

Lightweight Real-Time RGB-IR Pedestrian Detection Based on DoG Detail Enhancement and Region-Guided Attention

LinXi Zhong

Chongqing University of Technology

Zhaohui Xiao

1819969361a@gmail.com

Chongqing University of Technology

Article

Keywords: RGB-IR pedestrian detection, YOLOv10, Difference of Gaussian, Region-Guided Spatial Attention, real-time object detection

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10875416/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Object detection is a fundamental task in computer vision, and pedestrian detection is essential for intelligent surveillance, autonomous driving, and assisted perception. To address the limitations of single-modal RGB or infrared (IR) detection in complex environments, this paper proposes a lightweight real-time RGB-IR pedestrian detection method based on YOLOv10n. The proposed DRGF-YOLOv10n introduces a DualBranchStem structure to separately extract RGB texture-edge features and IR thermal-saliency cues. A Difference of Gaussian (DoG) operation is applied to the RGB branch to enhance local edges and contour details, while Region-Guided Spatial Attention (RG-SA) is embedded into the eighth layer to strengthen pedestrian-related regions and suppress background interference. Experiments on the CVC14 validation set show that the proposed method improves precision, recall, and mAP over the baseline while maintaining real-time performance. Therefore, DRGF-YOLOv10n is suitable as a post-stage high-confidence verification module for false-alarm-sensitive surveillance systems.

1 Introduction

Object detection aims to automatically localize all target objects in an image and identify the category of each object. Pedestrian detection is an important branch of object detection. Existing deep-learning-based object detection methods are commonly categorized into two-stage detectors, one-stage detectors, and Transformer-based detectors [1]. Among them, one-stage detectors directly perform classification and localization on feature maps, offering high inference efficiency and therefore stronger suitability for real-time image-processing tasks.

In recent years, the central direction of multimodal object detection has been to exploit complementary information from different sensors in order to improve robustness under complex environments. A single modality inevitably has limitations during object detection. Multimodal detection therefore integrates information from RGB, IR/thermal, depth, LiDAR, radar, and other sources, enabling more stable performance in adverse weather, night-time driving, intelligent surveillance, robotic navigation, and assisted perception. Recent surveys have also shown that multi-sensor-fusion-based object detection has been widely applied in autonomous driving, intelligent monitoring, robotics, and unmanned aerial systems. Its major advantage lies in compensating for the weaknesses of individual sensors in challenging environments. In complex scenes, RGB images are vulnerable to low illumination, backlighting, rain, fog, and cluttered backgrounds, which may weaken pedestrian boundaries and cause missed detections. IR images can highlight human targets based on thermal radiation differences, but they generally lack texture, boundary, and semantic details. Because RGB and IR imaging mechanisms are complementary, RGB-IR fusion has become an effective approach for improving the robustness of pedestrian detection in complex environments [2–4].

The RG-SA module used in this paper mainly performs region-guided aggregation on deep features. Although it can enhance target-region responses and suppress background interference, feature compression and regional selection may also weaken certain local edges and texture details. To address this issue, this paper introduces a Difference of Gaussian (DoG) detail-enhancement strategy during RGB preprocessing. The purpose is to strengthen edge, contour, and local high-frequency information in visible-light images, thereby providing richer detail cues for subsequent RGB-IR dual-branch fusion.

The main contributions of this paper are summarized as follows:

- A DualBranchStem structure is introduced to alleviate the limitations of RGB and IR single-modal detection in complex scenes. The structure separately exploits visible-light texture and edge information and infrared thermal-saliency information, thereby enhancing RGB-IR multimodal feature representation.
- RG-SA is embedded in the deep RGB branch to mitigate background interference and strengthen pedestrian-related feature responses.
- DoG detail enhancement is introduced during RGB-branch preprocessing to reduce detail degradation caused by background interference and to reinforce local edge information.

2 Related Work

2.1 Real-Time Object Detection and Pedestrian Detection

Object detection is a core task in computer vision that simultaneously performs target-category recognition and spatial localization. Recent surveys usually classify deep-learning-based object detection methods into two-stage detectors, one-stage detectors, and Transformer-based detectors [1]. Pedestrian detection is a representative task in this field and has been widely applied to intelligent surveillance, autonomous driving, assisted driving, and intelligent mobility-assistance systems. These real-time perception scenarios require the system to rapidly detect pedestrians in continuous video streams and respond promptly. Therefore, pedestrian detection is not only concerned with detection accuracy, but also imposes strict requirements on inference speed and model complexity. If the detection latency is excessive, a model with high accuracy may still fail to satisfy the safety-warning requirements of dynamic scenes.

The YOLO series is a representative family of real-time object detectors. YOLOv7 improves the speed-accuracy trade-off through training strategies and architectural optimization [5]. RT-DETR and YOLOv10 target end-to-end real-time detection and reduce post-processing overhead [6–7]. YOLOX, YOLOv6, and YOLOv9 further improve lightweight detection from the perspectives of anchor-free design, engineering deployment, and programmable gradient information [8–10]. Deformable DETR and DINO further demonstrate that sparse attention and denoising queries can improve convergence and localization in end-to-end detectors [11–12]. Considering the demand for real-time deployment, this paper adopts the lightweight YOLOv10n as the base detector.

2.2 RGB-IR Multimodal Pedestrian Detection

RGB-T/RGB-IR image analysis has covered detection, segmentation, re-identification, and tracking tasks [2]. Recent multispectral pedestrian detection studies emphasize modality complementarity, robustness against misalignment, and target-aware fusion. TFDet suppresses background-induced false detections through target-aware fusion [3]. MS-DETR improves multispectral pedestrian detection through loosely coupled fusion and modality-balanced optimization [4]. C2Former and ICAFusion exploit cross-modal attention to enhance complementary interactions between RGB and IR features [13–14]. In addition, Cross-Modality Fusion Transformer, Causal Mode Multiplexer, Hybrid Attention, and CCIFNet further improve pedestrian detection from the perspectives of cross-modal context, modality bias, and robust fusion [15–19]. In

contrast, this paper adopts a more lightweight dual-branch input design, which preserves modality-specific representations while reducing fusion complexity.

2.3 Detail Enhancement and Attention Mechanisms

Image detail enhancement can highlight edges, contours, and local textures, which is beneficial for weak-target detection. DoG can be regarded as a band-pass filtering operation that emphasizes high-frequency variations by subtracting Gaussian-smoothed results at different scales [20]. Attention mechanisms enhance important regions through feature reweighting. Coordinate Attention, RegionViT, BiFormer, Swin Transformer, MobileViT, EfficientViT, and VAN demonstrate that spatial position modeling, regional relationship modeling, and long-range dependency modeling can improve dense prediction performance [21–27]. The RG-SA module in this paper is inspired by regional-attention ideas, but it is applied to deep features in the RGB branch to reinforce pedestrian-related regions and suppress background interference.

