







## Dim-ObjectDet: Structure-Aware Network with iREMA Attention Mechanism and FRFM-Head Fusion for Dim Light Object Detection

Bin Li<sup>1\*</sup>, Shijie Zhou<sup>1</sup>, Junmin Xue<sup>1,2</sup>, Qinwei Yao<sup>3</sup>

<sup>1</sup> School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

<sup>2</sup> Operation Data Center, Postal Savings Bank of China, Beijing 100166, China

<sup>3</sup> Beijing Youan Hospital, Capital Medical University, Beijing 100069, China

Corresponding Author Email: [201912090910@std.uestc.edu.cn](mailto:201912090910@std.uestc.edu.cn)

Copyright: ©2025 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.420303>

### ABSTRACT

**Received:** 19 November 2024

**Revised:** 20 April 2025

**Accepted:** 3 May 2025

**Available online:** 30 June 2025

#### Keywords:

*dim light target detection, detection head, attention mechanism, detection feature extraction*

While deep learning-based object detection has achieved remarkable progress, existing methods exhibit significant performance degradation in dim-light environments due to three persistent challenges: (1) illumination bias in training data, where models optimized for normal lighting fail to generalize to dim conditions; (2) underutilization of latent features caused by data distribution shifts and noise interference; and (3) computational inefficiency in cascaded enhancement-detection frameworks, which limits real-time applicability. To address these issues, we propose Dim-ObjectDet, a novel low-light detection framework featuring three key innovations: (1) A multi-scale Feature Reassembling and Fusion Module (FRFM) detection head that dynamically integrates hierarchical features through adaptive channel attention, improving robustness to scale variations in dim environments; (2) An inverted Residual Efficient Multi-Scale Attention (iREMA) mechanism embedded in backbone and neck networks, synergizing local-global feature interactions to enhance noise-resistant representation learning; (3) A computationally optimized architecture combining Diverse Branch Blocks (DBB) and dynamic upsampling (DySample) to balance feature diversity and inference speed. Extensive experiments demonstrate state-of-the-art performance: Dim-ObjectDet achieves 7.0% higher accuracy and 7.3% mAP@0.5 improvement on the DarkFace dataset, along with 10.4% accuracy gain and 6.7% mAP@0.5 increase on ExDark. Practical validation in real-world bank data center scenarios confirms its efficacy in smoke detection under low-light conditions, highlighting its critical value for infrastructure security where reliable monitoring ensures operational stability of financial systems.

## 1. INTRODUCTION

Dim-light object detection, the task of identifying and localizing objects in poorly illuminated environments, plays a pivotal role in applications such as autonomous driving, video surveillance, and critical infrastructure monitoring. In financial sectors, bank data centers demand uninterrupted security monitoring to safeguard operational stability, necessitating robust detection of intrusions or anomalies (e.g., smoke, unauthorized personnel) under nighttime or dimly lit conditions [1]. Despite advancements in artificial intelligence-driven object detection, existing methods—including two-stage frameworks (e.g., R-CNN [2], Faster R-CNN [3]) and one-stage detectors (e.g., YOLO [4], SSD [5])—exhibit significant performance degradation in low-light scenarios. This limitation arises from three intrinsic challenges: (1) illumination bias in training data, where models optimized for well-lit environments fail to generalize to low-light conditions; (2) feature degradation caused by noise amplification, contrast reduction, and detail loss in dim images [6-8]; and (3) computational inefficiency in hybrid enhancement-detection

pipelines, which hinders real-time deployment [9, 10].

Current approaches to mitigate these issues focus on image enhancement, algorithm optimization, and multimodal fusion [11-13]. However, methods like histogram equalization [14] or Retinex-based techniques [15] often introduce artifacts, while deep learning solutions (e.g., EnlightenGAN [16]) prioritize enhancement over detection efficiency. Furthermore, cascaded frameworks combining enhancement and detection modules incur redundant computations, limiting their practicality for time-sensitive applications such as data center security. Consequently, there is an urgent need for lightweight, integrated solutions that balance accuracy, generalization, and computational efficiency in low-light environments.

In this work, we propose Dim-ObjectDet, a novel framework addressing these challenges through three key innovations:

(1) Feature Reassembling and Fusion Module (FRFM) Head: A multi-scale detection head leveraging adaptive channel attention to dynamically fuse features across resolutions, enhancing robustness to scale variations in low-light targets.

(2) Inverted Residual Efficient Multi-Scale Attention (iREMA): A hybrid mechanism combining inverted residual blocks (iRMB) and cross-dimensional attention to strengthen local-global feature interactions while suppressing noise.

(3) Computational Optimization: Strategic integration of Diverse Branch Blocks (DBB) and dynamic upsampling (DySample) to improve feature diversity and inference speed without sacrificing accuracy.

Our experiments demonstrate excellent performance on benchmark datasets (DarkFace, ExDark) and real-world bank data center scenarios. The framework's efficiency and accuracy advancements underscore its potential for real-time security systems in critical infrastructure.

## 2. RELATED WORK

Dim-light object detection aims to leverage computer vision technologies to automatically identify, locate, and track target objects in low-light environments through image or video data. The core challenge lies in extracting effective information from faint optical signals while mitigating the adverse effects of insufficient illumination to achieve precise perception and analysis. Its applications span multiple domains, including security surveillance, autonomous driving, military reconnaissance, and industrial inspection. For instance, in nighttime or dimly-lit environments such as warehouses, parking lots, and streets, dim-light object detection assists surveillance systems in recognizing personnel and vehicles, enabling timely identification of abnormal behaviors and safety hazards. In autonomous driving scenarios, vehicles must detect pedestrians, traffic signs, and other targets under low-light conditions (e.g., nighttime or tunnels) to ensure driving safety. In military operations, this technology enhances nighttime mission execution, while in industrial settings, it supports quality inspection during precision manufacturing and semiconductor production under controlled low-light conditions to maintain product qualification rates.

Despite its critical value, practical implementations of dim-light object detection face significant challenges. The low intensity of optical signals in dim environments amplifies image noise, and complex backgrounds further degrade detection accuracy. Most object detection algorithms, trained under normal illumination conditions, underperform in low-light scenarios. Additionally, insufficient illumination leads to loss of target details, rendering shape and texture features ambiguous or inconsistently represented across varying dim-light conditions. To address these issues, researchers have proposed a series of improvement schemes to enhance detection performance.

Early studies focused on image enhancement preprocessing to improve input quality. Traditional methods relied on physical imaging models for signal recovery: Histogram equalization [17, 18] enhances contrast by stretching grayscale distributions but risks local overexposure; Wavelet transform [19, 20] separates frequency-domain noise and signals but depends on basis function selection; Retinex theory-based methods [21-23] decompose illumination and reflection components by simulating human visual mechanisms but suffer from halo artifacts. Machine learning-based enhancement algorithms include deep autoencoders for natural low-light image enhancement [24, 25] and multi-branch CNNs for dark-to-bright image conversion, such as MBLLEN [13],

PENet (employing Laplacian pyramid decomposition for multi-resolution components) [26], and lightweight fast illumination adaptive transformers for restoring RGB images from low-light or overexposed conditions [15, 16].

Recent advancements prioritize computational efficiency and physical interpretability. Qiao and Chen [27] proposed a low-light enhancement method combining signal-to-noise ratio (SNR)-aware Transformers and convolutional models, adaptively adjusting long- and short-range operations based on regional SNR while suppressing noise in extremely low-SNR regions via a novel self-attention mechanism. Jiang et al. [28] introduced EnlightenGAN, an unsupervised generative adversarial network for low-light enhancement without paired training data. Ma et al. [29] developed a self-calibrated illumination learning framework using cascaded weight-sharing strategies and self-calibration modules for efficient and robust enhancement.

Algorithmic optimizations specifically targeting dim-light detection have also emerged. Gonzales et al. [30] explored multi-scale Retinex and color restoration algorithms, proposing a novel color constancy metric for enhancement evaluation. Lore et al. [31] designed a stacked sparse denoising autoencoder trained on synthetic low-light data to adaptively enhance images while avoiding oversaturation in high dynamic range scenarios. Yang et al. [32] developed LightingNet, an ensemble method integrating a Vision Transformer (ViT) subnetwork for local high-level feature extraction and a complementary learning subnetwork for global fine features via transfer learning.

