RESEARCH ARTICLE

# FID-YOLO: A pedestrian detection model integrating multispectral information in complex environments

**Di Yang**[1,2,3], **Xilong Zhang**[1,2,3], **Peng Wang**[1,2,3]*

**1** School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China, **2** Jilin Provincial Joint Key Laboratory of Big Data Science and Engineering, Changchun, China, **3** Jilin Provincial Science and Technology Innovation Center of Network Database Application Software, Changchun, China

* wangpeng@cust.edu.cn

Check for updates

OPEN ACCESS

**Citation:** Yang D, Zhang X, Wang P (2026) FID-YOLO: A pedestrian detection model integrating multispectral information in complex environments. PLoS One 21(3): e0342054. https://doi.org/10.1371/journal.pone.0342054

**Editor:** Lianmeng Jiao, Northwestern Polytechnical University, CHINA

**Received:** December 26, 2024

**Accepted:** January 17, 2026

**Published:** March 5, 2026

**Copyright:** © 2026 Yang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data availability statement:** All relevant code and resources supporting this study are publicly available in our GitHub repository https://github.com/ZhangXiLong1/FID-YOLO.

## Abstract

The advancement of pedestrian detection technology is of great importance for various applications such as intelligent driving, object tracking, and robot navigation. Many studies in this field have demonstrated that image quality significantly contributes to the precision of detection. However, unexpected factors such as adverse weather, occlusions, and scale variations, which extremely weaken the main features of the detected objects, leading to a decrease in detection accuracy. To address these problems, we propose a Feature-enriched Image Detection-YOLO (FID-YOLO), to improve pedestrian detection performance in complex environments by integrating visible and infrared light information. Specifically, we design an illumination-aware image fusion module for visible and infrared image information fusion to generate a new image within more information to enrich pedestrian features. Then, a cascaded feature aggregation module using reparameterization and channel shuffle is introduced to enhance the model's understanding and generalization capabilities for complex scenes. Furthermore, we exploit a scale-adaptive feature detection head for YOLO detector, which solves the problem of detecting small objects at varying object scales. Experiments on M3FD and LLVIP datasets demonstrate that FID-YOLO outperforms the benchmark models in pedestrian detection. Additionally, we validate the indispensability of each proposed module through ablation experiments.

## Introduction

Accurately detecting and localizing pedestrians in complex environments is a critical task in computer vision. The frequent occurrence of pedestrian traffic accidents has brought great pressure to traffic safety, high-precision pedestrian detection to meet the need in applications such as autonomous driving systems and road safety

**Funding:** Jilin Provincial Science and Technology Development Planned Project (Grant No. YDZJ202401621ZYTS) and Jilin Provincial Science and Technology Innovation Center of Network Database Application Software (Grant No. YDZJ202302CXJD027). Meanwhile, the funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

monitoring, is highly desiderated [1]. With the continuous advancement in pedestrian detection technology, it significantly contributes to the reduction of traffic accidents and the enhancement of pedestrian safety, which in turn increases road traffic efficiency and promotes the development of intelligent transportation systems.

The quality of the images directly impacts the accuracy of pedestrian detection [2]. The captured images are often affected by complex environmental factors, such as occlusions and small objects in low illumination conditions. Fig 1 respectively shows pedestrians under various challenging conditions: at night, in smoky environments, with mutual occlusions, and in low-resolution images captured on rainy days. Such complex conditions significantly reduce the accuracy of pedestrian detection. Therefore, optimization of pedestrian detection in these challenging environments is extremely crucial for enhancement of pedestrian detection performance.

Detecting pedestrians within dynamic and changeable environments is a major challenge in the field of computer vision. Over the past few decades, a variety of methods is employed to conduct diverse studies on object detection. Prior to the end-to-end detection, many researches constructed by traditional vision measurement, such as histogram of oriented gradients (HOG) [3,4], local binary pattern (LBP) [5] and aggregate channel features (ACF) [6], these methods based on hand crafted features and classifiers to detect objects. Although widely adopted in the early stages, however, the predictive accuracy critically depends on the quality of these hand-designed features. This often leads to poor generalization of the models when dealing with pedestrian objects in complex environment.

Currently, the advantages of deep learning are becoming increasingly prominent. Compared to traditional vision measurement methods, deep learning has proven highly effective at capturing the complex features and variable behavior of pedestrians, resulting in widespread use and significant benefits. Specifically, pedestrian detection approaches can be categorized into one-stage and two-stage methods. One-stage object detection methods simultaneously detect and classify objects in an image by using anchor boxes to predict bounding boxes and class labels directly. These methods have advantages in high efficiency but may sacrifice accuracy. Examples include SSD [7,8], YOLO [9–11], and RetinaNet [12]. Two-stage object detection methods first generate a set of proposals, typically using a Region Proposal Network (RPN), and then refine these proposals to predict final bounding boxes and class labels. Although these methods generally achieve higher accuracy, they often involve longer processing times. Such as R-CNN [13], SPPNet [14], Faster R-CNN [15], Mask R-CNN [16]. One-stage and two-stage methods each offer distinct trade-offs between speed and accuracy, making them applied to different application scenarios. On the basis of the foregoing, researchers have further improved object detection accuracy by incorporating auxiliary technologies, most previous researches have utilized image fusion [17], feature pyramid networks [18–21], transfer learning [22–24], and attention mechanisms [25,26] to augment deep learning methods in addressing the challenges of feature information loss, modality imbalance, and position offset. Specifically, the feature pyramid network is constructed by a multi-scale feature representation integrating feature maps at various levels, effectively enhancing the ability to detect pedestrians of different sizes. The top-down pathway




**Fig 1. Objects in complex environments (a) Nighttime small pedestrians detection (b) Smoky condition pedestrians (c) Nighttime occlusion detection (d) Rainy occluded object detection.**

https://doi.org/10.1371/journal.pone.0342054.g001

and lateral connections ensure an effective integration of high-level semantic information with low-level details, improving the precision of object localization. While transfer learning applies knowledge from pre-trained models to related but different tasks, accelerating model convergence and enhancing robustness. Additionally, the attention mechanism in pedestrian detection enhances the model's ability to focus on critical features, thereby boosting its detection performance. Despite significant progress in pedestrian detection, there are still some challenges that deserve to be emphasized:

Firstly, image fusion could lead to information loss and color distortion, which may prevent‘ accurately reflect details and features of the original images. Meanwhile, due to the low resolution of pedestrian images in low light conditions, the edges of pedestrian objects become blurred. Therefore, it is significant to develop methods that effectively restore object boundaries and improve overall image quality. Secondly, occlusion in pedestrian detection could occur in two forms: one is occlusion between pedestrians, and the other is occlusion of pedestrians by other objects. These two types of occlusions have varying degrees of impact on the detector's ability to detect targets, potentially leading to missed detection and reduced accuracy. Moreover, the distance between pedestrians and the camera could cause changes in pedestrian scale, low-resolution pedestrians at smaller scales are more prone to missed detection. Thus, improving detection accuracy for small-scale pedestrians is crucial.

To overcome these limitations, we propose a feature-enriched image detection-YOLO (FID-YOLO), a pedestrian detection framework. This model adopts the YOLO architecture to improve image quality, and minimize detection errors. By integrating visible and infrared light information, the proposed method outperforms traditional detection techniques that

rely on a single light source, thereby improving the precision of pedestrian detection. The main contributions of this work are summarized as follows:

- An illumination-aware image fusion module is designed to integrate infrared and visible light information, generating a composite image with deep semantic information and fine-grained details. It employs spatial and channel attention mechanisms to focus on crucial details in both visible and infrared images during the fusion process, achieving a more comprehensive pedestrian feature representation.
- A cascaded feature aggregation module is designed for multi-scale feature fusion, distributing the fused features across different scales. This module enables the model to capture object information more comprehensively, including partially occluded instances. By integrating features from multiple scales, the module improves the model's ability to recognize and localize occluded objects, ensuring robust detection even when visibility is compromised.
- A scale-adaptive feature detection head is proposed to capture object features at different scales more effectively, enhancing the detection of small objects. This module achieves task alignment in both label assignment and detection by using a feature extractor that learns interactive features from multiple convolutional layers. The joint features improve inter-task interaction.
- The proposed method is validated on the LLVIP and M3FD datasets. Ablation experiments are also conducted to demonstrate the effectiveness of the proposed modules.

## Related works

### Occlusion in pedestrian detection

Among numerous visual recognition tasks, occluded object detection is particularly critical, as it involves effectively recognizing and tracking objects that are partially obscured by others. Li et al. [27] addressed the issue of feature confusion caused by occlusions in UAV imagery by proposing an occlusion estimation module (OEM) for precise occlusion localization. The author implemented occlusion-guided detection through multi-task interactions, effectively resolving the occlusion challenges in object detection for UAV images. Hao [28] proposed an anchor-free infrared pedestrian detection algorithm that enhances the detection performance of multiscale and partially occluded objects by designing a cross-scale feature fusion module. Additionally, a hierarchical attention mapping module is constructed to increase the significance of pedestrian features in complex environments while suppressing background information. Shi et al. [29] proposed a Global-Local Awareness Detector (GLA-D) to extract scale variance feature information from input frames, addressing the scale variations of objects moving in the scene and the frequent occlusions caused by complex scenes. Additionally, they introduced Occlusion Awareness Data Association (OADA), which uses different metrics for high- and low-scoring detection frames to alleviate occlusion issues in tracking scenarios.

### Adverse weather pedestrian detection

Pedestrian detection in adverse weather conditions faces challenges such as poor image quality, uneven lighting, and low contrast. These factors could lead to a decline in detection performance and increase the risk of traffic accidents. To address these issues, researchers have conducted extensive studies. Liu et al. [30] employed six differentiable filters to automatically adjust parameters based on the brightness and weather information from the input image. These adjustments aim to mitigate adverse factors affecting the image and restore its underlying content, thereby enhancing detection performance. Additionally, a mixed dataset comprising normal and low-quality images was used to ensure that IA-YOLO can adaptively handle varying weather conditions. Luo et al. [31,32] addressed the challenge of degraded image quality in adverse weather conditions impacting object detection accuracy. Using the LaLM method, which enhances precision by minimizing the discrepancy between degraded and clean images at the prediction level rather than the image level itself,

the researchers achieved superior detection accuracy and inference speed in experimental evaluations on foggy, rainy, and low-light scenarios.

### Small object detection

Small object detection is an advanced field in computer vision that focuses on identifying and analyzing small objects in images with limited details. Qin et al. [33] separate the target from the background based on the diversity of morphological features and consider the continuity of target motion in the time domain by using the RX filter to extract the target trajectory in random projection. Experiments on various cluttered background sequences validate the proposed method's detection capability in the field of infrared small target detection. Wang et al. [34] developed a new lightweight network model named HV-YOLOv8, which enhances the accuracy of small object detection by incorporating residual structures within the convolutional modules and by introducing the variety of view group shuffle cross stage partial network (VOV-GSCSP) module. This approach significantly reduces the number of parameters and computational requirements. Moreover, relying solely on visible light often provides insufficient information for pedestrian detection, such as nighttime and small-scale objects. Cao et al. [35] introduce the LG-FAPF network, an end-to-end system that leverages locality-guided cross-modal feature aggregation and pixel-level fusion to learn robust pedestrian representations. The network effectively encodes local responses and mutual cues from various instances into a unified descriptor and fuses visible and thermal information for accurate detection. Wei et al. [36] combined UNet and YOLO for visible and infrared fusion, performing object detection by sharing visible light information. These methods enhance detection accuracy through image fusion techniques. However, the generated pseudo-color images have some stain-like areas that can block the field of view, which may have impact on the accuracy of object detection.

Most of the researches mentioned above have been proposed to improve detection accuracy, it is still challenging under complex and adverse weather conditions. Therefore, integrating visible and infrared light for object detection is anticipated to be a major focus in the field.

## Methodology

In order to address the increasingly challenging traffic environments, we propose a pedestrian detection model, as depicted in Fig 2. The model consists of two main parts: the illumination-aware image fusion module and the object detection module.

The illumination-aware image fusion module, which builds upon the progressive infrared and visible images fusion network to generate a fusion image within more feature information. It involves replacing the channel attention module in the cross-modality differential aware fusion (CMDAF) with the channel and spatial attention module (CSAM). This adjustment enables the model to more effectively capture main features while enhancing its ability to suppress background noise, aim to improve prediction accuracy and generalization capabilities. Additionally, the loss function has been optimized by replacing the original L1 norm with the Huber loss, which enhances training efficiency and stability. The image fusion process produces a single image with complementary information, enhancing pedestrian features for more effective pedestrian detection.

The object detection module is built on the YOLOv8 network, with a cascaded feature aggregation and extraction (CFAE) module integrated between the backbone and neck to optimize feature fusion across scales. The CFAE module combines shallow and deep features through channel shuffling, improving feature integration without adding unnecessary complexity or resource demands. A reparameterization operation is introduced to reduce computational and memory costs during inference. Additionally, a scale-adaptive feature detection head (SAFDH) is designed to use shared convolutions, reducing parameters while enabling the feature extractor to learn interactive features across tasks, boosting detection accuracy in complex environments.

