

Article

AQFS-Net: An Adaptive Quality-Aware Fusion and Saliency-Guided Network for Visible-Infrared Object Detection

Weijun Wu and Xufei Zhuang *

College of Intelligent Science and Technology, Inner Mongolia University of Technology,
Hohhot 010080, China; 202310201062@imut.edu.cn

* Correspondence: zxf@imut.edu.cn

Abstract

Object detection in real-world scenarios is often challenged by adverse visual conditions, such as low illumination, strong glare, and dense fog, which severely degrade visible-spectrum features and lead to missed detections, inaccurate localization, and reduced detection accuracy. To address these issues, this paper proposes AQFS-Net, a dual-modal fusion detection network for visible-infrared object detection. Built upon YOLOv13, AQFS-Net adopts a symmetric dual-branch backbone by incorporating infrared images, thereby exploiting the complementary information between the visible and infrared modalities. To alleviate the negative transfer caused by conventional static fusion strategies, an Adaptive Quality-Aware Fusion Module (AQFM) is designed to dynamically enhance informative features and suppress degraded information according to the modality-specific reliability of different regions. In addition, a Foreground-Aware Saliency Guidance (FASG) branch is introduced to guide the network to focus on target regions through foreground supervision, reducing interference from complex backgrounds. Experimental results on the public LLVIP and M3FD datasets show that the proposed method improves mAP@0.5 by 6.8 and 3.1 percentage points, respectively, compared with the baseline using only visible images. These results demonstrate the effectiveness of AQFS-Net in improving dual-modal fusion quality and detection performance under challenging visual conditions, providing a practical reference for visible-infrared object detection in complex illumination scenarios.

Keywords: optical degradation; multimodal fusion; visual sensing; YOLOv13; object detection; challenging illumination

Received: 16 June 2026

Revised: 13 July 2026

Accepted: 20 July 2026

Published: 21 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Object detection is a fundamental task in computer vision and has been widely applied in practical scenarios such as autonomous driving, industrial inspection, and intelligent surveillance [1–3]. Since the pioneering works of YOLO [4] and Fast R-CNN [5] in 2015, the performance of object detection has been continuously improved. In 2020, Carion et al. [6] introduced the Transformer architecture into object detection and proposed the end-to-end DETR framework. Subsequent studies, such as RT-DETR [7], further demonstrated the potential of Transformer-based detectors. However, most existing advanced detectors still rely heavily on favorable illumination conditions, and their perfor-

mance can degrade significantly under adverse imaging conditions such as low light, nighttime scenes, and backlighting [8–11]. In these environments, images often suffer from insufficient brightness, reduced contrast, and severe noise, which blur object features and lead to missed detections and inaccurate localization. These limitations restrict the practical deployment of detection systems in complex real-world scenarios.

To address these challenges, researchers have mainly explored several technical routes. The first route is image enhancement. For example, Zero-DCE [12] adjusts illumination through pixel-level curve mapping, but it may amplify noise or introduce color distortion under extremely dark conditions. Methods such as IA-YOLO [13] attempt to jointly optimize differentiable image processing modules and detection losses. However, it remains difficult to balance enhancement intensity with the semantic requirements of object detection, especially in scenes with complex dynamic ranges, where local over-enhancement or under-enhancement may occur. The second route is physics-based image restoration, such as reconstructing clear images from RAW-domain data to preserve more visual information. Nevertheless, these restore-then-detect pipelines usually incur high computational overhead [14,15]. Moreover, most restoration models are designed for specific degradation types, which limits their generalization ability and practicality under diverse real-world illumination conditions.

The third route is visible-infrared multimodal fusion, which has become an important direction for improving detection performance in complex illumination environments [16–18]. Compared with visible images, infrared images are less sensitive to illumination changes and can provide complementary information in dark or visually degraded scenes. However, existing fusion methods still face several key challenges, including cross-modal feature misalignment, heuristic fusion strategies, and insufficient computational efficiency [17,19–21]. These issues limit the performance gains that can be achieved through cross-modal complementary fusion.

Based on the above observations, this paper adopts YOLOv13 [22] as the baseline model and proposes a visible-infrared dual-modal fusion detection framework for complex illumination scenarios. The proposed framework aims to address the limitations of static fusion, cross-modal negative transfer, and computational redundancy in existing feature fusion methods. By reducing the common computational redundancy of dual-modal networks, the proposed method achieves a better balance among detection accuracy, environmental robustness, and computational cost.

In summary, our core contributions are as follows:

1. End-to-end dual-modal object detection framework: Based on the YOLOv13 architecture, a single-point fusion strategy is used to integrate alignment and enhancement mechanisms, reducing the computational redundancy of traditional multimodal fusion while achieving effective multimodal feature fusion with controllable computational cost.
2. Adaptive Quality-Aware Fusion Module (AQFM): By dynamically modeling cross-modal feature dependencies and introducing a quality evaluation mechanism, AQFM achieves fine-grained spatial alignment and adaptive fusion, alleviating the negative transfer caused by low-quality modalities.
3. Foreground-Aware Saliency Guidance (FASG) branch: An auxiliary task in the training stage is used to strengthen the discriminative feature representation of targets. This branch can be completely removed during inference, improving detection performance with no additional inference cost.
4. Experimental validation of the proposed method: Results on the LLVIP and M3FD datasets show that the proposed method exhibits good robustness and generalization ability under complex illumination conditions and complex scenarios, quantitatively

validating the feasibility of the proposed adaptive fusion and explicit supervision mechanisms for all-weather visual perception tasks.

2. Related Work

2.1. Single-Modality Object Detection

Driven by the development of convolutional neural networks (CNNs), object detection, as a fundamental task in computer vision, has made considerable progress. Early two-stage detectors, such as R-CNN [23], Fast R-CNN [5], and Faster R-CNN [24], established the paradigm of first extracting region proposals and then performing fine-grained classification, achieving relatively high detection accuracy. Subsequently, the YOLO series [4,25] reformulated object detection as a regression problem, enabling efficient real-time inference in industrial applications.

In recent years, Transformer-based architectures have become an alternative approach for object detection. DETR [6] removed non-maximum suppression (NMS) and anchor generation, introducing a fully end-to-end detection paradigm. Building on this, RT-DETR [7] further optimized the encoder capability and query selection strategy, demonstrating the potential of Transformers in real-time scenarios. Despite these advances, standard single-modality detectors still rely heavily on visible-light information. As a result, their detection performance may degrade under adverse conditions such as low illumination, dense fog, or complex backgrounds [26].

2.2. Object Detection Robustness Enhancement in Adverse Environments

To alleviate the decline in detection performance caused by image degradation, researchers have mainly explored two technical routes: image enhancement and joint optimization.

The first type of method focuses on restoring image quality before detection. Zero-DCE [12] models low-light enhancement as an image-specific curve estimation task, adjusting the dynamic range at the pixel level without requiring paired training data. Similarly, generative adversarial networks (GANs) have also been used to synthesize clear images from noisy inputs [27]. However, such enhancement-then-detection serial strategies are usually independent of downstream detection tasks. This separation may lead to suboptimal results, because visual artifacts introduced during enhancement, such as amplified noise, may mislead the detector [14], thereby reducing detection accuracy.

The second type of method directly integrates signal processing into the detection pipeline. IA-YOLO [13] proposes a differentiable image processing (DIP) module that is jointly optimized with the detector, allowing the network to learn enhancement parameters beneficial to detection. However, visible-light-based methods have inherent physical limitations and cannot effectively recover lost information when illumination is completely absent. This limitation indicates the need to introduce the infrared modality to provide complementary information and improve environmental robustness.

2.3. Object Detection Based on Visible-Infrared Fusion

Visible-infrared image fusion is an important strategy for improving detection performance in complex illumination environments. Visible images provide texture details, while infrared images capture thermal radiation information. The complementary characteristics of the two modalities are important for object detection under adverse conditions.

Early multimodal object detection methods mainly relied on simple feature concatenation strategies. For example, the two-stream network proposed by Liu et al. [28] directly fused aligned features to achieve basic multimodal detection capability. Based on this,

related studies [29] further improved feature representation by jointly optimizing detection and segmentation tasks. To address the imbalance in the contributions of different modalities, Li et al. [30] introduced an illumination-aware mechanism to dynamically adjust the weight allocation between visible and infrared modalities according to environmental brightness.

However, such methods usually rely on heuristic designs or fixed strategies. In complex illumination environments, the lack of dynamic discrimination of modality-specific feature quality may introduce background noise and cause negative transfer effects. In recent years, Transformer-based fusion methods have been increasingly explored. Fang et al. [16] used a global attention mechanism to achieve cross-modal feature interaction and improve information modeling. In addition, Zhou et al. [31] further improved detection performance through an inter-modal mutual promotion mechanism. Nevertheless, these methods usually have relatively high computational complexity and still have limitations in fine-grained modality alignment. Therefore, it remains difficult to suppress multimodal negative transfer while controlling computational redundancy, which limits their application in practical scenarios.

3. Methods

3.1. Overall Framework

The proposed dual-modal single-point supervised fusion detection network, termed AQFS-Net, is built upon the YOLOv13 baseline. As shown in Figure 1, the overall architecture consists of three main components: a symmetric dual-branch backbone network, an Adaptive Quality-Aware Fusion Module (AQFM), and a Foreground-Aware Saliency Guidance (FASG) branch. To provide a clearer understanding of the foundational architecture used in this study, the standard network topology of the single-stream YOLOv13 baseline is presented in Appendix A.

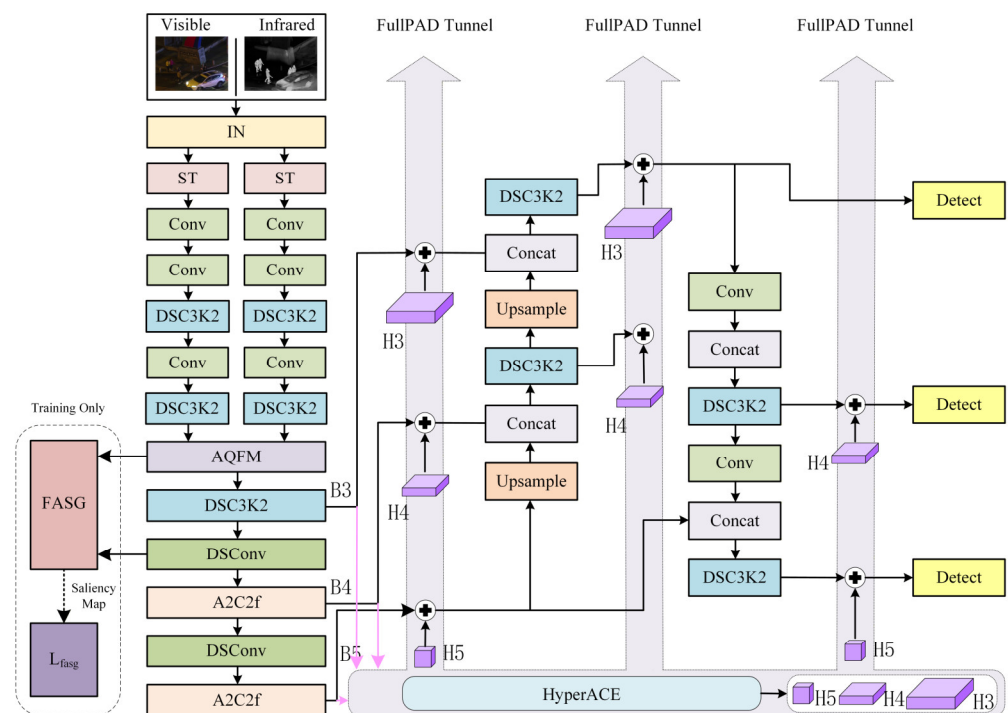


Figure 1. Overall topology of the AQFS-Net dual-modal object detection architecture. B_3 , B_4 , and B_5 denote the multi-scale output feature layers of the backbone network. AQFM performs intermediate-stage single-point adaptive feature fusion. The corresponding multi-scale aggregation levels

of the neck network are denoted as H_3 , H_4 , and H_5 , which correspond in spatial scale to the standard P_3 , P_4 , and P_5 levels commonly used in detection networks. The FASG branch is attached only during training for auxiliary supervision and is removed during inference.