Dataset and training strategy optimization are critical for improving generalization. Large-scale datasets encompassing diverse illumination conditions, scenarios, and objects are essential. Chen et al. [33] and Li et al. [34] constructed a low-light imaging dataset containing short-exposure low-light images and corresponding long-exposure references, proposing an end-to-end CNN to process raw sensor data, bypassing traditional imaging pipelines.

Multimodal fusion has emerged to overcome single-modality limitations. Infrared-visible fusion leverages thermal radiation to compensate for texture loss in dark regions, employing dual-stream networks with attention mechanisms for spatial alignment and channel-weighted fusion. LiDAR collaboration provides geometric constraints from 3D point clouds to aid 2D target localization. Recent efforts explore temporal fusion using optical flow features in video sequences to enhance single-frame robustness. However, semantic gaps from cross-modal heterogeneity, dynamic weight allocation, and computational efficiency remain challenges.

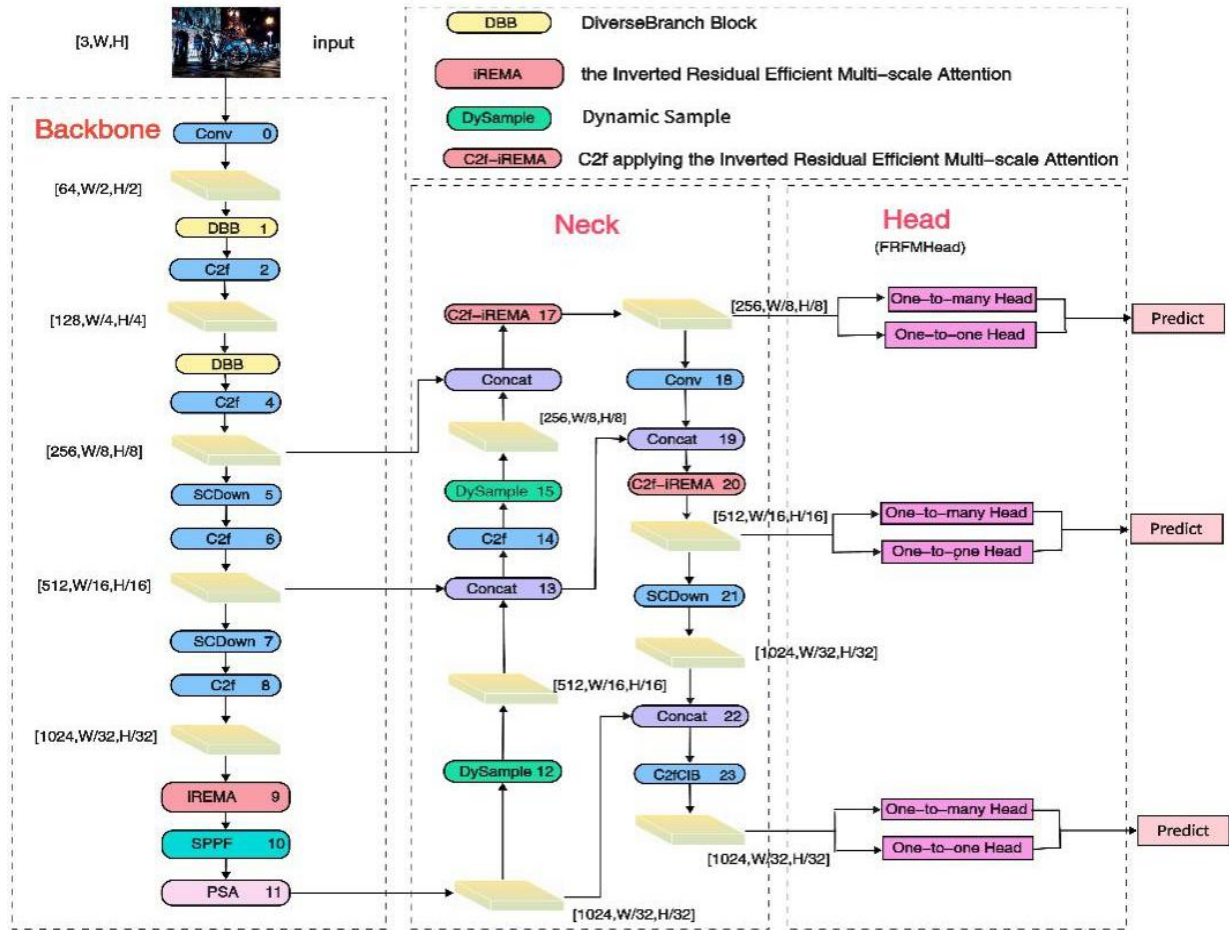
Despite progress, key limitations persist: 1) Cascaded enhancement-detection frameworks incur redundant computations, hindering real-time deployment; 2) Single-modality enhancement struggles in extreme low-light conditions; 3) Cross-modal fusion networks suffer from high parameter counts, limiting embedded system feasibility. Thus, balancing detection accuracy with computational efficiency remains a critical challenge in dim-light object detection.

## 3. OUR DIM-OBJECTDET METHOD

To address the dual challenges of feature degradation in low-light conditions and computational efficiency constraints, this study proposes the Dim-ObjectDet algorithm for low-light object detection. As illustrated in Figure 1, the algorithm

builds upon the YOLOv10 single-stage detection framework through synergistic innovations in feature representation enhancement and computational architecture optimization,

achieving balanced improvements in detection accuracy and inference efficiency.



**Figure 1.** Overall framework of the proposed Dim-ObjectDet

### 3.1 Overview

This study mainly achieves innovation in two key aspects.

First, at the feature reorganization level, we design a Feature Recombination and Fusion Module head (FRFM-head) that reconstructs multi-scale feature maps through convolutional-pooling joint operations. This module employs a lightweight architecture that significantly reduces GPU memory consumption while maintaining a moderate parameter increase. Constructed exclusively with fundamental operators (convolution and pooling), it eliminates additional computational overhead typically incurred by traditional preprocessing methods.

Second, for network architecture enhancement, we implement three key improvements in both backbone and neck components:

1) The proposed Inverted Residual Multi-scale Attention (iREMA) mechanism enhances dark region feature representation through convolutional-channel attention fusion, integrated into C2f modules to form C2f\_iREMA units;

2) Adoption of Diverse Branch Blocks (DBB) optimizes spatial convolution kernel combinations, enabling parallel extraction of fine-grained texture features through heterogeneous kernels;

3) Incorporation of the Dynamic Upsampling Strategy (DySample) replaces fixed interpolation functions with learnable weighting parameters, effectively reducing

upsampling latency while preserving feature continuity.

### 3.2 Technical details

#### 3.2.1 iREMA

To address blurred target features and low contrast in dim-light environments, this study proposes a Feature Recombination and Fusion Module Head (FRFMHead). Its core architecture comprises three components: Distribution Focal Loss (DFL), Pooling-ReLU-Convolutions Composite Module (PRC2), and the Feature Recombination and Fusion Module (FRFM).

The DFL loss dynamically adjusts the spatial sensitivity of the loss function to enhance the model's focus on edge features of dim-light targets. Given the predicted feature map  $P \in \mathbb{R}^{H \times W \times C}$  and ground-truth label  $Y$ , the loss function is defined as:

$$\mathcal{L}_{DFL} = - \sum_{i=1}^{H \times W} \sum_{c=1}^C w_i^{(c)} \cdot Y_i^{(c)} \log \sigma(P_i^{(c)}) \quad (1)$$

where,  $w_i^{(c)}$  is a weight coefficient dynamically calculated based on prediction confidence, and  $\sigma$  denotes the Sigmoid activation function. This design adaptively emphasizes low-response regions (e.g., dark smoke contours) to mitigate gradient sparsity issues in traditional loss functions under dim-

light scenarios.

The PRC2 module enhances feature robustness through multi-scale pooling and nonlinear mapping. As shown in Figure 2(a), the input feature map first extracts spatial statistical features via parallel max-pooling and average-pooling layers, followed by cross-interaction through dual convolutional paths. This structure fuses local extremum responses with global smooth features, effectively suppressing dark-light noise while preserving structural information of targets.

The FRFM module dynamically integrates multi-scale features to improve semantic consistency, as illustrated in Figure 2(b). For input multi-level feature pyramids, resolution alignment is performed via upsampling and downsampling operations, followed by channel attention-generated fusion weight matrices. Final features are aggregated through weighted summation. This mechanism enables adaptive selection of effective features across scales, significantly improving recall rates for multi-scale targets in dim-light smoke detection.