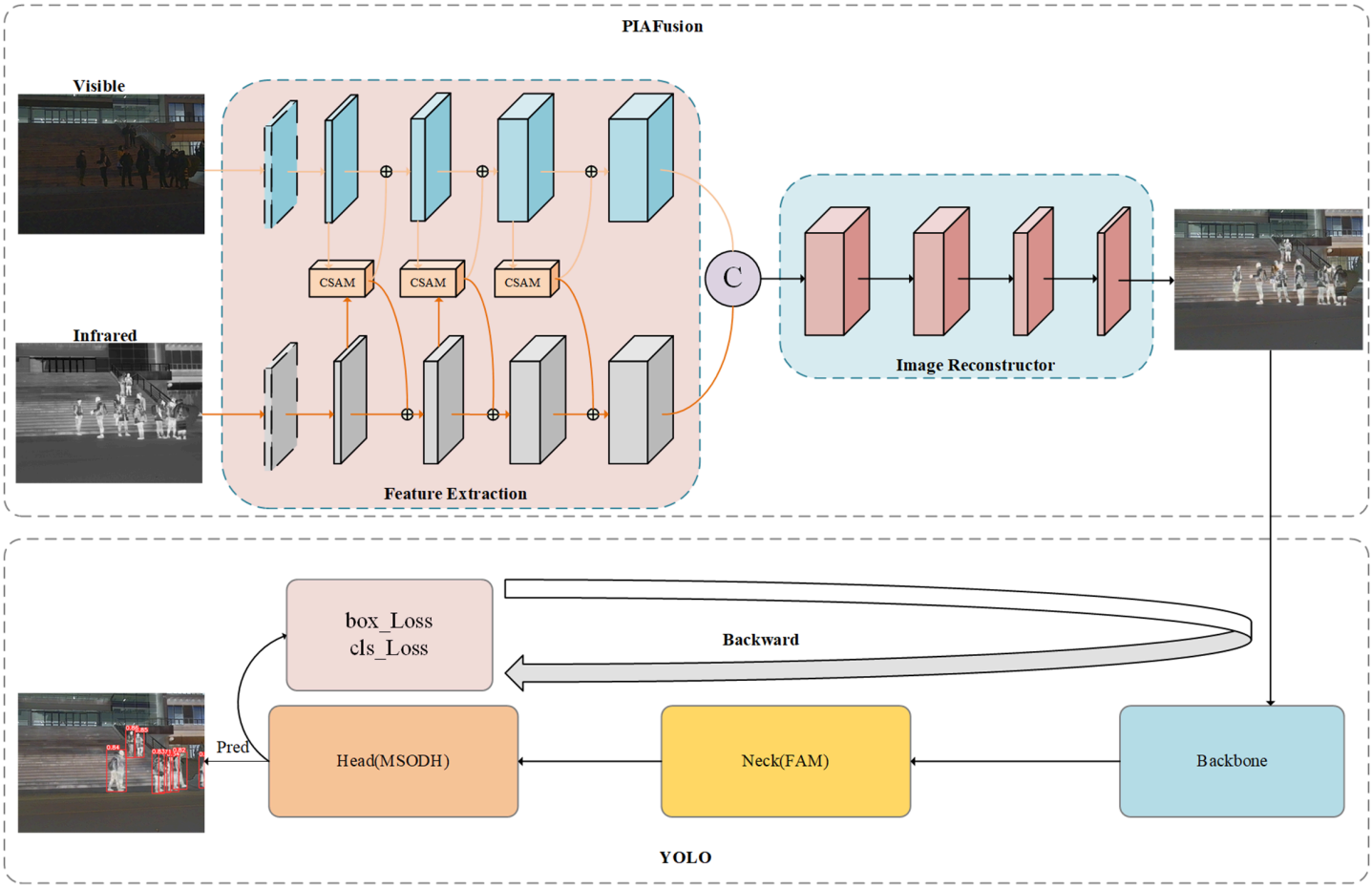



**Fig 2. The structure and workflow of the proposed FID-YOLO.**

https://doi.org/10.1371/journal.pone.0342054.g002

To be specific, Fig 3 is a detailed expansion of Fig 2, offering a more comprehensive illustration of the overall architecture of the pedestrian network. First of all, the progressive infrared and visible images as inputs are processed by the illumination-aware image fusion module to extract both deep and shallow features at multiple scales, then fused by differential computing to generate a fusion image with complementary features. Subsequently, the fusion image will be fed into object detection module, which consists of the backbone, the neck, and the head networks. In particular, the generated fusion image is inputted into backbone to learn multi-modal feature representations. Then, the obtained implicit representation is passed to the neck network for feature refinement and enhancement. Finally, the head network conducts the final pedestrian detection, outputting bounding boxes and classification results in complex environments.

### Illumination-aware image fusion module

Accurate detection of pedestrian objects primarily depends on the robustness of image features. However, existing methods mainly perform object detection based on single light source images, which can lead to blurred feature representations under low-light conditions, adverse weather, or occlusion. To preserve more comprehensive image features, we construct an illumination-aware image fusion module that integrates visible and infrared images. Visible images contain rich

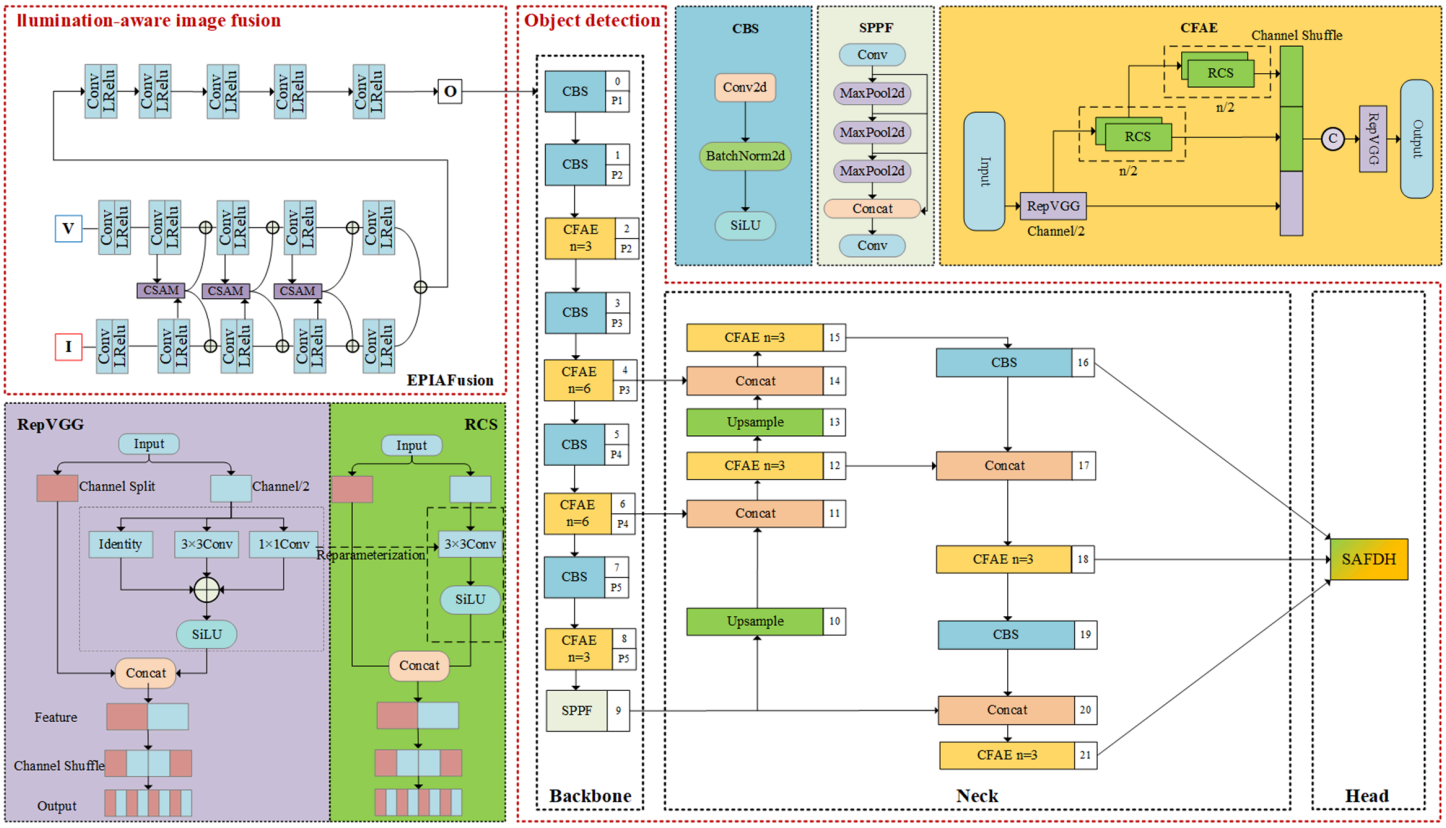



**Fig 3. Overall structure of infrared and visible light fusion object detection network.**

https://doi.org/10.1371/journal.pone.0342054.g003

texture and color information, whereas infrared images emphasize thermal targets. By connecting these two feature manifolds, the module enables complementary feature extraction, enhancing the distinction between background and object information while suppressing noise.

The proposed module employs an illumination-aware sub-network to incorporate environmental illumination as a prior for fusion, inspired by PIAFusion [37]. However, a critical theoretical limitation of PIAFusion is its reliance solely on channel attention mechanisms within its Cross-Modality Differential Aware Fusion (CMDAF). While channel attention effectively recalibrates the importance of different feature maps, it ignores the spatial distribution of features. In pedestrian detection, spatial details—such as limb positioning and edges—are crucial. To address this spatial information loss, we propose the Enhanced Progressive Infrared and Visible Image Fusion Network (EPIAFusion). The core innovation of EPIAFusion is the integration of a Channel and Spatial Attention Module (CSAM) into the differential fusion process. By simultaneously modeling inter-channel dependencies and inter-spatial relationships, CSAM allows the network to focus on the semantic content of the pedestrian while preserving the geometric details of the scene.

In EPIAFusion, a feature encoder is used to derive advanced features from both visible light and infrared light images, $I_{ir}$ represented as follows: $I_{vi}$ represents the visible light image, represents the infrared light image, $F_{vi}$ and $F_{ir}$ represent the visible light features and infrared light features respectively, and represent the feature extraction module. Fig 4 illustrates the structure of EPIAFusion. In EPIAFusion, a feature encoder is used to derive advanced features from both visible light and infrared light images, $I_{ir}$ represented as follows: $I_{vi}$ represents the visible light image, represents the infrared light image, $F_{vi}$ and $F_{ir}$ represent the visible light features and infrared light features respectively, and represent the feature

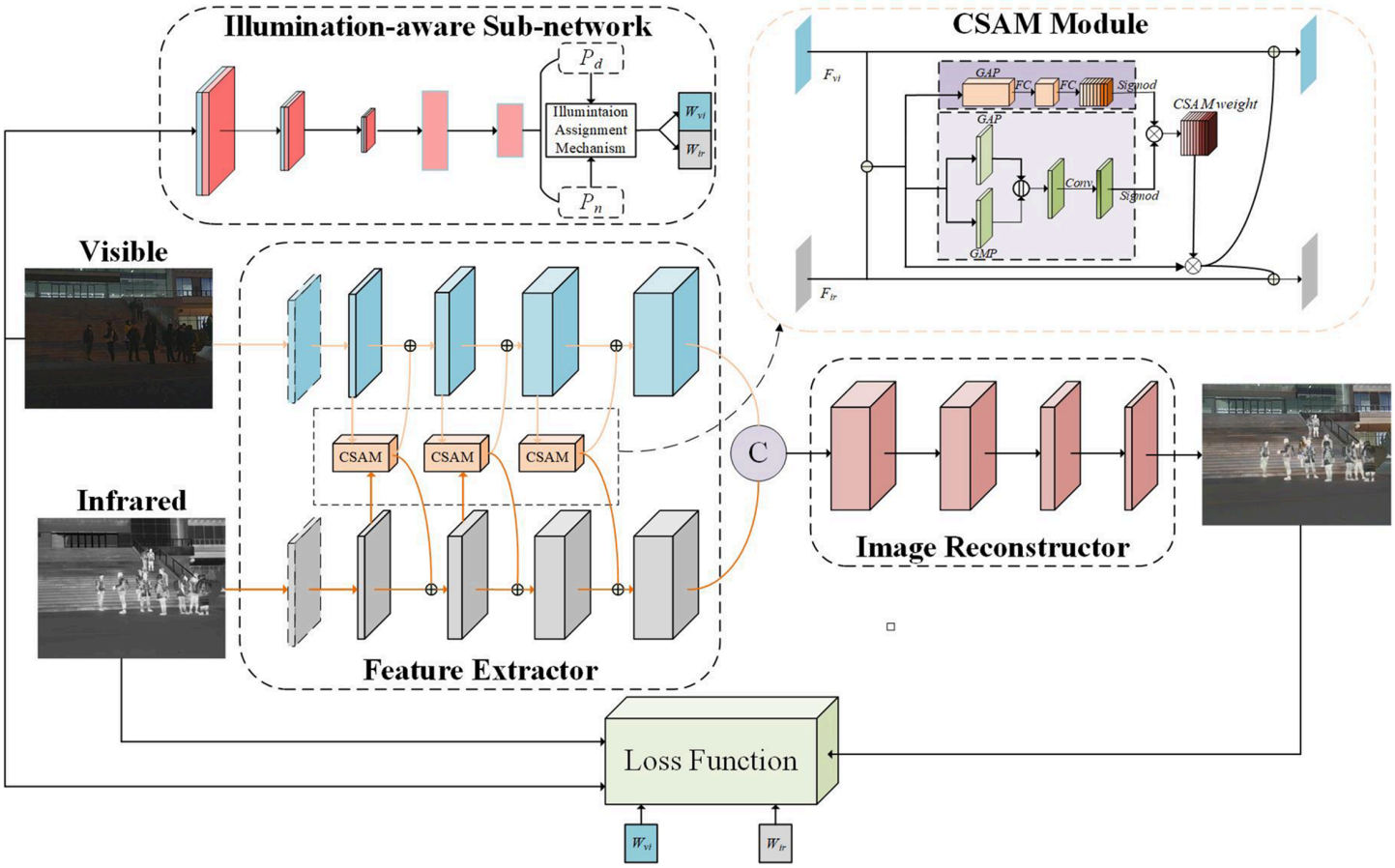



**Fig 4. The structure of EPIAFusion module.**

https://doi.org/10.1371/journal.pone.0342054.g004

extraction module.

