IDC-YOLO: Efficient Visible-Infrared Object Detection via Illumination-Guided Modulation and Difference-Aware Interaction

Xiao Ma, De Li, Jin-Chun Piao*
Department of Computer Science and Technology
Yanbian University, Yanji, China
*jcpiao@ybu.edu.cn

***Abstract*—Visible-infrared imagery provides complementary benefits that reduce the limitations of single-modality object detection, particularly under challenging illumination conditions. However, effectively exploiting cross-modal information while maintaining computational efficiency remains difficult for detector-oriented architectures. In this paper, we propose IDC-YOLO, a dual-stream visible-infrared object detection framework built upon a YOLO-style architecture. The framework processes visible and infrared inputs in parallel and introduces two modules to enhance cross-modal interaction. An Illumination-Guided Modulation (IGM) module implicitly estimates global scene illumination and adaptively regulates modality responses, improving robustness across varying lighting conditions. Meanwhile, a Difference-Aware Interaction Module (DAIM) explicitly models cross-modal structural discrepancies and enhances complementary texture and shape cues via residual difference-based interaction. To facilitate efficient feature fusion, a lightweight Channel Switching and Spatial Attention (CSSA) module is incorporated at the detector feature level with limited computational overhead. Experiments on the M3FD and FLIR benchmarks demonstrate that IDC-YOLO achieves a 5% improvement in mAP_{50} on the M3FD dataset and a 3.8% improvement on the FLIR dataset, compared to the baseline, while maintaining a favorable accuracy-efficiency trade-off. Further evaluations on an embedded platform verify its real-time inference capability. Furthermore, consistent performance gains across different YOLO backbones indicate good generalization capability of the proposed framework.**

***Index Terms*—Visible-infrared object detection, Illumination-guided modulation, Difference-aware interaction, Dual-stream architecture, Lightweight detection**

I. Introduction

Deep learning has significantly advanced object detection and enabled its deployment in applications such as autonomous driving and intelligent surveillance. However, most existing detectors rely on single-modality visible images, making their performance highly sensitive to image quality and illumination conditions. In real-world scenarios—including low-light environments, backlit scenes, adverse weather, and complex background occlusions—visible images often suffer from texture degradation, insufficient brightness, and amplified noise, which severely degrade detection accuracy [1].

As illustrated in Fig. 1, visible images provide rich texture details in well-lit regions, whereas infrared images preserve stable target perception in dark or occluded areas, motivating multimodal object detection under challenging illumination conditions.

Fig. 1. Visible and infrared image examples highlight their complementary characteristics.

Visible and infrared images exhibit distinct imaging characteristics: visible images capture detailed appearance information, while infrared images encode thermal radiation and remain reliable under low illumination [2]–[4]. Although their integration can substantially improve detection performance, effective multimodal fusion remains challenging. Simple feature concatenation often introduces redundancy due to modality discrepancies in texture, edge structure, and noise distribution [5]. Moreover, fixed fusion strategies lack adaptability to illumination variations, while attention-based or cross-layer interaction mechanisms frequently incur considerable computational overhead, limiting real-time applicability [6].

To address these issues, we propose a lightweight and robust visible-infrared object detection framework specifically designed for detector-oriented architectures.

The main contributions of this paper are summarized as follows:

1) An Illumination-Guided Modulation (IGM) module is proposed to implicitly estimate scene illumination and adaptively regulate visible and infrared modality responses under varying lighting conditions.
2) A Difference-Aware Interaction Module (DAIM) is designed to explicitly capture structural discrepancies between modalities, enabling complementary enhancement while suppressing redundant responses.
3) A lightweight dual-stream visible-infrared detection framework is presented, enabling efficient cross-modal interaction with low computational overhead and real-time embedded inference capability.

II. Related Work

Visible-infrared multimodal object detection improves robustness under challenging conditions such as low illumination, occlusion, and complex backgrounds by exploiting complementary information from heterogeneous modalities. Existing methods can be organized from two perspectives: fusion stage and technical design.

From the fusion-stage perspective, existing methods can be broadly categorized into pixel-level, feature-level, and decision-level approaches. Pixel-level fusion combines visible and infrared images directly in the image domain [7], but is sensitive to registration errors and may introduce artifacts. Decision-level fusion integrates detection results from independently processed modalities [8], preserving modality-specific representations at the cost of higher complexity. By contrast, feature-level fusion performs cross-modal interaction in the intermediate feature space [9], providing a better balance between accuracy and efficiency, and has therefore become the dominant paradigm in detector-oriented multimodal detection [10].