### 3.2.2 iREMA

To enhance global perception and local detail retention of target features in dim-light environments, this study proposes

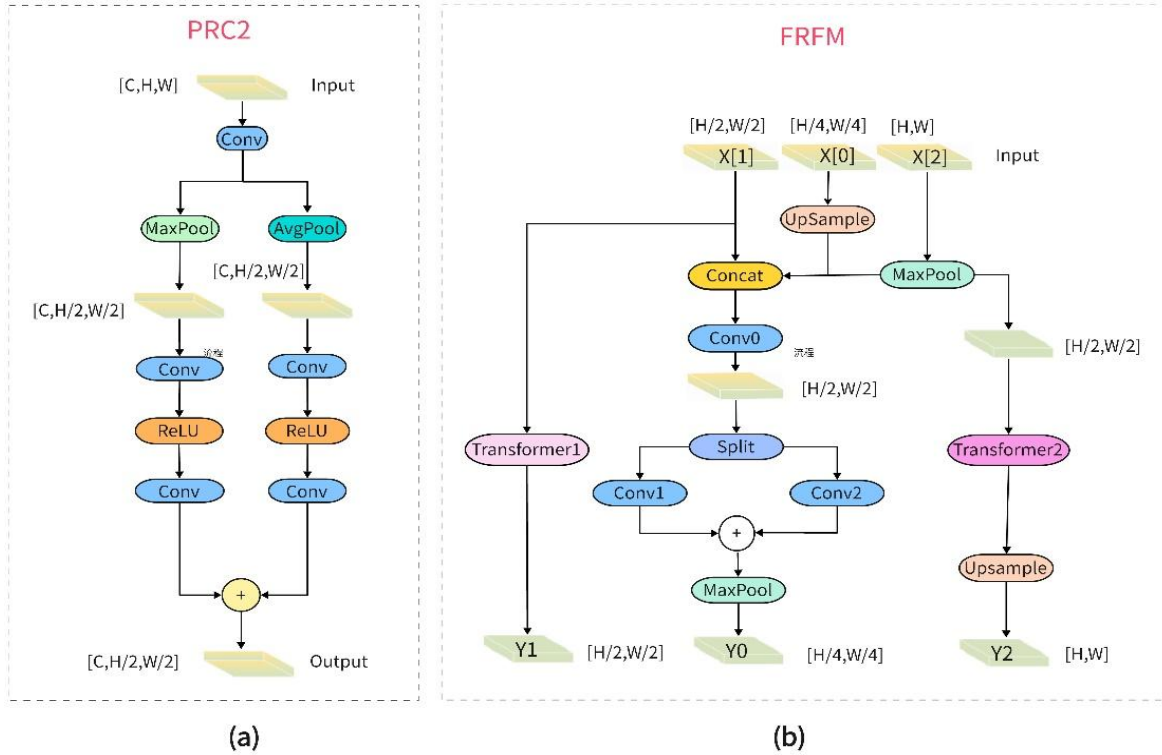
an iREMA mechanism. This approach synergistically integrates the dynamic modeling advantages of the Inverted Residual Mobile Block (iRMB) [35] with the cross-dimensional interaction capabilities of EMA Attention [23], forming a lightweight feature enhancement module. Its structure is illustrated in Figure 3.

The iRMB module combines the efficiency of CNN-based local feature modeling with the global dependency capture capability of Transformers. Specifically: Local Feature Extraction: Static convolution extracts local texture details. Dynamic Global Dependency Modeling: A triple-stage "expand-transform-compress" process integrates long-range spatial dependencies via Multi-Head Self-Attention (MHSA):

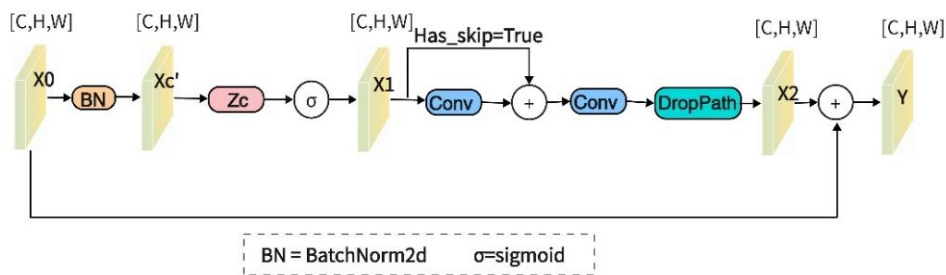
$$X_{\text{out}} = \text{Proj}_{\text{down}} \left( \mathcal{F}(\text{Proj}_{\text{up}}(X)) \right) + X \quad (2)$$

where,  $\text{Proj}_{\text{up}}$  and  $\text{Proj}_{\text{down}}$  are respectively the channel expansion and compression projection. The intermediate transformation layer contains parallel paths of static convolution and multi-head self-attention:

$$\mathcal{F}(X) = \text{Conv}(X) + \text{MHSA}(X) \quad (3)$$



**Figure 2.** The PRC2 and FRFM structures



**Figure 3.** The structure of the iREMA mechanism



EMA [36] is a lightweight attention module optimized for multi-scale features. Through the channel grouping reshaping and cross-dimensional interaction mechanism, the adaptive feature calibration of channel and space dimensions is realized while reducing the computational complexity. EMA adopts a dual-path parallel processing strategy. The global path extracts spatial statistical features through horizontal/vertical pooling to capture the macroscopic distribution of the target. The local path uses depth-wise separable convolution to enhance the detail response such as edge texture. The cross-scale feature interaction is realized by matrix dot product, and the spatially sensitive attention weight map is generated. This mechanism significantly improves the robustness of the model to low signal-to-noise ratio features by dynamically focusing on key regions.

iREMA builds a feature enhancement pathway by cascading iRMB with EMA modules. The input features were extended through the iRMB module to expand the channel dimension and fuse the local-global features. We then apply batch normalization to the intermediate features as follows:

$$\hat{X} = \gamma \cdot \frac{X - \mu}{\sigma} + \beta \quad (4)$$

where,  $\mu, \sigma$  are the batch statistics and  $\gamma, \beta$  are the learnable parameter. Then the spatial sensitive weight matrix is generated by the EMA module, and the enhanced features are output by the residual connection:

$$Y = \text{DropPath}(A \odot \hat{X}) + X \quad (5)$$

As shown in Figure 4, the iREMA module is embedded into the C2f base unit to construct the C2f-iREMA structure. The original C2f module fuses multi-scale features by cross-stage connection, and the iREMA mechanism is introduced to realize dynamic calibration of features after improvement. Specifically, the input features were extracted through the basic convolution layer to extract the primary features. Shallow fine layers are fused through series-parallel branch structures Section and deep semantic features; The iREMA module implements attention modulation on the fused features and outputs the optimized multi-scale feature pyramid.