In summary, existing studies offer three insights for RGB-IR pedestrian detection. First, a real-time detector should prioritize computational cost and deployment efficiency. Second, multimodal fusion should preserve the differential representations of RGB and IR rather than directly mixing the two modalities at the input stage. Third, an attention module should avoid excessively early or overly strong selection of shallow textures. Based on these observations, this paper adopts a lightweight combination of dual-branch input, RGB detail enhancement, and deep region-guided attention.

3 Method

This paper proposes a detail-enhanced region-guided fusion framework for pedestrian detection. The framework enhances pedestrian-region responses while suppressing interference from complex backgrounds. This section first presents the research motivation and architecture of each component in the DRGF framework and then describes the proposed modules in detail.

3.1 Research Motivation

3.1.1 Motivation for DualBranchStem

The original YOLOv10 is mainly designed for single-modal RGB image input and does not directly exploit the thermal-saliency information in IR images or combine the two modalities for target analysis. A simple channel-wise concatenation of RGB and IR images followed by standard convolution cannot sufficiently distinguish the feature differences between the two modalities. In our corresponding six-channel concatenation experiment, where RGB and IR each contributed three channels, the final precision, recall, and mAP were unsatisfactory. Therefore, this paper designs an RGB-IR dual-branch input module, namely DualBranchStem, to separately extract low-level features from RGB and IR images.

3.1.2 Motivation for DoG

RGB images contain relatively rich edge, texture, and contour structures, but local details may degrade under weak illumination and cluttered backgrounds. IR images provide stable thermal-saliency responses, but their texture and boundary details are usually limited. To strengthen the visible-light branch in perceiving

pedestrian boundaries, small-object contours, and weak-texture regions, this paper introduces a DoG detail-enhancement strategy at the RGB input stage.

3.1.3 Motivation for RG-SA

Standard self-attention models long-range dependencies by computing correlations among global tokens and can effectively capture global contextual information. However, its computational complexity increases rapidly with feature-map resolution, which restricts its application in high-resolution object detection. To reduce the cost of attention computation, recent methods such as window attention, deformable attention, regional attention, and linear attention optimize global attention from different perspectives [11, 22–24, 26]. To enhance target regions, suppress background interference, and reduce computational complexity, this paper introduces the Region-Guided Spatial Attention (RG-SA) module into the deep feature stage of YOLOv10.

3.2 Overall Framework

The overall architecture of DRGF is shown in Fig. 1. The proposed method consists of three main components: RGB data preprocessing, dual-branch channel feature extraction, and the region-guided attention mechanism in the RGB branch.

Based on YOLOv10n, the overall process is as follows. Paired RGB-IR inputs are first processed by DoG detail enhancement and DualBranchStem dual-branch feature extraction. The fused features are then fed into the YOLOv10n backbone and detection head, while RG-SA is introduced at the eighth-layer deep semantic position. This design jointly exploits the texture and edge advantages of RGB images and the thermal-saliency advantages of IR images, thereby improving pedestrian detection performance in complex environments with low computational overhead.

The design rationale is that RGB and IR modalities are clearly complementary in pedestrian detection. Accordingly, this paper constructs an RGB-IR dual-modal detection framework in which a dual-branch structure separately extracts modality-specific features and reduces modality interference caused by early shallow-feature mixing. Meanwhile, DoG is introduced to enhance local edges and contour details, and RG-SA is embedded into deep features to strengthen pedestrian-region representation. By combining the detail and semantic advantages of RGB images with the thermal-saliency advantage of IR images, the proposed method improves detection accuracy and robustness in complex scenes.

3.3 Architectural Modules

3.3.1 DualBranchStem Dual-Branch Input Module

DualBranchStem contains two parallel branches. The RGB branch processes visible-light images and extracts color, texture, edge, and contour

features. The IR branch processes infrared images and extracts thermal-saliency target information. The features from the two branches are then fused and passed into the subsequent YOLOv10 backbone network.

The training log shows that $\text{batch}['\text{img}'].shape = \text{torch.Size}(2 \times 6 \times 416 \times 416)$, RGB mean = 0.3287, and IR mean = 0.5908. Although the data are input as a six-channel tensor in each batch, the first layer of the model is not a standard single-branch six-channel convolution. Instead, DualBranchStem assigns the first three channels to the RGB branch and the last three channels to the IR branch, extracts features separately, and then fuses them. Therefore, this design is essentially different from simple single-branch channel concatenation. It also confirms that the current input is indeed six-channel RGB-IR data and that both RGB and IR modalities enter the model normally.

3.3.2 DoG Detail-Enhancement Module

Difference of Gaussian (DoG) extracts high-frequency details in an image by subtracting two Gaussian-smoothed results at different scales. It is formulated as follows:

$$\text{DoG}(I_{\text{rgb}}) = G_{\sigma 1}(I_{\text{rgb}}) - G_{\sigma 2}(I_{\text{rgb}}), \sigma 1 < \sigma 2$$

where I_{rgb} denotes the input RGB image, and $G_{\sigma 1}(\cdot)$ and $G_{\sigma 2}(\cdot)$ denote Gaussian filtering operations with different standard deviations. The Gaussian filter with a smaller scale preserves more local details, whereas the Gaussian filter with a larger scale represents a smoother background structure. Their difference highlights edges, contours, and local texture variations.

For RGB-IR pedestrian detection, this paper adopts residual DoG enhancement to avoid excessive changes to the original image distribution:

$$I_{\text{rgb_dog}} = I_{\text{rgb}} + \alpha \cdot \text{DoG}(I_{\text{rgb}})$$

where α is the enhancement coefficient, which is set to 0.025 in this paper. DoG is applied only to the RGB image, while the IR image remains unchanged. This strategy enhances visible-light edge and local structural information while preserving the thermal-saliency characteristics of the infrared image.

In this paper, DoG is not an internal convolutional module of the network; rather, it is an RGB detail-enhancement step at the data-preprocessing stage. After the original RGB image is input, DoG enhancement with $\alpha = 0.025$ is applied to enrich details and sharpen boundaries. The enhanced RGB image then enters the RGB branch of the dual-branch fusion module, whereas the IR image does not undergo detail enhancement.

