CAF-YOLO: YOLOv8-based Bimodal Pedestrian Detection Method Using Cross-Attention Fusion

Jianhong Li[1], Haoran Wu[1], Xiangjing Wei[1], Haiyi Huang[2], Yaojuan Wang[1], and Wenhui Zhang[1] (✉)

[1] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin, China
[2] School of Artificial Intelligence, Guilin University of Electronic Technology, Guilin, China
zhangwh@guet.edu.cn

Abstract. Visible-infrared bimodal pedestrian detection holds significant application value in scenarios such as intelligent surveillance and unmanned ground vehicles. However, visible images suffer performance degradation under low-light conditions, while infrared images lack detailed texture information. Moreover, existing fusion methods still have limitations in real-time performance, interaction efficiency, and lightweight design. Advancements to the YOLOv8 architecture, founded on cross-attention fusion, are introduced herein as CAF-YOLO to overcome these specific hurdles.The method designs a Cross-Attention Fusion (CAF) module to achieve feature complementarity and semantic alignment between the two modalities through bidirectional attention mechanism. Meanwhile, the Efficient Multi-scale Attention (EMA) component is integrated, aiming to bolster the model's representational power.Experiments on LLVIP and FLIR datasets show that CAF-YOLO achieves mAP@0.5 improvements of 9.1% and 1.2% over visible and infrared single-modal baselines on LLVIP, and 17.0% and 3.6% on FLIR, respectively. The model has only 11.74M parameters, significantly outperforming fusion methods with higher computational complexity. Ablation studies and visualization analysis verify the effectiveness of each module, and the model maintains robust detection capability in low-light and occlusion scenarios.

Keywords: Pedestrian detection, Bimodal fusion, Cross-attention, YOLOv8, Visible-infrared images

1 Introduction

Visible-infrared bimodal pedestrian detection holds significant application value in fields such as intelligent surveillance and unmanned ground vehicles. In practice, single-modal methods face major limitations. Visible-spectrum captures often suffer from severe distortion when collected in poorly lit or backlit environments, while infrared images, though illumination-invariant, lack texture details. Both issues lead to limited detection accuracy. The LLVIP dataset[5] and FLIR ADAS dataset[3] are currently two

benchmark datasets commonly used in bimodal pedestrian detection research, providing a unified platform for algorithm evaluation.

Existing bimodal fusion methods still lack efficiency in real-time performance and lightweight design. In response to these challenges, the CAF-YOLO framework is introduced. It features a Cross-Attention Fusion (CAF) component designed for dual-path modal interaction, alongside an Efficient Multi-scale Attention (EMA) unit that strengthens feature modeling capabilities. Such a mechanism achieves precise and instantaneous bimodal sensing of pedestrians.

The main contributions of this paper include:

1. Leveraging a Cross-Attention Fusion (CAF) unit, our approach strives to exploit how visible and infrared features complement each other via a bidirectional attention mechanism;
2. We design an improved C2f unit with EMA (C2f-EMA) to enhance feature representation capability;
3. We construct the CAF-YOLO model based on the YOLOv8 framework and conduct comprehensive experiments on the LLVIP and FLIR datasets aiming to demonstrate the robustness of our approach under poor illumination and shielded environments.

The structure of this paper is as follows: Section 2 contextualizes our study by evaluating existing methodologies in bimodal detection.; Section 3 details the CAF-YOLO model design; Section 4 presents experimental setup and results analysis; Section 5 concludes this study by synthesizing the primary insights while proposing potential avenues for subsequent investigations.

2 Related Work

The LLVIP dataset[5] and FLIR ADAS dataset[3] are currently two benchmark datasets commonly used in visible-infrared bimodal pedestrian detection research. LLVIP contains 30,976 images (15,488 pairs), focusing on pedestrian detection tasks with annotations containing only the pedestrian category. Most images are captured under low-light conditions. FLIR ADAS, released by FLIR Corporation, focuses on autonomous driving scenarios. Zhang et al.[12] first proposed a manually aligned version of the FLIR dataset comprising 5,142 synchronized RGB-IR image pairs, partitioned into sets for training (4,129 pairs) and evaluation (1,013 pairs). This dataset retains three main target categories: pedestrians, vehicles, and bicycles. These datasets provide a unified evaluation platform for tasks such as bimodal fusion and pedestrian detection, promoting the development of related algorithms.