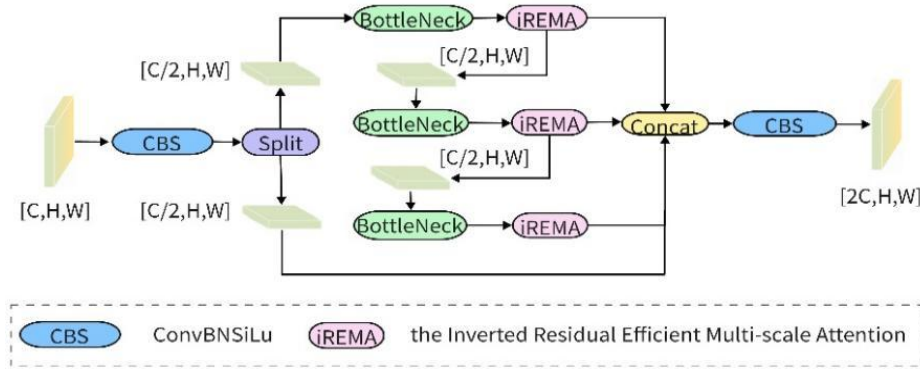


Figure 4. The structure of C2f-iREMA

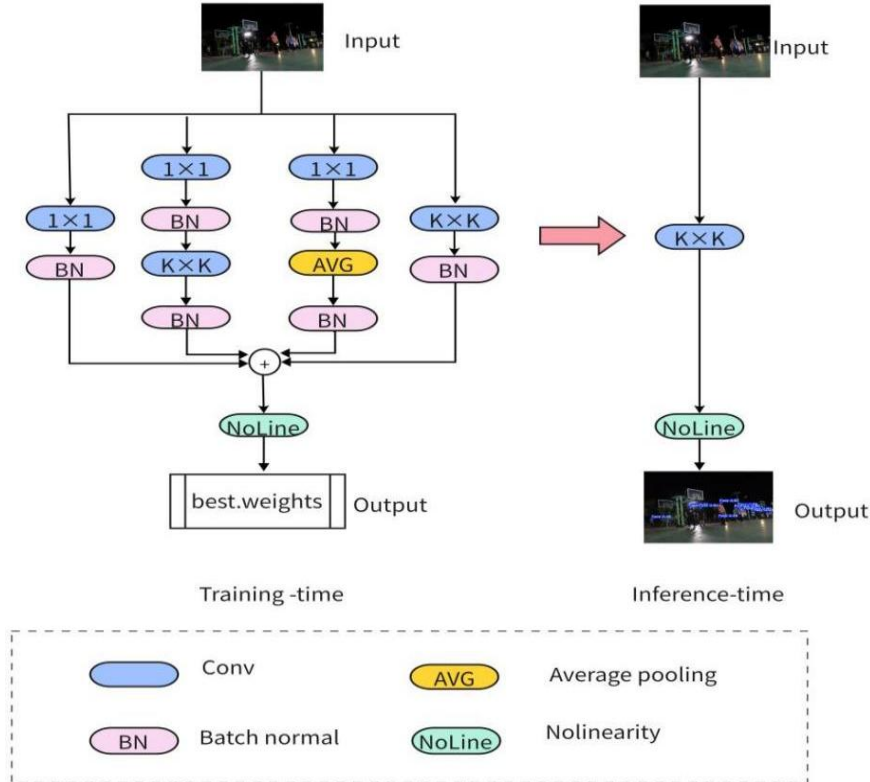


Figure 5. DBB architecture for training and inference

### 3.2.3 Convolution and upsampling feature extraction capability enhancement

To enhance the diversity of feature representation and computational efficiency in low-light scenarios, this study proposes a dual-path optimization strategy: introducing the Diverse Branch Block (DBB) [24, 37] into the backbone network to enhance the convolutional feature extraction capability, and introducing an upsampling module based on dynamic point resampling in the neck of the detection model. Efficiency is improved through a three-stage optimization mechanism.

The basic principle of DBB is to increase the complexity of the convolutional layer during the training phase by introducing convolutional branches of different sizes and structures to enrich the network's feature representation capability. DBB combines branches of different scales and complexities in structure, such as convolutional kernels of different sizes and average pooling operations, to enhance the feature representation capability of a single convolution, as shown in Figure 5. During the training phase, DBB adopts a complex branch structure, while during the inference phase, these branches can be equivalently transformed into a single convolutional layer to maintain efficient inference. DBB can also be inserted as a substitute for conventional convolutional layers in existing networks without altering the overall network structure.

The DBB module expands the representation space of the convolutional layer through a multi-branch structure, thereby enhancing the model's ability to capture the blurred edges and low-frequency textures of dim-light targets. Its mathematical essence is based on the linear additivity and homogeneity of the convolution operation, expressed as follows:

$$\begin{aligned}\mathcal{F}_1(X) + \mathcal{F}_2(X) &= (W_1 + W_2) * X \\ \alpha \mathcal{F}(X) &= (\alpha W) * X\end{aligned}\quad (6)$$

where,  $W$  represents the convolution kernel weights and  $\alpha$  is the scaling factor. Through the structural reparameterization technique, DBB maintains the exact same computational cost as the original single-branch convolution during the inference phase, while obtaining better feature representation capabilities through multi-branch learning during the training phase. In this study, DBB is used as a modular component to replace the standard convolutional layers in the backbone network, enhancing the gradient response intensity of dark area targets through multi-scale feature interaction.

To address the edge blurring issue of traditional upsampling operators in low-light scenarios, this study employs the dynamic point resampler DySample to optimize the resolution recovery of neck features [38]. The core of this approach involves three stages: dynamic offset prediction, adaptive range constraint, and lightweight interpolation calculation, which significantly improves the geometric fidelity of feature resolution recovery. Given the input feature map, the dynamic offset generation module learns local structural information through a lightweight linear layer and outputs the raw offset tensor. To limit the offset magnitude and prevent feature misalignment, a Sigmoid function and a learnable range factor are introduced to normalize the offsets: Finally, position-aware interpolation is performed on the input features based on the dynamic offsets. The calculation process can be expressed as:

$$\Delta = \text{Linear}(X) \quad (7)$$

$$\Delta_{\text{norm}} = \sigma(\Delta) \cdot \gamma \quad (8)$$

$$X_{\text{up}} = \sum_k w_k \cdot X(p_k + \Delta_{\text{norm}}) \quad (9)$$

where,  $w_k$  is the bilinear interpolation weight calculated from the distance of neighboring pixels, and is the preset regular grid coordinate. Compared with the traditional dynamic convolution method, Dysample abandons the complex kernel generation process and directly predicts the offset using a linear layer, reducing the computational complexity. At the same time, it alleviates the edge blurring effect by dynamically focusing on the target contour area.

## 4. EXPERIMENTS AND RESULTS ANALYSIS

To systematically evaluate the performance of the target detection framework based on multimodal fusion and low-light enhancement, this study takes YOLOv10 as the baseline model and conducts multi-dimensional verification experiments on three representative low-light target detection datasets: Smoke, DarkFace and ExDark. Firstly, cross-model comparison experiments are conducted to verify the effectiveness and advantages of the proposed method. Then, ablation experiments are carried out to verify the independent contributions of each module. Finally, further experiments are conducted to verify the proposed multimodal fusion method to fully explore the complementarity of different modal information and improve the target detection performance under low-light conditions.

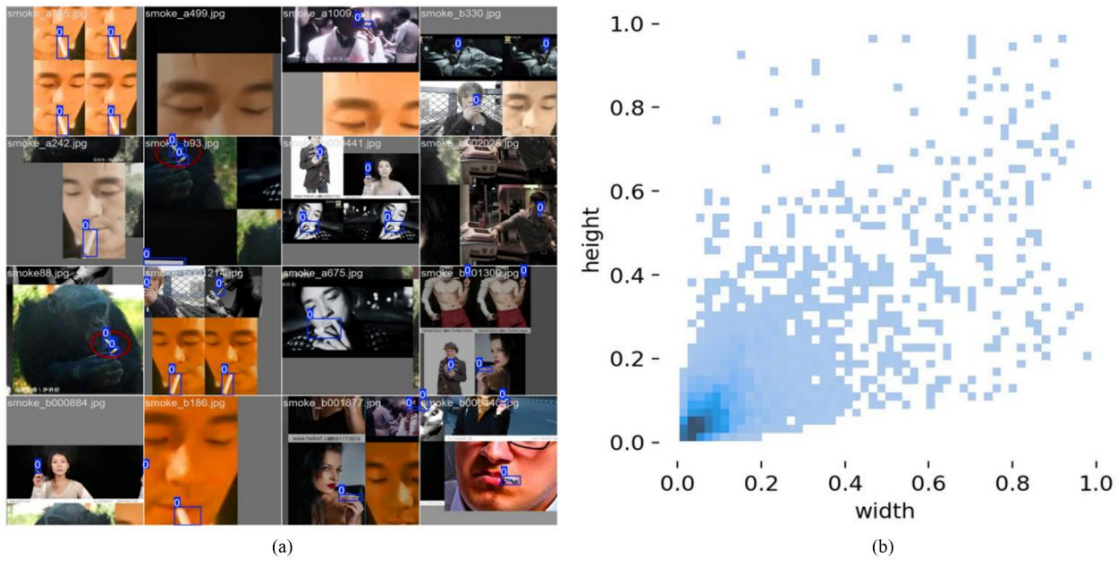
### 4.1 Experimental setup and datasets

This experiment uses a model based on the Ultralytics YOLOv10 architecture, deployed on a computing platform with NVIDIA Ampere architecture, equipped with an RTX 3090 24GB GDDR6X graphics card and an AMD EPYC 7R32 16-core CPU, and builds a mixed-precision training environment. The software environment is based on the PyTorch 2.0.1 framework, with CUDA 11.8 acceleration library and Python 3.10.12 programming language, and data pipeline optimization is achieved through TorchVision 0.15.2. During training, the SGD optimizer is used, with a momentum of 0.937, an initial learning rate of 0.01 combined with a cosine annealing strategy for dynamic adjustment, a weight decay coefficient of 0.0005, and a batch size of 16. The input images are uniformly scaled to a resolution of 640x640 during training, and the training period is set to 100 epochs.