From the technical-design perspective, more expressive fusion strategies mainly include attention-based, Transformer-based, and efficiency-oriented detector-friendly approaches. Spatial attention has been introduced to emphasize salient regions across modalities, while Transformer-based frameworks perform deeper cross-modal interaction at multiple feature levels [11]. However, these approaches often incur substantial computational overhead. Several works therefore focus on improving efficiency through partial interaction or detector-friendly designs, though fusion granularity and adaptability remain limited [12].

Despite these advances, existing methods still face three limitations: insufficient modeling of cross-modal structural discrepancies, limited adaptability to illumination variation, and high overhead in attention- or Transformer-based designs. These issues motivate the development of lightweight visible-infrared detection frameworks that explicitly model modality differences and adaptively regulate modality contributions.

III. Method

This paper proposes an illumination-guided and difference-aware cross-modal object detection framework, termed IDC-YOLO, to improve detection robustness under complex illumination conditions. The framework adopts a dual-stream architecture that processes visible and infrared inputs in parallel, enabling modality-specific feature extraction within a unified YOLO-style detector.

IDC-YOLO incorporates three key components. An Illumination-Guided Modulation (IGM) module is embedded into the backbone to implicitly estimate global scene illumination and adaptively regulate modality responses. A Difference-Aware Interaction Module (DAIM) is further introduced to explicitly model structural discrepancies between visible and infrared features, enhancing complementary information while suppressing redundant responses. In addition, a lightweight Channel Switching and Spatial Attention (CSSA) module [13] is integrated at the P3 feature level to facilitate efficient cross-modal feature fusion within a detector-oriented feature hierarchy. The overall architecture of IDC-YOLO and its key components are illustrated in Fig. 2.

A. Illumination-Guided Modulation Module

Visible image quality is highly sensitive to illumination variations, while existing fusion strategies often rely on coarse assumptions that lack adaptability. To address this issue, we propose a lightweight Illumination-Guided Modulation (IGM) module that estimates global scene illumination and adaptively regulates modality responses.

Specifically, IGM extracts a global scene-level illumination descriptor from the visible image and generates modality-wise modulation factors as bounded gain coefficients, rather than performing direct feature replacement or spatial reweighting. The overall structure of the IGM module is illustrated in Fig. 2(b).

A pixel-wise illumination intensity map $I_{\max} \in \mathbb{R}^{H\times W}$ is constructed as:

$$I_{\max}(x) = \max_{c\in\{R,G,B\}} I_c(x), \tag{1}$$

and fed into a lightweight convolutional regressor followed by Global Average Pooling to obtain a global illumination descriptor. The regressor adopts three 3×3 convolutions to progressively capture local illumination patterns with an enlarged receptive field, followed by a 1×1 convolution for compact feature projection with low computational overhead:

$$l = f_\theta(I_{\max}), \quad l \in [0, 1]. \tag{2}$$

The regressor is optimized implicitly through the end-to-end detection objective without additional supervision. Based on l, modality-wise modulation factors are defined as:

$$w_{vis} = l, \qquad w_{ir} = \max(\alpha, 1 - l), \tag{3}$$

where α denotes a minimum preservation factor for the infrared modality. These factors enable smooth illumination-aware adaptation while preventing excessive suppression of infrared cues. In our implementation, α is set to 0.3.

B. Difference-Aware Interaction Module

The Difference-Aware Interaction Module (DAIM) is designed to enhance complementary structural information between infrared and visible features by explicitly modeling cross-modal differences. Instead of direct feature fusion or attention-based reweighting, DAIM adopts a residual enhancement strategy that preserves modality-specific representations while selectively amplifying complementary cues. The overall architecture of DAIM is illustrated in Fig. 2(c).