The framework adopts a two-stream input strategy and receives aligned visible and infrared images separately. In the backbone stage, the two branches extract their respective shallow features (P_1 – P_3) in parallel. At the P_3 feature-map scale, AQFM performs modality alignment and adaptive interaction on the dual-stream features to generate a fused P_3 feature stream. Unlike traditional multi-scale or progressive fusion schemes, the proposed framework feeds the fused P_3 features into a shared backbone to further extract higher-level P_4 and P_5 semantic features. Finally, the P_3 features fused by AQFM and the P_4 and P_5 features extracted by the shared backbone are fed into the neck network for multi-scale aggregation. During training, the FASG branch is attached to the P_3 and P_4 feature levels, where an auxiliary task is used to guide the model to focus on target regions. This branch can be completely removed during inference, so it does not introduce additional inference cost.

3.2. Dual-Stream Backbone Feature Extraction

To preserve the original discriminative information of each modality before feature fusion, we construct a symmetric and parameter-independent dual-stream feature extraction pathway. Different from a traditional single-modality architecture, this stage focuses on dual-modal data access, modality decoupling, and modality-specific feature extraction.

3.2.1. Input Processing and Modality Decoupling

In the data input stage, we reconstruct the underlying data loading mechanism of YOLOv13. The aligned 3-channel visible image and 3-channel infrared image are concatenated along the channel dimension to form a 6-channel composite tensor $X_{in} \in R^{(B \times 6 \times H \times W)}$, which serves as the unified input of the network. As shown in Figure 2, to adapt to this input format, the framework introduces IN and ST modules at the beginning of the backbone network.

As the first layer of the entire network, the IN module receives the 6-channel composite input. In the current implementation, it acts as an identity mapping layer and passes the raw concatenated data to the subsequent stream-splitting module, providing a unified access point for the system. The ST module uses tensor slicing operations to separate the 6-channel features along the channel dimension into the visible stream corresponding to the first 3 channels and the infrared stream corresponding to the last 3 channels according to preset indices:

$$X_{vis} = X_{in}[:, :3, :, :] \quad X_{ir} = X_{in}[:, 3:, :, :]$$

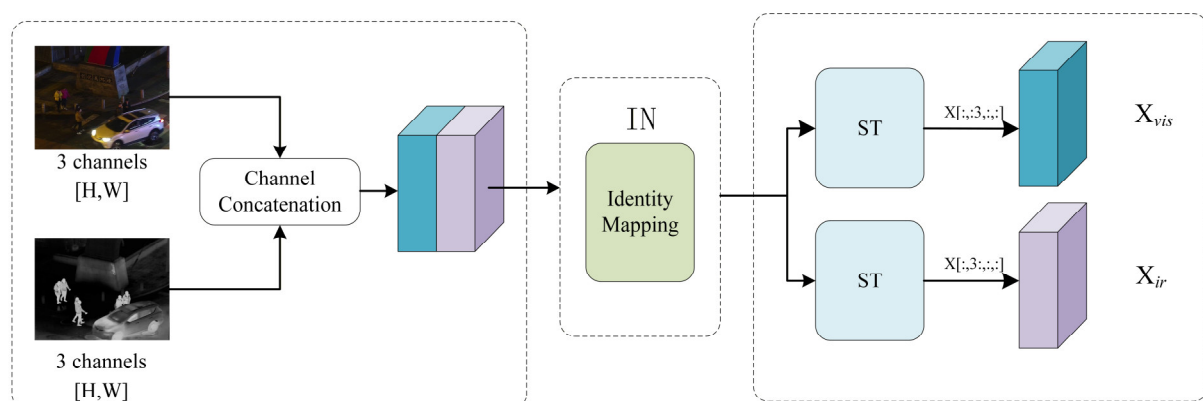


Figure 2. Topology of the input data processing and modality decoupling workflow. “Channel Concatenation” denotes the channel concatenation operation; IN represents the Identity Mapping module; ST represents the tensor slicing and stream-splitting module; X_{vis} and X_{ir} correspond to the independent visible and infrared streams generated after decoupling, respectively.

3.2.2. Modality-Specific Shallow Feature Extraction

After modality decoupling, the visible and infrared feature streams enter two symmetric parallel pathways for feature extraction from the P1 to P3 levels. To avoid semantic confusion between heterogeneous modalities during the shallow feature extraction stage, the two branches do not share any parameters before entering the fusion module. The visible pathway focuses on capturing high-resolution texture details, while the infrared pathway focuses on perceiving the thermal radiation contours of targets. Each branch is composed of alternating standard convolutional (Conv) layers and DSC3k2 modules specific to the YOLOv13 backbone. This independent feature extraction design enables AQFM to directly obtain modality-specific features at the P3 layer.

3.3. Key Module Design

To alleviate negative transfer in multimodal feature fusion and feature degradation under complex illumination conditions, we design two related modules. AQFM is designed to achieve adaptive cross-modal quality awareness and feature fusion, while the FASG branch introduces foreground-aware saliency supervision during training to guide the backbone network to extract more discriminative feature representations. The two modules work from the perspectives of forward feature interaction and backward gradient constraints, respectively, jointly enhancing cross-modal feature fusion.

3.3.1. Adaptive Quality-Aware Fusion Module (AQFM)

After entering the module, the modality-specific P_3 features from the two streams, X_{vis}^{P3} and X_{ir}^{P3} , are first fed into two sets of shallow residual blocks with independent parameters and without batch normalization layers so as to explicitly preserve the original discriminative features of each modality during nonlinear mapping:

$$X_{vis}^{res} = \mathcal{F}_{res1}(X_{vis}^{P3}) \quad X_{ir}^{res} = \mathcal{F}_{res2}(X_{ir}^{P3})$$

where X_{vis}^{P3} and X_{ir}^{P3} denote the input visible and infrared modality-specific feature tensors at the P_3 layer, respectively. $\mathcal{F}_{res1}$ and $\mathcal{F}_{res2}$ represent two residual block operations with non-shared parameters. Since visible and infrared sensors capture fundamentally different optical properties, such as detailed textures and thermal radiation profiles, respectively, employing independent parameters allows each parallel branch to exclusively optimize its modality-specific features without cross-modal interference. Each residual block consists of two cascaded standard convolutional layers with 3×3 kernels and ReLU activation functions. X_{vis}^{res} and X_{ir}^{res} denote the shallow residual feature tensors generated after the mapping, respectively.

To quantify the reliability of each modality-specific feature in the current scene, the module introduces a cross-modal dynamic quality evaluation mechanism. Specifically, this mechanism receives the visible and infrared feature tensors X_{vis}^{res} and X_{ir}^{res} processed by the preceding residual blocks, both with dimensions of $[B, C, H_{P3}, W_{P3}]$, where B denotes the training batch size, $C = 64$ denotes the number of feature channels, and H_{P3} and W_{P3} denote the height and width of the feature maps, respectively. The two feature tensors are first concatenated along the channel dimension to form a composite tensor with dimensions of $[B, 2C, H_{P3}, W_{P3}]$. Then, a global average pooling operation $\text{GAP}(\cdot)$ is used to compress the spatial dimensions of the features, generating a global statistical descriptor with dimensions of $[B, 2C, 1 \times 1]$. To establish a modality confidence mapping

from this descriptor, the module feeds it into a lightweight quality evaluation head, denoted as QualityHead, which consists of two 1×1 convolutional layers. The descriptor is first processed by the first evaluation convolution with the weight tensor $\mathcal{W}1 \in \mathbb{R}^{H_{dim} \times 2C \times 1 \times 1}$, reducing the channel dimension to the hidden dimension $H_{dim} = \max(C/4, 16) = 16$. After a nonlinear mapping introduced by the Gaussian Error Linear Unit (GELU) activation function, it is further processed by the second evaluation convolution with the weight tensor $\mathcal{W}2 \in \mathbb{R}^{2 \times H_{dim} \times 1 \times 1}$ to output the quality evaluation scores of the two modalities. Finally, the Sigmoid activation function $\sigma(\cdot)$ constrains the output scores to the interval $[0, 1]$, and the channel-splitting operation $\text{Split}(\cdot)$ separates them along the channel axis into two independent adaptive spatially broadcast scalar weights, $q_{vis} \in \mathbb{R}^{B \times 1 \times 1 \times 1}$ and $q_{ir} \in \mathbb{R}^{B \times 1 \times 1 \times 1}$.

$$q_{vis}, q_{ir} = \text{Split} \left(\sigma \left(\mathcal{W}_2 * \text{GELU}(\mathcal{W}_1 * \text{GAP}([X_{vis}^{res}, X_{ir}^{res}])) \right) \right)$$

After quality weighting, the module uses a spatial broadcasting mechanism to multiply the predicted scalar quality weights element-wise with the shallow feature-enhanced visible feature tensor X_{vis}^{res} and infrared feature tensor X_{ir}^{res} , respectively. This step aims to adaptively suppress ineffective information in degraded modalities, such as visible-light background noise under extremely dark conditions, and outputs the reweighted feature streams $X_{vis}^{weighted}$ and $X_{ir}^{weighted}$. The adaptive quality weighting process is formulated as follows:

$$X_{vis}^{weighted} = q_{vis} \cdot X_{vis}^{res} \quad X_{ir}^{weighted} = q_{ir} \cdot X_{ir}^{res}$$

Subsequently, the two weighted feature streams are fused along the channel dimension and mapped nonlinearly by a joint representation generator, producing the modality-shared feature ($X_s \in \mathbb{R}^{B \times C \times H_{P3} \times W_{P3}}$), where B denotes the training batch size, $C = 64$ denotes the number of feature channels, and H_{P3} and W_{P3} denote the height and width of the feature maps, respectively. As shown in Figure 3, the overall topology of AQFM is divided into two parallel streams. The upper stream contains a series of spatial-domain modality-specific operations, including modality-specific feature enhancement, cross-attention computation, and bidirectional cross-gating. The lower stream uses the CMCA shared feature refinement branch to reconstruct and refine the feature distribution of X_s .