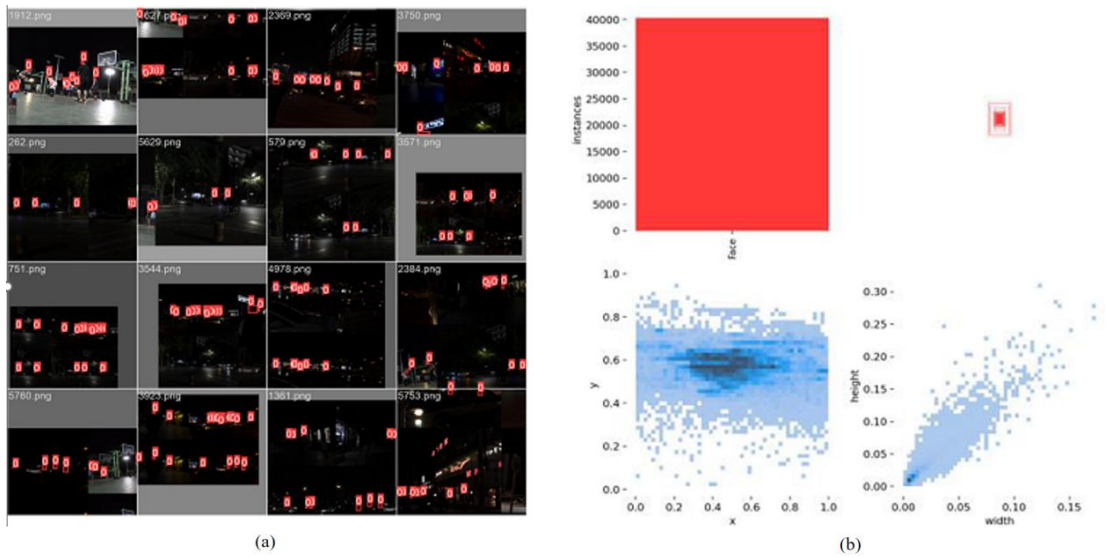
The low-light target detection algorithm based on the FRFM fusion mechanism uses the Smoke, DarkFace [39], and ExDark [40] datasets to build a multi-dimensional verification system.

(1) The Smoke dataset focuses on the financial security scene and covers smoking scenes under normal and low-light backgrounds, as shown in Figure 6. It contains 5731 images, with 4585 images in the training set and 1146 images in the validation set.

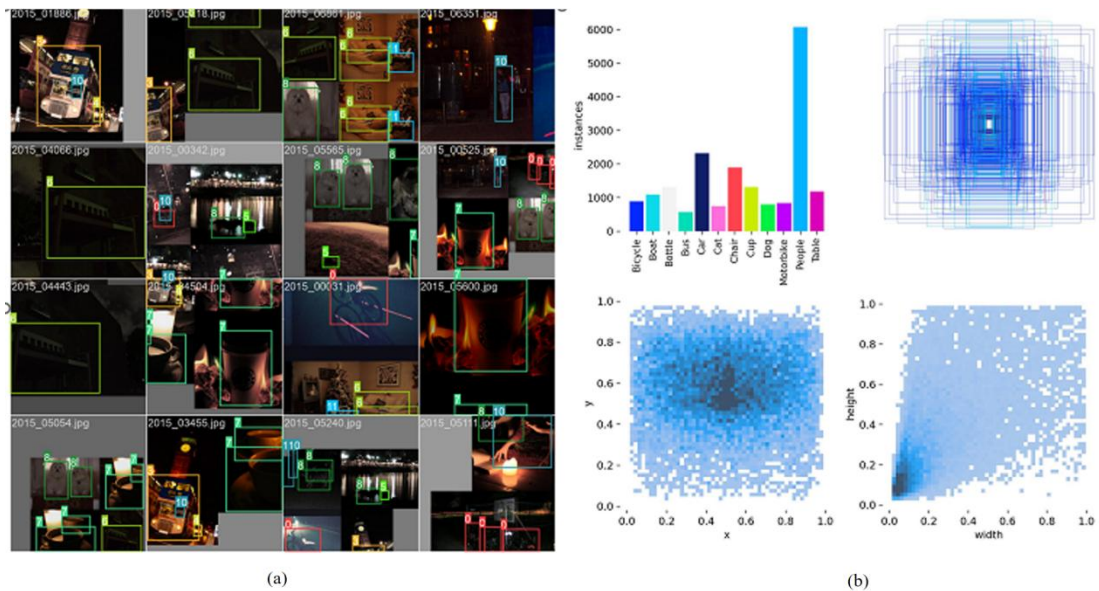
(2) The DarkFace dataset is specifically designed for low-light face detection and is mainly captured at night in teaching buildings, streets, bridges, overpasses, and parks, as shown in Figure 7. It contains 6000 low-light images in the real world, with 4800 images in the training set and 1200 images in the validation set. The image resolutions range from 640×48 pixels to 1920×1080 pixels.



**Figure 6.** Mosaic augmentation and distribution of detected objects on the Smoke dataset



**Figure 7.** Mosaic augmentation and detection object distribution on the DarkFace dataset



**Figure 8.** Mosaic augmentation and detection object distribution on the ExDark dataset

(3) ExDark, as a benchmark dataset for low-light detection, is specifically constructed for low-light target detection, as shown in Figure 8. It contains a total of 7363 labeled images, with 5891 images for training and 1472 images for validation. It includes 12 types of targets and covers 10 different lighting conditions from extremely low-light environments to night environments.

## 4.2 Comparative experiments and performance advantages

The dim light target detection algorithm based on the Feature Relationship Fusion Module (FRFM) proposed in this chapter has been systematically verified on the bank operation and maintenance scene simulation dataset Smoke, as well as the representative datasets DarkFace and ExDark in the field of dim light scene target detection. Firstly, a comparative experiment was conducted on the Smoke dataset, and the experimental data are shown in Table 1.

**Table 1.** Comparison experiments on the Smoke dataset

Models	Precision	Recall	mAP@0.5	mAP@0.5:0.95
YOLOv10n	0.776	0.733	0.811	0.455
YOLOv10n-ours	0.808	0.741	0.835	0.491
YOLOv10s	0.826	0.804	0.878	0.532
YOLOv10s-ours	0.851	0.823	0.899	0.554

From the table, it can be seen that the precision, recall, mAP@0.5, and mAP@0.5:0.95 of the improved model have all increased. Among them, the improvement in mAP@0.5 and mAP@0.5:0.95 is 3% and 8% compared to the YOLOv10n baseline model, and 2.4% and 4.1% compared to the YOLOv10s baseline model, respectively. This proves the effectiveness of the improved model in enhancing detection accuracy under different IoU thresholds.

Based on the above experiments, further comparative experiments were conducted on representative low-light target detection datasets such as DarkFace and ExDark. The experimental results are shown in Table 2. The experiments indicate that the innovation of the target detection model in dim lighting scenarios is not only applicable to specific face target detection tasks but also effectively increases the relevant detection indicators for multi-target detection tasks. From the data in the table, it can be seen that the improved model has a significant improvement over the baseline models. On the DarkFace dataset, mAP@0.5 is improved by 7.1% and 4% compared to the original YOLOv10n and YOLOv10s,

respectively, and mAP@0.5:0.95 is improved by 13.1% and 8%, respectively, indicating that the improvement strategy effectively enhances the detection accuracy at high IoU thresholds. On the more complex ExDark dataset, mAP@0.5 (0.492) is improved by 6.6% and 4.1% compared to the original versions, and mAP@0.5:0.95 is improved by 6.5% and 3.4%, respectively, indicating that the improvement strategy can also effectively enhance the detection accuracy at different IoU thresholds in more complex scenarios.