Bimodal fusion seeks to extensively exploit the cross-modal synergies from visible and infrared modalities. The MCFF module[2] adaptively fuses color stream and thermal stream features according to lighting conditions, exploring feature transmission methods at different fusion stages. GAFF[13] introduces an attention-based multispectral feature fusion method that dynamically weights and fuses multispectral features through mutual-modal and self-spectral attention components, significantly improving detection accuracy while maintaining low computational cost. The CSSA module[1]

achieves efficient fusion under computationally constrained scenarios through lightweight channel switching and spatial attention operations. M2I2HA[11] constructs a multimodal perception network based on hypergraph theory, capturing intra-modal global high-order relationships through hypergraph enhancement modules and achieving cross-modal feature alignment and fusion through hypergraph fusion modules. These works demonstrate that an effective cross-modal interaction mechanism is key to improving bimodal detection performance.

Attention mechanisms, as an important means of enhancing feature representation, have become increasingly prevalent across diverse computer vision applications over the past few years. SENet[4] proposed a channel attention mechanism where inter-channel correlations are modeled to achieve adaptive refinement of feature intensity, laying the foundation for subsequent attention models. The EMA framework[6] further aggregates multi-scale information through cross-dimensional interactions, preserving rich channel information while reducing computational overhead, and performs excellently in image classification and object detection tasks. Zhu et al.[14] introduced the EMA mechanism into the YOLOv8 framework, proposing the C2f-Faster-EMA structure that replaces the original BlockNeck with FastBlock and inserts EMA modules before them, enhancing the model's feature representation capability.

Transformer-based architectures show strong potential in cross-modal fusion. By leveraging the Transformer framework, the Cross-modal Fusion Transformer (CFT) [7] executes concurrent intra-modal and inter-modal integration through self-attention. This approach excels at resolving the complex dependencies between thermal and RGB data, significantly improving multispectral object detection performance. GM-DETR[10] proposes a General Multispectral Detection Transformer, originating a Modality-specific Feature Interaction (MSFI) module to capture sophisticated RGB and IR data, while incorporating a Cross-modal Scale Feature Fusion (CMSF) unit to harmonize these representations to achieve multi-scale cross-modal fusion. The capacity for capturing extensive contextual relationships inherent in the Transformers provides a new solution for cross-modal feature alignment.

Building on YOLOv8, we introduce two key improvements. First, we design a fine-grained C2f-EMA module that applies multi-scale attention after each BlockNeck. Second, we design a Cross-Attention Fusion (CAF) component, which aims to align and complement visible and infrared features via bidirectional attention.

3 Method

3.1 Overall Architecture

The overall architecture of CAF-YOLO is shown in **1**Fig.**1**. The model takes visible and infrared images as bimodal inputs, each passing through independent feature extraction branches. Each branch is based on YOLOv8's backbone network (CSPDarknet) for feature extraction, generating multi-scale feature maps. We insert CAF modules at multiple FPN levels to enable bidirectional interaction between modalities. In the neck, we replace the original C2f blocks with C2f-EMA to strengthen feature representation.

Detailed internal configurations of the CAF and EMA units are presented in Fig. 2a and Fig. 2b, respectively. The entire network adopts an end-to-end training approach, with the output layer maintaining YOLOv8's detection head design to predict bounding boxes, class confidence, and object presence probability.