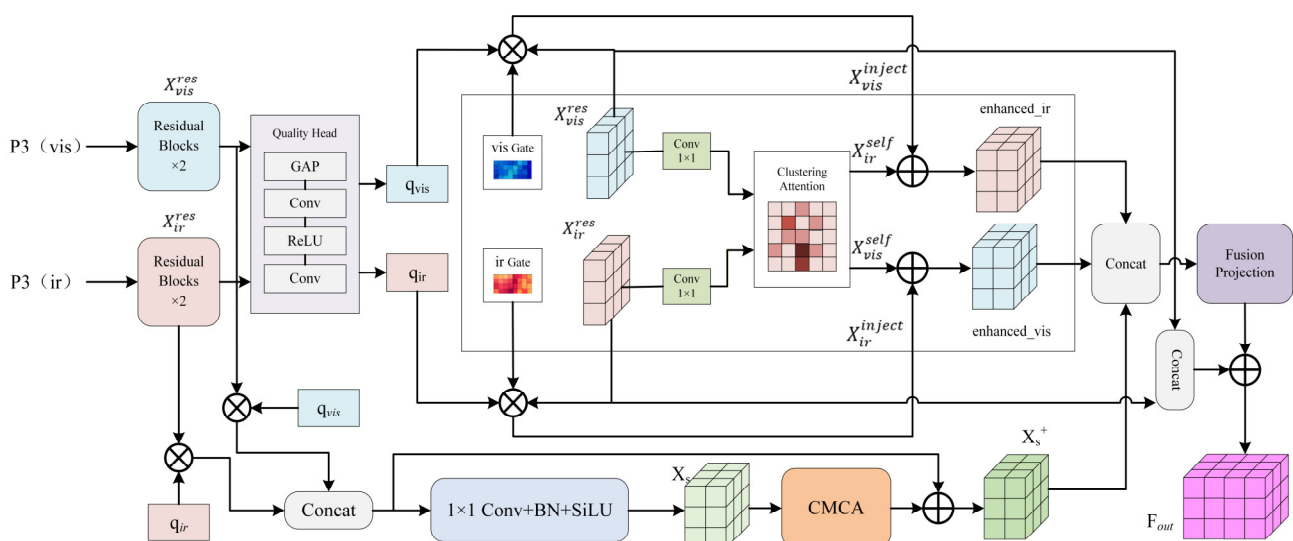


Figure 3. Internal topology of the Adaptive Quality-Aware Fusion Module (AQFM). The overall architecture is divided into upper and lower parallel streams: the upper stream performs modality-specific spatial attention computation and bidirectional heterogeneous feature injection; the lower

stream represents the refinement of the shared feature X_s ; “Fusion Projection” denotes the feature projection layer used for channel normalization.

To overcome the limitation of insufficient latent feature representation in traditional methods, the latent feature X_s is then fed into the lower CMCA refinement branch, as shown in Figure 4. Specifically, this mechanism is driven by the local concatenation of the shared feature X_s and independent modality-specific features. Through parameter-independent clustering convolutions and projections, it separately computes latent class-center features that represent the common semantics of the two modalities in the channel space. The latent feature X_s is concatenated with the infrared feature X_{ir} and projected by a clustering convolution to generate the first cross-modal shared class-center matrix $C_{shared1} \in \mathbb{R}^{B \times C \times N}$, where $N = H_{P3} \times W_{P3}$ denotes the number of flattened spatial pixels. Symmetrically, the latent feature X_s is concatenated with the visible feature X_{vis} to generate the second cross-modal shared class-center matrix $C_{shared2} \in \mathbb{R}^{B \times C \times N}$. The two class-center features containing multimodal clustering commonalities are concatenated and fed into the inverse clustering reconstruction operation $\text{Unclustering}(\cdot)$. A 1×1 convolutional projection is then used to map the features back to the original latent space. At the output end, the reconstructed feature is added to the original latent feature X_s through a skip connection without additional transformation, finally producing the refined cross-modal shared feature $X_s^+ \in \mathbb{R}^{B \times C \times H_{P3} \times W_{P3}}$. The mathematical expression of the lower shared feature refinement and reconstruction process is as follows:

$$X_s^+ = \text{Unclustering}(\text{Concat}(C_{shared1}, C_{shared2})) + X_s$$

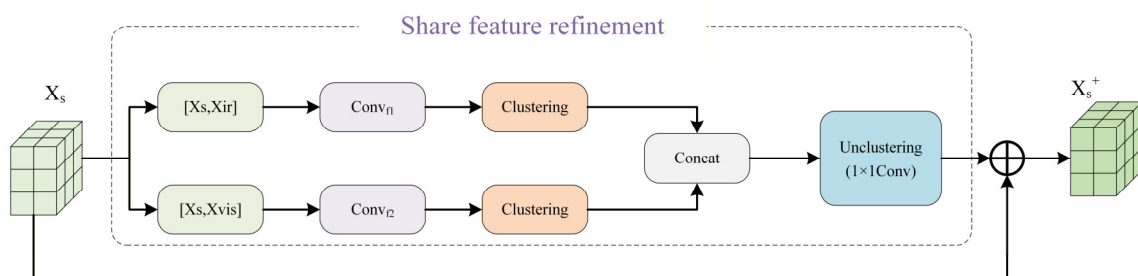


Figure 4. Detailed architecture of the shared feature refinement (CMCA) branch. This branch shows the internal implementation of the lower CMCA module in Figure 3. “Clustering” and “Unclustering” represent the clustering projection and inverse clustering reconstruction mapping of latent common semantics in the feature space, respectively. A skip connection is used at the end to maintain stable deep gradient propagation.

While the lower branch computes the class centers, the latent class-center matrices $C_{shared1}$ and $C_{shared2}$ derived from the lower refinement branch are passed upward as guiding information to drive the cross-modal clustering attention interaction in the upper stream. This mechanism models the semantic dependencies between heterogeneous channels through a cross-modal correlation matrix. Taking the visible branch as an example, batch matrix multiplication is performed between the infrared-to-shared class-center matrix $C_{shared1}$ delivered from the lower branch and the transposed visible feature matrix $V_{vis}^T \in \mathbb{R}^{B \times N \times C}$ obtained through value projection convolution. A scaling factor $\sqrt{d}$ is introduced, where $d = 64$ is the feature channel dimension, to prevent the matrix multiplication results from falling into the saturation region of the Softmax function when the channel dimension is large. Finally, the channel-level cross-correlation attention matrix $A_{vis} \in \mathbb{R}^{B \times C \times C}$ is generated through the Softmax function. This matrix is then multiplied with the value-projected features to output the visible attention feature tensor $X_{vis}^{att} \in$

$\mathbb{R}^{B \times C \times H_{P3} \times W_{P3}}$ required by the upper enhancement stream. The infrared attention stream X_{ir}^{att} is computed in a symmetric manner. The mathematical expression of the upper cross-modal correlation attention interaction process is as follows:

$$A_{vis} = \text{Softmax}\left(\frac{C_{shared1} \cdot V_{vis}^T}{\sqrt{d}}\right) X_{vis}^{att} = A_{vis} \cdot V_{vis}$$

After obtaining the attention-enhanced streams X_{vis}^{att} and X_{ir}^{att} , the module adds them element-wise to the modality-specific shallow residual features X_{vis}^{res} and X_{ir}^{res} , which are directly mapped by the preceding convolutional layers and both have dimensions of $[B, C, H_{P3}, W_{P3}]$. The resulting visible and infrared self-enhanced feature tensors are denoted as X_{vis}^{self} and X_{ir}^{self} , respectively:

$$X_{vis}^{self} = X_{vis}^{att} + X_{vis}^{res} \quad X_{ir}^{self} = X_{ir}^{att} + X_{ir}^{res}$$

To further refine the multimodal feature flow, the module introduces a bidirectional cross-gating mechanism. The relatively stable spatial distribution of the infrared modality is used to generate the spatial mask gate g_{vis} , which suppresses background noise in the visible branch. Meanwhile, the rich texture and edge information of the visible modality is used to generate the structural guidance gate g_{ir} , which enriches the local boundary representation of the infrared branch. In this process, the concatenated features are fed into the gating branches and processed by standard 3×3 convolutional kernels $\mathcal{W}_{gate1}, \mathcal{W}_{gate2} \in \mathbb{R}^{1 \times 2C \times 3 \times 3}$, producing the final single-channel spatial gating masks $g_{vis}, g_{ir} \in \mathbb{R}^{B \times 1 \times H_{P3} \times W_{P3}}$:

$$g_{vis} = \sigma(\mathcal{W}_{gate1} * [X_{vis}^{self}, X_{ir}^{res}]) \quad g_{ir} = \sigma(\mathcal{W}_{gate2} * [X_{ir}^{self}, X_{vis}^{res}])$$

The gating masks are applied to the shallow residual features of the opposite modality through Hadamard products and are combined with the scalar quality evaluation scores obtained from the quality evaluation stage to generate the cross-modal injection feature streams $X_{vis}^{inject}, X_{ir}^{inject} \in \mathbb{R}^{B \times C \times H_{P3} \times W_{P3}}$:

$$X_{vis}^{inject} = q_{ir} \cdot g_{vis} \odot X_{ir}^{res} \quad X_{ir}^{inject} = q_{vis} \cdot g_{ir} \odot X_{vis}^{res}$$

The module performs a triple concatenation along the channel dimension by combining the modality-specific representations enhanced by bidirectional injection in the upper stream, namely $X_{vis}^{self} + X_{vis}^{inject}$ and $X_{ir}^{self} + X_{ir}^{inject}$, with the shared refined feature X_s^+ output by the lower refinement branch. Then, a channel-level fusion projection layer is used to normalize the number of channels to $2C$, producing the projected multimodal fused feature $F_{fused} \in \mathbb{R}^{B \times 2C \times H_{P3} \times W_{P3}}$:

$$F_{fused} = \text{Proj}\left(\text{Concat}(X_{vis}^{self} + X_{vis}^{inject}, X_{ir}^{self} + X_{ir}^{inject}, X_s^+)\right)$$

Finally, to maintain stable deep gradient backpropagation and reduce the semantic degradation of heterogeneous features after multiple nonlinear projections, the module directly adds the fused feature stream F_{fused} to the shallow independent residual stream calculated in the first stage without additional transformation. The final output tensor of AQFM is denoted as $F_{out} \in \mathbb{R}^{B \times 2C \times H_{P3} \times W_{P3}}$:

$$F_{out} = F_{fused} + \text{Concat}(X_{vis}^{res}, X_{ir}^{res})$$

3.3.2. Foreground-Aware Saliency-Guided (FASG) Auxiliary Branch

To alleviate the limitation that the final detection loss provides sparse constraints on the intermediate fusion module during training, we introduce a Foreground-Aware Saliency Guidance (FASG) auxiliary branch, as shown in Figure 5. This branch serves as an intermediate single-point explicit supervision mechanism and is directly attached to the output end of the fusion module. It is designed to provide direct backward gradient con-

straints for the fusion module through pixel-level saliency reconstruction of foreground regions, thereby guiding the fusion of cross-modal heterogeneous features.

