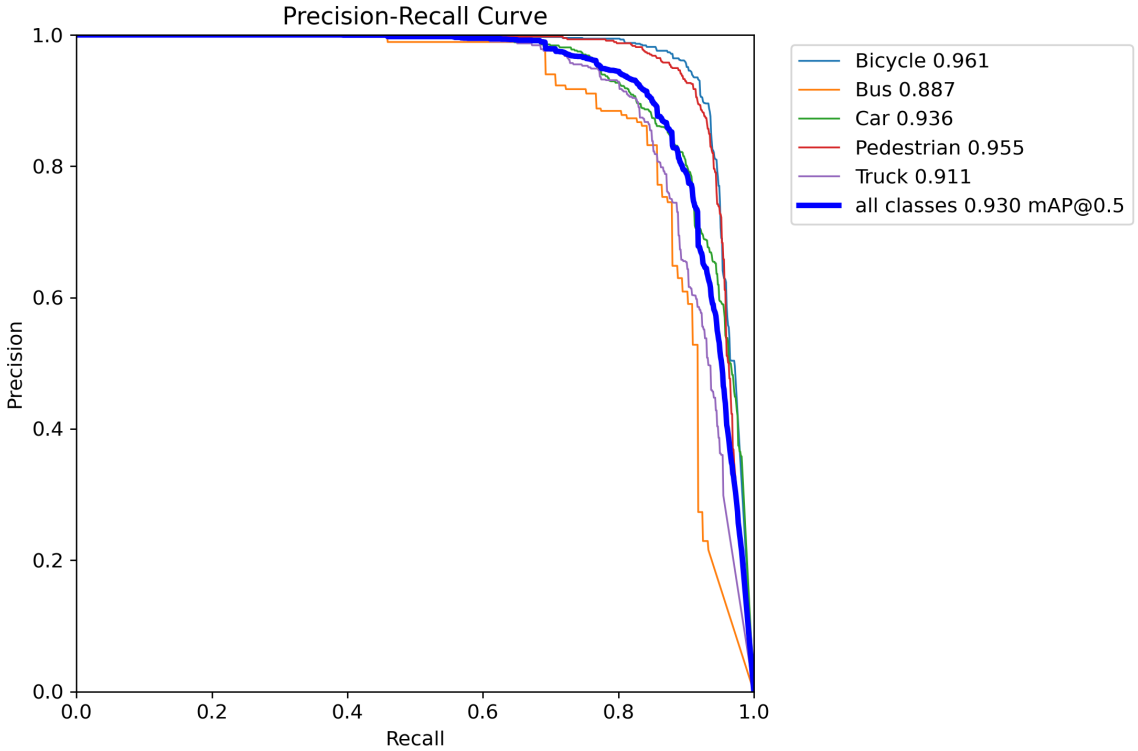


Figure 8. Precision-recall of RES-YOLO on our dataset.

4.2. Performance Comparison with Mainstream Network Models

To further evaluate the performance of the RES-YOLO model in infrared object detection, this study conducted systematic comparisons with mainstream YOLO series models, including YOLOv6s, YOLOv8n, YOLOv8l, YOLOv9t [28], YOLOv10n [29], and YOLOv11n [30]. The results for mAP, Precision, Recall, and FPS are summarized in Table 6. RES-YOLO achieved the highest mAP values on

 doi:10.20944/preprints202601.0503.v1

both datasets, reaching 87.98% and 93.04%, respectively. It also maintained consistently high Precision and Recall scores. Notably, its performance advantage was more pronounced on the self-collected pseudo-color dataset, where enhanced texture details contributed to a comprehensive lead in detection accuracy.

Table 6. Comparison results of mainstream models.