Given infrared and visible feature maps F_{ir} and F_{vis}, DAIM first extracts modality-specific structural representations using fixed Sobel operators. The horizontal and vertical gradient kernels are defined as:

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}, \tag{4}$$

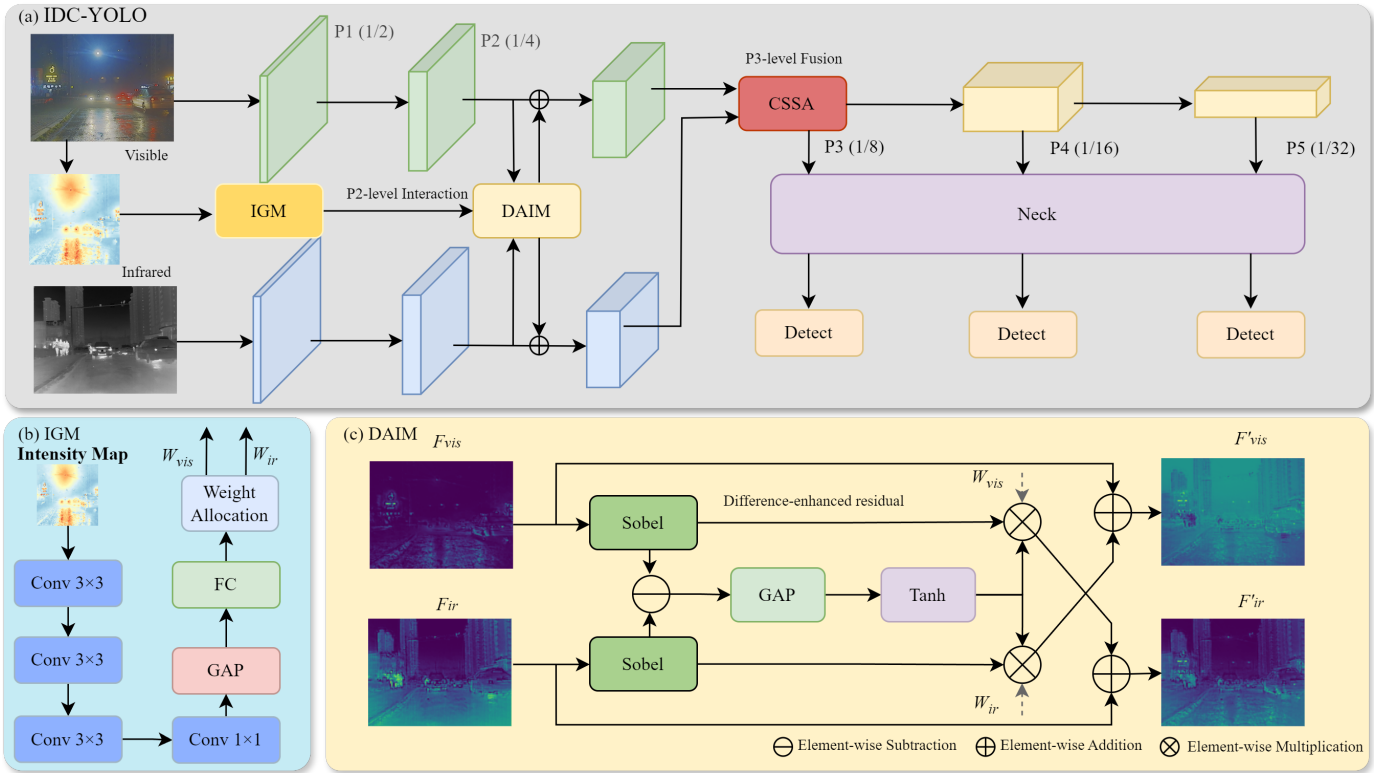

Fig. 2. Overall architecture of IDC-YOLO. The framework employs a dual-stream backbone to extract visible and infrared features and integrates illumination-guided modulation and difference-aware interaction for adaptive cross-modal representation learning. A lightweight CSSA module is applied at the P3 feature level for efficient feature fusion, followed by a standard YOLO11 neck and detection heads for multi-scale object detection.

and applied to each modality to obtain first-order gradient responses:

$$E_m = |F_m * K_x| + |F_m * K_y|, \quad m \in \{ir, vis\}, \tag{5}$$

where E_m denotes the edge response of modality m.

To characterize modality-specific discrepancies, DAIM computes signed difference maps between edge responses:

$$D_m = E_m - E_n, \quad (m, n) \in \{(ir, vis), (vis, ir)\}. \tag{6}$$

The signed formulation preserves relative dominance between modalities, enabling asymmetric enhancement for each branch. The enhanced feature map is obtained through a residual difference-based formulation:

$$F'_m = F_m \oplus E_m \otimes \tanh(\mathrm{GAP}(D_m)) \otimes w_m, \tag{7}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling and $\tanh(\cdot)$ bounds the modulation magnitude. The scalar term $\tanh(\mathrm{GAP}(D_m))$ reflects the overall structural discrepancy between modalities, while w_m denotes an optional modality-wise modulation factor provided by the illumination-guided module.