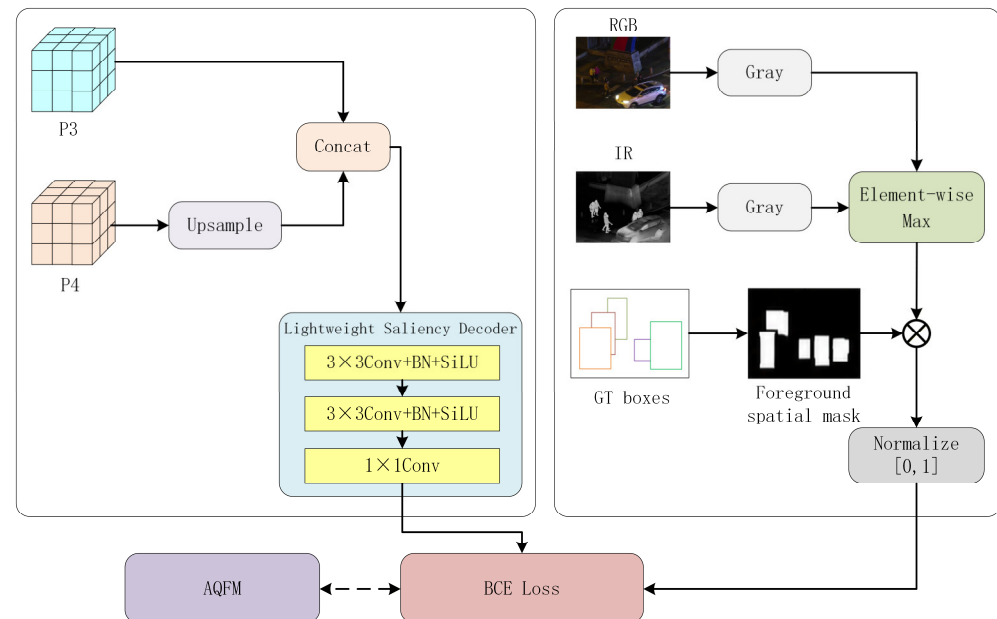


Figure 5. Flowchart of forward decoding and supervised ground-truth generation in the Foreground-Aware Saliency Guidance (FASG) branch. The left side shows the forward process of multi-scale feature fusion and the lightweight saliency decoder. The right side illustrates the backward supervision logic, where a foreground spatial mask is dynamically constructed based on the input dual-modal images and ground-truth boxes (GT boxes) and then filtered to generate the target ground-truth map M_{target} .

The FASG branch receives the core fused feature tensor F_{out} directly output by the AQFM and the high-level semantic feature tensor X_{shared}^{P4} extracted by the shared backbone network in parallel, where X_{shared}^{P4} represents the P_4 level and $C_2 = 128$. The branch first upsamples the high-level feature map X_{shared}^{P4} to the P_3 scale through bilinear interpolation to achieve spatial alignment. It is then concatenated with F_{out} along the channel axis to form a composite tensor with dimensions of $[B, C_1 + C_2, H_{P3}, W_{P3}]$, which is fed into a lightweight saliency decoder. The decoder gradually compresses the feature channels from high dimensions to 64 and 32 through two 3×3 convolutional layers. Finally, a 1×1 convolutional layer outputs a single-channel unactivated saliency prediction map $M \in \mathbb{R}^{B \times 1 \times H_{P3} \times W_{P3}}$. The mathematical expression of this multi-scale feature decoding process is as follows, where $\mathcal{F}_{decoder}$ represents the cascaded forward mapping function of the lightweight saliency decoder, and the output M denotes the feature activation intensity:

$$M = \mathcal{F}_{decoder} \left(\text{Concat} \left(F_{out}, \text{Interpolate}(X_{shared}^{P4}) \right) \right)$$

To ensure that the auxiliary backward gradient can guide the AQFM to extract effective complementary foreground information and to avoid interference with the fusion constraints from infrared background thermal noise, such as heated road surfaces, or sudden high-frequency noise in visible images, a foreground-aware spatial mask mechanism is introduced at the supervision end. This mechanism directly uses the original dual-modal input images $I_{vis}, I_{ir} \in \mathbb{R}^{B \times 3 \times H \times W}$ as the source, where H and W denote the original height and width of the input images. The two images are first averaged along the channel axis to obtain single-channel grayscale images I_{vis}^{gray} and I_{ir}^{gray} , and then fused by a pixel-wise maximum operation to generate the initial grayscale map $I_{max} =$

$\text{Max}(I_{vis}^{gray}, I_{ir}^{gray}) \in \mathbb{R}^{B \times 1 \times H \times W}$. Meanwhile, the system uses the ground-truth boxes to dynamically generate a binary foreground spatial mask $S_{mask} \in \mathbb{R}^{B \times 1 \times H \times W}$, where pixels inside the ground-truth boxes are assigned a value of 1.0, and pixels in non-target background regions are assigned a value of 0.0. After I_{max} and S_{mask} are multiplied element-wise to remove background noise, sample-wise min-max normalization is applied to constrain the confidence values to the interval $[0,1]$. Finally, the normalized map is downsampled to the P_3 resolution to generate the final saliency-guided target ground-truth map $M_{target} \in \mathbb{R}^{B \times 1 \times H_{P3} \times W_{P3}}$. Its mathematical expression is as follows, where $\min(\cdot)$ and $\max(\cdot)$ denote the extremum scalars of the valid foreground region and $\epsilon = 10^{-6}$ is a predefined small scalar used to prevent division by zero.

$$M_{target} = \text{Downsample} \left(\frac{(I_{max} - \min) \odot S_{mask}}{\max - \min + \epsilon} \right)$$

The forward flow of the FASG branch is controlled by the self-training conditional gate. During training, the prediction map M and the ground-truth map M_{target} are jointly fed into the binary cross-entropy loss function for optimization. The backward gradient directly acts on the output end of the intermediate fusion module, thereby adjusting the parameters of AQFM and improving feature fusion. During inference, this branch is completely bypassed, and the internal upsampling and decoding convolutions are not executed. This mechanism ensures that no additional floating-point operations are introduced during inference, enabling explicit enhancement of multimodal fused features with no additional inference cost.

3.4. Loss Function

To achieve the joint optimization of object detection accuracy and feature discriminative ability, the proposed framework adopts a multi-task joint loss function for end-to-end training. The total loss L_{total} consists of the main object detection loss L_{det} and the foreground-aware saliency guidance loss L_{fasg} , with the weighting coefficient λ_{fasg} set to 0.02.

$$L_{total} = L_{det} + \lambda_{fasg} L_{fasg}$$

The selection of this value is based on preliminary exploratory experiments and gradient magnitude balance analysis. Experimental results show that if λ_{fasg} is set too low, the backward gradient constraint provided by the auxiliary branch is insufficient, making it difficult to guide the backbone network to focus its feature responses on target regions. Conversely, if λ_{fasg} is set too high, since the foreground guidance loss performs pixel-level dense supervision over the full image of size $H \times W$, the initial gradient flow may excessively dominate the update of network parameters, thereby interfering with the convergence process of the main detection task. Setting $\lambda_{fasg} = 0.02$ achieves a better balance between feature reshaping and the detection task, as demonstrated in Table 1.

Table 1. Sensitivity Analysis of λ_{fasg} on the LLVIP and M3FD Datasets.

λ_{fasg}	LLVIP		M3FD	
	mAP@0.5	mAP@0.5:0.95	mAP@0.5	mAP@0.5:0.95
0.05	0.955	0.613	0.825	0.558
0.02	0.957	0.630	0.846	0.587
0.001	0.952	0.607	0.828	0.560

The detection loss follows the optimization scheme of YOLOv13 and consists of three parts: classification loss (L_{cls}), bounding box regression loss (L_{box}), and distribution focal loss (L_{dfl}).

$$L_{det} = \lambda_{cls}L_{cls} + \lambda_{box}L_{box} + \lambda_{dfl}L_{dfl}$$

The foreground-aware saliency guidance loss (L_{fasg}) is used to impose pixel-level constraints on the saliency prediction map M generated by the auxiliary branch. Since the prediction map M output by the lightweight decoder consists of unactivated logits, the proposed method adopts binary cross-entropy loss with logits to improve numerical stability of the gradients.

$$L_{fasg} = -\frac{1}{N} \sum_{i=1}^N \left[M_{target}^{(i)} \cdot \ln(\sigma(M^{(i)})) + (1 - M_{target}^{(i)}) \cdot \ln(1 - \sigma(M^{(i)})) \right]$$

where N denotes the total number of pixels in the feature maps of the current input batch, namely $N = B \times H_{p3} \times W_{p3}$, $M^{(i)}$ denotes the predicted logit value at the i -th pixel, and $\sigma(\cdot)$ denotes the standard Sigmoid activation function; $M_{target}^{(i)}$ denotes the saliency ground truth at the corresponding position after foreground spatial mask filtering and normalization, with a value range of $[0,1]$. Through the explicit constraint of this loss function, the backward gradient directly acts on the output end of the intermediate fusion module, encouraging the backbone network to strengthen its response to real target regions during feature extraction while suppressing irrelevant environmental interference such as infrared thermal noise, thereby improving the fusion quality of cross-modal heterogeneous features.

4. Experimental Studies

4.1. Experimental Setup

The proposed network is implemented using the PyTorch (version 2.4.0) deep learning framework. All training and testing tasks are conducted on a high-performance computing server equipped with an NVIDIA RTX 3090 GPU. To ensure a consistent experimental environment, the software configuration includes Ubuntu 22.04, Python 3.9, and CUDA 12.0. The main hardware configuration includes an Intel (R) Xeon (R) Silver 4214 CPU @ 2.20GHz and 64 GB of RAM. All input images are uniformly resized to 640×640 . During training, the Mosaic data augmentation strategy is adopted to alleviate model overfitting. The training batch size is set to 32, and a cosine annealing strategy is used for learning rate adjustment to promote stable convergence.

Considering the differences between the LLVIP and M3FD datasets in terms of scene distribution, number of categories, and object characteristics, different hyperparameter configurations, such as training epochs and learning rates, are adopted for the two datasets to promote sufficient convergence, as detailed in Table 2.

Table 2. Hyperparameter Configurations for Different Datasets.

Dataset	Batch	Initial LR	Cosine LR	Optimizer	Epochs
LLVIP	32	0.0005	Yes	AdamW	100
M3FD	32	0.001	Yes	AdamW	300

To ensure the reliability and comparability of the experimental results, ablation studies and comparative experiments on the same dataset use the same training hyperparameters.