3.3.3 RG-SA Region-Guided Spatial Attention Module

The RG-SA module is introduced into the RGB branch to enhance the regional contextual representation capability of visible-light features. It should be noted that the RG-SA used in this paper is a lightweight convolutional region-reweighting module designed for RGB-IR pedestrian detection, rather than the recursive generalization attention structure in Recursive Generalization Transformer [28]. Its core idea is to jointly enhance pedestrian-related region responses and suppress complex background interference through local-region context, spatial attention, and channel attention at the eighth-layer deep semantic feature stage of YOLOv10n.

As shown in Fig. 3, the proposed RG-SA module mainly consists of three parts: the Region Guidance Module (RGM), query-key-value mapping, and region-guided attention computation. Given an input feature $X_{\text{in}} \in$

$R^{H \times W \times C}$, RGM first performs local-region compression and semantic aggregation on the input feature, producing representative regional features X_r and a region guidance map M . Here, H and W denote the height and width of the original feature map; h and w denote the compressed regional scale; and C_r denotes the reduced channel dimension. The guidance map M describes the correspondence between global positions and local regions, thereby providing prior constraints for subsequent attention computation.

Specifically, RG-SA first generates low-resolution regional features from RGB features through the Region Guidance Module. The original RGB features are then used to generate Query, while the regional features are used to generate Key and Value. In this way, the attention relationship between global RGB features and regional semantic features is established. This design strengthens the RGB branch in modeling target edges, textures, and contextual structures with relatively low computational cost. Compared with standard global self-attention, RG-SA uses low-resolution regional features to generate Key and Value, thereby reducing the token-interaction scale while retaining global contextual perception. Because the module is placed at the deep feature stage, the input features already contain stronger semantic information, enabling more effective enhancement of pedestrian-related regions rather than shallow texture noise.

Let the input feature of the RGB branch be:

$$X_{rgb} \in R^{B \times C \times H \times W}$$

The regional feature obtained by RGM is:

$$X_r = RGM(X_{rgb}), X_r \in R^{B \times C_r \times h \times w}$$

Query, Key, and Value are then generated as:

$$Q = W_Q(X_{rgb}), K = W_K(X_r), V = W_V(X_r)$$

where Q is derived from the original RGB feature, while K and V are derived from the region-guided feature. The region-guided attention is computed as:

$$A_{rg} = \text{Softmax}(QK^T / \sqrt{C_r})$$

The attention output is:

$$Y = A_{rg} V, X_{rgb}' = W_O(Y)$$

Finally, a residual connection is applied:

$$X_{rgb_out} = X_{rgb} + X_{rgb}'$$

Equivalently, the module can be written as:

$$X_{rgb_out} = X_{rgb} + W_O \left(\text{Softmax} \left(W_Q X_{rgb} \cdot (W_K RGM(X_{rgb}))^T / \sqrt{C_r} \right) W_V RGM(X_{rgb}) \right)$$

4 Experiments

4.1 Experimental Settings and Visualization Analysis

The proposed method is mainly evaluated on the publicly available CVC-14 RGB-IR pedestrian dataset, which can be accessed from the Computer Vision Center at <https://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset/>. Each sample consists of paired RGB images, IR images, and shared annotations in YOLO format. The input size is uniformly set to 416×416 . The experiments are implemented using PyTorch 2.8.0 + cu129 and the Ultralytics YOLO framework on an NVIDIA RTX 5060 Laptop GPU. Training is initialized with YOLOv10n pretrained weights, the batch size is set to 2, and the main training schedule contains 30 epochs. To avoid strong data augmentation damaging the edge structures enhanced by DoG, HSV, Mosaic, MixUp, Copy-Paste, Erasing, and other strong augmentations are disabled, while only mild geometric perturbations are retained.

To ensure comparability among different models, all core ablation experiments use the same input size, number of training epochs, batch size, optimizer settings, and validation protocol. The single-modal baselines only change the input modality. The dual-modal models gradually introduce DualBranchStem, DoG, and RG-SA under the same data split, so that the effects of each module on Precision, Recall, and mAP can be observed. In addition, practical pedestrian detection requires not only detection accuracy but also real-time responsiveness.

As shown in Fig. 4, the feature visualization results indicate that RG-SA can perform regional reweighting on deep feature responses. It suppresses irrelevant background responses to a certain extent and strengthens the region representation associated with pedestrian detection. Because the pedestrian targets in the visualization are small, the figure is mainly used to demonstrate the reweighting effect of RG-SA on deep feature responses, rather than to provide precise target localization.

To further analyze the influence of RG-SA on internal feature representation, this paper visualizes block-wise feature correlations for the baseline model and the model with RG-SA. Specifically, the output responses of multiple feature blocks are extracted, global average pooling is used to obtain block-level feature vectors, and cosine similarity is calculated between any two blocks. In Fig. 5, brighter colors indicate stronger correlations between two feature blocks. After RG-SA is introduced, the distribution of block-wise correlations changes significantly: the response associations between certain key semantic layers are strengthened, while some redundant correlations are weakened. The block-wise response-gain curve on the right further shows that RG-SA has a relatively small influence on shallow features, whereas it produces more obvious response changes in middle and high-level feature blocks. This suggests that RG-SA mainly acts on deep stages with richer semantic information, helping enhance pedestrian-region representation and reduce the influence of complex backgrounds on detection results.

4.2 Baseline Models and Module Ablation Experiments

To clarify the influence of different input forms on detection performance, this paper additionally sets up RGB-only, IR-only, direct RGB-IR concatenation, and dual-branch input baselines. RGB-only and IR-only are used to analyze single-modal detection capability; RGB-IR concat is used to verify the limitation of simple six-

channel concatenation; and DualBranchStem is used to verify the effectiveness of the dual-branch input structure.

As shown in Table 1, compared with single-modal RGB input, the dual-modal dual-branch structure significantly improves mAP50 and mAP50-95, indicating that RGB texture-edge information and IR thermal-saliency information are complementary. When simple six-channel concatenation is compared with dual-branch input fusion, all indicators of the dual-branch input are clearly better. Simple channel concatenation mixes modalities from the first layer, causing the convolution kernel to simultaneously process all modal channels and potentially ignore modality-specific properties. By contrast, dual-branch input allows the feature activations of each branch to be observed separately and enables analysis of the contribution of each modality.