$$\{F_{vi}, F_{ir}\} = \{E_F(I_{vi}), E_F(I_{ir})\} \tag{1}$$

To effectively fuse this differential information, the CSAM is applied as a compensation mechanism. Theoretically, simple addition of features introduces noise; CSAM acts as a dynamic filter. Eq (2) elucidates the principle of differential decomposition, where $F^i_{vi}$ and $F^i_{ir}$ denotes the features extracted by the i-th convolutional layer. By employing the channel and spatial attention module (CSAM) to compensate, the differential information between visible and infrared light can be effectively reconciled.

$$\begin{aligned} F^i_{vi} &= \frac{F^i_{vi}+F^i_{vi}}{2} + \frac{F^i_{ir}-F^i_{ir}}{2} = \frac{F^i_{vi}+F^i_{ir}}{2} + \frac{F^i_{vi}-F^i_{ir}}{2} \\ F^i_{ir} &= \frac{F^i_{ir}+F^i_{ir}}{2} + \frac{F^i_{vi}-F^i_{vi}}{2} = \frac{F^i_{ir}+F^i_{vi}}{2} + \frac{F^i_{ir}-F^i_{vi}}{2} \end{aligned} \tag{2}$$

Therefore, the CSAM module is defined as shown in Eq (3), where $\oplus$ represents element-wise summation, $\odot$ represents channel-wise multiplication, $\delta$, *GAP* and *GMP* respectively denote the sigmoid function, global average pooling, and global max pooling operations. By applying the sigmoid function to scale the generated weights to the range [0, 1],

the results from combining channel attention and spatial attention are added as modality compensation information to the original features, that mitigates the impact of background noise to outstand the main features for object detection.

To assign different weights to images for day and night condition, the light perception process in the light perception subnetwork is defined as follows: $P_d$ and $P_n$ represent the likelihood of whether an image is taken during the day or night. Since visible light images provide more detailed information about lighting conditions, the illumination probability is derived from these visible light images. This light probability is then used to calculate the light perception weights $w_{vi}$ for visible light and $w_{ir}$ for infrared.

$$\begin{aligned} \hat{F}_{vi}^{i} &= F_{ir}^{i} \oplus \delta\left(GAP\left(F_1\right)\right) \odot \delta\left(Conv\left[GAP\left(\bar{F}_1\right), GMP\left(\hat{F}_1\right)\right]\right) \odot F_1 \\ \hat{F}_{ir}^{i} &= F_{ir}^{i} \oplus \delta\left(GAP\left(F_2\right)\right) \odot \delta\left(Conv\left[GAP\left(\bar{F}_2\right), GMP\left(\hat{F}_2\right)\right]\right) \odot F_2 \end{aligned} \tag{3}$$

$$\begin{aligned} F_1 &= F_{vi}^{i} - F_{ir}^{i} \\ F_2 &= F_{ir}^{i} - F_{vi}^{i} \end{aligned} \tag{4}$$

$$\{P_d, P_n\} = N_{IA}\left(I_{vi}\right) \tag{5}$$

$$\begin{aligned} W_{ir} &= \frac{P_n}{P_d + P_n} \\ W_{vi} &= \frac{P_d}{P_d + P_n} \end{aligned} \tag{6}$$

Standard fusion networks often rely on L1 or L2 norms, both of which exhibit inherent limitations. The L2 norm is highly sensitive to outliers and tends to produce blurred edges, whereas the L1 norm may lead to unstable gradient updates when the error approaches zero. To improve training stability and accelerate convergence, we replace the original L1 norm with the Huber loss. The Huber loss offers a robust trade-off by exhibiting quadratic behavior for small errors, which ensures smooth differentiability and fine-grained parameter updates, while transitioning to linear behavior for large errors, thereby reducing sensitivity to outliers such as extreme pixel intensity variations. The loss function for the image fusion network is formulated as a weighted combination of illumination loss, auxiliary intensity loss, and texture detail loss, as shown in Eq (4):

$$\mathcal{L}_{fusion} = \lambda_1 \cdot \mathcal{L}_{illum} + \lambda_2 \cdot \mathcal{L}_{aux} + \lambda_3 \cdot \mathcal{L}_{texture} \tag{7}$$

The improved light perception loss calculation formula is shown in Eq (8), where $\mathcal{L}_{\text{Huber}}^{vi}$ and $\mathcal{L}_{\text{Huber}}^{ir}$ represent the intensity losses for visible and infrared light, respectively.

$$\mathcal{L}_{\text{illum}} = W_{vi} \cdot \mathcal{L}_{\text{Huber}}^{vi} + W_{ir} \cdot \mathcal{L}_{\text{Huber}}^{ir} \tag{8}$$

The intensity loss can balance the differences in pixel values between the pre-fusion and post-fusion images. Therefore, the intensity loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{Huber}}^{\text{ir}} &= \begin{cases} \frac{1}{2}\left(I_f - I_{ir}\right)^2 & \text{for } |I_f - I_{ir}|_2 \le \delta \\ \delta \cdot \left(||I_f - I_{ir}||_2 - \frac{1}{2}\delta^2\right) & \text{otherwise} \end{cases} \\ \mathcal{L}_{\text{Huber}}^{vi} &= \begin{cases} \frac{1}{2}\left(I_f - I_{vi}\right)^2 & \text{for } |I_f - I_{vi}|_2 \le \delta \\ \delta \cdot \left(||I_f - I_{vi}||_2 - \frac{1}{2}\delta^2\right) & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

In the context of defining the intensity loss within an image fusion network, the terms $||I_f - I_{vi}||_2$ and $||I_f - I_{ir}||_2$ represent the Euclidean distances between the fused image and the respective visible light image and infrared image. The parameter $\delta$ is the threshold for the Huber loss function.

Relying solely on light intensity is insufficient to maintain an optimal intensity distribution in the fused image. Therefore, an auxiliary intensity loss is defined as follows:

$$\mathcal{L}_{aux} = \frac{1}{HW}||I_f - \max(I_{ir}, I_{vi})||_1 \tag{10}$$

To preserve detailed textures in the fused image, a texture detail loss is definded. The gradient operator $\nabla$ is used to capture the texture information within the image, where the gradients are computed using the Sobel operator.

$$\mathcal{L}_{texture} = \frac{1}{HW}|||\nabla I_f| - max(|\nabla I_{vi}|, |\nabla I_{ir}|)||_1 \tag{11}$$

### Cascaded feature aggregation and extraction module

To address the feature dilution effect caused by heterogeneous interference in multi-spectral fusion and enhance the capture of pedestrian semantics in complex backgrounds, the Cascaded Feature Aggregation and Extraction (CFAE) module is designed and constructed, as illustrated in Fig 5. The module combines structural reparameterization with a multi-level feature aggregation mechanism, achieving a balance between representational expression and computational efficiency by equilibrating deep semantic extraction and the preservation of original spatial features.

Within the CFAE module, the RepVGG unit serves as the fundamental feature extraction operator. As depicted in Fig 6, this unit utilizes a multi-branch topology during the training phase to capture diverse spatial gradients and smooth the loss landscape of complex multi-spectral data. Its training output, $F_{train}$, is defined as follows:

$$F_{train}(x) = \mathrm{BN}(x * K_3) + \mathrm{BN}(x * K_1) + \mathrm{BN}(x) \tag{12}$$

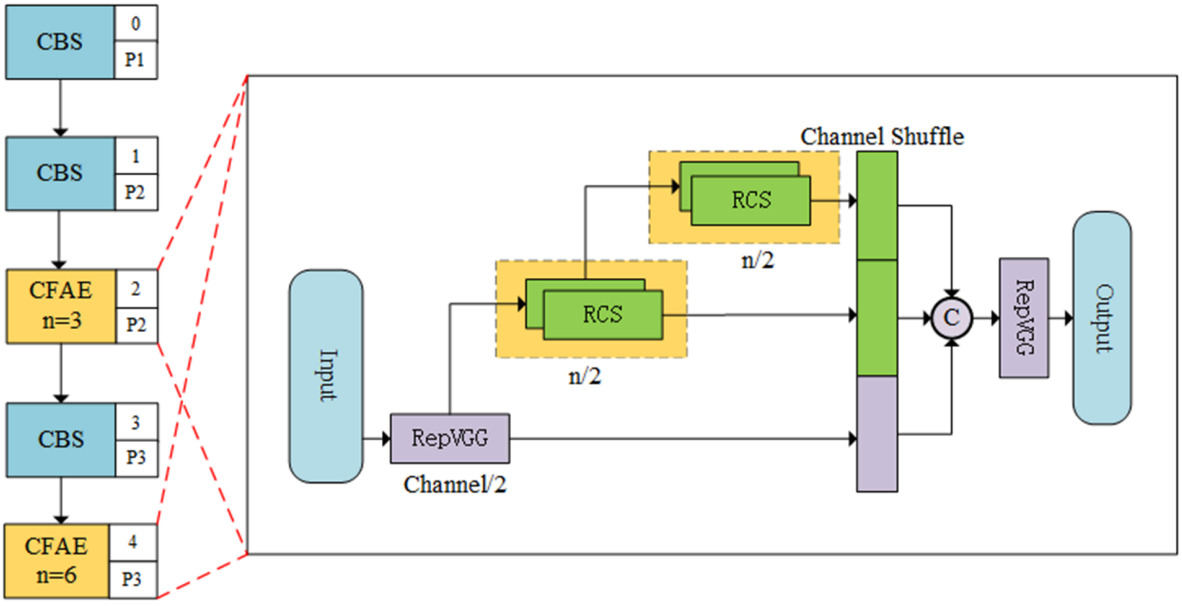



**Fig 5. The structure of the CFAE module.**

https://doi.org/10.1371/journal.pone.0342054.g005

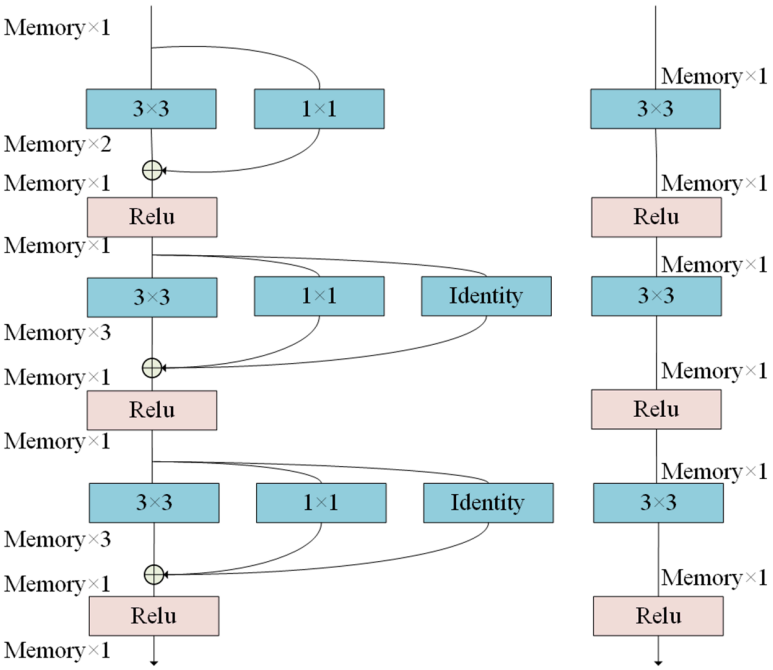



**Fig 6. The structure of RepVGG module.**

https://doi.org/10.1371/journal.pone.0342054.g006

where $K_3$ and $K_1$ represent the kernel parameters for the $3 \times 3$ and $1 \times 1$ convolutions, respectively. In the inference phase, these branches are collapsed into an equivalent singular kernel, $K_{eq} = f_{rep}(K_3, K_1, I)$, through structural reparameterization. This transformation ensures semantic abstraction while maintaining $O(1)$ memory complexity, addressing the computational burden of multi-spectral models during deployment.