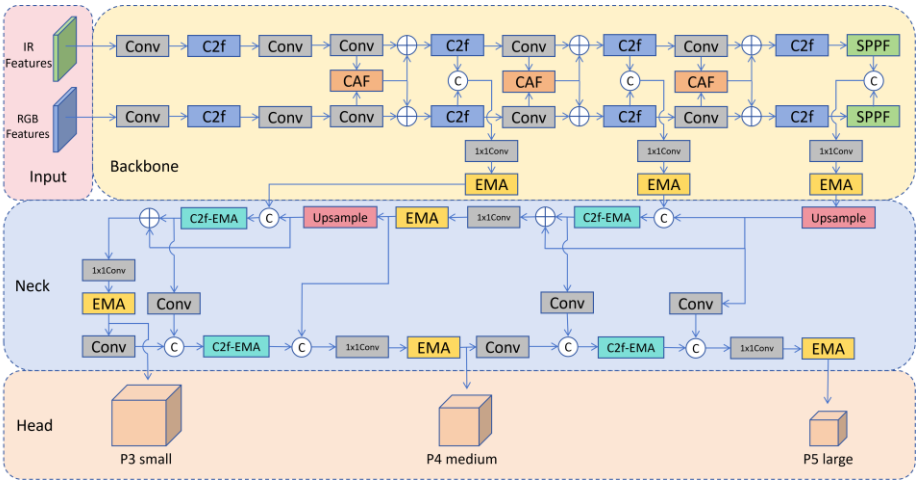


Fig. 1. Overall Architecture of CAF-YOLO

3.2 Cross-Attention Fusion (CAF)

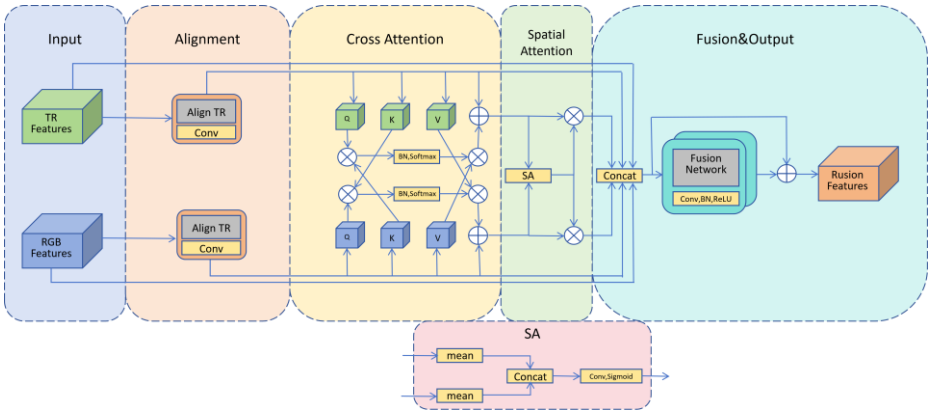
The Cross-Attention Fusion (CAF) module is a core component of CAF-YOLO, designed to achieve effective information complementarity between visible and infrared modalities. CAF inherits the self-attentional paradigm established by the Transformer[8], achieving semantic alignment and feature enhancement between modalities through bidirectional cross-attention. As shown in Fig.2a, the forward computation of the CAF module includes the following steps:

Given input features $X \in \mathbb{R}^{B\times C\times H\times W}$, they are split into visible features F_v and infrared features F_i. To ensure dimensional consistency across both modalities, we initially employ 1×1 convolutions for feature alignment.

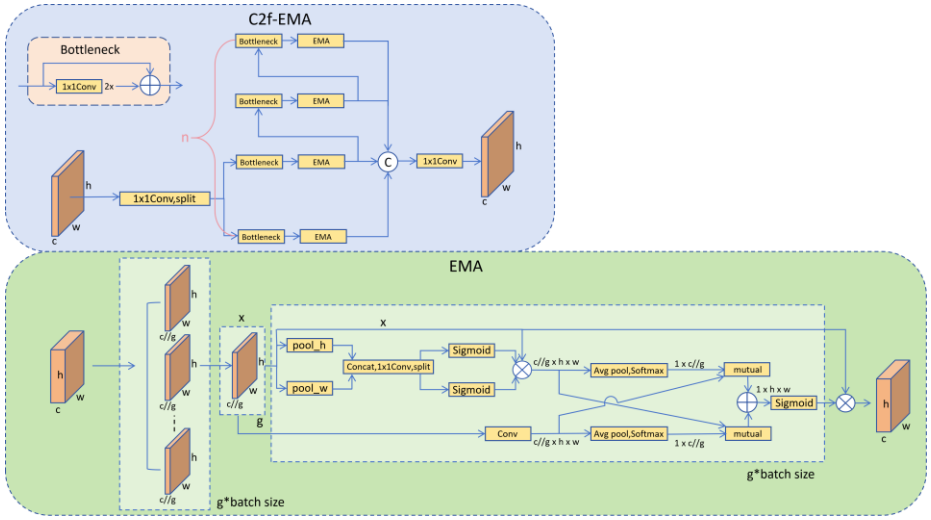
Leveraging the self-attention framework from Transformer, Query, Key, and Value projections are defined for each modality. The visible modality's queries Q_v retrieve semantic information from the infrared modality, while the infrared modality's queries Q_i retrieve semantic information from the visible modality, achieving bidirectional cross-modal interaction:

$$Q_v = W_q^v F_v^a, \quad K_i = W_k^i F_i^a, \quad V_i = W_v^i F_i^a \tag{1}$$

$$Q_i = W_q^i F_i^a, \quad K_v = W_k^v F_v^a, \quad V_v = W_v^v F_v^a \tag{2}$$



(a) CAF Architecture Diagram



(b) C2f-EMA and EMA Architecture Diagram. Where pool_h denotes adaptive average pooling along height direction, pool_w denotes adaptive average pooling along width direction, and Avg pool denotes global adaptive average pooling.

Fig. 2. Architecture Diagrams of CAF Module and C2f-EMA Module

After flattening the features, cross-modal correlation is computed through scaled dot-product attention:

$$A_{v\to i} = Softmax\left(\frac{Q_v K_i^T}{\sqrt{d}}\right), A_{i\to v} = Softmax\left(\frac{Q_i K_v^T}{\sqrt{d}}\right) \tag{3}$$

where the scaling factor $\sqrt{d}$ prevents gradient vanishing. The attention matrix $A_{v\to i}$ represents the semantic correlation strength between visible and infrared features at

spatial positions, and Softmax normalization ensures adaptive focusing on relevant regions.

Attention weights are used to aggregate features from the other modality, generating cross-modal compensation features:

$$C_{v\to i} = A_{v\to i}V_i, \quad C_{i\to v} = A_{i\to v}V_v \tag{4}$$

After reshaping the compensation features and adding them to the original features, enhanced features are obtained:

$$F_v^e = F_v^a + \gamma \cdot Reshape(C_{i\to v}), \quad \mathrm{F}_i^e = F_i^a + \gamma \cdot Reshape(C_{v\to i}) \tag{5}$$

where γ represents a scaling coefficient that is learned during training. This design enables the model to achieve cross-modal information complementarity through the cross-attention mechanism: for example, clear thermal radiation features in infrared images can supplement blurred information in low-light areas of visible images, thereby improving feature discriminability.

To further focus on important regions, a spatial attention module is introduced to refine the enhanced features. First, channel average pooling is performed on the enhanced features. After concatenating the results, Spatial weights W_s are derived by applying a 7×7 convolutional layer followed by a Sigmoid non-linearity. These weights reflect the importance of different spatial positions. Multiplying them element-wise with the enhanced features yields spatially refined features F_v^s and F_i^s.

Finally, the original aligned features, enhanced features, and spatially refined features are concatenated, fused through 1×1 convolution for dimensionality reduction, then passed through batch normalization and ReLU activation. Finally, a residual connection adds them to the original input features, ensuring effective gradient propagation.

3.3 Efficient Multi-scale Attention (EMA)

The Efficient Multi-scale Attention (EMA) module captures multi-scale contextual information through a decomposed pooling strategy while preserving rich channel information. hown in Fig. 2b, the EMA unit starts its process by dividing input features into several channel-wise groups, followed by extracting spatial context via adaptive average pooling across the vertical and horizontal axes.

$$X_h = AdaptiveAvgPool_{H\times 1}(X), \quad X_w = AdaptiveAvgPool_{1\times W}(X) \tag{6}$$

where X_h is the pooling result along the height direction (pool_h) and X_w is the pooling result along the width direction (pool_w). After concatenating them, cross-channel interaction is performed through a 1×1 convolution, then split to obtain enhanced horizontal features H and vertical features W. Spatial attention weights are generated through GroupNorm and Sigmoid gating mechanism, while local features are extracted through 3×3 convolution.

The EMA module employs a dual-path cross-attention mechanism: global adaptive average pooling (Avg pool) is applied separately to the spatial attention features and local features to obtain channel descriptors, which are then used to generate attention weights through Softmax. The two paths of features enhance each other through matrix multiplication, finally fusing into a unified attention weight map. The resulting weight map is applied to scale the original features via element-wise multiplication, achieving multi-scale feature enhancement with low computational overhead.