This study further verified the performance advantages of the proposed YOLOv10 series models in object detection tasks through comparative analysis. The experiment selected the current mainstream single-stage detector RT-DETR-ResNet50 as the control model, which achieved excellent performance on the COCO dataset. However, from the perspective of engineering deployment, the computational complexity (86MB) and parameter scale (126GFLOPs) of this model are significantly higher than the lightweight YOLOv10 architecture proposed in this paper, which makes it face higher computing power requirements and energy consumption constraints in practical applications.

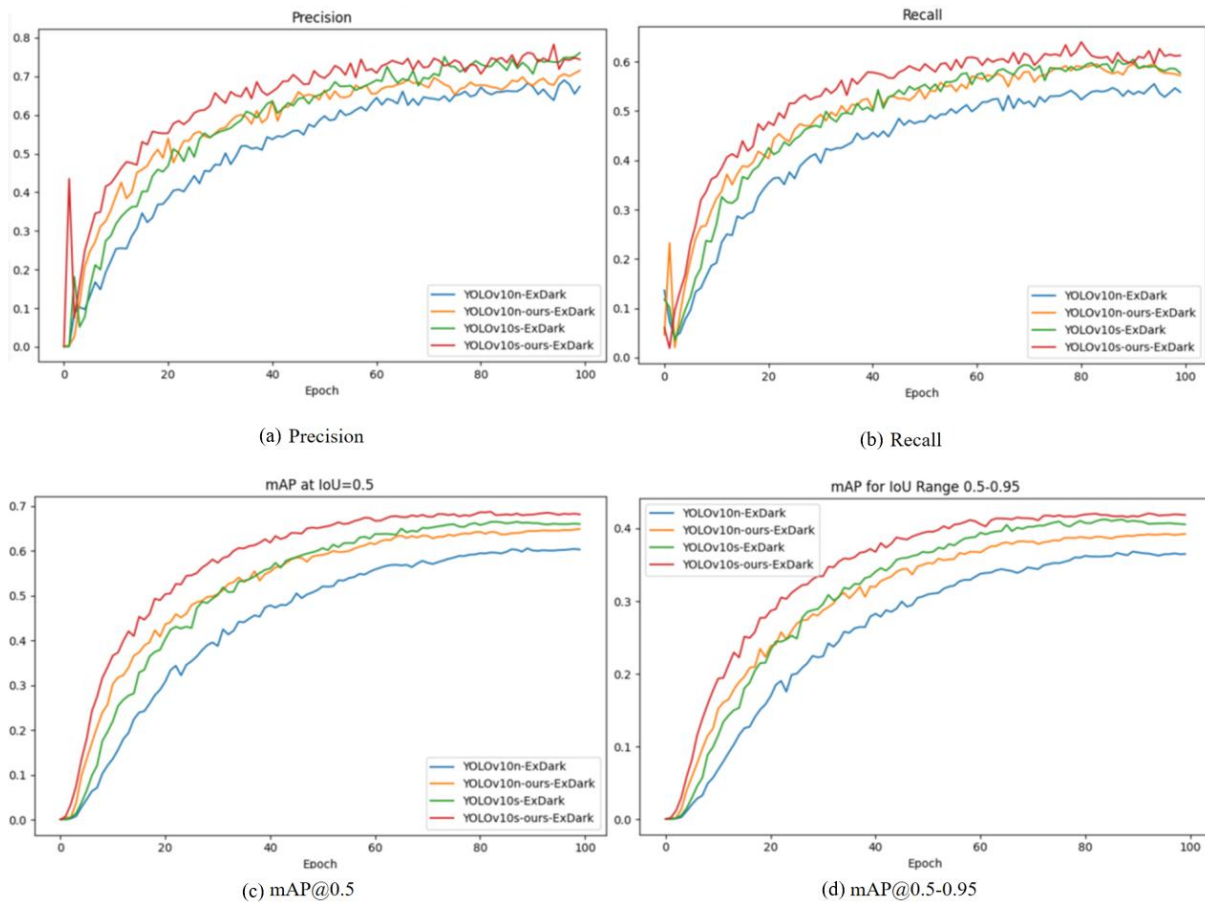
To systematically evaluate the learning characteristics of the model, this study conducted comparative experiments on the DarkFace and ExDark two low-light object detection benchmark datasets. As shown in Figure 9 and Figure 10, the training curves indicate that within 100 training cycles on the single-object DarkFace dataset, the improved model shows significant advantages over the baseline YOLOv10 series in core metrics such as precision, recall, mAP@0.5, and mAP@0.5:0.95. The gap between the corresponding curves gradually expands and then remains stable as the training iterations progress. Notably, on the multi-object ExDark dataset, although the trend of the four metrics is similar to that of DarkFace, the improved model not only has a more distinct curve separation from the baseline model but also surpasses the YOLOv10s version with a larger parameter scale in key metrics such as precision and mAP@0.5. This cross-level performance breakthrough confirms that the improved algorithm enhances the detection accuracy while improving the model's generalization ability, especially in dim-light scenarios.

Figure 11 shows the detection results of the original YOLOv10s model and the improved model with the innovative strategy applied in this study on the DarkFace, ExDark, and Smoke datasets under the same experimental conditions and parameters. (a) is the original image, (b) is the result of the original model (YOLOv10s), and (c) is the result of the improved YOLOv10s model.

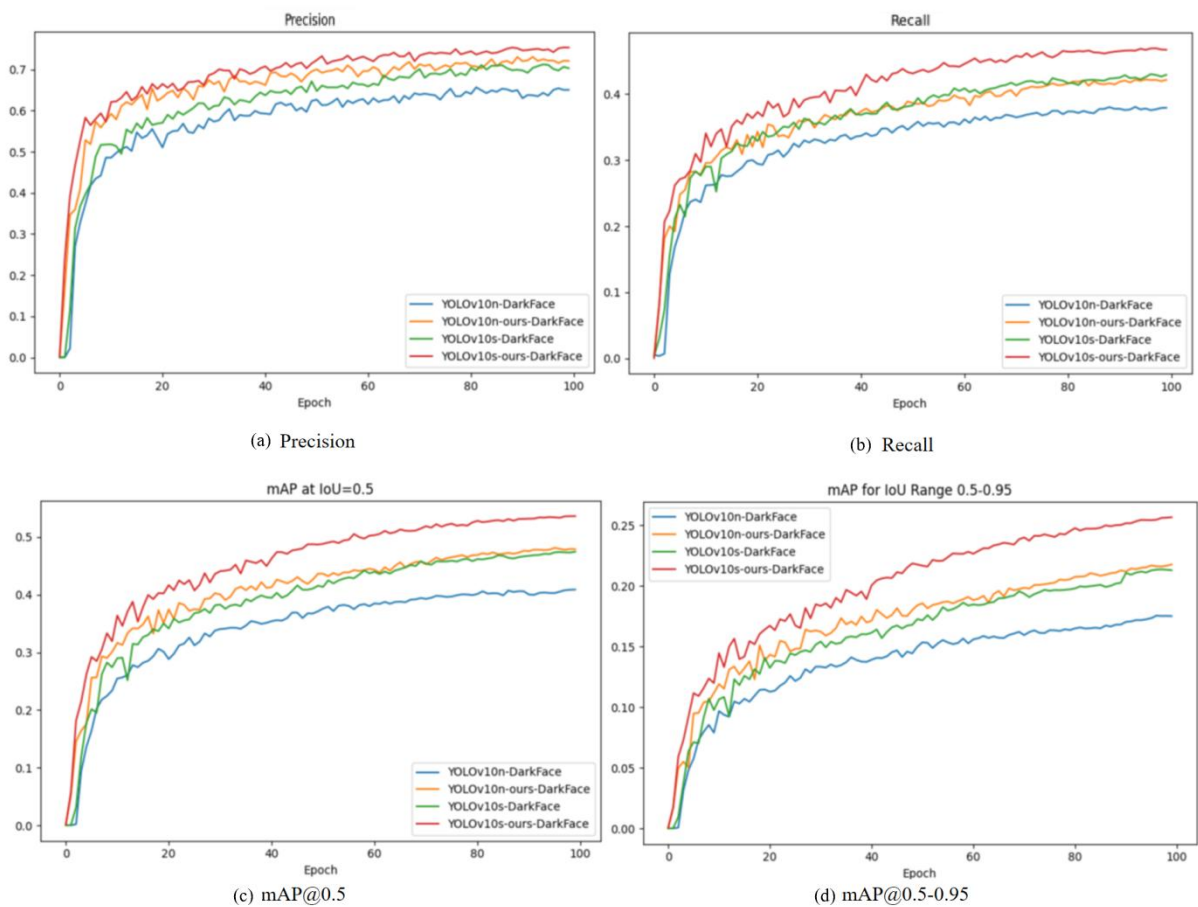
**Table 2.** Comparison experiments on DarkFace and ExDark datasets