4.2. Datasets and Evaluation Metrics

To evaluate the proposed method under different scenarios and object categories, we conduct experiments on two multimodal object detection datasets, LLVIP and M3FD.

LLVIP is a visible-infrared pedestrian detection dataset for low-light environments, containing more than 15,000 strictly aligned visible and infrared image pairs, most of which were captured in nighttime scenes. This dataset is widely used to evaluate multimodal detection performance under low-illumination conditions. In terms of dataset splitting, the official LLVIP dataset provides only training and test sets, without a public validation set. To ensure a rigorous experimental protocol and effective model training, the original official test set is used as an independent external test set only for final model evaluation. Meanwhile, the original training set is randomly split into new training and validation sets at a ratio of 8:2.

M3FD is a multimodal dataset for general object detection tasks. It contains multiple object categories and more complex environmental conditions, making it suitable for evaluating the generalization ability of multimodal detection models in complex scenes. Since M3FD does not provide an official standard split, all 4200 paired infrared-visible images are divided into training, validation, and test sets at a ratio of 6:2:2 using stratified random sampling with a fixed random seed. The detailed dataset statistics and split configurations are summarized in Table 3.

Table 3. Dataset Statistics and Split Details.

Dataset	Description	Classes	Training Set	Validation Set	Test Set
LLVIP	Low-light scene	1	9620	2405	3463
M3FD	Complex scenes	6	2520	840	840

For evaluation metrics, mAP@0.5, mAP@0.75, and mAP@0.5:0.95 are adopted as the main performance indicators. In addition, precision, recall, the number of parameters, and computational cost are reported to provide a more comprehensive evaluation of model performance.

5. Results and Analysis

5.1. Comparative Experiments

To evaluate the effectiveness of the proposed method, we conduct systematic comparisons with single-modality detectors, image enhancement methods, and representative multimodal fusion methods on two public multimodal object detection datasets, LLVIP and M3FD.

The single-modality detectors mainly include YOLOv8 [32], a single-stage anchor-free architecture commonly used in industrial deployment; YOLO11 [33], which introduces multi-scale joint attention to balance detection speed and accuracy; and YOLOv13 [22], the baseline model adopted in this work.

The cascaded image enhancement and object detection schemes include Zero-DCE [12], a low-light enhancement algorithm based on non-parametric zero-reference curve mapping, and OneRestore [34], a general image restoration framework based on a multi-branch network for handling composite scene degradations. The enhanced images processed by these two methods are then fed into the YOLOv13 detector for detection.

The end-to-end multimodal feature fusion networks include RT-DETR-Fusion [35], a real-time object detector based on an anchor-free Transformer architecture; YOLO-MIF [36], which introduces a multi-channel information complementarity mechanism to improve edge perception efficiency; DEYOLO [37], a cross-modal object detector using a semantic-spatial bidirectional collaborative enhancement strategy; and YOLOv11-RGBT [35], which extends the network topology to dual-stream inputs.

As shown by the experimental results on the LLVIP dataset in Table 4, the infrared single-modality detector achieves higher overall performance than the visible-light

model. Under the IR modality, YOLOv13 obtains an mAP@0.5 of 0.947 and an mAP@0.5:0.95 of 0.618, whereas the RGB modality only reaches 0.889 and 0.509, respectively. This indicates that in extremely low-light environments, the texture information in visible images is severely degraded, while the infrared modality can still preserve target thermal radiation features because it does not rely on ambient illumination.

However, although the infrared modality has better stability under low-light conditions, it still has certain limitations in representing edge details and target contours. Therefore, when relying only on a single infrared modality, further improvement in detection accuracy is often limited. In comparison, the proposed dual-modal fusion framework uses the texture details of the visible modality and the thermal radiation response of the infrared modality to obtain a more complementary feature representation.

Specifically, the proposed method achieves an mAP@0.5 of 0.957 and an mAP@0.5:0.95 of 0.630 on the LLVIP dataset. Compared with the YOLOv13-IR baseline, the improvements are 1.0 and 1.2 percentage points, respectively. Compared with YOLOv13-RGB, the improvements reach 6.8 and 12.1 percentage points, respectively. In particular, under a higher IoU threshold, the mAP@0.75 of the proposed method reaches 0.719, which is higher than those of the other compared methods. This suggests that the proposed model improves not only detection recall but also localization accuracy and bounding box regression quality.

Furthermore, compared with existing multimodal methods, the proposed method also achieves better results. Compared with YOLO-MIF, the proposed method improves mAP@0.5:0.95 by 2.8 percentage points. Compared with DEYOLO and YOLOv11-RGBT, the improvements are 4.9 and 18.3 percentage points, respectively. It is worth noting that some multimodal models even perform worse than the single-modality infrared detector. This suggests that simple feature concatenation or static fusion strategies may be affected by noise interference. As a result, the fused features may be disturbed by low-quality modalities, leading to negative transfer and reduced detection accuracy.

This performance difference can be attributed in part to the dynamic quality-aware mechanism of AQFM. AQFM does not directly fuse the two modalities in a static manner. Instead, it first evaluates the quality of different modality features and dynamically adjusts their contribution weights. When visible images suffer from severe noise or texture loss in nighttime scenes, the module suppresses the interference of low-quality RGB features in the fusion space. In regions with rich texture information, the detailed information from the visible branch is retained.

In addition, the proposed method introduces a computational cost of 14.5 GFLOPs. Although this is slightly higher than that of YOLO-MIF and YOLOv11-RGBT, it is still much lower than the 194.0 GFLOPs of RT-DETR-Fusion, while achieving higher detection accuracy. This indicates that the proposed method achieves a better balance between detection performance and computational cost.

Table 4. Performance Comparison of Different Methods on the LLVIP Dataset.

Setting	Model	Modality	Precision	Recall	mAP@0.5	mAP@0.75	mAP@0.5:0.95	Params	GFLOPs
Single	YOLOv8n	RGB	0.908	0.814	0.891	0.535	0.509	3,005,843	8.1
		IR	0.955	0.897	0.953	0.672	0.603	3,005,843	8.1
	YOLO11n	RGB	0.938	0.831	0.899	0.529	0.51	2,582,347	6.3
		IR	0.944	0.907	0.956	0.664	0.601	2,582,347	6.3
	YOLOv13n	RGB	0.918	0.819	0.889	0.531	0.509	2,448,090	6.2
		IR	0.957	0.881	0.947	0.697	0.618	2,448,090	6.2
Enhanced	Zero-DCE + YOLOv13	RGB	0.895	0.83	0.897	0.508	0.497	2,527,510	38.6
	OneRestore +	RGB	0.887	0.81	0.878	0.502	0.491	8,428,090	78.9

YOLOv13									
Multimodal	RT-DETR-Fusion	RGB + IR	0.867	0.804	0.873	0.457	0.467	66,329,443	194
	YOLO-MIF	RGB + IR	0.944	0.865	0.938	0.687	0.602	3,011,187	8.9
	DEYOLO	RGB + IR	0.928	0.842	0.926	0.649	0.581	6,008,143	16.8
	YOLOv11-RGBT	RGB + IR	0.818	0.764	0.824	0.442	0.447	3,784,883	9.3
	AQFS-Net	RGB + IR	0.96	0.897	0.957	0.719	0.63	3,231,375	14.5

To further evaluate the generalization ability of the model, we conduct experiments on the more complex M3FD dataset, as shown in Table 5. Unlike LLVIP, which mainly focuses on single-category pedestrian detection at night, M3FD includes more complex illumination conditions, such as daytime, nighttime, low light, and strong glare, as well as multiple object categories. Therefore, it can better reflect the adaptability of the model in real-world scenarios.

The experimental results show that, under complex scenarios, most fusion models suffer from more obvious performance degradation. For example, the mAP@0.5:0.95 of YOLO-MIF decreases to 0.519, and that of YOLOv11-RGBT further decreases to 0.508. This indicates that cross-modal inconsistencies become more pronounced in complex illumination environments, making it difficult for traditional fusion methods to stably handle dynamic illumination and background interference.

In comparison, the proposed method still achieves an mAP@0.5 of 0.846 and an mAP@0.5:0.95 of 0.587, representing improvements of 3.1 and 4.6 percentage points over the YOLOv13-RGB baseline, respectively. Meanwhile, its recall reaches 0.782, which is higher than those of the compared methods. This indicates that the proposed method can preserve useful target information more stably and reduce missed detections in complex illumination environments and multi-category scenarios.

It is worth noting that the proposed method shows larger performance gains on M3FD than on LLVIP, especially in terms of mAP@0.75 and mAP@0.5:0.95. This suggests that AQFM and FASG can not only enhance dual-modal feature fusion but also improve the localization quality of the model. This improvement is mainly attributed to the foreground saliency supervision imposed by the FASG branch on target regions during training. Such supervision encourages the model to focus more on real target regions rather than background thermal noise or local texture interference, thereby improving the discriminative ability of the fused features.

Table 5. Performance Comparison of Different Methods on the M3FD Dataset.

Setting	Model	Modality	Precision	Recall	mAP@0.5	mAP@0.75	mAP@0.5:0.95	Params	GFLOPs
Single	YOLOv8n	RGB	0.836	0.736	0.805	0.561	0.538	3,006,818	8.1
		IR	0.879	0.712	0.792	0.55	0.522	3,006,818	8.1
	YOLO11n	RGB	0.849	0.733	0.809	0.561	0.54	2,583,322	6.3
		IR	0.886	0.692	0.786	0.546	0.524	2,583,322	6.3
	YOLOv13n	RGB	0.858	0.74	0.815	0.569	0.541	2,449,065	6.2
		IR	0.847	0.724	0.791	0.553	0.525	2,449,065	6.2
Enhanced	Zero-DCE + YOLOv13n	RGB	0.88	0.728	0.809	0.559	0.537	2,528,485	38.6
	OneRestore + YOLOv13n	RGB	0.843	0.752	0.814	0.562	0.541	8,429,065	78.9
Multimodal	RT-DETR-Fusion	RGB + IR	0.757	0.729	0.79	0.508	0.492	66,339,718	194
	YOLO-MIF	RGB + IR	0.835	0.715	0.788	0.547	0.519	3,157,344	8.9

DEYOLO	RGB + IR	0.855	0.716	0.799	0.565	0.53	6,000,422	16.6
YOLOv11-RGBT	RGB + IR	0.825	0.706	0.781	0.532	0.508	3,785,858	9.3
AQFS-Net	RGB + IR	0.89	0.782	0.846	0.618	0.587	3,232,350	14.5

Based on the experimental results on the two datasets, the proposed method achieves better detection performance under complex illumination conditions and shows relatively stable performance in multi-category and multi-scene environments. These results indicate that the proposed method has certain advantages in terms of robustness and generalization ability.