Dataset	Methods	mAP/%	P/%	R/%	FPS
FLIR	YOLOv6s	84.19	84.06	79.28	333
	YOLOv8l	84.60	85.43	78.93	303
	YOLOv8n	83.89	85.28	79.60	311
	YOLOv9t	84.49	84.33	80.12	116
	YOLOv10n	85.02	85.80	79.88	201
	YOLOv11n	85.34	85.11	81.23	231
	RES-YOLO	**87.98**	**87.27**	**82.97**	**305**
Ours	YOLOv6s	88.02	87.43	82.57	333
	YOLOv8l	88.12	87.56	82.44	303
	YOLOv8n	88.17	88.12	83.68	311
	YOLOv9t	88.86	88.21	84.01	116
	YOLOv10n	89.24	89.01	84.13	201
	YOLOv11n	89.33	88.74	85.19	231
	RES-YOLO	**93.04**	**92.28**	**87.97**	**305**

Figure 9 shows that on the FLIR dataset, the mAP differences among models are relatively small, and the Precision and Recall curves exhibit greater fluctuations. This indicates that in standard grayscale thermal image scenarios, most models' performance tends to converge, leaving limited room for marginal improvements through structural optimizations. In contrast, on our self-collected pseudo-color dataset, the differences in model capabilities are more pronounced, highlighting the positive impact of pseudo-color images in enhancing target distinction and reducing background interference. The smaller fluctuations in the Precision curve on our dataset also indirectly confirm that colored thermal images provide more consistent feature distributions, helping models maintain stable detection performance in complex scenarios.

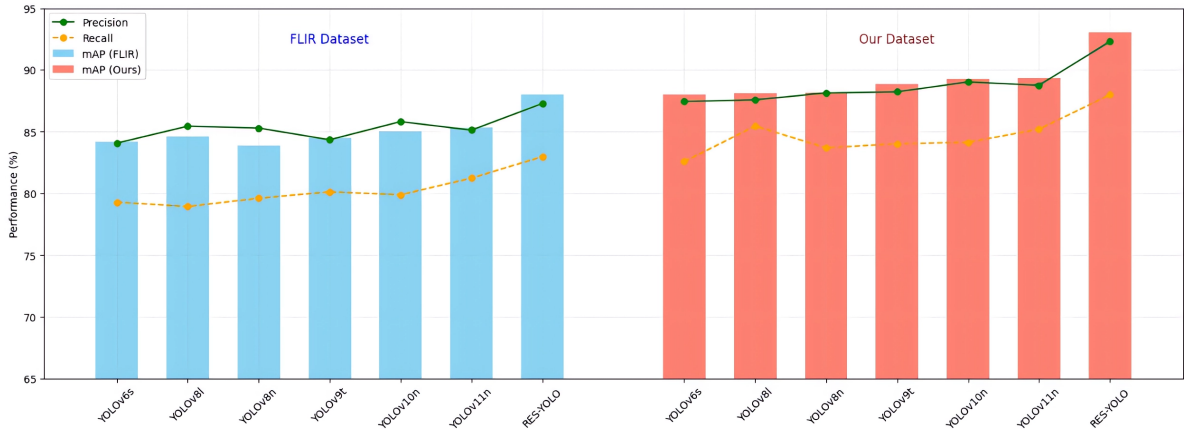


Figure 9. Comparison of mAP, Precision, and Recall across Models on FLIR and Our Dataset.

4.3. Comparison and Generalization Experiments

Using trained weights, inference was run on test images and results visualized. Figure 10 shows that in complex scenes with small targets, weak thermal contrast, or occlusion, YOLOv8n often yields

 doi:10.20944/preprints202601.0503.v1

false positives and misses, especially at long range or low contrast. RES-YOLO, however, accurately detects pedestrians, vehicles, and bicycles, demonstrating better robustness and generalization. This improvement stems from the RFAConv and EMA modules enhancing fine-grained features, along with the richer pseudo-color data providing more detailed edge, texture, and thermal information. Together, these factors enable better target learning and background suppression, resulting in more stable and precise detection performance.