CFAE integrates the Shuffle RepVGG (SR) structure, which corresponds to the cascaded RCS component shown in Fig 7. The SR unit employs a channel-splitting strategy, partitioning the input feature $X$ into a primary stream $X_{prim}$ and an identity stream $X_{cons}$:

$$F_{SR} = \text{Shuffle}\left(\text{Concat}(X_{cons}, \Phi_{Rep}(X_{prim}))\right) \tag{13}$$

where $\Phi_{Rep}$ denotes the RepVGG operator applied to the primary path. This configuration establishes a controlled residual path for deep networks, ensuring that the raw spatial information of pedestrian targets contained in the identity stream $X_{cons}$ is preserved during the cascading process. Subsequent channel shuffle operations, as detailed in Fig 7, complete the cross-channel information interaction between the disparate streams.

Furthermore, as shown in Fig 5, CFAE adopts a cascaded One-Shot Aggregation (OSA) mechanism. It integrates the initial projection feature $x_1$ and the successive outputs of the SR units, $x_2$ and $x_3$, through a concatenation operation:

$$F_{CFAE} = \Phi_{out}(\text{Concat}(x_1, x_2, x_3)) \tag{14}$$

where $\Phi_{out}$ represents the final projection operator. This cascaded design constructs an ensemble of multi-scale receptive fields capable of aggregating implicit semantic features from varying depths. This mechanism improves the reconstruction of partially occluded pedestrians and limits inference-time computational overhead through reparameterization.

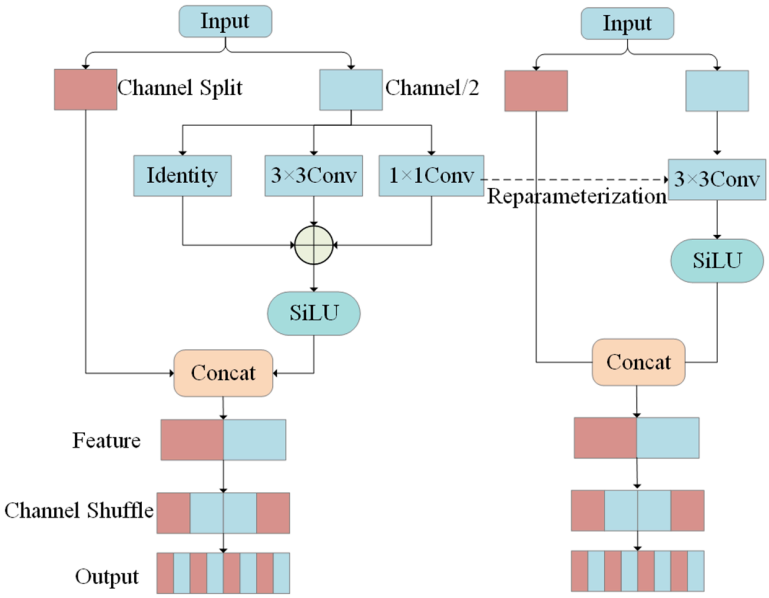



**Fig 7. The structure of RCS module.**

https://doi.org/10.1371/journal.pone.0342054.g007

### Scale-adaptive feature detection head

To overcome the inherent limitations of rigid feature extraction in multi-scale pedestrian detection, we propose a Scale-Adaptive Feature Detection Head (SAFDH), as illustrated in Fig 8. Unlike conventional detection heads that decouple classification and localization into parallel but homogeneous branches, SAFDH is formulated from a task–scale interaction perspective. Specifically, it explicitly models how object scale and task objectives jointly influence feature representation, which is particularly critical under multi-spectral feature fusion where spatial inconsistency and scale ambiguity frequently coexist.

A key design principle of SAFDH is to establish a stable and scale-consistent shared representation prior to task specialization. To this end, we introduce a Shared Normalization Convolution (SNC) module, implemented as share_conv, which combines shared convolution with Group Normalization (GN) [38]. From a theoretical standpoint, GN mitigates feature distribution shifts caused by illumination and modality variations, while parameter sharing constrains both tasks to operate on a unified semantic manifold. This design not only reduces computational redundancy but also provides a well-conditioned feature space for subsequent task-aware decomposition.

Recent studies, such as TOOD [39], have demonstrated that modeling task interaction can improve detection performance by aligning classification and localization features within a shared interaction space.However, directly enforcing task alignment may obscure the fundamentally different optimization requirements of the two tasks, especially in multi-scale pedestrian detection scenarios where localization is highly scale-sensitive while classification emphasizes semantic consistency. Instead of directly aligning tasks through a single interaction space, SAFDH decomposes task learning into two scale-adaptive and asymmetric pathways. We construct two independent Adaptive Task Decomposition (ATD) branches, denoted as cls_decomp and reg_decomp, which dynamically reweight shared features conditioned on global contextual statistics. The task-specific transformation is formulated as

$$F_{\text{task}} = \Psi_{\text{ATD}}\left(F_{\text{shared}}, \text{Pool}\left(F_{\text{shared}}\right)\right), \tag{15}$$

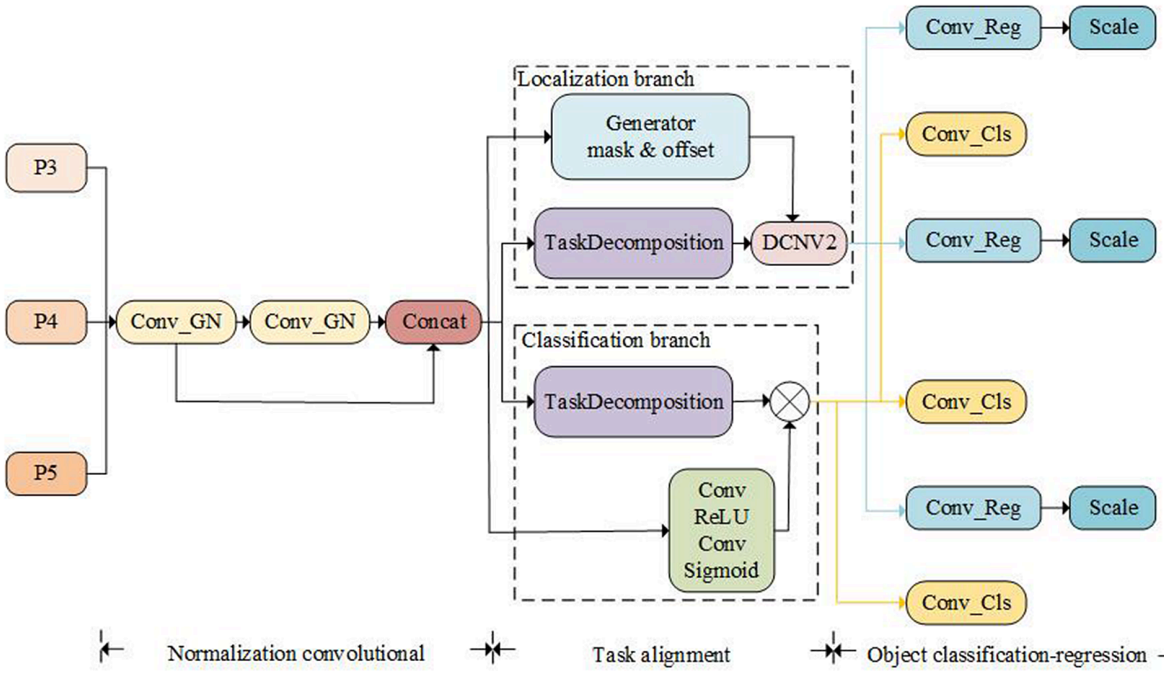



**Fig 8. The structure of SAFDH module.**

https://doi.org/10.1371/journal.pone.0342054.g008

where $F_{shared}$ represents the SNC output, and Pool(·) encodes global scale information. This formulation reflects the inherently different optimization requirements of localization and classification, in which localization emphasizes scale-dependent geometric boundaries, whereas classification focuses on scale-invariant semantic discrimination. By decoupling these objectives through ATD, SAFDH reduces task interference and enables scale-aware feature specialization, rather than enforcing uniform task alignment.

Furthermore, SAFDH introduces an asymmetric spatial alignment strategy tailored to the geometric nature of pedestrian localization. An Offset and Mask Mechanism (OMM) is employed to generate pixel-wise offsets and modulation masks, which guide a Dynamic Deformable Convolutional Network v2 (DyDCNv2) [40] exclusively in the regression branch. Spatial misalignment has a more pronounced impact on localization accuracy, whereas introducing excessive spatial deformation in the classification branch may undermine semantic consistency.

## Experiment setup and result analysis

### Dataset and model training details

In subsequent experiments, we validate the FID-YOLO model using two public datasets: M3FD [41] and LLVIP [42].

(1) M3FD Dataset: The M3FD Dataset, The M3FD Dataset, created by Peng Cheng Laboratory, contains 4,200 pairs of aligned infrared and visible light images across four challenging scenarios: daytime, overcast, nighttime, and occlusion. It includes 33,603 object annotations in six categories (People, Car, Bus, Motorcycle, Truck, and Lamp). The dataset's diversity in pixel variations, lighting, seasons, and weather makes it valuable for training and evaluating object detection in fused images.

(2) LLVIP Dataset: The LLVIP Dataset focuses on low-light object detection with 15,488 pairs of visible and infrared images (30,976 images in total) from 24 nighttime and 2 daytime scenes. It provides temporally and spatially aligned

**Table 1. Experimental environments and model parameter configuration.**

| Name | Version |
|---|---|
| Operating System | Ubuntu 22.04.4 LTS |
| CPU | 13th Gen Intel(R) Core(TM) i9-13900K |
| GPU | NVIDIA GeForce RTX 4090 24G |
| Python | 3.8 |
| PyTorch | 2.3.0 |
| CUDA | 12.1 |
| Input size | 640×640 |
| Batch size | 32 |
| Initial learning rate | 0.01 |
| Final learning rate | 0.1 |

https://doi.org/10.1371/journal.pone.0342054.t001

image pairs, with annotations for objects detectable in infrared under dim lighting, making it essential for low-light detection research.

The experimental configuration is detailed in Table 1. For the experiments conducted in this study, the Stochastic Gradient Descent (SGD) optimizer is employed to train the YOLO model with input images resized to 640×640 pixels. The batch size is configured at 32, with an initial learning rate of 0.01. 16 threads are allocated for data loading. Additionally, mosaic data augmentation, a technique used to increase the robustness of the model by augmenting the training data, is deactivated during the final ten epochs of the training regimen.

### Evaluation metrics

We evaluate the effectiveness of the model on the M3FD and LLVIP datasets by a suite of commonly used performance metrics, including mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5 (mAP@0.5), mean Average Precision across the IoU range from 0.5 to 0.95 (mAP@0.5-0.95), and the F1 score. Furthermore, this study evaluates the computational consumption of the model by examining the quantity of parameters.

### Experimental comparisons

We compare our model with the YOLO series, including YOLOv5n [43], YOLOv6n [44], YOLOv7-tiny [45], YOLOv8n [46], YOLOv8s [46], YOLO-MIF-n [47], and ICA-Fusion-n [48], RGBT-YOLO [49], DAMSDet [50], WaveMamba [51], LASFNet [52] on LLVIP and M3FD datasets to validate the performance of our model, in which considering the computational comsuption, we choose medium and small models within the YOLO series, where the '-n', '-s' and '-m' respectively represent nano, small, and medium models.

In addition, to further evaluate the cross-dataset generalization capability of the proposed method, we conduct additional experiments on the WiderPerson dataset. WiderPerson is a representative dense pedestrian detection benchmark, whose images are primarily collected from the Internet and exhibit high scene complexity, including cluttered backgrounds and watermark interference. The dataset defines five categories for pedestrian-related objects: (1) pedestrians, (2) riders, (3) partially occluded persons, (4) person-like objects, and (5) dense crowd regions with indistinguishable individual boundaries. Since our objective is to assess generalization performance for pedestrian detection across datasets, categories 1–4 are unified as a single "pedestrian" class for training and evaluation, while category 5 is excluded due to the absence of clear instance-level annotations. This preprocessing strategy is consistent with common practices in existing dense pedestrian detection studies.