Table 1
Comparison of baseline models and input forms

Model	Input form	Comparison purpose	P	R	mAP50	mAP50-95
YOLOv10n-RGB	RGB only	Visible-light single-modal baseline	0.2510	0.1460	0.1240	0.0516
YOLOv10n-IR	IR only	Infrared single-modal baseline	0.4330	0.3185	0.3125	0.1270
YOLOv10n RGB-IR concat	6-channel concat	Simple channel-concatenation baseline	0.8950	0.0996	0.4397	0.1780
RGB-IR dual baseline	DualBranchStem	Dual-branch input-fusion baseline	0.8854	0.1265	0.4711	0.1995
Ours	DualBranchStem + DoG + RG-SA	Final model	0.9204	0.1912	0.5231	0.2361

As shown in Table 2, RG-SA improves mAP50 from 0.4711 to 0.5166 and mAP50-95 from 0.1995 to 0.2203, indicating that region-guided attention enhances target-region representation. After adding DoG, Recall increases, suggesting that edge-detail enhancement is beneficial for weak-target detection. However, if strong augmentation is retained, local background textures may also be amplified. After strong data augmentation is removed, the model achieves the best Precision, mAP50, and mAP50-95, demonstrating the complementarity between DoG and RG-SA.

Table 2
Main module ablation experiments

RGB-IR Baseline	DoG	RG-SA	W/o Strong Aug.	Precision	Recall	mAP50	mAP50-95
✓				0.8854	0.1265	0.4711	0.1995
✓		✓		0.9020	0.0732	0.5166	0.2203
✓	✓	✓		0.9130	0.1375	0.4714	0.1994
✓	✓	✓	✓	0.9204	0.1912	0.5231	0.2361

4.3 Analysis of RG-SA Location and Fusion Strategies

Table 3 shows that stacking RG-SA at more layers does not necessarily lead to better performance. Adding RG-SA simultaneously to layers 4, 6, and 8 may improve Precision or Recall in certain cases, but both mAP50 and mAP50-95 are lower than those of the single-layer structure at layer 8. Shallow features contain abundant local textures; therefore, premature regional selection may suppress useful information. By contrast, the semantic information at layer 8 is more sufficient and is more suitable for regional enhancement. Variants such as L-SA, MCR, and PCDF-lite do not bring stable mAP improvements, suggesting that overly strong local attention may introduce false detections.

Table 3
Comparison of RG-SA placement and fusion strategies

Model	P	R	mAP50	mAP50-95
Layer-8 RG-SA + DoG	0.9204	0.1912	0.5231	0.2361
Full RG-SA at layers 4/6/8, epoch 25	0.9303	0.0875	0.4642	0.1834
Full RG-SA at layers 4/6/8, epoch 30	0.8921	0.1373	0.4336	0.1630
Dual + deep L-SA	0.8956	0.1104	0.4654	0.1904
Dual + L-SA + RG-SA	0.8724	0.1087	0.4715	0.1942
MCR + RG-SA + img-size 512	0.4990	0.2290	0.2910	0.1320
PCDF-lite + RG-SA	0.2760	0.3250	0.2370	0.0948

4.4 Comparison with Representative Fusion Strategies

In addition to internal ablation studies, this paper further compares the proposed method with existing RGB-IR/RGB-T pedestrian detection methods from the perspectives of method design and deployment characteristics. It should be noted that different studies vary in data split, evaluation protocol, and training configuration. Therefore, this section does not directly compare absolute metrics, but focuses on differences in methodological design.

Table 4
Explanatory comparison with existing RGB-IR methods

Method	Source / publication	Category	Input form	Resolution	P	mAP50	mAP50-95	Model Size / MB
Add Fusion	Baseline strategy	Simple feature fusion	RGB-IR	416×416	0.7320	0.6050	0.2740	32.62
Gated Fusion	GMU / gated multimodal fusion inspired, 2017	Dynamic gated fusion	RGB-IR	416×416	0.7120	0.5930	0.2620	32.81
TFDet-style Fusion	arXiv 2023 / TNNLS 2025	Target-aware fusion	RGB-IR	416×416	0.7530	0.6130	0.2780	32.81
CBAM-6ch Fusion	ECCV 2018	Channel-spatial attention fusion	6-channel concat	416×416	0.9043	0.4932	0.2093	32.63
SE-6ch Fusion	CVPR 2018 / TPAMI journal version	Channel-attention fusion	6-channel concat	416×416	0.9189	0.4465	0.1867	32.62
Ours	This paper	DoG + RG-SA lightweight fusion	RGB-IR	416×416	0.9204	0.5231	0.2361	32.01

As shown in Table 4, many existing methods adopt complex cross-modal attention or Transformer-based fusion structures to improve detection performance, whereas this paper emphasizes lightweight real-time deployment. The proposed method preserves modality differences at the input stage through the dual-branch structure, enhances RGB details through DoG, and introduces RG-SA only at the deep semantic feature stage. Therefore, it achieves favorable real-time detection performance with relatively low parameter and computational costs. The proposed method obtains the highest Precision, reaching 0.9204, which is higher than SE-6ch Fusion (0.9189) and CBAM-6ch Fusion (0.9043). This indicates that DoG detail enhancement and RG-SA region-guided attention further improve the reliability of target detection. Meanwhile, the method achieves mAP50 of 0.5231 and mAP50-95 of 0.2361, outperforming CBAM-6ch Fusion and SE-6ch Fusion under the reported settings. This suggests that, compared with ordinary channel attention or channel-spatial attention, the proposed method provides better overall detection capability while maintaining high precision. The method is particularly suitable for false-alarm-sensitive RGB-IR surveillance tasks, especially as a post-stage high-confidence target verification module for reducing false alarms in complex backgrounds and improving alarm reliability.

4.5 Computational Complexity, Inference Latency, and Post-Stage Verification Deployment

To further verify the deployment feasibility of the proposed method in real-time surveillance scenarios, this paper analyzes DRGF-YOLOv10n in terms of the number of parameters, computational complexity, model storage size, single-frame inference latency, and inference frame rate. All real-time tests are conducted with an input size of $6 \times 416 \times 416$ and a batch size of 1 on an NVIDIA RTX 5060 Laptop GPU. To reduce the effect of random fluctuations, the test first performs multiple warm-up iterations to stabilize the GPU inference state and then records the average time of multiple forward passes.

From the perspective of model architecture, the proposed method is based on YOLOv10n and only introduces the lightweight DualBranchStem at the input stage, DoG detail enhancement before the RGB branch, and the RG-SA module at the eighth-layer deep semantic feature stage. DualBranchStem separately extracts shallow RGB and IR features and introduces only a small number of additional parameters. DoG belongs to the image detail-enhancement operation at the data-preprocessing stage and does not serve as an internal learnable convolutional layer or participate in backpropagation; thus, it does not significantly increase the number of model parameters. RG-SA operates only on deep low-resolution feature maps, which incurs lower computational overhead than introducing global attention into shallow high-resolution features. Therefore, the proposed method enhances RGB-IR multimodal feature representation while maintaining low computational complexity.