Datasets	Models	Precision	Recall	mAP@0.5	mAP@0.5:0.95
DarkFace	YOLOv10n	0.649	0.38	0.409	0.175
	YOLOv10n-ours	0.701	0.392	0.438	0.198
	YOLOv10s	0.699	0.429	0.473	0.214
	YOLOv10s-ours	0.731	0.436	0.492	0.231
	RT-DETR-ResNet50	0.567	0.389	0.408	0.160
ExDark	YOLOv10n	0.656	0.556	0.604	0.368
	YOLOv10n-ours	0.724	0.579	0.644	0.392
	YOLOv10s	0.736	0.594	0.658	0.409
	YOLOv10s-ours	0.764	0.602	0.685	0.423
	RT-DETR-ResNet50	0.706	0.567	0.627	0.389

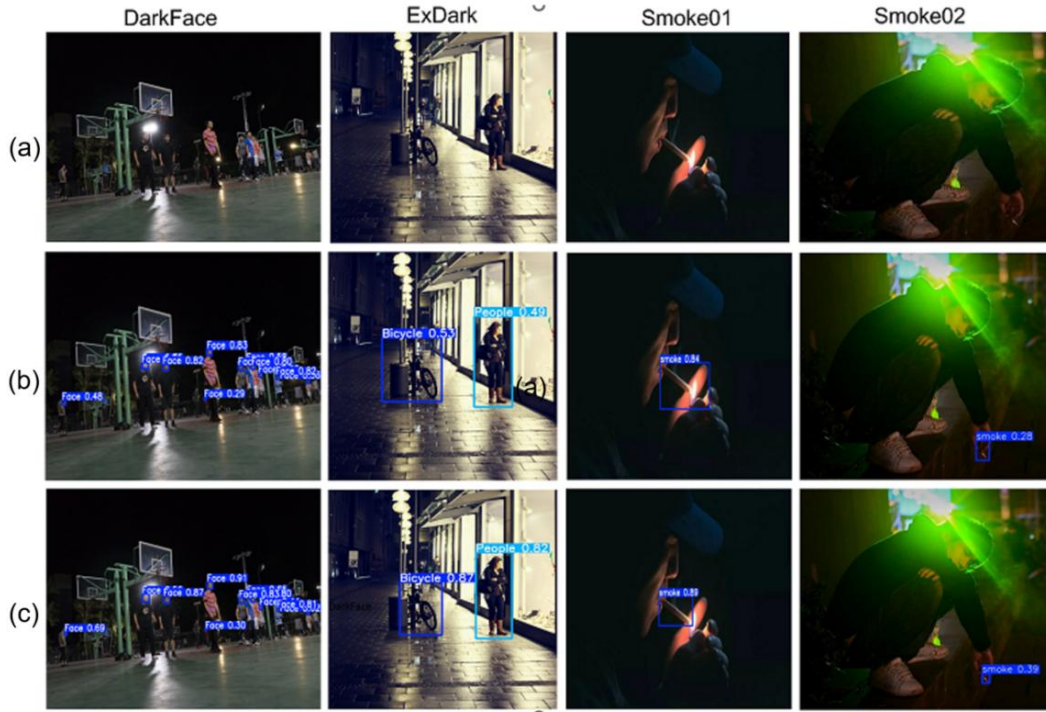




**Figure 9.** Comparison of training data on the ExDark



**Figure 10.** Comparison plot of the training data on the DarkFace



**Figure 11.** Detection comparison of the original model and the innovative model

By comparing the detection results on Darkface, it can be seen that this model can detect more targets more accurately in low-light scenarios, such as distant faces. The detection accuracy for the same target is also higher. The experimental results on ExDark show that in the dimly lit scenes with multiple targets, the detection accuracy of this model has been effectively improved, especially for bicycles and pedestrians. On the Smoke01 dataset, (b) is the detection result of YOLOv10S with an accuracy of 0.84; (c) is the improved detection result with an accuracy of 0.89. The Smoke02 dataset is a situation with dim lighting and small targets. The accuracies of the original model (b) and the improved model (c) are 0.28 and 0.39 respectively. In the above scenarios, the detection effect of the innovative model is better.

#### 4.3 Ablation experiment and module performance verification

The dim light target detection algorithm based on the FRFM fusion mechanism adopts YOLOv10n as the baseline framework and quantitatively evaluates the innovations of each module through the control variable method. The ablation experiment results on the DarkFace dataset are shown in Table 3.

In the Table, B, C, D, and E correspond to the experiments conducted by applying the DBB, DySample, iREMA, and FRFMhead innovative strategies respectively to the baseline model. As can be seen from the table, the DBB strategy enhances the feature extraction capability with a slight increase in computational cost. Compared with the baseline model, all indicators have significantly improved; mAP@0.5 and mAP@0.5:0.95 have increased by 5.4% and 5.7% respectively. DySample and iREMA maintain approximately the same computational cost, with mAP@0.5 increasing by about 2.4% and 2.2% respectively. FRFMhead innovates the detection head, accompanied by a significant increase in computational cost, and the precision, mAP@0.5, and mAP@0.5:0.95 indicators have significantly improved, with increases of 7.4%, 6.1%, and 12% respectively, while the recall rate has slightly increased. F, G, H, I, and J respectively integrate various innovative strategies into the baseline model, further improving the detection indicators. Among them, J integrates the four innovative methods into the original model, and the detection indicator values are significantly improved. The precision, mAP@0.5, and mAP@0.5:0.95 indicators are the best, but the recall rate is lower than that of B, F, G, and I. Considering all aspects of the indicators comprehensively, J is ultimately selected as the model innovation scheme.

**Table 3.** Ablation experiments

Models	Y	D	S	R	F	Params (M)	FLOPs (G)	Precision	Recall	mAP@0.5	mAP@0.5:0.95
A	√					5.7	8.2	0.649	0.38	0.409	0.175
B	√	√				5.8	9.1	0.674	0.396	0.431	0.185
C	√		√			5.8	8.2	0.66	0.389	0.419	0.178
D	√			√		5.9	8.4	0.657	0.384	0.418	0.177
E	√				√	27	16.4	0.697	0.385	0.434	0.196
F	√	√	√			5.9	9.1	0.664	0.402	0.434	0.19
G	√	√		√		6.0	9.2	0.662	0.396	0.43	0.184
H	√		√	√		6.0	8.4	0.657	0.392	0.424	0.182
I	√	√	√	√		6.0	9.2	0.682	0.396	0.436	0.193
J	√	√	√	√	√	28	17.9	0.701	0.392	0.438	0.198

Note: Header fields Y=YOLOv10n, D=DBB, S=Dysampl, R=iREMA, F=FRFHead

## 5. CONCLUSIONS

The dim light target detection algorithm Dim-ObjectDet addresses the issues of feature degradation and computational efficiency bottlenecks in low-light environments by constructing a multi-level feature enhancement and multi-modal fusion framework. The feature reorganization and fusion module head (FRFMHead) integrates the distribution caustic loss (DFL), the pooling-activation-convolution composite module (PRC2), and the feature reorganization and fusion module (FRFM). It enhances the contrast of low-light targets through loss function constraints and dynamic feature reorganization. The inverse residual efficient multi-scale attention mechanism (iREMA) combines the dynamic modeling capability of the inverse residual mobile block (iRMB) with the cross-dimensional interaction characteristics of the EMA attention, and is embedded in the C2f unit to form the C2f-iREMA structure, achieving the collaborative optimization of local and global features. The dual-path optimization strategy introduces diverse branch blocks (DBB) in the backbone network to enhance feature extraction capabilities, and uses a dynamic point resampling upsampling module in the neck to improve computational efficiency through a three-stage optimization process. The lightweight dynamic sampling module (Dysample) directly predicts offsets using a linear layer to reduce computational complexity and alleviate the edge blurring effect.