The residual formulation preserves the original feature representation via identity connections and allows DAIM to operate independently or in combination with illumination-guided modulation.

C. Cross-Modal Feature Fusion via CSSA

The Channel Switching and Spatial Attention (CSSA) module is adopted as a lightweight cross-modal fusion component at the P3 feature level of the dual-stream YOLO framework to enable efficient feature interaction under a detector-oriented hierarchy.

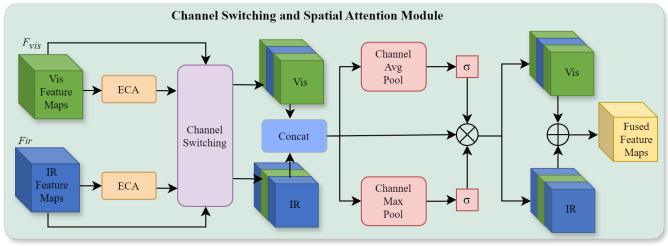

Fig. 3. Structure of Channel Switching and Spatial Attention Module.

As shown in Fig. 3, CSSA employs channel switching and spatial attention. Specifically, Efficient Channel Attention (ECA) [14] is used to estimate channel-wise importance, and channels with low contribution are replaced by their counterparts from the other modality to enhance cross-modal complementarity while preserving modality-specific characteristics. The channel switching operation is defined as:

$$X'_{m,c} = \begin{cases} X_{m,c}, & \omega_{m,c} \geq k, \\ X_{m',c}, & \omega_{m,c} < k, \end{cases} \tag{8}$$

TABLE I
PERFORMANCE COMPARISON ON M3FD DATASET

Methods	Backbone	mAP_{50}	mAP	Car	Bus	Lamp	People	Truck	Motorcycle
YOLOv11-RGBT (Baseline) [10]	YOLO11	84.7	58.5	92.9	91.9	83.6	84.6	80.5	74.4
MMFN [15]	YOLOv5	86.2	57.4	93.2	92.1	87.6	83.0	87.4	73.7
KCDNet [16]	YOLOv5	83.2	56.3	90.9	88.3	80.3	83.3	72.1	**84.1**
EI2Det [17]	YOLOv5	86.2	55.5	91.4	91.6	84.7	82.8	**89.3**	77.2
RI-YOLO [18]	YOLOv8	83.6	56.6	91.6	88.5	75.7	84.1	87.8	73.5
UniFusOD [19]	YOLOv8	86.9	59.4	**95.0**	94.1	88.6	83.4	82.4	77.8
Ours (YOLOv5)	YOLOv5	88.5	60.5	94.4	**96.1**	90.6	89.6	82.3	78.1
Ours (YOLOv8)	YOLOv8	89.1	62.1	94.2	95.1	91.0	89.8	85.3	79.4
Ours	YOLO11	**89.7**	**62.9**	94.7	95.3	**91.6**	**90.6**	86.1	79.8

where X_m denotes the feature map of modality $m \in \{ir, vis\}$, ω_m represents the channel-wise importance weights produced by ECA, m' denotes the counterpart modality, and k is a channel-switching threshold set to the mean channel importance within each feature map.

After channel switching, features from both modalities are concatenated and passed through a spatial attention module to refine spatially salient responses with minimal computational overhead. CSSA replaces the original P3-level concatenation-based fusion operation in the dual-stream network, avoiding channel expansion and reducing redundant computation.

IV. EXPERIMENTS

A. Experimental Environment and Datasets

Experimental Setup. All experiments are conducted using PyTorch on a workstation with a single NVIDIA Quadro RTX 5000 GPU and an input resolution of 640×640. Unless otherwise specified, all models are trained for 300 epochs with a batch size of 16 and an initial learning rate of 0.01. The implementation is based on Python 3.10 and CUDA 11.6. For fair comparison and efficiency considerations, small-scale YOLO variants are adopted throughout the experiments, including YOLOv5s, YOLOv8s, and YOLO11s, as YOLO-style detectors provide a strong accuracy-efficiency trade-off for real-time detection.