5.2. Ablation Studies

To further evaluate the contribution of each component, we conduct ablation studies on the LLVIP and M3FD datasets, as detailed in Tables 6 and 7. The experiments use the single-modality YOLOv13 detector as the baseline and progressively introduce the dual-modal fusion structure, the AQFM, and the FASG auxiliary supervision branch to analyze their effects on the overall detection performance.

Table 6. Ablation Results on the LLVIP Dataset.

Model	Modality	Precision	Recall	mAP@0.5	mAP@0.75	mAP@0.5:0.95	Params	GFLOPs
YOLOv13n	RGB	0.918	0.819	0.889	0.531	0.509	2,448,090	6.2
	IR	0.957	0.881	0.947	0.697	0.618	2,448,090	6.2
Dual-modal (Concat)	RGB + IR	0.946	0.836	0.927	0.636	0.572	2,483,386	7.0
Dual-modal + AQFM	RGB + IR	0.949	0.88	0.947	0.703	0.608	3,065,262	14.5
Dual-modal + AQFM + FASG	RGB + IR	0.96	0.897	0.957	0.719	0.63	3,231,375	14.5

Table 7. Ablation Results on the M3FD Dataset.

Model	Modality	Precision	Recall	mAP@0.5	mAP@0.75	mAP@0.5:0.95	Params	GFLOPs
YOLOv13n	RGB	0.858	0.74	0.815	0.569	0.541	2,449,065	6.2
	IR	0.847	0.724	0.791	0.553	0.525	2,449,065	6.2
Dual-modal (Concat)	RGB + IR	0.891	0.738	0.829	0.575	0.555	2,484,361	7.0
Dual-modal + AQFM	RGB + IR	0.87	0.773	0.838	0.608	0.572	3,066,237	14.5
Dual-modal + AQFM + FASG	RGB + IR	0.89	0.782	0.846	0.618	0.587	3,232,350	14.5

When only simple feature concatenation is used for dual-modal fusion, the model already shows a clear improvement over the single-modality baseline using only one modality. On the LLVIP dataset, mAP@0.5:0.95 increases from 0.509 to 0.572; on the M3FD dataset, it increases from 0.541 to 0.555. This indicates that complementary information exists between the visible and infrared modalities. In particular, under complex illumination conditions, infrared thermal radiation information can help compensate for the information loss caused by visible texture degradation.

However, the improvement brought by simple concatenation is still limited, because this strategy cannot distinguish the quality differences in different modalities in different scenarios. When one modality is severely degraded, direct fusion may introduce redundant noise, thereby limiting further performance improvement.

After introducing the AQFM, the model performance further improves. On the LLVIP dataset, mAP@0.75 increases from 0.636 to 0.703, and mAP@0.5:0.95 increases to 0.608. On the M3FD dataset, mAP@0.75 increases from 0.575 to 0.608. These results indicate that AQFM improves the fusion quality of cross-modal features.

It is worth noting that the improvement brought by AQFM on the high-IoU metric mAP@0.75 is more evident than that on mAP@0.5. This suggests that the module not only improves detection performance but also contributes to more accurate bounding box localization.

After further adding the FASG auxiliary supervision branch, the model performance is further improved. On the LLVIP dataset, mAP@0.5:0.95 increases from 0.608 to 0.630; on the M3FD dataset, it increases from 0.572 to 0.587. Meanwhile, the recall values increase to 0.897 and 0.782, respectively, indicating that FASG improves the model's perception of target regions.

Overall, the ablation results show that the modules have a complementary relationship. The dual-modal structure provides the basis for complementary information, AQFM alleviates the limitations of static fusion and reduces the negative transfer caused by low-quality modalities, and FASG further improves the target-aware ability of the fused features. With these components, the model achieves more stable object detection performance under complex illumination conditions.

5.3. Statistical Validation

To assess the stability and reliability of the proposed AQFS-Net, we conducted a statistical validation comprising repeated training runs.

First, to verify that the performance improvements are not heavily dependent on a specific random seed, AQFS-Net was trained independently four times using different random initialization states. As summarized in Table 8, the model exhibits stable training performance across both datasets. On the LLVIP dataset, the model achieves a mean mAP@0.5 of 0.953 with a relatively small standard deviation of ± 0.004 . Similarly, on the M3FD dataset, the mean mAP@0.5 reaches 0.839 ± 0.005 . The maximum standard deviation observed across all evaluation metrics is 0.010. This indicates that the proposed AQFS-Net is insensitive to initial random states and can maintain stable convergence. It further confirms that the model performance reported in this paper is reliable and reproducible, rather than stemming from an accidentally favorable parameter initialization.

Table 8. Detection Performance of AQFS-Net under Different Random Seeds on the LLVIP and M3FD Datasets.

Random Seed	LLVIP mAP@0.5	LLVIP mAP@0.5:0.95	M3FD mAP@0.5	M3FD mAP@0.5:0.95
Seed 1	0.957	0.630	0.846	0.587
Seed 2	0.955	0.610	0.838	0.567
Seed 3	0.949	0.615	0.838	0.574
Seed 4	0.952	0.624	0.834	0.564
Mean $\pm$ Std	0.953 ± 0.004	0.620 ± 0.009	0.839 ± 0.005	0.573 ± 0.010

Second, to confirm that the performance gains over existing multimodal models are supported by statistical tests, we computed the 95% confidence intervals and p -values for the performance deltas Δ mAP. To ensure a highly rigorous evaluation, the two top-performing multimodal architectures from the comparative experiments, DEYOLO and YOLO-MIF, were selected as the representative comparative methods for this significance testing. As shown in Table 9, AQFS-Net yields reliable positive performance shifts. On both the LLVIP and M3FD datasets, compared to these two top-performing multimodal methods, the proposed method achieves relatively higher performance metrics. All p -values strictly satisfy $p < 0.001$, which indicates statistical significance and rejects the null hypothesis. Specifically, on the M3FD dataset, the 95% confidence intervals for Δ mAP@0.5 against DEYOLO and YOLO-MIF are [0.0339, 0.0591] and [0.0419, 0.0737], respectively,

both of which strictly do not contain zero. Similarly, on the LLVIP dataset, the model shows stable positive performance deltas, where the $\Delta mAP@0.5:0.95$ metric increases by +0.049 and +0.028 against the two comparative methods, respectively. These statistical testing results indicate that the performance improvements of AQFS-Net under complex illumination conditions are mathematically supported.

Table 9. Statistical Significance Testing of AQFS-Net Against Comparative Methods.

Dataset	Baseline	$\Delta mAP@0.5$	95% CI	<i>p</i> -Value	$\Delta mAP@0.5:0.95$	95% CI	<i>p</i> -Value
LLVIP	DEYOLO	+0.031	[0.0263, 0.0366]	<0.001	+0.049	[0.0440, 0.0548]	<0.001
	YOLO-MIF	+0.019	[0.0141, 0.0244]	<0.001	+0.028	[0.0223, 0.0345]	<0.001
M3FD	DEYOLO	+0.047	[0.0339, 0.0591]	<0.001	+0.057	[0.0470, 0.0661]	<0.001
	YOLO-MIF	+0.058	[0.0419, 0.0737]	<0.001	+0.068	[0.0575, 0.0776]	<0.001

5.4. Qualitative Visual Analysis

To more intuitively demonstrate the practical applicability of the proposed method under complex illumination conditions, this section provides a qualitative visual analysis of the experimental results. Since the M3FD dataset covers more complex all-weather conditions, including daytime, nighttime, low light, strong glare, and dense fog, and contains more diverse object categories with more evident multimodal inconsistencies, all qualitative visualization experiments in this section are conducted on the M3FD test set. The analysis is conducted from two aspects: qualitative comparison of object detection results and ablation analysis of intermediate feature activation heatmaps.

As shown by the single-modality detection results in the first two columns of Figure 6, single-stream detectors have certain limitations in detection robustness under degraded environmental conditions. Under nighttime low illumination and strong oncoming glare, the visible-only detector (RGB) misses some nearby pedestrians and moving vehicles due to local overexposure or underexposure in low-light regions. In contrast, in daytime dense fog or high-background-radiation environments, the infrared-only detector (IR) may confuse distant low-contrast traffic targets with background contours because of the lack of color and boundary texture details, resulting in localization deviations. When handling these scenarios, the dual-stream static concatenation scheme (Concat) tends to introduce redundant modality features in regions with direct strong light and road surface reflections because it does not distinguish feature quality, leading to more false detections. Even YOLO-MIF and DEYOLO, which show relatively good performance in the comparative experiments, still have limitations such as loose bounding boxes, low confidence scores, and missed or false detections for some multi-category targets when facing small-object occlusion, dense fog, or complex lighting conditions. In comparison, the proposed method shows relatively stable target detection under complex illumination conditions. For distant buses occluded by dense fog, pedestrians under low illumination, and vehicles under direct glare, the predicted bounding boxes of the proposed method are relatively compact, with better localization accuracy and classification confidence. This indicates that the proposed method can integrate complementary dual-modal features and reduce information conflicts between heterogeneous modalities, thereby improving the overall performance of all-weather multi-object detection.