The relevant experiments were conducted on the dim-light single-modal datasets of Smoke, DarkFace, and ExDark, which are representative of the bank operation and maintenance site, to systematically verify the proposed method. These advancements validate the framework's efficacy in real-world scenarios, particularly in bank data center smoke detection, where reliable low-light monitoring ensures operational safety. Future work will focus on lightweight deployment via neural architecture search (NAS) and quantization-aware training, aiming to reduce model complexity by 20–40% while retaining detection accuracy for edge-device applications.

## REFERENCES

- [1] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580-587. <https://doi.org/10.1109/CVPR.2014.81>
- [2] Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1440-1448. <https://doi.org/10.1109/ICCV.2015.169>
- [3] Ren, S., He, K., Girshick, R., Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137-1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
- [4] He, K., Zhang, X., Ren, S., Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9): 1904-1916. <https://doi.org/10.1109/TPAMI.2015.2389824>
- [5] Kang, S.H., Park, J.S. (2023). Aligned matching: Improving small object detection in SSD. *Sensors*, 23(5): 2589. <https://doi.org/10.3390/s23052589>
- [6] Fu, C.Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C. (2017). DSSD: Deconvolutional single shot detector. *Conference on Computer Vision and Pattern Recognition*, arXiv:1701.06659. <https://doi.org/10.48550/arXiv.1701.06659>
- [7] Li, Z., Yang, L., Zhou, F. (2017). FSSD: Feature fusion single shot multibox detector. *Computer Vision and Pattern Recognition*, arXiv preprint arXiv:1712.00960. <https://doi.org/10.48550/arXiv.1712.00960>
- [8] Redmon, J., Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 7263-7271. <https://doi.org/10.1109/CVPR.2017.690>
- [9] Redmon, J., Divvala, S., Girshick, R., Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 779-788. <https://doi.org/10.1109/CVPR.2016.91>
- [10] Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-7475.
- [11] Zimmerman, J.B., Pizer, S.M., Staab, E.V., Perry, J.R., McCartney, W., Brenton, B.C. (1988). An evaluation of the effectiveness of adaptive histogram equalization for contrast enhancement. *IEEE Transactions on Medical Imaging*, 7(4): 304-312. <https://doi.org/10.1109/42.14513>
- [12] Kim, Y.T. (1997). Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE Transactions on Consumer Electronics*, 43(1): 1-8. <https://doi.org/10.1109/30.580378>
- [13] Chen, S.D., Ramli, A.R. (2003). Minimum mean brightness error bi-histogram equalization in contrast enhancement. *IEEE Transactions on Consumer Electronics*, 49(4): 1310-1319. <https://doi.org/10.1109/TCE.2003.1261234>
- [14] Ding, X.M. (2010). The study of image enhancement algorithm based on wavelet transform. Anhui University.
- [15] Zhou, C., Yin, A., Li, Z., Huang, Y. (2016). Partial differential equation based image edge enhancement and its application in measuring shaft. In *2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, Xi'an, China, pp. 972-978. <https://doi.org/10.1109/URAI.2016.7734121>
- [16] Kim, J.H., Kim, J.H., Jung, S.W., Noh, C.K., Ko, S.J. (2011). Novel contrast enhancement scheme for infrared image using detail-preserving stretching. *Optical Engineering*, 50(7): 077002. <https://doi.org/10.1117/1.3597639>
- [17] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 2117-2125. <https://doi.org/10.1109/CVPR.2017.106>
- [18] Guo, C., Fan, B., Zhang, Q., Xiang, S., Pan, C. (2020). AugFPN: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

- Seattle, WA, USA, pp. 12595-12604. <https://doi.org/10.1109/CVPR42600.2020.01261>
- [19] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, pp. 8759-8768. <https://doi.org/10.1109/CVPR.2018.00913>
- [20] Deng, J., Pan, Y., Yao, T., Zhou, W., Li, H., Mei, T. (2020). Single shot video object detector. IEEE Transactions on Multimedia, 23: 846-858. <https://doi.org/10.1109/TMM.2020.2990070>
- [21] Gong, Y., Yu, X., Ding, Y., Peng, X., Zhao, J., Han, Z. (2021). Effective fusion factor in FPN for tiny object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, pp. 1160-1168. <https://doi.org/10.1109/WACV48630.2021.00120>
- [22] Akyon, F.C., Altinuc, S.O., Temizel, A. (2022). Slicing aided hyper inference and fine-tuning for small object detection. In 2022 IEEE international conference on image processing (ICIP), Bordeaux, France, pp. 966-970. <https://doi.org/10.1109/ICIP46576.2022.9897990>
- [23] Bosquet, B., Cores, D., Seidenari, L., Brea, V.M., Mucientes, M., Del Bimbo, A. (2023). A full data augmentation pipeline for small object detection based on generative adversarial networks. Pattern Recognition, 133: 108998. <https://doi.org/10.1016/j.patcog.2022.108998>
- [24] Khan, S.A., Hussain, S., Yang, S. (2020). Contrast enhancement of low-contrast medical images using modified contrast limited adaptive histogram equalization. Journal of Medical Imaging and Health Informatics, 10(8): 1795-1803. <https://doi.org/10.1166/jmihi.2020.3196>
- [25] Dhal, K.G., Das, A., Ray, S., Gálvez, J., Das, S. (2021). Histogram equalization variants as optimization problems: A review. Archives of Computational Methods in Engineering, 28: 1471-1496. <https://doi.org/10.1007/s11831-020-09425-1>
- [26] Liu, J., Ma, Y., Meng, X., Zhang, S., Liu, Z., Song, Y. (2024). Enhancing intrinsic image decomposition with transformer and Laplacian pyramid network. Traitement du Signal, 41(1): 511. <https://doi.org/10.18280/ts.410146>
- [27] Qiao, S., Chen, R. (2024). Progressive feature fusion for SNR-aware low-light image enhancement. Journal of Visual Communication and Image Representation, 100: 104148. <https://doi.org/10.1016/j.jvcir.2024.104148>
- [28] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Wang, Z. (2021). EnlightenGAN: Deep light enhancement without paired supervision. IEEE Transactions on Image Processing, 30: 2340-2349. <https://doi.org/10.1109/TIP.2021.3051462>
- [29] Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z. (2022). Toward fast, flexible, and robust low-light image enhancement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, pp. 5637-5646. <https://doi.org/10.1109/CVPR52688.2022.00555>
- [30] Gonzales, A.M., Grigoryan, A.M. (2015). Fast Retinex for color image enhancement: Methods and algorithms. In Proceedings Volume 9411, Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2015, pp. 94110F. <https://doi.org/10.1117/12.2083546>
- [31] Lore, K.G., Akintayo, A., Sarkar, S. (2017). LLNet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition, 61: 650-662. <https://doi.org/10.1016/j.patcog.2016.06.008>
- [32] Yang, S., Zhou, D., Cao, J., Guo, Y. (2023). LightingNet: An integrated learning method for low-light image enhancement. IEEE Transactions on Computational Imaging, 9: 29-42. <https://doi.org/10.1109/TCI.2023.3240087>
- [33] Chen, C., Chen, Q., Xu, J., Koltun, V. (2018). Learning to see in the dark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, pp. 3291-3300. <https://doi.org/10.1109/CVPR.2018.00347>
- [34] Li, C., Guo, C., Han, L., Jiang, J., Cheng, M.M., Gu, J., Loy, C.C. (2021). Low-light image and video enhancement using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12): 9396-9416. <https://doi.org/10.1109/TPAMI.2021.3126387>
- [35] Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J. (2022). Decoupled knowledge distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Louisiana, USA, pp. 11953-11962. <https://doi.org/10.1109/CVPR52688.2022.01165>
- [36] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C. (2017). Learning efficient convolutional networks through network slimming. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, pp. 2736-2744. <https://doi.org/10.1109/ICCV.2017.298>
- [37] Ding, X., Zhang, X., Han, J., Ding, G. (2021). Diverse branch block: Building a convolution as an inception-like unit. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, pp. 10886-10895. <https://doi.org/10.1109/CVPR46437.2021.01074>
- [38] Liu, W., Lu, H., Fu, H., Cao, Z. (2023). Learning to upsample by learning to sample. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, pp. 6027-6037. <https://doi.org/10.1109/ICCV51070.2023.00554>
- [39] DARK FACE: Face Detection in Low Light Condition, 2019. <https://flyywh.github.io/CVPRW2019LowLight>.
- [40] Exclusively Dark Image Dataset, 2022. <https://github.com/cs-chan/Exclusively-Dark-Image-Dataset>.