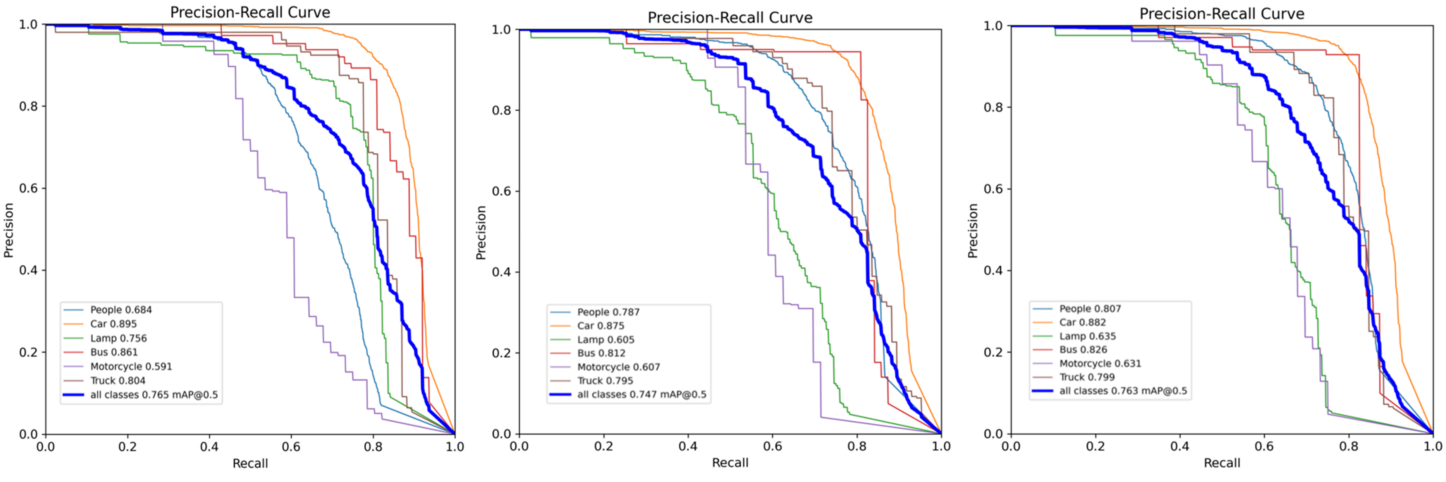
**Comparative experiment on M3FD dataset.** Table 2 shows the performance of the FID-YOLO-n model on the M3FD dataset compared to benchmark models. Our model outperforms YOLOv5-n, YOLOv6-n, YOLOv7-tiny, and YOLOv8-n in all metrics except for parameter scale. This indicates that the EPIAFusion module effectively enriches feature information, improving detection accuracy at the cost of computational resources. In visible light detection, our model achieves a

**Table 2. The experimental effects of different models on the M3FD dataset.**

| Network architecture | Input type | Pedestrian | mAP@0.5 | mAP@0.5-0.95 | F1 score | Param(M) | GFLOPs |
|---|---|---|---|---|---|---|---|
| YOLOv5-n [43] | Visible | 0.633 | 0.742 | 0.471 | 0.73 | 2.6 | 4.6 |
| YOLOv5-n [43] | Infrared | 0.790 | 0.728 | 0.457 | 0.72 | 2.6 | |
| YOLOv6-n [44] | Visible | 0.630 | 0.699 | 0.446 | 0.70 | 4.7 | 11.4 |
| YOLOv6-n [44] | Infrared | 0.758 | 0.690 | 0.441 | 0.69 | 4.7 | |
| YOLOv7-tiny [45] | Visible | 0.701 | 0.754 | 0.475 | 0.74 | 6.2 | 13.8 |
| YOLOv7-tiny [45] | Infrared | 0.753 | 0.701 | 0.455 | 0.70 | 6.2 | |
| YOLOv8-n [46] | Visible | 0.712 | 0.756 | 0.491 | 0.74 | 3.2 | 8.7 |
| YOLOv8-n [46] | Infrared | 0.787 | 0.747 | 0.488 | 0.74 | 3.2 | |
| YOLOv8-s [46] | Visible | 0.750 | 0.805 | 0.534 | 0.79 | 11.2 | 28.6 |
| YOLOv8-s [46] | Infrared | 0.825 | 0.779 | 0.506 | 0.77 | 11.2 | |
| YOLO-MIF-n [47] | Mixed | 0.792 | 0.786 | 0.514 | 0.77 | 3.3 | 12.3 |
| ICAFusion-n [48] | Mixed | 0.795 | 0.791 | 0.479 | 0.77 | 3.0 | 25.8 |
| FID-YOLO-n (Ours) | Mixed | 0.801 | 0.803 | 0.535 | 0.79 | 8.1 | 22.4 |

https://doi.org/10.1371/journal.pone.0342054.t002

pedestrian detection precision of 0.801, with a mAP@0.5 just 0.2% lower than YOLOv8-s. In infrared detection, despite a 2.4% lower precision, our mAP@0.5 improves by 2.6% compared to YOLOv8-s. YOLOv8-s achieves high precision by increasing parameter scale, but our model, with 27.7% fewer parameters, maintains a balance between detection accuracy and computational cost. Compared to the YOLO-MIF and ICAFusion models, designed for mixed light input, FID-YOLO achieves superior performance in most metrics, confirming the effectiveness of the EPIAFusion module for integrating visible and infrared images. Our model also achieves optimal mAP@0.5-0.95 and F1 scores. As observed from the Fig 9(a), 9(b), 9(c) represent the baseline model under visible light and infrared light, and the P-R curve of FID-YOLO, respectively. The area enclosed by the P-R curve and the coordinate axes is larger for FID-YOLO, indicating that, compared to the baseline model, FID-YOLO is better at balancing the identification of positive classes and the accuracy of predictions across different thresholds. Fig 10 shows FID-YOLO's advantage in object detection in complex environments on the M3FD dataset, further validating the model's effectiveness.




(a) (b) (c)

**Fig 9. M3FD Dataset FID-YOLO and other models P-R curve comparison result.**

https://doi.org/10.1371/journal.pone.0342054.g009

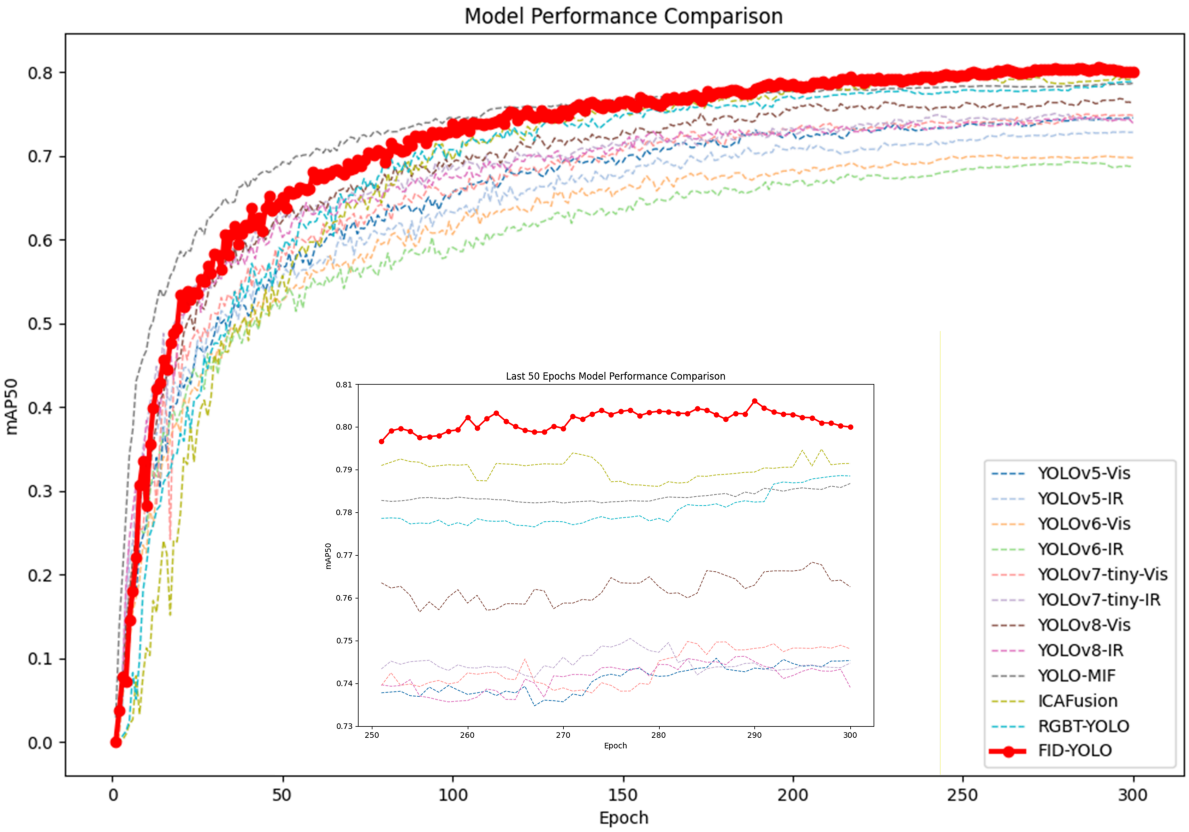



**Fig 10. M3FD Dataset FID-YOLO and Other models mAP comparison result.**

https://doi.org/10.1371/journal.pone.0342054.g010

**Comparative experiment on LLVIP dataset.** Fig 11 shows the P-R curve comparison between FID-YOLO and the baseline model. The results indicate that FID-YOLO outperforms the baseline model in terms of precision and recall across different thresholds. Fig 12 and Table 3 show the results of benchmark models on the LLVIP dataset using inputs from different light sources. Pedestrian detection performance on infrared images significantly outperforms that of visible light images, as infrared images highlight object contours in low-light or dark conditions, while visible light often obscures object features. However, infrared images may miss environmental details. The integration of the EPIAFusion network improves detection performance, with a 4.5% increase in mAP@0.5 compared to YOLOv8 with visible light input, a 12.1% improvement in mAP@0.5-0.95, and a 3% increase in the F1 score. Compared to the infrared model, the improvement is 1.6% in mAP and 2% in F1 score.

In comparison with existing multispectral detection approaches, YOLO-MIF and WaveMamba achieve competitive performance in terms of mAP@0.5; however, they exhibit relatively lower mAP@0.5–0.95, suggesting limited localization accuracy under stricter IoU thresholds. DAMSDet demonstrates robust detection capability but incurs significantly higher computational cost, as reflected by its large parameter size and GFLOPs. LASFNet achieves a favorable balance between accuracy and model complexity, yet its overall performance remains slightly inferior to the proposed method under comprehensive evaluation metrics. Overall, FID-YOLO maintains strong detection accuracy while preserving moderate computational complexity, and consistently performs well across multiple evaluation metrics. These results demonstrate the effectiveness and robustness of the proposed approach for multispectral pedestrian detection under low-light