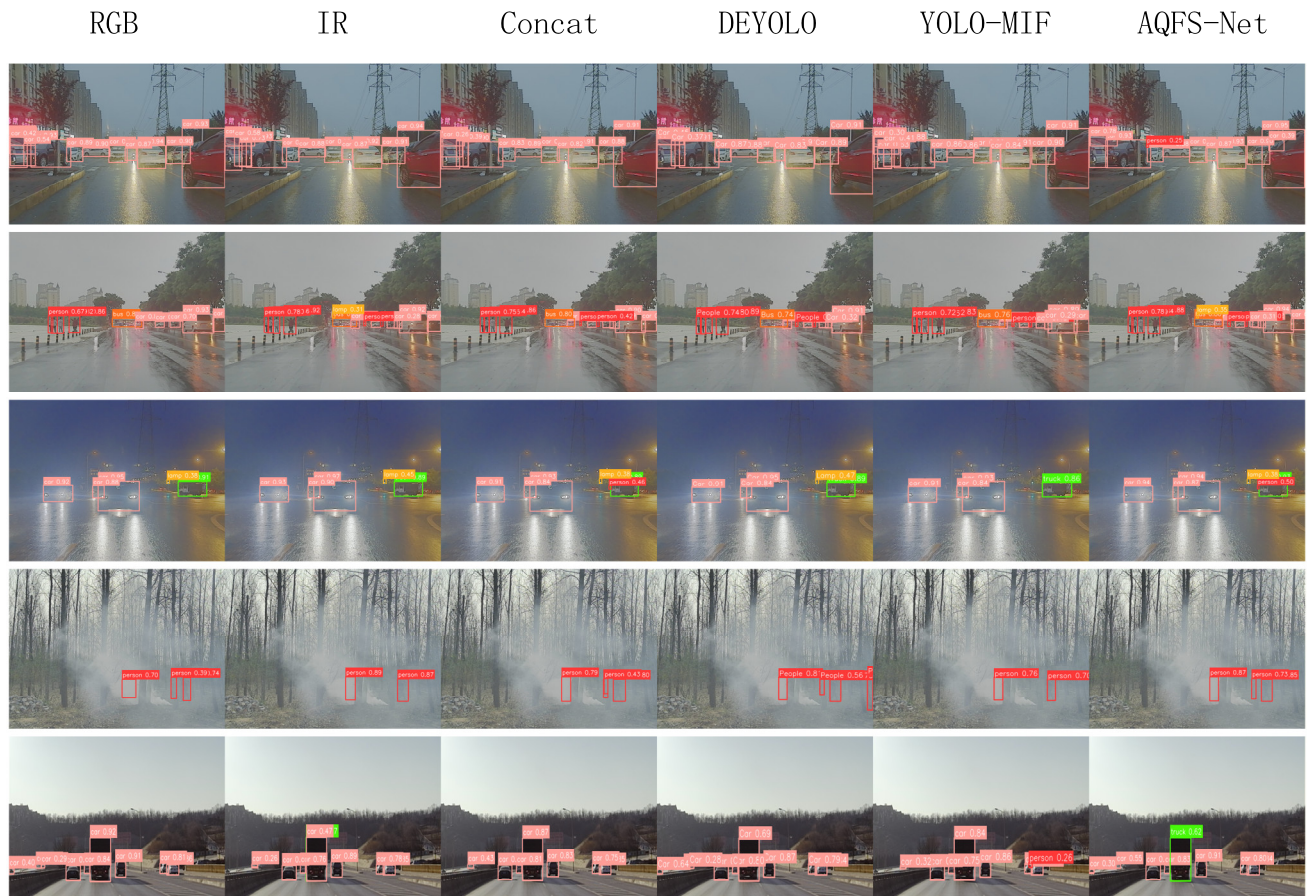


Figure 6. Qualitative comparison of object detection results under typical adverse conditions on the M3FD test set. From left to right, the columns correspond to the visible single-modality baseline (RGB), infrared single-modality baseline (IR), dual-stream static concatenation (Concat), DEYOLO, YOLO-MIF, and AQFS-Net. The visualized results include bounding box predictions and classification confidences in real degraded scenarios, such as nighttime low illumination, strong glare interference, and dense fog.

To further analyze the specific roles of AQFM and FASG from the perspectives of feature flow and computational mechanism, we use Grad-CAM to visualize and compare the activation heatmaps of the intermediate fusion-layer features.

As shown by the local details in the third column of Figure 7, since the standard static concatenation scheme does not filter the feature quality of the input modalities, the activation regions in the feature maps show a certain degree of scattering. In rainy-night scenes and areas with mixed strong light, the high-activation regions of the static concatenation features are distributed over strong reflection areas on wet road surfaces, surrounding building walls, and vehicle headlight glare. In dense fog scenes, the heatmap energy is also distributed over non-target fog regions, which weakens the discriminative ability of target features due to environmental interference. In comparison, the feature fusion mechanism of the proposed method shows clearer target-oriented activation regions. Across different test scenes, background radiation noise, glare scattering, and fog interference are suppressed to some extent, resulting in lower activation weights in background areas. Correspondingly, the feature activation tends to concentrate on the regions and geometric boundaries of real vehicles, pedestrians, and distant low-contrast targets.

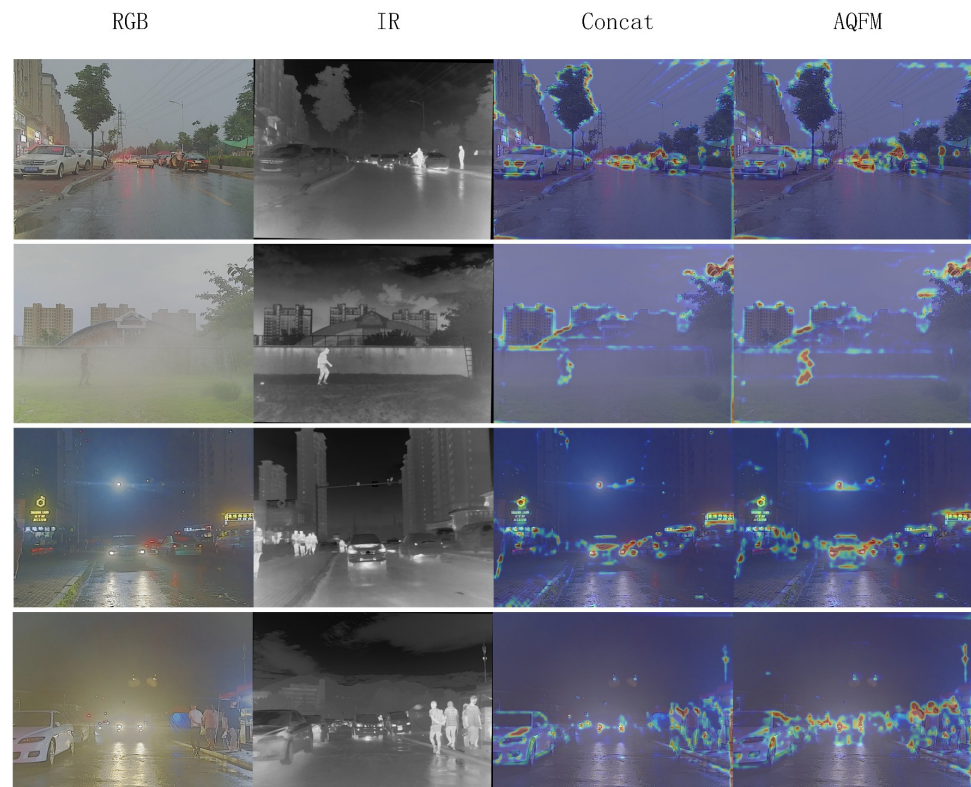


Figure 7. Visualization of feature activation heatmaps in the fusion layer. The first two columns show the paired input visible (RGB) and infrared (IR) images, respectively. The third column shows the activation heatmaps of the dual-stream static concatenation (Concat) feature layer, and the fourth column shows the core feature activation heatmaps obtained after applying the proposed feature fusion mechanism.

This improvement in feature distribution is consistent with the improvements in the quantitative detection metrics discussed above. After applying AQFM, the high-activation regions in the feature maps are less affected by non-target rainy-road reflections, glare scattering, and fog interference. Meanwhile, the FASG auxiliary branch attached during training introduces supervision constraints through the foreground spatial mask, encouraging the network activation to concentrate on the core regions and geometric boundaries of real vehicles and pedestrians. This improves the discriminative ability of the fused features between targets and the background.

6. Discussion

6.1. Physical and Optical Considerations

Although AQFS-Net is evaluated using semantic detection metrics, its architecture is motivated by the physical differences between visible and infrared sensing. Visible images mainly encode reflected-light information and provide texture, color, and boundary details, but their feature reliability can decrease under low illumination, glare, overexposure, and fog scattering. Infrared images capture thermal radiation responses and are less dependent on ambient illumination, but they may also contain thermal background interference, weak boundaries, and sensor-related noise. Therefore, the proposed dual-stream structure aims to preserve modality-specific optical information before fusion, allowing the network to exploit the complementary physical properties of the two sensing modalities.

Within AQFM, the QualityHead provides a feature-level mechanism for estimating the relative reliability of visible and infrared features under different imaging conditions.

When visible features are degraded by illumination-related noise, glare, or contrast loss, the module is expected to reduce their negative influence during fusion. When infrared features are affected by thermal clutter or weak target-background contrast, their contribution can also be adaptively adjusted. In this sense, the QualityHead does not directly model physical sensor parameters, but it provides a data-driven way to reflect reliability variations caused by sensor noise and optical degradation in the feature space.

Cross-modal inconsistencies may arise from sensor-view differences, local registration deviations, lens response variations, or imaging-condition changes, causing spatial responses from the two modalities to be partially inconsistent. Instead of directly concatenating such heterogeneous features, AQFM introduces cross-modal correlation attention and bidirectional cross-gating to model inter-modal dependencies and selectively exchange complementary information. This design allows the fused representation to retain target-related responses from both modalities while reducing interference from inconsistent or redundant modality-specific features. In addition, FASG provides foreground-aware supervision during training, encouraging the fused features to focus on target-related optical responses while suppressing non-target responses caused by road-surface reflections, thermal backgrounds, glare, or fog. From the perspective of photonics and optical engineering, the proposed architecture provides a learning-based feature fusion strategy for exploiting complementary visible and infrared sensing information under degraded imaging conditions and may provide a reference for multimodal optical perception algorithm design.

6.2. Limitations and Failure Cases

While AQFS-Net demonstrates robustness across various illumination and weather conditions, it exhibits limitations under extreme, compounded environmental degradations involving simultaneous nighttime darkness, wet road surface reflections, and severe headlight glare. Figure 8 illustrates two representative failure cases. In Scenario 1, the model generates a false positive by misclassifying an unknown heat source, which possesses a misleading thermal signature, as a pedestrian. In Scenario 2, severe visible and infrared reflections and artifacts corrupt the information, leading to the omission of distant, small pedestrian targets. These failures suggest that under extreme optical and thermal disturbances, current fusion mechanisms may struggle to disambiguate target features from noise artifacts. Future work will focus on developing more fine-grained cross-modal feature learning and fusion mechanisms to suppress compound degradations. Furthermore, exploring the integration of sensors that are robust to optical and thermal interference, such as millimeter-wave radar, will be considered to establish a more comprehensive multi-modal perception system. In addition, from an architectural perspective, while recent state-of-the-art multimodal Transformer architectures exhibit strong capabilities in capturing global contextual dependencies, we opted for a CNN-based YOLO architecture to maintain a better balance between local feature extraction efficiency and overall computational cost. Exploring the integration of Transformer-based self-attention mechanisms into AQFS-Net to further enhance global cross-modal alignment without introducing excessive computational overhead remains a promising direction for future research.



Figure 8. Representative failure cases of AQFS-Net under extreme conditions. (**Top row**): False positive caused by a misleading heat source. (**Bottom row**): False negative caused by severe optical and thermal reflections.

7. Conclusions

This paper constructs and implements AQFS-Net, an adaptive quality-aware fusion and saliency-guided visible-infrared dual-modal object detection network based on the single-stage detector YOLOv13. The proposed network adopts a dual-branch backbone to extract visible and infrared modality-specific features, uses AQFM to perform cross-modal quality-aware fusion, and introduces the FASG auxiliary supervision branch to enhance the model's ability to focus on target regions during training. The FASG branch can be removed during inference without introducing additional inference cost.

Experimental results on the public LLVIP and M3FD datasets show that the proposed method achieves further improvements over the better-performing single-modality baseline while maintaining manageable computational cost. Under complex illumination conditions and multi-scene environments, the proposed method also outperforms the compared multimodal methods, indicating that it can alleviate the negative transfer effect in conventional dual-modal fusion methods.

Future work will focus on more lightweight cross-modal dynamic fusion mechanisms and adaptive modality selection strategies under uncertain degradation conditions, aiming to improve the generalization ability and deployment efficiency of the model in more complex real-world scenarios.