For embedded deployment evaluation, inference efficiency is further measured on an NVIDIA Jetson Xavier NX platform with 21 TOPS of AI computing capability, running JetPack 5.1.1 and TensorRT 8.5.2, with batch size set to 1 under both FP32 and FP16 precision.

Datasets. Experiments are conducted on two multispectral object detection benchmarks, M3FD [20] and FLIR [21]. M3FD consists of strictly aligned visible-infrared image pairs covering six object categories and is randomly divided into training and validation sets with an 8:2 ratio. For FLIR, the aligned version is used in this work, which contains 5,142 paired samples with a standard train-test split and three object categories: *person*, *car*, and *bicycle*. Unless otherwise specified, all references to FLIR in this paper refer to this aligned version.

Evaluation Metrics. Detection performance is evaluated using mean Average Precision (mAP). mAP_{50} is reported at an IoU threshold of 0.50, while $mAP_{50–95}$ denotes the average over IoU thresholds from 0.50 to 0.95. Unless otherwise specified, mAP refers to $mAP_{50–95}$.

YOLOv5s, YOLOv8s, and YOLO11s are evaluated on the M3FD dataset, while YOLO11s is adopted for FLIR due to its dataset scale and evaluation focus.

B. Comparative Experiments

1) Comparison on the M3FD Dataset: Table I compares the proposed method with the YOLOv11-RGBT baseline [10] and recent visible-infrared object detection approaches on the M3FD dataset, where methods are grouped by backbone for fair comparison. The best and second-best results are highlighted in bold and underlined, respectively.

Across YOLOv5, YOLOv8, and YOLO11 backbones, the proposed method consistently outperforms representative multispectral detectors, achieving the best overall performance with 62.9% mAP and 89.7% mAP_{50} under the YOLO11 backbone. These consistent gains across different backbones indicate good generalization with respect to network design. Moreover, notable improvements are observed on illumination-sensitive categories such as *People*, *Bus*, and *Lamp*, suggesting that the proposed modules enhance robust cross-modal feature representation.

TABLE II
PERFORMANCE COMPARISON ON FLIR DATASET

Model	mAP_{50}	mAP	Bicycle	Person	Car
Baseline [10]	76.8	40.6	61.4	82.6	86.4
DATFF [22]	80.2	-	**68.9**	82.9	88.8
MMI-Det [23]	79.8	40.5	-	-	-
MMFN [15]	**80.8**	41.7	65.5	85.7	**91.2**
EI2Det [17]	80.2	-	66.3	84.9	89.4
Ours	80.6	**42.1**	65.6	**86.0**	90.3

2) Comparison on the FLIR Dataset: Table II presents the performance comparison on the FLIR dataset. Compared with representative multispectral detection methods, the proposed approach achieves the highest mAP of 42.1%, indicating improved localization accuracy under stricter IoU thresholds.

While the gain in mAP_{50} is relatively moderate, the consistent improvement in mAP suggests more precise bounding box localization, which is particularly relevant for FLIR scenarios with large illumination variations. In addition, the proposed

method achieves the best AP on the *Person* category and competitive performance on *Car* and *Bicycle*, indicating effective cross-modal feature discrimination.

3) Efficiency Analysis: Table III summarizes the accuracy and inference efficiency of different models on the NVIDIA Jetson Xavier NX platform. Under both FP32 and FP16 precision, the proposed method consistently outperforms the baseline in detection accuracy.

TABLE III
ACCURACY AND INFERENCE EFFICIENCY ON JETSON XAVIER NX

Model	Precision (TensorRT)	mAP	FPS (Inference)
Baseline [10]	FP32	58.3	26.4
	FP16	58.2	45.7
Ours	FP32	62.6	20.9
	FP16	**62.4**	**37.6**

TensorRT FP16 inference significantly accelerates both models compared to FP32 with only negligible accuracy degradation. Despite the additional computational overhead, the proposed method maintains real-time performance on the Jetson Xavier NX platform, demonstrating a favorable accuracy–efficiency trade-off for embedded deployment.