3.4 C2f-EMA Module

The C2f-EMA module integrates the EMA mechanism into the C2f structure of YOLOv8, as shown in Fig.2b. Specifically, C2f-EMA adds an EMA module after each Bottleneck unit in C2f, forming a series structure of "Bottleneck+EMA". This design allows the network to maintain the advantages of C2f's cross-stage partial connections while enhancing feature representation capability through EMA. With this design, C2f-EMA effectively improves feature expression capability in the neck network while maintaining a lightweight architecture.

3.5 Network Details and Training Strategy

CAF-YOLO is improved based on the YOLOv8n architecture. Its network configuration is as follows: the backbone uses CSPDarknet, with visible and infrared branches sharing structure but having independent parameters. Input size is 640×640, and output feature maps at three scales of 80×80, 40×40, and 20×20 are generated. CAF modules are deployed at P3, P4, and P5 levels of the FPN (corresponding to feature map sizes of 32×32, 16×16, and 8×8, with channel numbers 128, 256, and 512). Since the computational complexity of the cross-attention mechanism is proportional to the square of the spatial resolution, deploying it at shallower layers would significantly increase computational overhead, which is not conducive to model real-time performance. The P3-P5 feature map sizes are already small enough to strike a trade-off between precision and computational economy. while maintaining low computational cost. The neck network replaces the original C2f modules with C2f-EMA modules, enhancing feature representation capability through the series structure of Bottleneck and EMA. The loss function adopts the standard YOLOv8 format, which is a weighted sum of bounding box loss (CIoU), classification loss (BCE), and object presence loss (BCE), with weight coefficients $\lambda_{box} = 7.5$, $\lambda_{cls} = 0.5$, and $\lambda_{obj} = 1.0$, respectively.

Training Strategy: We trained all models from scratch, adopting with a batch size of 16 and a pixel resolution of 640×640. The optimization process employed an SGD optimizer with an initial learning rate of 10^{-2}, which decayed to a final factor of 10^{-2} by the end of the training. Weight decay was consistently maintained at 5×10^{-4}. Regarding the schedule, the LLVIP and FLIR datasets were trained for 120 and 80 epochs, respectively. Furthermore, a robust data augmentation pipeline was implemented, encompassing random erasing (0.4), horizontal flipping (0.5), and geometric transformations such as 10% translation and 0.5x scaling, alongside HSV color space adjustments.

Table 1. Experimental Environment Configuration

Name	Parameter
OS	Ubuntu 22.04
GPU	NVIDIA RTX 3090 (24GB) × 1
CPU	AMD EPYC 7402 24-Core
Memory	90GB
Python	3.10.18

Table 2. Main Training Parameter Settings

Param	Value
Epochs	120 (LLVIP), 80 (FLIR)
Batch size	16
Img size	640
lr0	0.01
lrf	0.01

4 Experiments and Results Analysis

4.1 Experimental Setup

Datasets and Evaluation Metrics Experiments are conducted using the LLVIP (Low-Light Visible-Infrared Paired) dataset[5] and the FLIR dataset (manually aligned version proposed by Zhang et al.[12]). We follow the official split of the LLVIP dataset, using the official test set for final evaluation. A random seed of 42 is set, with the training data partitioned into training and validation subsets at a 9:1 ratio. For FLIR, we use the official training and test sets without additional splitting, retaining three target classes: pedestrians, vehicles, and bicycles. The original LLVIP resolution is 1280×1024, while FLIR is 640×512, both are resized to 640×640 during training.

The proposed model is evaluated using a combination of precision and efficiency benchmarks, as detailed in Table X. These include mean Average Precision at various IoU thresholds (mAP_{50}, mAP_{75}) alongside model complexity indicators (Params and GFLOPs).

Implementation Details Table **1** and Table **2** detail the experimental setup and the primary training hyperparameters, respectively.

All models adopt the same training strategy: SGD optimizer with initial learning rate 0.01, weight decay 5×10^{-4}, batch size 16, 120 training epochs for LLVIP and 80 epochs for FLIR, and input image size 640×640. Data augmentation includes random horizontal flipping, translation, scaling, color space adjustments, and random erasing. All models are initialized randomly and optimized without reliance on any pre-existing weights.