Experimental results show that DRGF-YOLOv10n has 2.894M parameters, 1.908 GFLOPs, and a model size of 32.01 MB. Its average single-frame inference latency is 31.468 ms, corresponding to an inference frame rate of 31.78 FPS. The frame rate is computed as:

$$FPS = 1000 / T_{latency}$$

where $T_{latency}$ denotes the average single-frame inference latency in milliseconds. When $T_{latency} = 31.468$ ms, the inference frame rate is approximately 31.78 FPS, satisfying the basic processing requirement of 25–30 FPS for common real-time video surveillance. These results indicate that although DoG detail enhancement and RG-SA region-guided attention are introduced, the overall model remains lightweight and feasible for deployment in real-time surveillance systems.

It should be noted that the proposed method is not positioned as a high-recall detector that independently undertakes the entire detection process. Instead, it is more suitable as a post-stage high-confidence verification module in false-alarm-sensitive surveillance systems. The deployment process can be described as follows. First, a front-stage candidate detector runs on RGB-IR video streams with a relatively low confidence threshold to generate as complete a set of pedestrian candidate regions as possible, thereby ensuring high recall. Next, the candidate regions or corresponding frames are input into the DRGF-YOLOv10n verification module. The RGB local contour details are enhanced through DoG, and the deep pedestrian-related region responses are strengthened through RG-SA. Finally, the system filters false positives according to the high-confidence detection results output by DRGF-YOLOv10n, retaining only reliable pedestrian targets and triggering alarms.

In summary, DRGF-YOLOv10n achieves a favorable balance among model size, computational complexity, and inference speed. Its 2.894M parameters and 31.78-FPS inference speed demonstrate real-time deployment potential. At the same time, the post-stage high-confidence verification positioning reasonably

explains the high-Precision and low-Recall performance characteristics of the model, making it more suitable for false-alarm suppression and alarm-result refinement in complex surveillance scenarios.

4.6 Result Discussion

Overall, the performance improvement of the proposed method mainly comes from three aspects. First, the dual-branch input structure provides independent low-level feature-extraction paths for RGB and IR, allowing texture details and thermal-saliency information to be preserved before fusion. Second, DoG explicitly enhances high-frequency edges in RGB images, helping compensate for the limited texture information of IR images. Third, RG-SA performs region guidance at the deep semantic stage, encouraging the model to focus more on pedestrian targets rather than background textures.

As shown in Fig. 6, DRGF-YOLOv10n demonstrates better pedestrian detection capability than the ordinary DualBranch model under different illumination conditions. In all four examples, the ordinary DualBranch model fails to produce effective detection results, with $TP = 0$ and the number of false negatives equal to the number of ground-truth objects. This indicates that simple RGB-IR dual-branch fusion is insufficient for addressing small targets, weak textures, and low contrast in complex scenes. By contrast, DRGF-YOLOv10n effectively detects the main pedestrian targets. Across the four groups of examples, it detects nine true targets without introducing false positives. These results show that DoG enhancement improves the edge and contour representation of the RGB branch, while the RG-SA attention mechanism further strengthens pedestrian-related regions and suppresses background interference, thereby improving detection robustness at night and under complex backgrounds.

At the same time, DRGF-YOLOv10n still suffers from certain missed detections, especially in distant, small-scale, occluded, or crowded pedestrian scenes, where some weak targets cannot be accurately recognized. This indicates that while the model maintains a low false-positive rate, its ability to recall difficult samples still requires improvement. Overall, DRGF-YOLOv10n is more suitable as a post-stage high-confidence verification module in false-alarm-sensitive surveillance systems. After a front-stage detector generates candidate targets, it can further filter false positives and improve the reliability of final detection and alarm results.

The Recall of the proposed model still has room for improvement. This is related to scale variation, occlusion, and cross-modal alignment errors in the CVC14 dataset, and it also indicates that single-layer regional attention alone cannot completely solve missed detections. Future work may combine more stable cross-modal alignment, sample resampling, and lightweight dynamic fusion strategies to further improve recall.

5 Conclusion

This paper proposes a lightweight RGB-IR pedestrian detection method based on DoG detail enhancement and RG-SA region-guided attention. The method uses DualBranchStem to separately extract RGB texture-edge features and IR thermal-saliency features, applies DoG to enhance local RGB contours, and introduces RG-SA into the eighth deep layer of YOLOv10n to strengthen pedestrian target regions. The proposed method improves detection reliability while maintaining lightweight real-time performance.

The original DRGF-YOLOv10n exhibits high Precision and low Recall in the experiments, indicating that the model tends to output high-confidence targets rather than detect all potential targets. Therefore, this paper does not position it as a high-recall detector that independently handles the full detection pipeline. Instead, it is more suitable as a post-stage verification module or candidate-target refinement module in surveillance systems. In practical monitoring systems, as shown in Fig. 7, a front-stage detector usually generates candidate targets with a low threshold to ensure high recall. DRGF-YOLOv10n can then perform secondary verification on candidate regions and filter false detections caused by background textures, shadows, heat sources, railings, or non-human objects, thereby improving the credibility of final alarm results. This cascade detection mode, namely “front-stage recall and post-stage refinement,” is suitable for scenarios with high false-alarm costs, such as industrial-park intrusion detection, nighttime warehouse monitoring, railway and power-facility perimeter security, and dangerous-area proximity warnings. In these scenarios, frequent false alarms impose considerable manual-review costs and generate invalid alerts. Therefore, the post-stage model should focus more on detection precision and alarm reliability. The high-Precision characteristic of DRGF-YOLOv10n can meet the requirement for a low false-alarm rate in such scenarios.

As shown in Fig. 8, when the confidence threshold increases from 0.05 to 0.30, Precision gradually increases, whereas Recall continuously decreases and the Miss Rate correspondingly rises. This indicates that a higher confidence threshold can effectively filter low-confidence prediction boxes and reduce false positives, but it may also suppress some true targets and increase the missed-detection rate. The F1-score reaches a higher value at lower thresholds, suggesting that a lower threshold should be used when a balance between Precision and Recall is required. In false-alarm-sensitive surveillance alarm scenarios, however, the confidence threshold can be appropriately increased so that the model outputs more reliable detections. Therefore, DRGF-YOLOv10n is more suitable for deployment in the post-stage verification phase of a cascaded surveillance system. The front-stage detector generates candidate targets with a lower threshold to ensure high recall, whereas post-stage DRGF-YOLOv10n refines these candidates with a higher confidence threshold, thereby reducing false alarms and improving the reliability of alarm results.

Declarations

Author Contribution

Zhong Linxi was responsible for drafting the manuscript and completing all sections of the main text. All authors reviewed and approved the final manuscript.