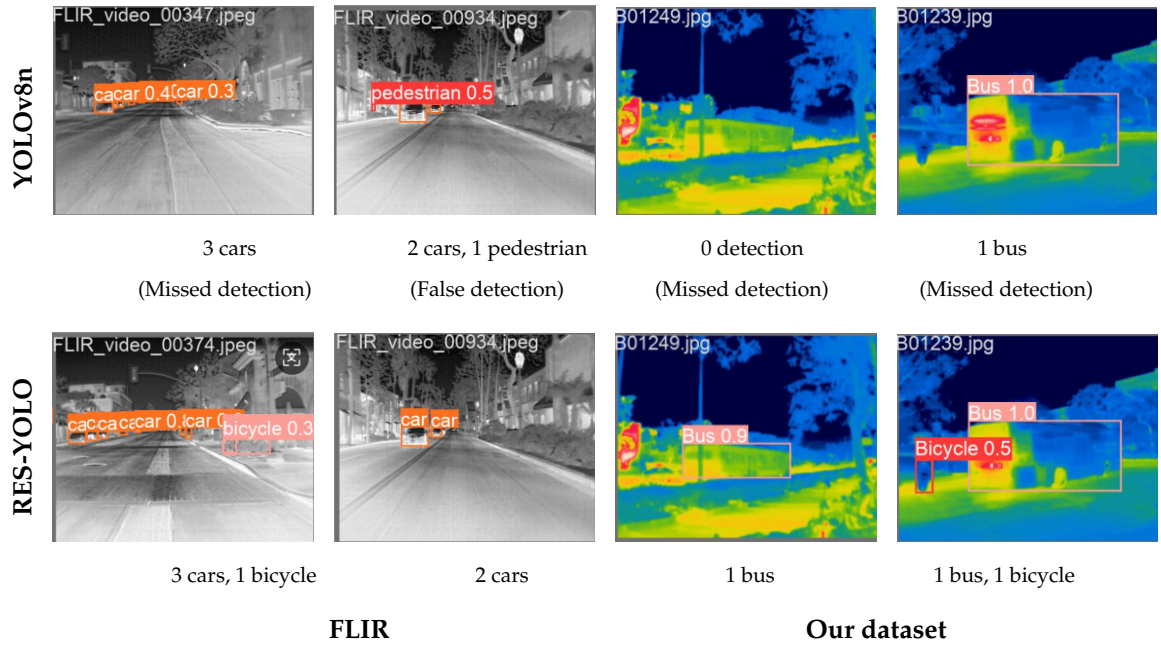


Figure 10. Comparison of detection performance on FLIR and Our dataset.

5. Conclusions

In this work, we addressed the challenges of infrared traffic object detection, including low image resolution, weak thermal contrast, and poor robustness in detecting small and long-range targets. To overcome these issues, we proposed an enhanced detection framework, RES-YOLO, based on the lightweight YOLOv8n model. The framework incorporates AI-driven techniques such as Receptive Field Adaptive Convolution (RFAConv) to dynamically adjust the receptive field for improved multi-scale perception, integrates Efficient Multi-scale Attention (EMA) to strengthen feature representation in complex backgrounds, and employs the Scylla-IoU (SIoU) loss to optimize bounding box regression, improving localization accuracy and accelerating convergence.

In addition to the architectural modifications, this study also developed a pseudo-color infrared image dataset, enhancing the image's dimensionality by incorporating richer texture and contrast features compared to conventional white-hot images. The dataset effectively improves the distinguishability of distant and small targets, providing a more reliable benchmark for evaluating infrared detection algorithms. Extensive experiments on both the FLIR dataset and a self-constructed dataset demonstrated that RES-YOLO consistently outperforms the original YOLOv8n baseline, achieving accuracy gains of 4.9% and 5.5%, respectively, while maintaining real-time inference capability. The results highlight the potential of AI-based models, particularly those utilizing receptive field adjustment, efficient attention mechanisms, and optimized loss functions, to enhance the performance of lightweight detection models in complex infrared scenarios.

From the perspective of AI applications in intelligent systems, this method offers a robust and efficient solution for traffic object detection under low-visibility conditions, which is crucial for enhancing the environmental perception capabilities of intelligent transportation systems and autonomous vehicles. By improving the accuracy of infrared object detection while preserving real-time performance,

 doi:10.20944/preprints202601.0503.v1

RES-YOLO supports the advancement of AI-driven perception frameworks, paving the way for more reliable and efficient intelligent systems.

Author Contributions: Conceptualization, J.D. and K.Z.; Methodology, J.D. and K.Z.; Software, J.D.; Validation, J.D.; Formal analysis, J.D.; Investigation, J.D.; Resources, J.D.; Data curation, J.D.; Writing—original draft, J.D.; Writing—review & editing, K.Z.; Supervision, K.Z.; Funding acquisition, K.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Changzhou Science & Technology Program (grant number CJ20240036).