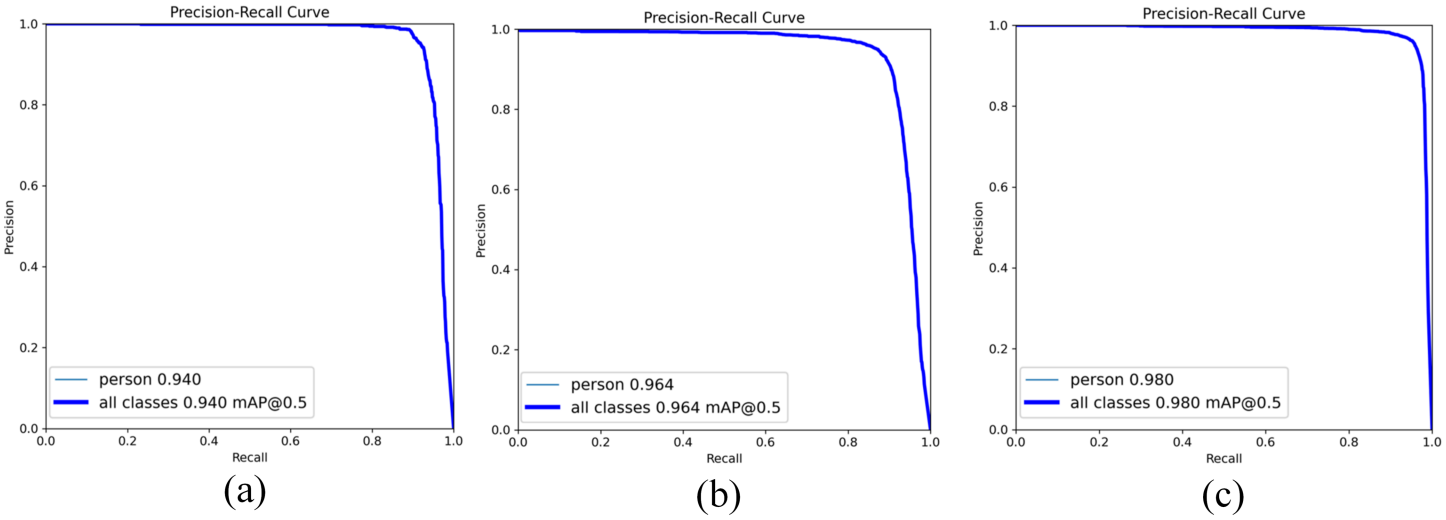



**Fig 11. LLVIP dataset FID-YOLO and other models P-R Curve comparison result.**

https://doi.org/10.1371/journal.pone.0342054.g011

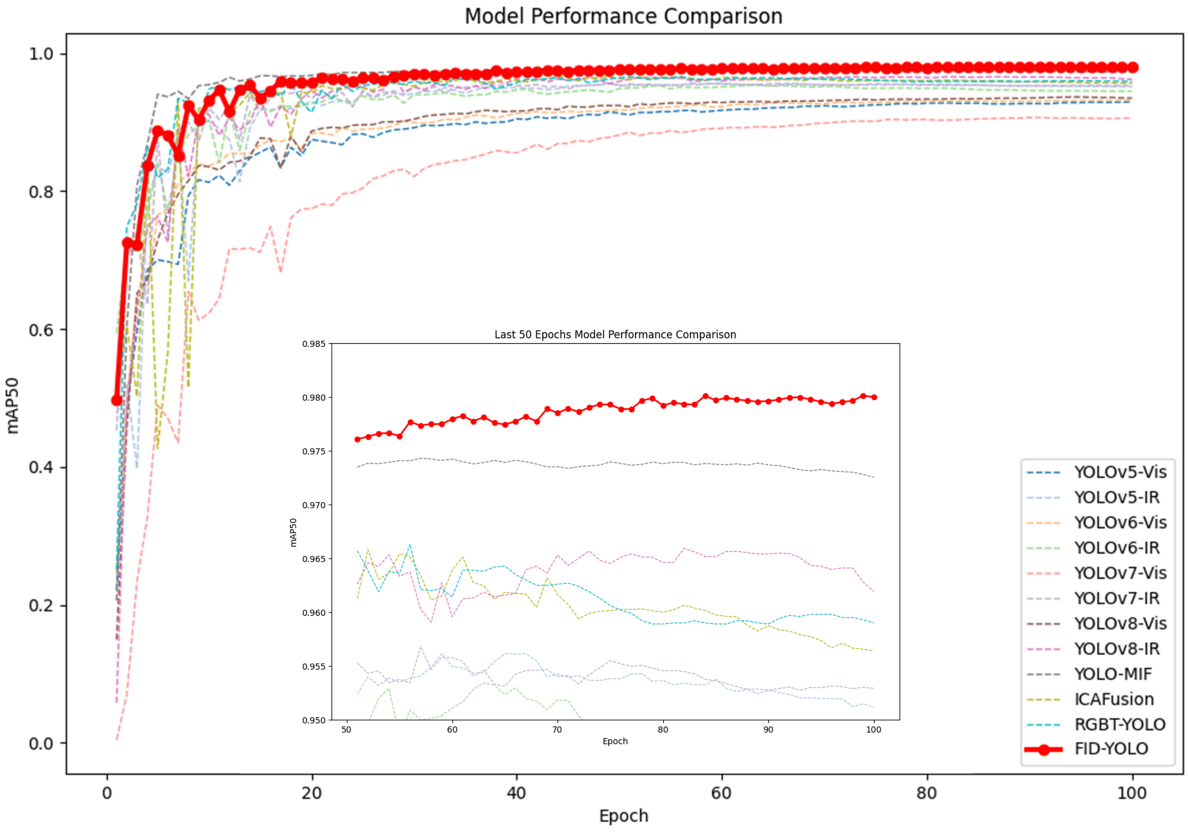



**Fig 12. LLVIP dataset FID-YOLO and other models mAP comparison result.**

https://doi.org/10.1371/journal.pone.0342054.g012

**Table 3. The experimental effects of different models on the LLVIP dataset.**

| Network architecture | Input type | mAP@0.5 | mAP@0.5-0.95 | F1 score | Param(M) | GFLOPs |
|---|---|---|---|---|---|---|
| YOLOv5-m [43] | Visible | 0.929 | 0.544 | 0.89 | 20.8 | 64.2 |
| | Infrared | 0.953 | 0.615 | 0.90 | | |
| YOLOv6-m [44] | Visible | 0.932 | 0.552 | 0.90 | 34.9 | 85.8 |
| | Infrared | 0.952 | 0.637 | 0.87 | | |
| YOLOv7-tiny [45] | Visible | 0.905 | 0.467 | 0.87 | 6.0 | 13.8 |
| | Infrared | 0.958 | 0.633 | 0.91 | | |
| YOLOv8-m [46] | Visible | 0.935 | 0.552 | 0.90 | 25.8 | 78.9 |
| | Infrared | 0.964 | 0.684 | 0.91 | | |
| YOLO-MIF-m [47] | Mixed | 0.974 | 0.684 | 0.91 | 26.6 | 120.0 |
| ICAFusion-m [48] | Mixed | 0.765 | 0.618 | 0.94 | 34.5 | 96.32 |
| RGBT-YOLO-m [49] | Mixed | 0.959 | 0.600 | 0.92 | 33.4 | 109.82 |
| DAMSDet [50] | Mixed | 0.915 | 0.553 | 0.92 | 79.21 | 173.76 |
| WaveMamba [51] | Mixed | 0.983 | 0.660 | 0.95 | 69.1 | 128.91 |
| LASFNet [52] | Mixed | 0.974 | 0.676 | 0.94 | 7.7 | - |
| FID-YOLO-m (Ours) | Mixed | 0.980 | 0.673 | 0.93 | 28.3 | 89.39 |

https://doi.org/10.1371/journal.pone.0342054.t003

conditions. In addition, Fig 12 illustrates the training curves of mAP@0.5 for different models on the LLVIP dataset, further confirming the stability and effectiveness of the proposed method.

**Comparative generalization on WiderPerson dataset.** The WiderPerson dataset is a representative dense crowd pedestrian detection dataset. The images are primarily collected from the internet, and many samples contain visible watermarks, which increases the realism and complexity of the scenes. The dataset defines five annotated categories for pedestrian-related objects: normal pedestrians, cyclists, partially occluded human bodies, human-like objects and indistinguishable dense crowds. To align with the objective of evaluating cross-dataset generalization in pedestrian detection, we retained categories 1 to 4 and merged them into a single “pedestrian” class, while category 5 (dense crowd regions) was excluded because it does not provide clear individual-level bounding boxes and is therefore unsuitable for standard pedestrian detection. This preprocessing strategy is consistent with common practice in dense pedestrian detection studies.

Furthermore, since WiderPerson contains only visible-light images and does not involve multi-spectral or cross-modal image pairs, the experiments on this dataset were conducted using only the YOLO detection branch of our model. The proposed EPIAFusion module was not employed in this setting, as image fusion is not applicable. This setup enables a fair evaluation of the generalization capability of the learned detection representations under a significantly different data distribution.

The experimental results are summarized in Table 4. Compared with the baseline YOLOv8-n, our FID-YOLO achieves consistent improvements across multiple evaluation metrics. Specifically, FID-YOLO improves mAP@0.5 from 0.714 to 0.727, mAP@0.5–0.95 from 0.429 to 0.437, and F1-score from 0.68 to 0.69, while maintaining comparable precision and recall. Notably, these gains are achieved with only a modest increase in computational cost (GFLOPs and parameters), indicating that the performance improvement does not stem from excessive model complexity. These results demonstrate that the proposed feature representation and detection head learned from multi-spectral training data retain strong discriminative capability when transferred to a dense, single-modality pedestrian dataset with substantially different data

**Table 4. Cross-dataset generalization performance on the WiderPerson dataset.**

| Network architecture | mAP@0.5 | mAP@0.5-0.95 | F1 | Precision | Recall | GFLOPs | IT | Param(M) |
|---|---|---|---|---|---|---|---|---|
| YOLOv8-n | 0.714 | 0.429 | 0.68 | 0.78 | 0.61 | 8.2 | 3.7 | 3.1 |
| FID-YOLO-n | 0.727 | 0.437 | 0.69 | 0.78 | 0.63 | 8.7 | 3.9 | 3.9 |

https://doi.org/10.1371/journal.pone.0342054.t004

characteristics. This confirms that the performance gains are not overfitted to M3FD or LLVIP, but instead reflect improved generalization ability under diverse pedestrian detection scenarios.

### Ablation experiments

This study conducts ablation experiments on both the M3FD and LLVIP datasets using YOLOv8 as the base model to assess the contribution of each module within FID-YOLO. The findings from these ablation studies are delineated in Tables 5 and 6, which provide a comprehensive view of the impact of each module on the overall performance of the FID-YOLO model. Tables 7 and 8 present the ablation results for the CSAM components and CFAE configuration choices on the M3FD and LLVIP datasets, respectively. The CSAM module consistently improves mAP compared to the baseline CMADF, indicating that integrating channel and spatial attention helps the fusion network better preserve both semantic and geometric information. The CFAE and SAFDH modules provide additional gains over using only RepVGG or TOOD, validating their effectiveness in multi-scale feature extraction and task-specific feature decomposition. Across both datasets, combining CSAM with the downstream CFAE and SAFDH modules yields the best performance, demonstrating that each component contributes meaningfully to the overall accuracy.

We conducted 30 epochs of image fusion training and evaluation in LLVIP and M3FD dataset to observe the loss curves for $\delta = 0.5, 1, 2$, as well as the original L1 loss. As shown in the Figs 13 and 14, the model achieves the lowest

**Table 5. M3FD ablation experiment results.**

| Baseline | PIAFusion | EPIAFusion | CFAE | SAFDH | mAP@0.5 | mAP@0.5-0.95 | F1 score |
|---|---|---|---|---|---|---|---|
| | | | | | 0.712 | 0.491 | 0.74 |
| | ✓ | | | | 0.765 | 0.493 | 0.76 |
| | | ✓ | | | 0.774 | 0.501 | 0.77 |
| | | | ✓ | | 0.735 | 0.500 | 0.76 |
| YOLOv8n | | | | ✓ | 0.729 | 0.512 | 0.76 |
| | | ✓ | ✓ | | 0.791 | 0.520 | 0.77 |
| | | | ✓ | ✓ | 0.741 | 0.524 | 0.78 |
| | | ✓ | | ✓ | 0.799 | 0.522 | 0.79 |
| | | ✓ | ✓ | ✓ | 0.801 | 0.535 | 0.79 |

https://doi.org/10.1371/journal.pone.0342054.t005

**Table 6. LLVIP ablation experiment results.**

| Baseline | PIAFusion | EPIAFusion | CFAE | SAFDH | mAP@0.5 | mAP@0.5-0.95 | F1 score |
|---|---|---|---|---|---|---|---|
| | | | | | 0.935 | 0.552 | 0.90 |
| | ✓ | | | | 0.962 | 0.664 | 0.94 |
| | | ✓ | | | 0.970 | 0.672 | 0.94 |
| | | | ✓ | | 0.940 | 0.565 | 0.91 |
| YOLOv8m | | | | ✓ | 0.941 | 0.566 | 0.91 |
| | | ✓ | ✓ | | 0.977 | 0.670 | 0.95 |
| | | | ✓ | ✓ | 0.946 | 0.571 | 0.92 |
| | | ✓ | | ✓ | 0.972 | 0.671 | 0.95 |
| | | ✓ | ✓ | ✓ | 0.980 | 0.673 | 0.95 |

https://doi.org/10.1371/journal.pone.0342054.t006

**Table 7. Ablation Experiments on CSAM Components and CFAE Configuration Choices (M3FD Dataset).**

| Components | mAP@0.5 | mAP@0.5-0.95 | Param(M) | GFLOPs |
|---|---|---|---|---|
| CMADF | 0.769 | 0.491 | 7.9 | 21.7 |
| CSAM | 0.776 | 0.512 | 8.2 | 22.4 |
| RepVGG | 0.762 | 0.497 | 3.0 | 8.4 |
| TOOD | 0.721 | 0.501 | 3.3 | 9.1 |

https://doi.org/10.1371/journal.pone.0342054.t007

**Table 8. Ablation Experiments on CSAM Components and CFAE Configuration Choices (LLVIP Dataset).**

| Components | mAP@0.5 | mAP@0.5-0.95 | Param(M) | GFLOPs |
|---|---|---|---|---|
| CMADF | 0.959 | 0.631 | 7.9 | 21.7 |
| CSAM | 0.964 | 0.649 | 8.2 | 22.4 |
| RepVGG | 0.937 | 0.557 | 3.0 | 8.4 |
| TOOD | 0.942 | 0.561 | 3.3 | 9.1 |

https://doi.org/10.1371/journal.pone.0342054.t008

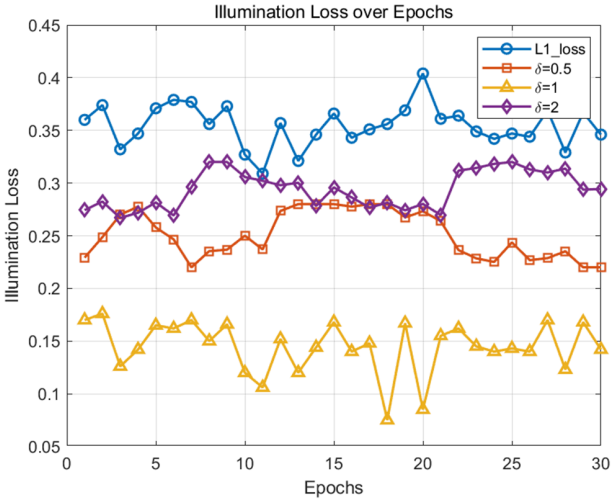



**Fig 13. LLVIP dataset illumination loss over epoch.**

https://doi.org/10.1371/journal.pone.0342054.g013

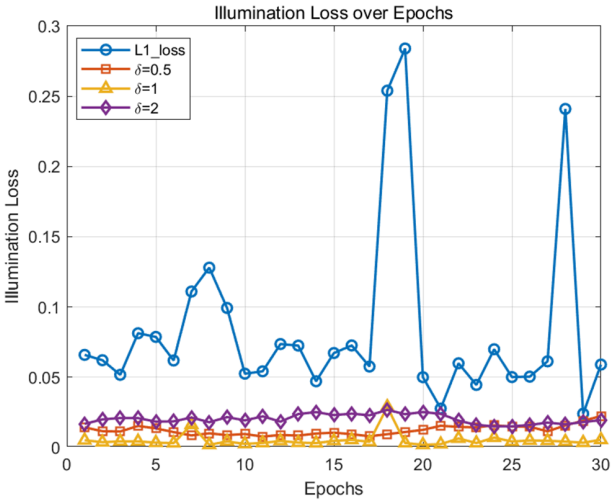



**Fig 14. M3FD dataset illumination loss over epoch.**

https://doi.org/10.1371/journal.pone.0342054.g014

loss when $\delta = 1$, which also outperforms the original L1 loss. This demonstrates that $\delta = 1$ provides an optimal balance between robustness and sensitivity during training.