Acknowledgements

The authors thank the School of Computer Science and Engineering, Chongqing University of Technology, for providing academic support for this work.

Funding

This work was supported by the National Natural Science Foundation of China [grant number 62472059] and the Science and Technology Commission of the Central Military Commission [grant number 2025CGJ017].

Conflict of interest

The author declares no competing interests.

Data availability

The CVC-14 RGB-IR pedestrian dataset analyzed during the current study is publicly available from the Computer Vision Center at <https://adas.cvc.uab.es/elektra/enigma-portfolio/cvc-14-visible-fir-day-night-pedestrian-sequence-dataset/>. Additional data generated during the current study are available from the corresponding author upon reasonable request.

Code availability

The implementation code is available from the author upon reasonable request.

References

1. Sun, Y., Sun, Z. & Chen, W. The evolution of object detection methods[J]. *Eng. Appl. Artif. Intell.* **133**, 108458 (2024).
2. Song, K. et al. RGB-T image analysis technology and application: A survey[J]. *Eng. Appl. Artif. Intell.* **120**, 105919 (2023).
3. Zhang, X. et al. TFDet: Target-aware fusion for RGB-T pedestrian detection[J]. *IEEE Trans. Neural Networks Learn. Syst.* **36** (7), 12413–12426 (2025).
4. Xing, Y. et al. MS-DETR: Multispectral pedestrian detection transformer with loosely coupled fusion and modality-balanced optimization[J]. *IEEE Trans. Intell. Transp. Syst.* **25** (12), 20628–20642 (2024).
5. 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[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 7464–7475. (2023).
6. Zhao, Y. et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 16965–16974. (2024).
7. Wang, A. et al. YOLOv10: Real-time end-to-end object detection[C]//Advances in Neural Information Processing Systems. **37**: 107984–108011. (2024).
8. Ge, Z. et al. YOLOX: Exceeding YOLO series in 2021[EB/OL]. (2021). arXiv:2107.08430.
9. Li, C. et al. YOLOv6: A single-stage object detection framework for industrial applications[EB/OL]. arXiv:2209.02976, 2022.
10. Wang, C. Y., Yeh, I. H. & Liao, H. Y. M. YOLOv9: Learning what you want to learn using programmable gradient information[C]//European Conference on Computer Vision. Cham: Springer, : 1–21. (2024).
11. Zhu, X. et al. Deformable DETR: Deformable transformers for end-to-end object detection[C]//International Conference on Learning Representations. (2021).
12. Zhang, H. et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection[C]//International Conference on Learning Representations. (2023).

13. Yuan, M. & Wei, X. C2Former: Calibrated and complementary transformer for RGB-infrared object detection[J]. *IEEE Trans. Geosci. Remote Sens.* **62**, 1–12 (2024).
14. Shen, J. et al. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection[J]. *Pattern Recogn.* **145**, 109913 (2024).
15. Fang, Q., Han, D. & Wang, Z. Cross-modality fusion transformer for multispectral object detection[EB/OL]. (2021). arXiv:2111.00273.
16. Kim, T. et al. Causal mode multiplexer: A novel framework for unbiased multispectral pedestrian detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 26784–26793. (2024).
17. Rathinam, A. et al. Hybrid attention for robust RGB-T pedestrian detection in real-world conditions[J]. *IEEE Rob. Autom. Lett.* **10** (1), 319–326 (2025).
18. Yan, C. et al. Cross-modality complementary information fusion for multispectral pedestrian detection[J]. *Neural Comput. Appl.* **35** (14), 10361–10386 (2023).
19. Fang, Q. & Wang, Z. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery[J]. *Pattern Recogn.* **130**, 108786 (2022).
20. Sun, R. et al. Survey of image edge detection[J]. *Front. Signal. Process.* **2**, 826967 (2022).
21. Hou, Q., Zhou, D. & Feng, J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 13713–13722. (2021).
22. Chen, C. F., Panda, R., Fan, Q. & RegionViT Regional-to-local attention for vision transformers[C]//International Conference on Learning Representations. (2022).
23. Zhu, L. et al. BiFormer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. : 10323–10333. (2023).
24. Liu, Z. et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. : 10012–10022. (2021).
25. Mehta, S., Rastegari, M. & MobileViT Light-weight, general-purpose, and mobile-friendly vision transformer[C]//International Conference on Learning Representations. (2022).
26. Cai, H. et al. EfficientViT: Lightweight multi-scale attention for high-resolution dense prediction[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. : 17302–17313. (2023).
27. Guo, M. H. et al. Visual attention network[J]. *Comput. Visual Media.* **9** (4), 733–752 (2023).
28. Chen, Z. et al. Recursive generalization transformer for image super-resolution[C]//International Conference on Learning Representations. (2024).