Data Availability Statement: The study used open data, and the self-constructed data is included in the article. Due to confidentiality requirements related to the data and the project, the dataset cannot be made publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hou, F.; Zhang, Y.; Zhou, Y.; Zhang, M.; Lv, B.; Wu, J. Review on infrared imaging technology. *Sustainability* **2022**, *14*, 11161.
2. Ibarra-Castanedo, C.; Gonzalez, D.; Klein, M.; Pilla, M.; Vallerand, S.; Maldague, X. Infrared image processing and data analysis. *Infrared physics & technology* **2004**, *46*, 75–83.
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
4. Ullah, A.; Xie, H.; Farooq, M.O.; Sun, Z. Pedestrian detection in infrared images using fast RCNN. In Proceedings of the 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2018, pp. 1–6.
5. Li, B.; Xiao, C.; Wang, L.; Wang, Y.; Lin, Z.; Li, M.; An, W.; Guo, Y. Dense nested attention network for infrared small target detection. *IEEE Transactions on Image Processing* **2022**, *32*, 1745–1758.
6. Dai, Y.; Wu, Y.; Zhou, F.; Barnard, K. Asymmetric contextual modulation for infrared small target detection. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 950–959.
7. Zhang, M.; Zhang, R.; Yang, Y.; Bai, H.; Zhang, J.; Guo, J. ISNet: Shape matters for infrared small target detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 877–886.
8. Liu, F.; Guan, S.; Yu, K.; Gong, H. Infrared target detection based on the fusion of Mask R-CNN and image enhancement network. In Proceedings of the 2022 China Automation Congress (CAC). IEEE, 2022, pp. 2011–2016.
9. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* **2020**.
12. Sohan, M.; Sai Ram, T.; Rami Reddy, C.V. A review on yolov8 and its advancements. In Proceedings of the International Conference on Data Intelligence and Cognitive Informatics. Springer, 2024, pp. 529–545.
13. Lin, Z.; Huang, M.; Zhou, Q. Infrared small target detection based on YOLO v4. In Proceedings of the Journal of Physics: Conference Series. IOP Publishing, 2023, Vol. 2450, p. 012019.
14. Li, S.; Li, Y.; Li, Y.; Li, M.; Xu, X. Yolo-firi: Improved yolov5 for infrared image object detection. *IEEE access* **2021**, *9*, 141861–141875.
15. Zhan, W.; Zhang, C.; Guo, S.; Guo, J.; Shi, M. EGISD-YOLO: Edge guidance network for infrared ship target detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2024**, *17*, 10097–10107.
16. Tang, Y.; Xu, T.; Qin, H.; Li, J. IRSTD-YOLO: An Improved YOLO Framework for Infrared Small Target Detection. *IEEE Geoscience and Remote Sensing Letters* **2025**.
17. Farooq, M.A.; Corcoran, P.; Rotariu, C.; Shariff, W. Object detection in thermal spectrum for advanced driver-assistance systems (ADAS). *IEEE Access* **2021**, *9*, 156465–156481.
18. Monnier, G.F. A review of infrared spectroscopy in microarchaeology: Methods, applications, and recent trends. *Journal of Archaeological Science: Reports* **2018**, *18*, 806–823.

 doi:10.20944/preprints202601.0503.v1

19. Zhang, X.; Liu, C.; Yang, D.; Song, T.; Ye, Y.; Li, K.; Song, Y. RFAConv: Innovating spatial attention and standard convolutional operation. *arXiv preprint arXiv:2304.03198* **2023**.
20. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2023, pp. 1–5.
21. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv preprint arXiv:2205.12740* **2022**.
22. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.
23. Zhang, Y.; Guo, Z.; Wu, J.; Tian, Y.; Tang, H.; Guo, X. Real-time vehicle detection based on improved yolo v5. *Sustainability* **2022**, *14*, 12274.
24. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976* **2022**.
25. Li, H.; Xiong, P.; An, J.; Wang, L. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180* **2018**.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
27. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 13713–13722.
28. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 1–21.
29. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems* **2024**, *37*, 107984–108011.
30. Khanam, R.; Hussain, M. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725* **2024**.