The EPIAFusion module shows the most significant improvement, capturing the textural details of objects from visible light and feature information from infrared light, while preserving environmental lighting through light perception. As a

result, mAP increased by 1.3% on the M3FD dataset and 3.5% on the LLVIP dataset compared to using only visible light input. The CFAE module enhances the network's sensitivity to occluded objects by integrating features with distinct receptive fields, resulting in mAP improvement of 1.8% and 0.5%. The SAFDH module, through task alignment and shared convolutions, not only reduced the number of parameters but also elevated the mAP by 1.3% and 0.6%, respectively. Ablation studies confirm that the proposed modules have significantly improved average accuracy on both the M3FD and LLVIP datasets compared to the benchmark model.

## Result visualization

**M3FD visualization analysis.** Fig 15 shows some prediction results of the models that performed well on the M3FD dataset. Undetected objects are marked with yellow circles, while false positives are indicated by green circles, and multiple detections are highlighted with gold circles. In the first set of images, YOLOv8 shows a significant number of missed detections in both visible and infrared light, while our model and YOLO-MIF perform better. The second and third sets of images were captured under nighttime obstruction conditions, where FID-YOLO exhibits higher confidence with no missed detections or false positives. In the final set of images, which involves detecting small, occluded pedestrian objects, YOLOv8, YOLO-MIF, and ICAFusion all experience issues with multiple detections, while our model successfully detects all targets, demonstrating its superior detection capability in complex environments.

**LLVIP visualization analysis.** Fig 16 shows a subset of prediction results for the well-performing models on the LLVIP dataset. False positives are indicated by orange circular markers, while missed detections are represented by green circles and multiple detections are highlighted with gold circles. In the initial set of imagery, YOLOv8 was unable to accurately identify the object features due to the constraints of the nocturnal environment, resulting in false detections. In the

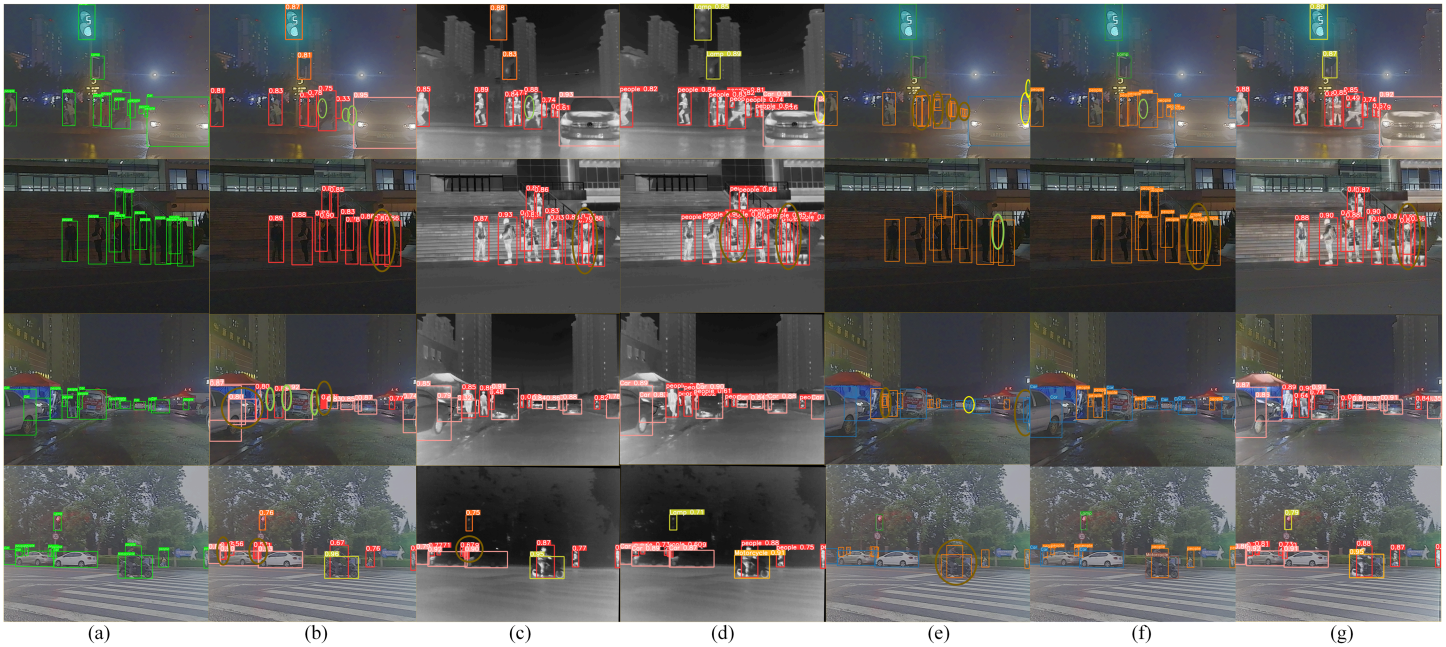



**Fig 15. Partial display of detection results on the M3FD dataset (a) M3FD dataset ground truth (b) the prediction results of YOLOv8 using the visible light Input (c) the prediction results of YOLOv8 using the infrared light Input (d) the prediction results of YOLO-MIF (e) the prediction results of ICAFusion (f) the prediction results of RGBT-YOLO (g) the prediction results of FID-YOLO.**

https://doi.org/10.1371/journal.pone.0342054.g015

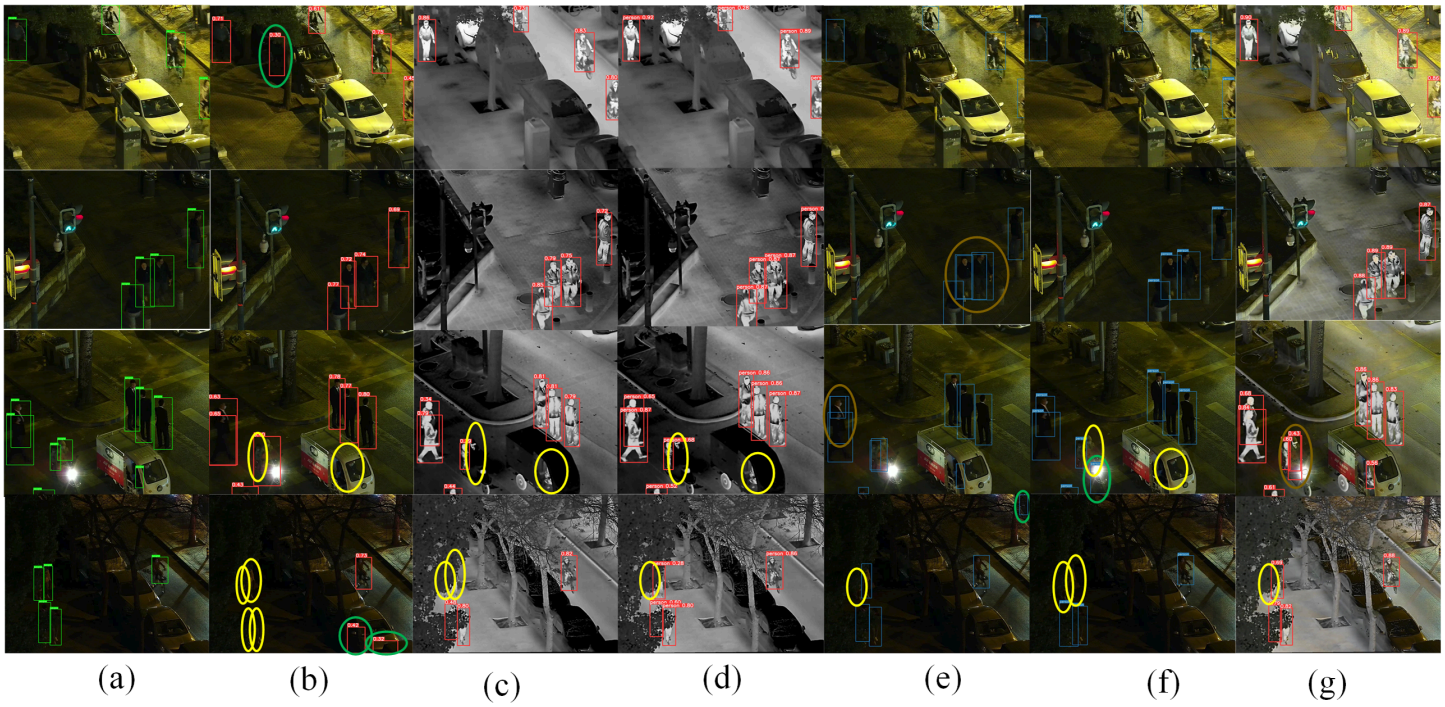



**Fig 16. Partial display of detection results on the LLVIP dataset (a)LLVIP dataset ground truth (b) the prediction results of YOLOv8 using the visible light Input (c) the prediction results of YOLOv8 using the infrared light Input (d) the prediction results of YOLO-MIF (e) the prediction results of ICAFusion (f) the prediction results of RGBT-YOLO (g) the prediction results of FID-YOLO.**

https://doi.org/10.1371/journal.pone.0342054.g016

second set of images, other models demonstrated good detection performance; however, there were differences in overall confidence compared to FID-YOLO, and the ICAFusion model produced duplicate detections. The third set of photographs revealed that other models failed to identify certain objects, particularly smaller items such as courier vehicles and motorcycles under visible light and infrared conditions. Although our model produced a single duplicate detection, its overall performance remained superior to that of the other models. In the final set of images, our model still maintained good detection performance, with only one missed detection, while the other models exhibited significant issues with both missed and false detections. This demonstrates that our model has better detection capabilities under low-light conditions compared to other models.

Fig 17 shows the object feature heatmaps on the LLVIP dataset. Compared to YOLOv8, the heatmaps of FID-YOLO more accurately align with the detected objects. In the initial set of imagery, under poor lighting conditions, the network's output focuses on the background and other irrelevant areas, while FID-YOLO shows greater focus on the relevant objects.

## Conclusions

This study introduces FID-YOLO, an object detection approach that integrate visible and infrared imagery. FID-YOLO integrates an advanced image fusion technique with the YOLOv8 architecture, creating a synergy of information from both visible and infrared light. The process begins with the EPIAFusion module, which extracts and maps features from infrared and visible light into a unified feature space. Subsequently, the resulting fused image is processed by the adapted YOLOv8 framework. In this framework, the CFAE feature aggregation module integrates features from deeper and shallower layers, enhancing the model's ability to distinguish occluded pedestrians from their surroundings. Finally, SAFDH

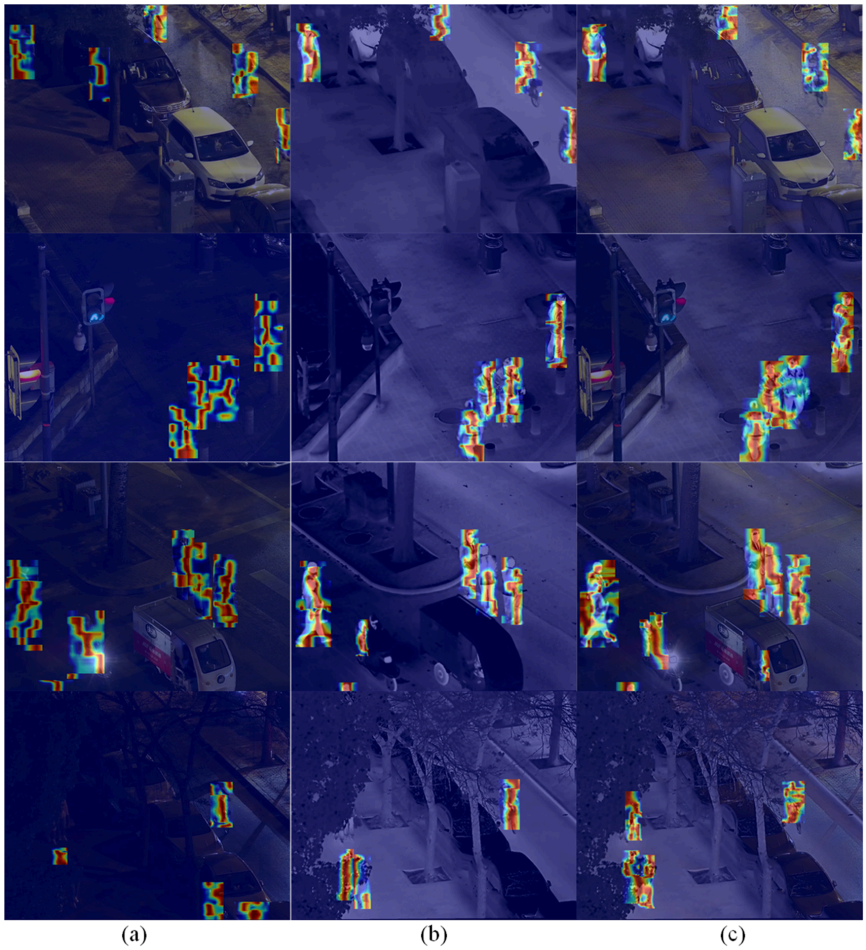



**Fig 17. Partial detection effect heatmaps on the LLVIP dataset (a) YOLOv8 visible light input (b) YOLOv8 infrared light input (c) FID-YOLO.**

https://doi.org/10.1371/journal.pone.0342054.g017

captures object characteristics across varying scales, improving the model's feature representation and addressing the challenge of detecting small objects in low-resolution images. Comparative experiments and ablation studies conducted on the M3FD and LLVIP datasets validate the effectiveness and feasibility of the proposed method for Pedestrian detection.