Figures

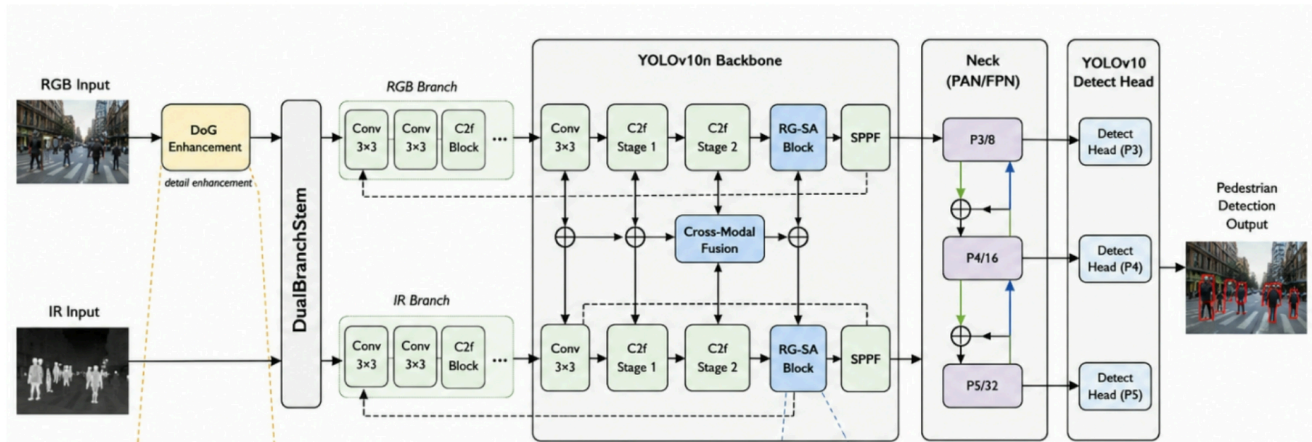


Figure 1. Overall workflow of the proposed RGB-IR pedestrian detection method

Figure 1

See image above for figure legend

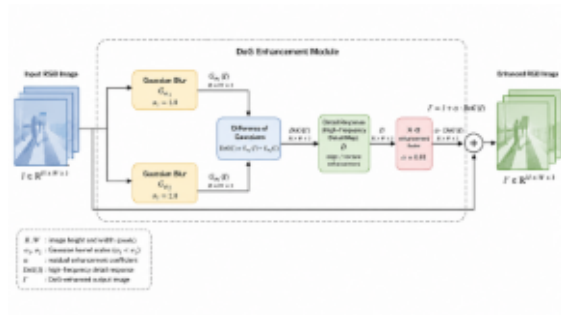


Figure 2

Workflow of the DoG detail-enhancement module

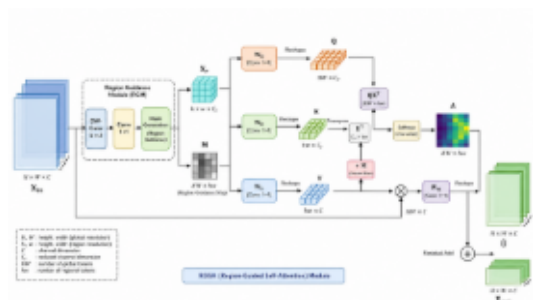


Figure 3

Workflow of the RG-SA module

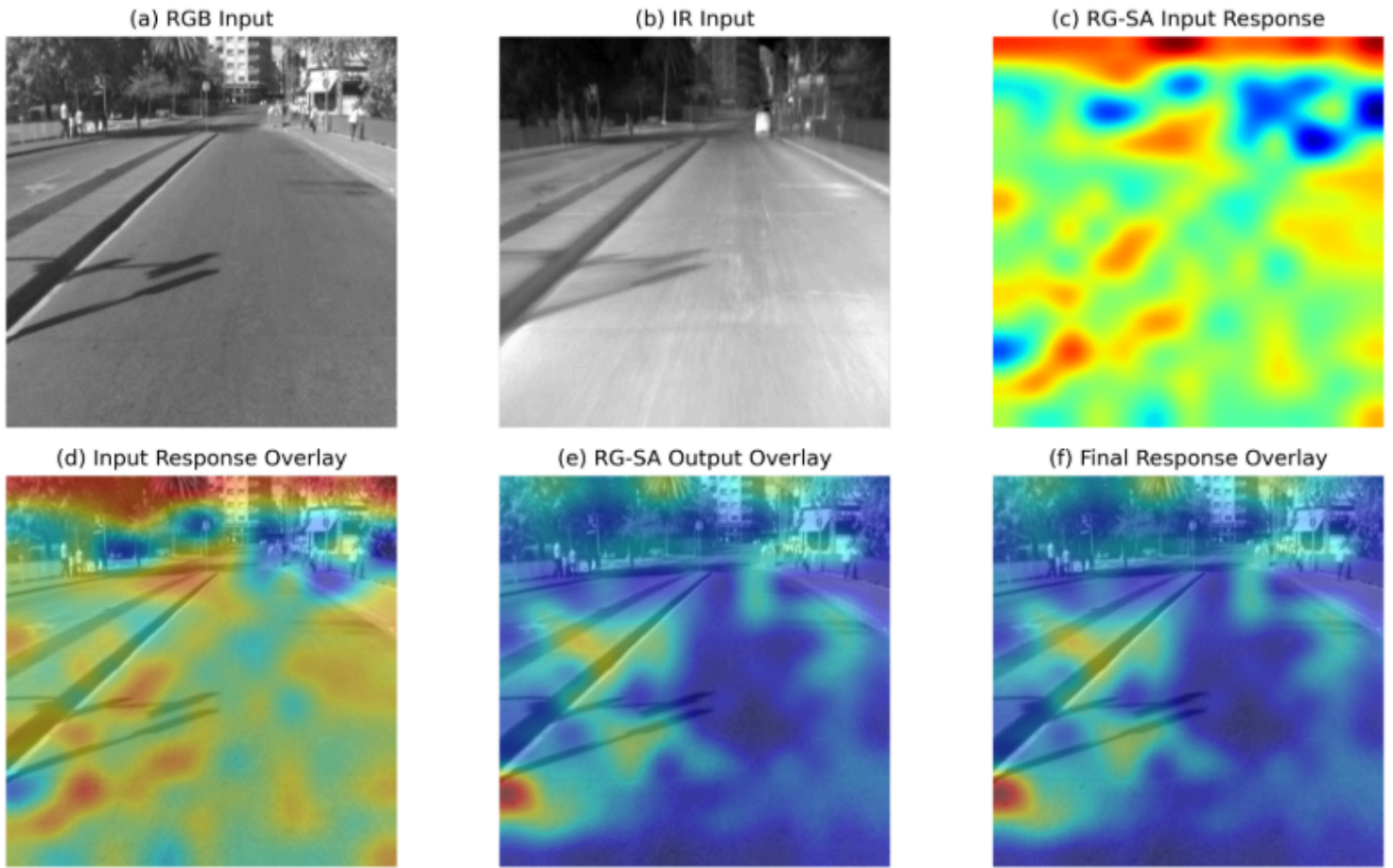


Figure 4

Feature responses of the RG-SA module