Table 3. Performance Comparison of Various Models on LLVIP and FLIR Test Sets

Model	LLVIP	FLIR	Params	GFLOPs
	mAP@0.5 (%)	mAP@0.5(%)	(M)	(G)
YOLOv8n(RGB)	86.7	56.9	3.01	8.1
YOLOv8n(IR)	94.6	70.3	3.01	8.1
IWFC-Net[9]	91.1	-	-	-
CSSA[1]	94.3	79.2	69.1[a]	69.134[a]
M2I2HA(YOLOv8s)[11]	95.9	72.5	37.6	68.7
GAFF[13]	-	72.9	23.77	-
CFR_3[12]	-	72.93	-	-
GM-DETR[10]	97.4	83.9	70	176
CFT[7]	97.5	78.7	206.03	224.4
CAF-YOLO(Ours)	95.8	73.9	11.74	22.6

[a]**:**The parameters and FLOPs of the CSSA module are reproduced using the same experimental setup as our CAF-YOLO model.

Note: '-' indicates metrics not reported in the original paper.

4.2 Main Results

Table**3** shows the performance comparison of various models on LLVIP and FLIR test sets. The infrared single-modal (YOLOv8n(IR)) achieves 94.6% mAP@0.5 on LLVIP, significantly outperforming the visible single-modal (YOLOv8n(RGB)) at 86.7%, indicating that the infrared modality has stronger detection capability under low-light conditions. GM-DETR[10] and CFT[7] achieve the highest accuracy of 97.4% and 97.5% on LLVIP, respectively, while CSSA[1] achieves good performance of 79.2% on FLIR.

The proposed CAF-YOLO model achieves 95.8% mAP@0.5 on LLVIP and 73.9% on FLIR. Although CAF-YOLO is slightly lower than CFT and GM-DETR on LLVIP, its parameter count is only 11.74M and GFLOPs is only 22.6, significantly lower than these high-accuracy models, reflecting a balance between lightweight design and efficient fusion. On FLIR, CAF-YOLO performs comparably to M2I2HA and still has a large gap from GM-DETR, but its parameter count is significantly smaller than both M2I2HA and GM-DETR. In summary, CAF-YOLO strikes an optimal balance between detection precision and model compactness, rendering it highly effective for resource-constrained environments.

Table 4. Ablation study results analysis (Improvement 1: Introduce CAF module to achieve bimodal feature complementarity and semantic alignment through bidirectional attention mechanism; Improvement 2: Replace traditional ADD operation with Concat+1×1Conv+EMA fusion structure to achieve efficient bimodal feature interaction; Improvement 3: Deploy C2f-EMA module in neck network and enhance feature representation capability by embedding EMA attention mechanism)

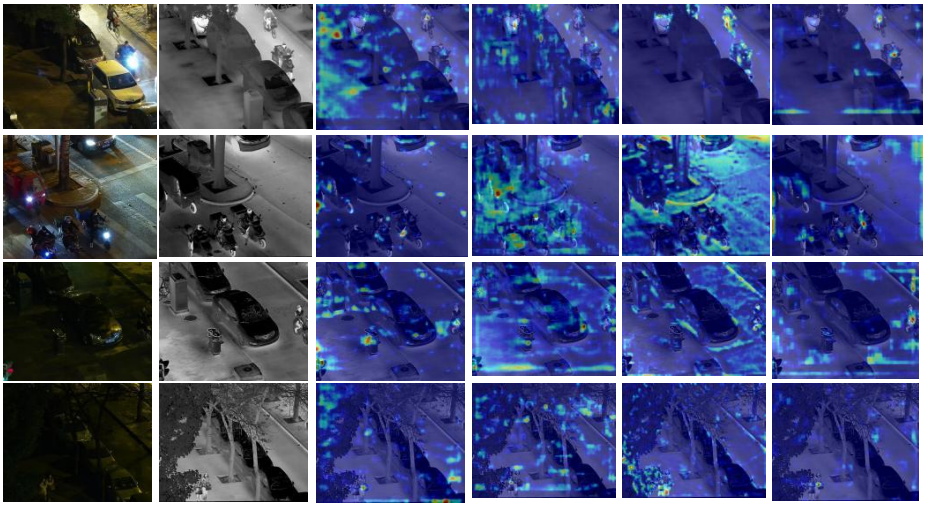
Configuration			Params&GFLops		LLVIP		FLIR	
Imp1	Imp2	Imp3	Params(M)	GFLOPs	mAP@0.5	mAP@0.75	mAP@0.5	mAP@0.75
			4.28	11.2	0.945	0.664	0.697	0.326
✓			8.80	19.5	0.947	0.702	0.701	0.342
	✓		4.45	11.6	0.954	0.671	0.725	0.327
		✓	4.28	11.2	0.955	0.656	0.709	0.333
	✓	✓	7.21	14.3	0.954	0.676	0.700	0.337
✓	✓		8.97	19.8	0.957	0.689	0.713	0.340
✓		✓	8.80	19.5	0.958	0.693	0.721	0.353
✓	✓	✓	11.74	22.6	0.958	0.703	0.739	0.351