Author Contributions: Conceptualization, W.W. and X.Z.; methodology, W.W.; software, W.W.; validation, W.W.; formal analysis, W.W.; investigation, W.W.; resources, X.Z.; data curation, W.W.; writing—original draft preparation, W.W.; writing—review and editing, W.W. and X.Z.; visualization, W.W.; supervision, X.Z.; project administration, X.Z.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Inner Mongolia University of Technology College Student Innovation and Entrepreneurship Training Program under grant number 2025026010, and the Science and Technology Plan Projects of Inner Mongolia Autonomous Region under grant number 2020GG0104.

Institutional Review Board Statement: Not applicable. This study used publicly available datasets and did not involve new data collection from human participants or animals.

Informed Consent Statement: Not applicable. This study used publicly available datasets and did not involve the collection of new data from human participants.

Data Availability Statement: The datasets analyzed in this study are publicly available. The LLVIP dataset is available at <https://bupt-ai-cz.github.io/LLVIP/> (accessed on 2 July 2026), and the M3FD dataset is available at <https://github.com/JinyuanLiu-CV/TarDAL> (accessed on 2 July 2026). No new datasets were created in this study. The custom code underlying this article is available on reasonable request to the corresponding author due to ongoing related research. To facilitate future benchmarking, the hardware and software requirements are detailed in Section 4.1.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

To provide a clearer understanding of the foundational architecture used in this study, the standard network topology of the single-stream YOLOv13 baseline is presented in Figure A1. Additionally, the official implementation and source code of this baseline can be accessed at <https://github.com/iMoonLab/yolov13> (accessed on 2 July 2026).

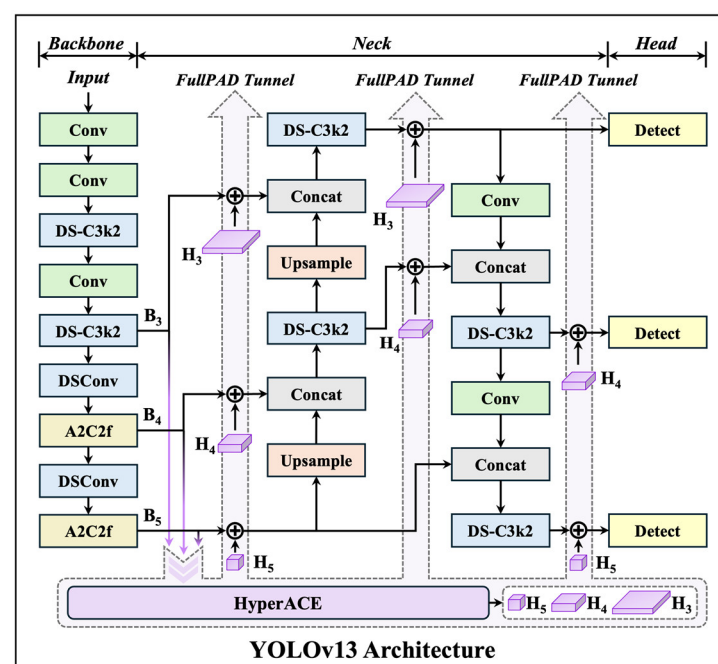


Figure A1. Official network architecture of the baseline YOLOv13 model.

References

1. Nikolaidis, S.; Koukaras, P. Vision and Multimodal Perception for Autonomous Driving: Deep Learning Architectures, Tasks, and Sensor Fusion. *World Electr. Veh. J.* **2026**, *17*, 277. <https://doi.org/10.3390/wevj17060277>.
2. Wang, Q.; Wang, M.; Sun, J.; Chen, D.; Shi, P. Review of Surface-Defect Detection Methods for Industrial Products Based on Machine Vision. *IEEE Access* **2025**, *13*, 90668–90697. <https://doi.org/10.1109/ACCESS.2025.3571297>.
3. Xie, X.; Zhan, Y.; Ma, Z.; Liu, H.; Chen, Z.; Luo, X.; Xu, X. Distributed Learning for Privacy-Preserving Semi-Supervised Video Anomaly Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 9161–9174. <https://doi.org/10.1109/TCSVT.2025.3550803>.
4. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788. <https://doi.org/10.1109/cvpr.2016.91>.
5. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448. <https://doi.org/10.1109/iccv.2015.169>.

6. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Computer Vision—ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.
7. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974. <https://doi.org/10.1109/cvpr52733.2024.01605>.
8. Wang, W.; Wu, X.; Yuan, X.; Gao, Z. An Experiment-Based Review of Low-Light Image Enhancement Methods. *IEEE Access* **2020**, *8*, 87884–87917. <https://doi.org/10.1109/access.2020.2992749>.
9. Guo, J.; Ma, J.; García-Fernández, Á.F.; Zhang, Y.; Liang, H. A survey on image enhancement for Low-light images. *Heliyon* **2023**, *9*, e14558. <https://doi.org/10.1016/j.heliyon.2023.e14558>.
10. Galea, S.; Seychell, D.; Bugeja, M. A Survey of Intelligent transportation systems based Modern Object Detectors Under Night-time Conditions. In *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 3–5 December 2020; pp. 265–270. <https://doi.org/10.1109/iciss49785.2020.9316076>.
11. Tiwari, A.K.; Pattanaik, M.; Sharma, G.K. Low-light DETECTION TRansformer (LDETR): Object detection in low-light and adverse weather conditions. *Multimed. Tools Appl.* **2024**, *83*, 84231–84248. <https://doi.org/10.1007/s11042-024-19087-x>.
12. Guo, C.; Li, C.; Guo, J.; Loy, C.C.; Hou, J.; Kwong, S.; Cong, R. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 14–19 June 2020; pp. 1777–1786. <https://doi.org/10.1109/cvpr42600.2020.00185>.
13. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence, Virtual*, 22 February–1 March 2022; pp. 1792–1800. <https://doi.org/10.1609/aaai.v36i2.20072>.
14. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3291–3300. <https://doi.org/10.1109/cvpr.2018.00347>.
15. Chang, M.; Feng, H.; Xu, Z.; Li, Q. Low-Light Image Restoration with Short- and Long-Exposure Raw Pairs. *IEEE Trans. Multimed.* **2022**, *24*, 702–714. <https://doi.org/10.1109/tmm.2021.3058586>.
16. Fang, Q.; Han, D.; Wang, Z. Cross-Modality Fusion Transformer for Multispectral Object Detection. *SSRN Electron. J.* **2022**, 4227745. <https://doi.org/10.2139/ssrn.4227745>.
17. Zhang, X.; Zhang, X.; Wang, J.; Ying, J.; Sheng, Z.; Yu, H.; Li, C.; Shen, H. TFDet: Target-Aware Fusion for RGB-T Pedestrian Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 13276–13290. <https://doi.org/10.1109/tnnls.2024.3443455>.
18. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. <https://doi.org/10.1016/j.patcog.2023.109913>.
19. Wang, R.; Zhou, Z.; Li, S.; Zhang, Z. Advances and challenges in infrared-visible image fusion: A comprehensive review of techniques and applications. *Artif. Intell. Rev.* **2026**, *59*, 18. <https://doi.org/10.1007/s10462-025-11426-0>.
20. Sun, C.; Zhang, C.; Xiong, N. Infrared and Visible Image Fusion Techniques Based on Deep Learning: A Review. *Electronics* **2020**, *9*, 2162. <https://doi.org/10.3390/electronics9122162>.
21. Song, K.; Zhao, Y.; Huang, L.; Yan, Y.; Meng, Q. RGB-T image analysis technology and application: A survey. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105919. <https://doi.org/10.1016/j.engappai.2023.105919>.
22. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
23. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 23–28 June 2014; pp. 580–587. <https://doi.org/10.1109/cvpr.2014.81>.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. <https://doi.org/10.1109/tpami.2016.2577031>.
25. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. <https://doi.org/10.1109/cvpr.2017.690>.
26. Gallagher, J.E.; Oughton, E.J. Surveying You Only Look Once (YOLO) Multispectral Object Detection Advancements, Applications, and Challenges. *IEEE Access* **2025**, *13*, 7366–7395. <https://doi.org/10.1109/access.2025.3526458>.
27. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134. <https://doi.org/10.1109/cvpr.2017.632>.

28. Liu, J.; Zhang, S.; Wang, S.; Metaxas, D. Multispectral Deep Neural Networks for Pedestrian Detection. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 73.1–73.12. <https://doi.org/10.5244/c.30.73>.
29. Li, C.; Song, D.; Tong, R.; Tang, M. Multispectral Pedestrian Detection via Simultaneous Detection and Segmentation. In Proceedings of the British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018; p. 225. Available online: <http://bmvc2018.org/contents/papers/0738.pdf> (accessed on 2 July 2026).
30. Li, C.; Song, D.; Tong, R.; Tang, M. Illumination-aware faster R-CNN for robust multispectral pedestrian detection. *Pattern Recognit.* **2019**, *85*, 161–171. <https://doi.org/10.1016/j.patcog.2018.08.005>.
31. Zhou, K.; Chen, L.; Cao, X. Improving Multispectral Pedestrian Detection by Addressing Modality Imbalance Problems. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 787–803. https://doi.org/10.1007/978-3-030-58523-5_46.
32. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 11 June 2026).
33. Jocher, G.; Qiu, J. Ultralytics YOLO11. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 11 June 2026).
34. Guo, Y.; Gao, Y.; Lu, Y.; Zhu, H.; Liu, R.W.; He, S. OneRestore: A Universal Restoration Framework for Composite Degradation. In Proceedings of the European Conference on Computer Vision (ECCV), Milan, Italy, 29 September–4 October 2024; pp. 255–272. https://doi.org/10.1007/978-3-031-72655-2_15.
35. Wan, D.; Lu, R.; Fang, Y.; Lang, X.; Shu, A.; Chen, J.; Shen, S.; Xu, T.; Ye, Z. YOLOv11-RGBT: Towards a Comprehensive Single-Stage Multispectral Object Detection Framework. *arXiv* **2025**, arXiv:2506.14696. <https://doi.org/10.48550/arXiv.2506.14696>.
36. Wan, D.; Lu, R.; Hu, B.; Yin, J.; Shen, S.; Xu, T.; Lang, X. YOLO-MIF: Improved YOLOv8 with Multi-Information fusion for object detection in Gray-Scale images. *Adv. Eng. Inform.* **2024**, *62*, 102709. <https://doi.org/10.1016/j.aei.2024.102709>.
37. Chen, Y.; Wang, B.; Guo, X.; Zhu, W.; He, J.; Liu, X.; Yuan, J. DEYOLO: Dual-Feature-Enhancement YOLO for Cross-Modality Object Detection. In Proceedings of the International Conference on Pattern Recognition (ICPR), Kolkata, India, 1–5 December 2024; pp. 236–252. https://doi.org/10.1007/978-3-031-78447-7_16.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.