Although FID-YOLO performs well in pedestrian detection, it has some limitations. Model efficiency (parameters and FLOPs) can be further improved, especially for resource-constrained devices. Fusion may degrade performance in cases of very low-quality infrared images or severe misalignment between modalities. Moreover, the model is trained on well-aligned datasets, limiting generalization to non-aligned scenarios. Future work will focus on lightweight design, adaptive fusion, faster inference, and handling non-aligned data to enhance accuracy, efficiency, and robustness.

## Supporting information

**S1 Data. Experimental results and evaluation data supporting the findings of this study.**
(XLSX)

## Author contributions

**Conceptualization:** Di Yang, Xilong Zhang.

**Data curation:** Xilong Zhang.

**Formal analysis:** Di Yang.

**Funding acquisition:** Peng Wang.

**Investigation:** Xilong Zhang.

**Methodology:** Di Yang, Xilong Zhang.

**Project administration:** Di Yang.

**Resources:** Di Yang, Peng Wang.

**Software:** Di Yang.

**Supervision:** Di Yang, Peng Wang.

**Validation:** Di Yang, Xilong Zhang.

**Writing – original draft:** Xilong Zhang.

**Writing – review & editing:** Di Yang, Xilong Zhang.

## References


1. Jain DK, Zhao X, González-Almagro G, Gan C, Kotecha K. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. Information Fusion. 2023;95:401–14.
2. Dollár P, Wojek C, Schiele B, Perona P. Pedestrian detection: an evaluation of the state of the art. IEEE Trans Pattern Anal Mach Intell. 2012;34(4):743–61. https://doi.org/10.1109/TPAMI.2011.155 PMID: 21808091
3. Pang Y, Yuan Y, Li X, Pan J. Efficient HOG human detection. Signal Processing. 2011;91(4):773–81. https://doi.org/10.1016/j.sigpro.2010.08.010
4. Barbu T. Pedestrian detection and tracking using temporal differencing and HOG features. Computers & Electrical Engineering. 2014;40(4):1072–9. https://doi.org/10.1016/j.compeleceng.2013.12.004
5. Wang X, Han TX, Yan S. An HOG-LBP human detector with partial occlusion handling. In: 2009 IEEE 12th International Conference on Computer Vision. 2009. p. 32–9. https://doi.org/10.1109/iccv.2009.5459207
6. Abadir KM, Caggiano G, Talmain G. Nelson–Plosser revisited: the ACF approach. Journal of Econometrics. 2013;175(1):22–34. https://doi.org/10.1016/j.jeconom.2013.02.006
7. Luo Q, Ma H, Tang L, Wang Y, Xiong R. 3D-SSD: Learning hierarchical features from RGB-D images for amodal 3D object detection. Neurocomputing. 2020;378:364–74. https://doi.org/10.1016/j.neucom.2019.10.025
8. Li Y, Dong H, Li H, Zhang X, Zhang B, Xiao Z. Multi-block SSD based on small object detection for UAV railway scene surveillance. Chinese Journal of Aeronautics. 2020;33(6):1747–55. https://doi.org/10.1016/j.cja.2020.02.024
9. Hong Y, Pan R, Su J, Li M. Infrared image detection of defects in lightweight solar panels based on improved MSRCR and YOLOv8n. Infrared Physics & Technology. 2024;141:105473. https://doi.org/10.1016/j.infrared.2024.105473
10. Zhou J, Zhang B, Yuan X, Lian C, Ji L, Zhang Q, et al. YOLO-CIR: the network based on YOLO and ConvNeXt for infrared object detection. Infrared Physics & Technology. 2023;131:104703. https://doi.org/10.1016/j.infrared.2023.104703
11. Shi Y, Wang N, Guo X. YOLOV: making still image object detectors great at video object detection. AAAI. 2023;37(2):2254–62. https://doi.org/10.1609/aaai.v37i2.25320
12. Liu J, Jia R, Li W, Ma F, Abdullah HM, Ma H, et al. High precision detection algorithm based on improved RetinaNet for defect recognition of transmission lines. Energy Reports. 2020;6:2430–40. https://doi.org/10.1016/j.egyr.2020.09.002
13. Turan M, Almalioglu Y, Araujo H, Konukoglu E, Sitti M. Deep EndoVO: a recurrent convolutional neural network (RCNN) based visual odometry approach for endoscopic capsule robots. Neurocomputing. 2018;275:1861–70. https://doi.org/10.1016/j.neucom.2017.10.014
14. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16. https://doi.org/10.1109/TPAMI.2015.2389824 PMID: 26353135
15. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision. 2015. p. 1440–8.

16. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. p. 2961–9.
17. Wang P, Wu J, Fang A, Zhu Z, Wang C. Multi-spectral image fusion for moving object detection. Infrared Physics & Technology. 2024;141:105489. https://doi.org/10.1016/j.infrared.2024.105489
18. Guo F, Wang Y, Qian Y. Real-time dense traffic detection using lightweight backbone and improved path aggregation feature pyramid network. Journal of Industrial Information Integration. 2023;31:100427. https://doi.org/10.1016/j.jii.2022.100427
19. Lan R, Zhang Y, Xie L, Wu Z, Liu Y. BEV feature exchange pyramid networks-based 3D object detection in small and distant situations: a decentralized federated learning framework. Neurocomputing. 2024;583:127476. https://doi.org/10.1016/j.neucom.2024.127476
20. Wang Q, Zhou L, Yao Y, Wang Y, Li J, Yang W. An interconnected feature pyramid networks for object detection. Journal of Visual Communication and Image Representation. 2021;79:103260. https://doi.org/10.1016/j.jvcir.2021.103260
21. Zhao J, Zhu H, Niu L. BiTNet: a lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network. Journal of King Saud University - Computer and Information Sciences. 2023;35(8):101670. https://doi.org/10.1016/j.jksuci.2023.101670
22. Gong Y, Luo J, Shao H, Li Z. A transfer learning object detection model for defects detection in X-ray images of spacecraft composite structures. Composite Structures. 2022;284:115136. https://doi.org/10.1016/j.compstruct.2021.115136
23. Cao X, Wang Z, Yan P, Li X. Transfer learning for pedestrian detection. Neurocomputing. 2013;100:51–7. https://doi.org/10.1016/j.neucom.2011.12.043
24. Wang J-T, Yan G-L, Wang H-Y, Hua J. Pedestrian recognition in multi-camera networks based on deep transfer learning and feature visualization. Neurocomputing. 2018;316:166–77. https://doi.org/10.1016/j.neucom.2018.07.063
25. Zhou W, Chen Y, Lei J, Yu L, Zhou X, Luo T. Boundary-aware pyramid attention network for detecting salient objects in RGB-D images. Digital Signal Processing. 2021;111:102975. https://doi.org/10.1016/j.dsp.2021.102975
26. Hou Z, Yang C, Sun Y, Ma S, Yang X, Fan J. An object detection algorithm based on infrared-visible dual modal feature fusion. Infrared Physics & Technology. 2024;137:105107. https://doi.org/10.1016/j.infrared.2023.105107
27. Li X, Diao W, Mao Y, Gao P, Mao X, Li X, et al. OGMN: occlusion-guided multi-task network for object detection in UAV images. ISPRS Journal of Photogrammetry and Remote Sensing. 2023;199:242–57. https://doi.org/10.1016/j.isprsjprs.2023.04.009
28. Hao S, Gao S, Ma X, An B, He T. Anchor-free infrared pedestrian detection based on cross-scale feature fusion and hierarchical attention mechanism. Infrared Physics & Technology. 2023;131:104660. https://doi.org/10.1016/j.infrared.2023.104660
29. Shi L, Zhang Q, Pan B, Zhang J, Su Y. Global-local and occlusion awareness network for object tracking in UAVs. IEEE J Sel Top Appl Earth Observations Remote Sensing. 2023;16:8834–44. https://doi.org/10.1109/jstars.2023.3308042
30. Liu W, Ren G, Yu R, Guo S, Zhu J, Zhang L. Image-adaptive YOLO for object detection in adverse weather conditions. AAAI. 2022;36(2):1792–800. https://doi.org/10.1609/aaai.v36i2.20072
31. Luo P, Nie J, Xie J, Cao J, Zhang X. Localization-aware logit mimicking for object detection in adverse weather conditions. Image and Vision Computing. 2024;146:105035. https://doi.org/10.1016/j.imavis.2024.105035
32. Jeon H, Kim H, Kim D, Kim J. PASS-CCTV: proactive anomaly surveillance system for CCTV footage analysis in adverse environmental conditions. Expert Systems with Applications. 2024;254:124391. https://doi.org/10.1016/j.eswa.2024.124391
33. Qin H, Han J, Yan X, Zeng Q, Zhou H, Li J, et al. Infrared small moving target detection using sparse representation-based image decomposition. Infrared Physics & Technology. 2016;76:148–56. https://doi.org/10.1016/j.infrared.2016.02.003
34. Wang W, Meng Y, Li S, Zhang C. HV-YOLOv8 by HDPconv: better lightweight detectors for small object detection. Image and Vision Computing. 2024;147:105052. https://doi.org/10.1016/j.imavis.2024.105052
35. Cao Y, Luo X, Yang J, Cao Y, Yang MY. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. Information Fusion. 2022;88:1–11. https://doi.org/10.1016/j.inffus.2022.06.008
36. Wei J, Su S, Zhao Z, Tong X, Hu L, Gao W. Infrared pedestrian detection using improved UNet and YOLO through sharing visible light domain information. Measurement. 2023;221:113442. https://doi.org/10.1016/j.measurement.2023.113442
37. Tang L, Yuan J, Zhang H, Jiang X, Ma J. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware. Information Fusion. 2022;83–84:79–92. https://doi.org/10.1016/j.inffus.2022.03.007
38. Wu Y, He K. Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 3–19.
39. Feng C, Zhong Y, Gao Y, Scott MR, Huang W. TOOD: Task-aligned One-stage Object Detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021. p. 3490–9. https://doi.org/10.1109/iccv48922.2021.00349
40. Zhu X, Hu H, Lin S, Dai J. Deformable ConvNets V2: more deformable, better results. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 9308–16.
41. Liu J, Fan X, Huang Z, Wu G, Liu R, Zhong W, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 5802–11.
42. Jia X, Zhu C, Li M, Tang W, Zhou W. LLVIP: a visible-infrared paired dataset for low-light vision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 3496–504.
43. Jocher G. YOLOv5. 2021.

44. Li C, Li L, Jiang H, Weng K, Geng Y, Li L. YOLOv6: a single-stage object detection framework for industrial applications. arXiv preprint 2022. https://arxiv.org/abs/2209.02976
45. Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. p. 7464–75. https://doi.org/10.1109/cvpr52729.2023.00721
46. Ocher G, et al. YOLOv8. 2023.
47. Wan D, Lu R, Hu B, Yin J, Shen S, xu T, et al. YOLO-MIF: improved YOLOv8 with multi-information fusion for object detection in gray-scale images. Advanced Engineering Informatics. 2024;62:102709. https://doi.org/10.1016/j.aei.2024.102709
48. Shen J, Chen Y, Liu Y, Zuo X, Fan H, Yang W. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection. Pattern Recognition. 2024;145:109913. https://doi.org/10.1016/j.patcog.2023.109913
49. Qingyun F, Zhaokui W. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. Pattern Recognition. 2022;130:108786. https://doi.org/10.1016/j.patcog.2022.108786
50. Guo J, Gao C, Liu F, Meng D, Gao X. Damsdet: Dynamic adaptive multispectral detection transformer with competitive query selection and adaptive feature fusion. In: European Conference on Computer Vision. 2024. p. 464–81.
51. Zhu H, Dong W, Yang L, Li H, Yang Y, Ren Y, et al. WaveMamba: Wavelet-Driven Mamba Fusion for RGB-infrared object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025. p. 11219–29.
52. Hao L, Xu L, Liu C, Dong Y. LASFNet: a lightweight attention-guided self-modulation feature fusion network for multimodal object detection. arXiv preprint 2025. https://doi.org/arXiv:250621018