4.3 Ablation Study

To assess the efficacy of various architectural elements in the CAF-YOLO model, this paper designs a systematic ablation study to investigate the impact of the Cross-Attention Fusion (CAF) module, neck feature fusion improvement, and C2f-EMA structure optimization on model performance. Through the control variable method, the independent contribution and combined effect of each module are evaluated separately: for the CAF module, the baseline is set without enabling this module; for the feature fusion method, simple feature addition (ADD) is used as the baseline; for the C2f-EMA module, the standard C2f structure is used as the baseline. The experimental configuration is shown in Table**4**, where "✓" signifies the incorporation of the respective component, and blank indicates not enabled.

From the ablation study results in Table**4**, it can be seen that each module makes a significant contribution to model performance. The CAF module improves mAP@0.75 by 3.8 percentage points on LLVIP and 1.6 percentage points on FLIR, indicating its effectiveness in achieving cross-modal semantic alignment, especially under strict IoU thresholds. The bimodal fusion improvement achieves a significant 2.8 percentage point increase in mAP@0.5 on FLIR, demonstrating that this improved structure can integrate bimodal information more effectively. The C2f-EMA structural improvement achieves approximately 1 percentage point mAP@0.5 enhancement on both datasets, reflecting its enhancement of feature representation capability. All combinations of improvements yield positive gains, and the complete model achieves optimal results on both datasets, verifying the rationality of CAF-YOLO's component design and the effectiveness of the overall architecture.

4.4 Visualization Analysis



(a) Visible (b) Infrared (c) RGB (d) IR (e) CAF (f) Full

Fig. 3. Heatmap visualization of CAF-YOLO on LLVIP dataset. The first two rows are daytime scenes, and the last two rows are nighttime scenes. Each column from left to right shows: visible original image, infrared original image, heatmap generated by YOLOv8-RGB, heatmap generated by YOLOv8-IR, heatmap after CAF module, and heatmap of the final full model. Visualization results indicate that the CAF module can effectively fuse bimodal information, and the heatmap of the full model shows more concentrated feature response regions.

For a comprehensive insight into the operational mechanism of CAF-YOLO, we implement Grad-CAM to visualize the model's attentional focus via heatmaps under varying conditions. Grad-CAM generates heatmaps that reflect the model's decision-making basis by calculating the gradient of the target class relative to the feature maps and performing weighted summation, then overlaying the heatmaps with the original images.

The CAF module, through its cross-attention mechanism, can better focus attention on human targets, achieving semantic alignment and enhancement of RGB and IR features. Visualization results show that in daytime scenes, single-modal models (YOLOv8-RGB and YOLOv8-IR) either misjudge or inadequately attend to the two pedestrians in the upper right corner, while the CAF module can accurately focus attention on these two pedestrian regions, achieving effective complementarity of bimodal information. In nighttime scenes, the RGB single-modal model erroneously focuses attention on a trash can, the IR single-modal model focuses on car parts, while the CAF module successfully shifts attention to the pedestrians at the right edge, demonstrating its ability of cross-modal semantic alignment.

The EMA modules enhance important channels and positions while suppressing irrelevant information, thereby improving the model's discriminative capability. Grad-CAM heatmap comparison shows that the EMA model has higher activation peaks within true targets and weaker background activation. For example, in daytime scenes,

after the effect of CAF, the EMA module further focuses attention on the human body; in nighttime scenes, although the CAF module can already attend to pedestrians, the EMA module further concentrates attention on key parts of the pedestrian body, reducing interference from background noise.

In summary, visualization analysis demonstrates that the CAF module achieves semantic alignment between RGB and IR modalities, and the EMA module enhances important features while suppressing irrelevant information. The synergistic work of both modules improves the reliability of the detector in sophisticated real-world contexts..