C. Ablation Experiments

TABLE IV
ABLATION EXPERIMENTS ON M3FD DATASET

DAIM	IGM	CSSA	mAP_{50}	mAP	Params (M)	GFLOPs
–	–	–	84.7	58.5	9.84	28.5
✓	–	–	87.9	61.4	9.91	35.4
–	–	✓	86.1	59.8	9.77	27.6
✓	✓	–	88.3	61.9	9.93	35.9
✓	–	✓	88.9	62.1	9.82	35.0
✓	✓	✓	**89.7**	**62.9**	9.85	34.6

1) Component-wise Ablation on M3FD: To evaluate the contribution of each component, ablation experiments are conducted on the M3FD dataset under identical training and evaluation settings. The results are reported in Table IV.

DAIM brings the largest performance gain among all components, improving mAP_{50} from 84.7% to 87.9% and mAP from 58.5% to 61.4%. This shows that modeling cross-modal discrepancies helps improve feature complementarity.

Applying CSSA alone leads to consistent performance gains with negligible complexity. The IGM module, as a modulation mechanism, is evaluated in combination with other components, and combining DAIM with IGM further improves performance.

It should be noted that GFLOPs increased by 21.4%, primarily due to the added complexity of DAIM and IGM. Despite this, the computational overhead remains manageable, and the accuracy-efficiency trade-off is favorable, especially in complex illumination conditions.

Overall, the best results are achieved when all components are integrated, reaching 89.7% mAP_{50} and 62.9% mAP, highlighting the synergy among difference-aware interaction, illumination-guided modulation, and lightweight fusion.

TABLE V
IMPACT OF HYPERPARAMETER α ON PERFORMANCE

Hyperparameter α	mAP_{50}	mAP
0.0	87.7	61.8
0.1	87.6	61.5
0.2	88.2	61.7
0.3	**88.3**	**61.9**
0.4	86.9	61.2
0.5	86.3	60.9

2) Effect of the Preservation Factor α: A hyperparameter ablation study is conducted to examine the influence of the preservation factor α in the illumination-guided modulation module. The results are reported in Table V.

When $\alpha = 0$, infrared features may be overly suppressed under strong illumination, leading to degraded performance. Increasing α to small values (e.g., 0.1 or 0.2) partially preserves infrared cues and yields moderate improvements. The best performance is achieved at $\alpha = 0.3$, indicating a favorable balance between visible information and infrared structural robustness. Further increasing α results in a slight performance decline. Accordingly, α is fixed at 0.3 in all subsequent experiments.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT FUSION METHODS

Fusion Method	mAP	Params (M)	GFLOPs
Concat	61.9	9.93	35.9
ICAFusion [24]	62.6	12.23	49.9
LIFAdd [25]	61.4	10.48	42.7
CSSA	**62.9**	**9.85**	**34.6**

3) Comparison of Fusion Methods: To evaluate different feature fusion strategies, CSSA is compared with representative modality-level fusion operators at the same fusion position. All methods are inserted at the P3 layer of the dual-stream YOLO11s framework, replacing the original fusion operation under identical training and evaluation settings.

Concat denotes a reference mid-level fusion strategy that directly concatenates features, whereas *ICAFusion* and *LIFAdd* are plug-and-play fusion modules that perform feature fusion using attention-based or illumination-guided weighting. These methods are selected for their functional similarity to CSSA and compatibility with detector-oriented architectures.

As shown in Table VI, CSSA achieves the most favorable accuracy–efficiency trade-off, maintaining competitive detection accuracy with fewer parameters and lower computational overhead. Therefore, CSSA is adopted as the fusion operator in our framework.

D. Qualitative Results

Fig. 4 presents qualitative comparisons between the baseline method and IDC-YOLO on the M3FD dataset under varying illumination conditions. In the first row, IDC-YOLO produces more complete detections for distant pedestrians and vehicles under low illumination. In the second and third rows, it more reliably localizes small or partially obscured pedestrians in

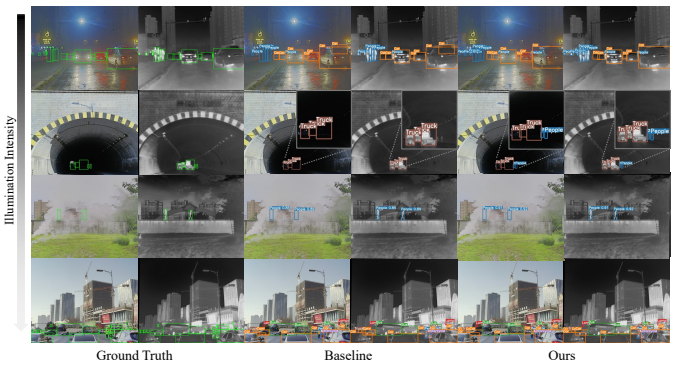

Fig. 4. Qualitative results on the M3FD dataset under varying illumination conditions.

dark or low-contrast regions. In the fourth row, it better distinguishes densely distributed objects such as cars, buses, and motorcycles while preserving clearer object boundaries.