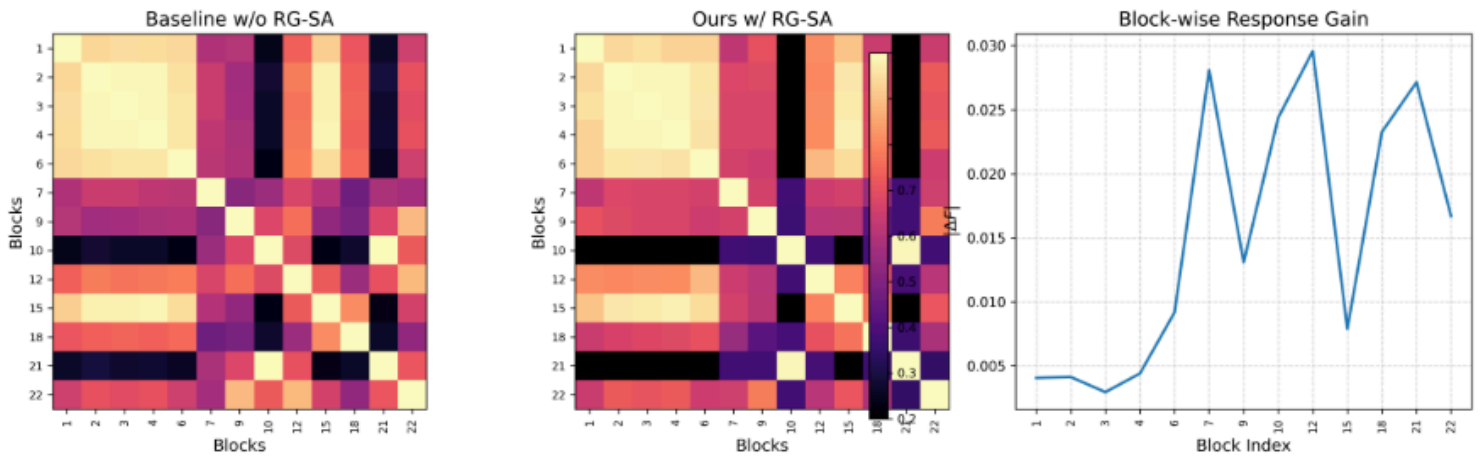


Figure 5

Visualization of block-wise feature correlations and response changes before and after introducing RG-SA



Figure 6

Visualization comparison between DRGF-YOLOv10n and the DualBranch model on the CVC14 dataset

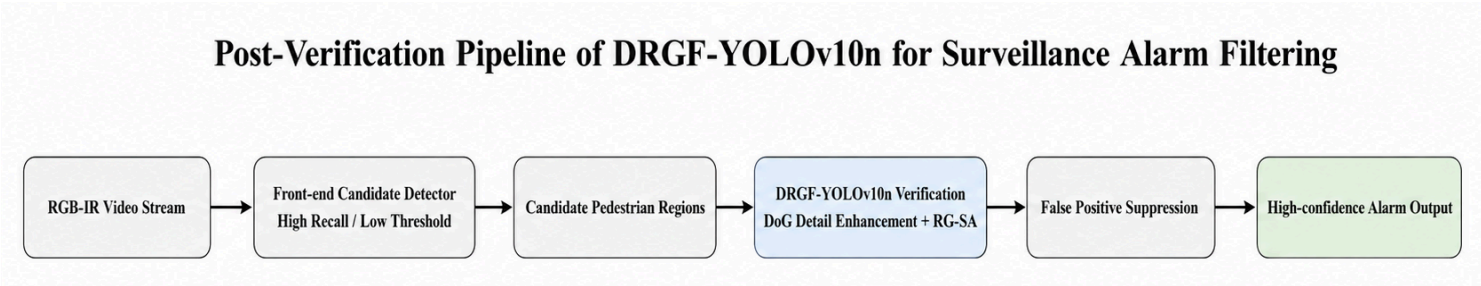


Figure 7

Application-scenario workflow

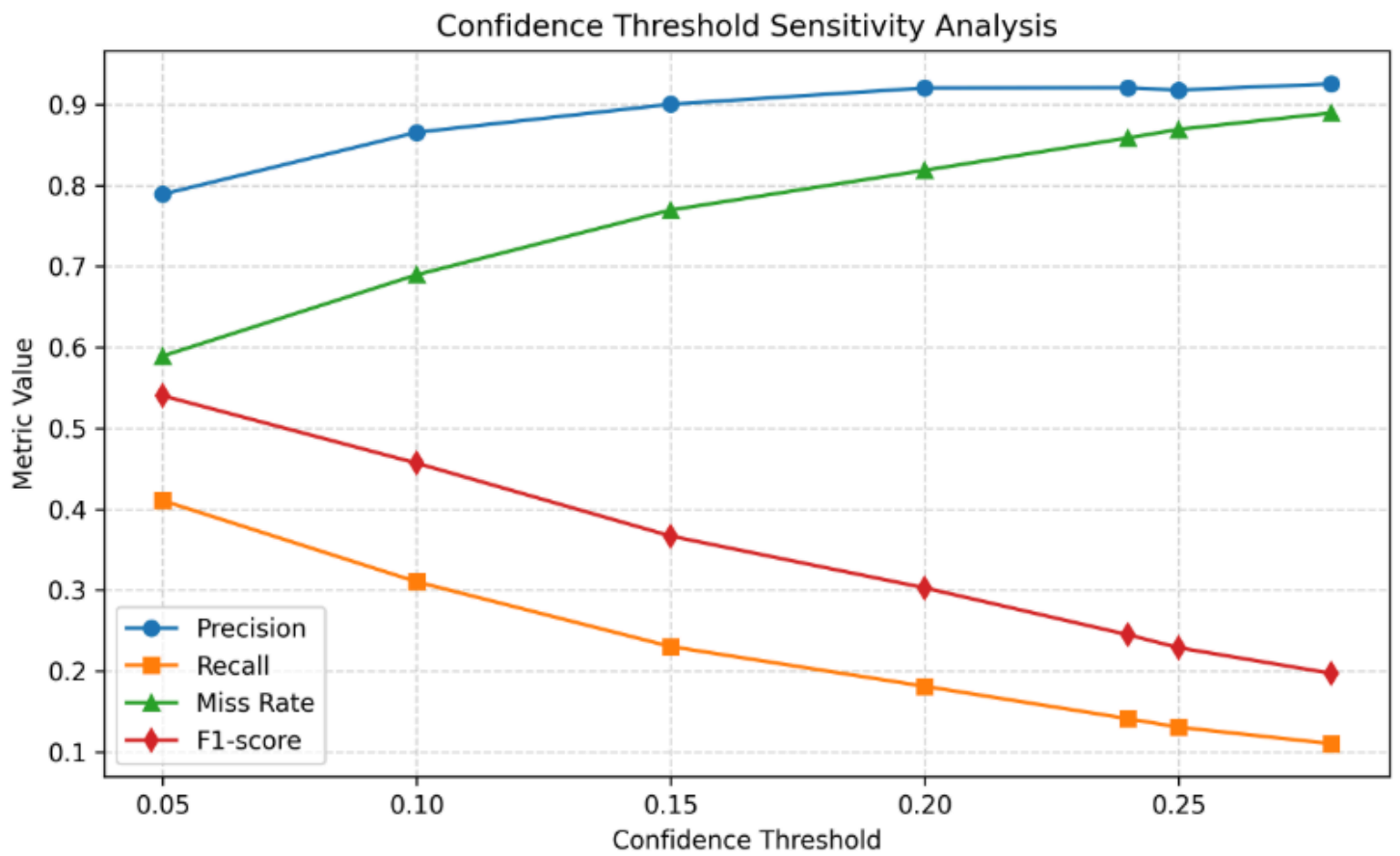


Figure 8

Confidence-threshold sensitivity