5 Conclusion

The CAF-YOLO model achieves 95.8% mAP@0.5 on the LLVIP dataset, representing improvements of 9.1 percentage points over the YOLOv8 visible single-modal baseline and 1.2 percentage points over the YOLOv8 infrared single-modal baseline. On the FLIR dataset, CAF-YOLO reaches 73.9% mAP@0.5, with improvements of 17.0 and 3.6 percentage points over the visible and infrared single-modal baselines, respectively. Meanwhile, CAF-YOLO has only 11.74M parameters and 22.6 GFLOPs, significantly lower than other high-accuracy models, attaining a synergy between detection precision and inference speed.

By utilizing Cross-Attention Fusion (CAF) and Efficient Multi-scale Attention (EMA) modules, this research proposes an accurate and efficient bimodal pedestrian detection method. These attention mechanisms help improve detection accuracy, simplify model architecture, and enhance the performance of intelligent surveillance and autonomous driving systems. Future work can explore more efficient attention mechanisms, extend to other bimodal detection tasks, and evaluate the model's adaptability under different conditions by testing on diverse scenarios and datasets to ensure its reliability in practical applications.

Acknowledgments. This work was supported by the College Student Innovation and Entrepreneurship Training Program of Guilin University of Electronic Technology (No. G202510595482).

Disclosure of Interests. The authors have no relevant financial or non-financial interests to disclose.

References

1. Cao, Y., Bin, J., Hamari, J., Blasch, E., Liu, Z.: Multimodal object detection by channel switching and spatial attention. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 403–411 (2023). https://doi.org/10.1109/CVPRW59228.2023.00046
2. Cao, Z., Yang, H., Zhao, J., Guo, S., Li, L.: Attention fusion for one-stage multispectral pedestrian detection. Sensors 21(12), 4184 (2021)

3. FLIR Systems: Flir adas dataset: Thermal dataset for algorithm training. https://oem.flir.com/solutions/automotive/adas-dataset-form/ (2018)
4. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
5. 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. pp. 3496–3504 (2021)
6. Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z.: Efficient multi-scale attention module with cross-spatial learning. In: ICASSP 2023-2023 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 1–5. IEEE (2023)
7. Qingyun, F., Dapeng, H., Zhaokui, W.: Cross-modality fusion transformer for multispectral object detection. arXiv preprint arXiv:2111.00273 (2021)
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
9. Wang, Z., Xie, W., Wen, J.: Pedestrian detection via visible-infrared bimodal image fusion. Software Guide 21(12), 174–181 (2022), https://kns.cnki.net/kcms2/article/abstract?v=Z3g7xh44_ZjYX0dwfGdUF8YFRyjJx4YQ-DiTEAN7p75PdU1OR3h7elJkoVydfvB56hd-fRd-UHj-sGGBvR4mySO9qs5PB-gXJt3Wi_HefA2KhjUUWaxuorewG2b5IexJ4c8-M-84jb0VyLdC8-ISJyRnE-vFkbv6Mmv8YtjuQTfyA84-7ZqMegg==&uniplatform=NZKPT&language=CHS, in Chinese
10. Xiao, Y., Meng, F., Wu, Q., Xu, L., He, M., Li, H.: Gm-detr: Generalized muiltispectral detection transformer with efficient fusion encoder for visible-infrared detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 5541–5549 (2024). https://doi.org/10.1109/CVPRW63382.2024.00563
11. Yang, X., Liu, Y., Pan, W., Chu, G., Zhang, J., Zhao, J., Man, Z., Cao, X.: M2i2ha: Multimodal object detection based on intra- and inter-modal hypergraph attention (2026), https://arxiv.org/abs/2601.14776
12. Zhang, H., Fromont, E., Lefevre, S., Avignon, B.: Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In: 2020 IEEE International Conference on Image Processing (ICIP). pp. 276–280 (2020). https://doi.org/10.1109/ICIP40778.2020.9191080
13. ZHANG, H., FROMONT, E., LEFEVRE, S., AVIGNON, B.: Guided attentive feature fusion for multispectral pedestrian detection. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 72–80 (2021). https://doi.org/10.1109/WACV48630.2021.00012
14. Zhu, J., Hu, T., Zheng, L., Zhou, N., Ge, H., Hong, Z.: Yolov8-c2f-faster-ema: an improved underwater trash detection model based on yolov8. Sensors 24(8), 2483 (2024)