V. Conclusion

This paper presents IDC-YOLO, a lightweight visible-infrared object detection framework for robust detection under complex illumination conditions. The dual-stream architecture preserves modality-specific characteristics while enabling cross-modal interaction. An illumination-guided modulation module regulates modality responses according to scene lighting, while a difference-aware interaction strategy models cross-modal structural discrepancies, enhancing feature complementarity with low computational overhead. Experiments on two multispectral benchmarks demonstrate that IDC-YOLO achieves a favorable accuracy–efficiency trade-off and generalizes well across varying conditions, yielding a 5% improvement in mAP_{50} on M3FD and a 3.8% improvement on FLIR over the baseline. Evaluations on an embedded platform further verify real-time inference capability, achieving 37.6 FPS under FP16 precision and highlighting its practicality for resource-constrained scenarios. Future work will investigate finer-grained illumination evaluation, modality misalignment, and further lightweight design.

Acknowledgment

This work was supported in part by the Jilin Provincial Nature Science Foundation, China (No. YDZJ202201ZYTS566); in part by Education Department of Jilin Province of China (No. JJKH20240682KJ); and in part by National Natural Science Foundation of China (No. 62062064).

References

[1] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 8115, 2023.

[2] Z. Hou, C. Yang, Y. Sun, S. Ma, X. Yang, and J. Fan, "An object detection algorithm based on infrared-visible dual modal feature fusion," *Infrared Physics & Technology*, vol. 137, p. 105107, 2024.

[3] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, p. 10535, 2023.

[4] N. Bustos, M. Mashhadi, S. K. Lai-Yuen, S. Sarkar, and T. K. Das, "A systematic literature review on object detection using near infrared and thermal images," *Neurocomputing*, vol. 560, p. 126804, 2023.

[5] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, X. Fan, and R. Liu, "Infrared and visible image fusion: From data compatibility to task adaption," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[6] X. Huang and G. Ma, "Cross-modality object detection based on detr," *IEEE Access*, 2025.

[7] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 5906, 2023.

[8] Z. Cao, H. Yang, J. Zhao, S. Guo, and L. Li, "Attention fusion for one-stage multispectral pedestrian detection," *Sensors*, vol. 21, no. 12, p. 4184, 2021.

[9] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based for co-saliency detection framework," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, p. 2473, 2018.

[10] D. Wan, R. Lu, Y. Fang, X. Lang, S. Shu, J. Chen, S. Shen, T. Xu, and Z. Ye, "Yolov11-rgbt: Towards a comprehensive single-stage multispectral object detection framework," *arXiv preprint arXiv:2506.14696*, 2025.

[11] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.

[12] Y. Shao, Q. Huang *et al.*, "Mod-yolo: Multispectral object detection based on transformer dual-stream yolo," *Pattern Recognition Letters*, vol. 183, p. 26, 2024.

[13] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, p. 403.

[14] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 11534, 2020.

[15] F. Yang, B. Liang, W. Li, and J. Zhang, "Multidimensional fusion network for multispectral object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[16] H. Wang, S. Qu, Z. Qiao, and X. Liu, "Kcdnet: Multimodal object detection in modal information imbalance scenes," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[17] K. Hu, Y. He, Y. Li, J. Zhao, S. Chen, and Y. Kang, "Ei 2 det: Edge-guided illumination-aware interactive learning for visible-infrared object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[18] Y. Chen, B. Wang, W. Zhu, and J. Yuan, "Rgb-ir yolo combining modality-specific reconstruction and information integration," in *2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*. IEEE, 2024, p. 2045.

[19] X. Xiang, G. Zhou, B. Niu, Z. Pan, L. Huang, W. Li, Z. Wen, J. Qi, and W. Gao, "Infrared-visible image fusion meets object detection: Towards unified optimization for multimodal perception," *Remote Sensing*, vol. 17, no. 21, p. 3637, 2025.

[20] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, p. 5802.

[21] H. Zhang, E. Fromont, S. Lefevre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," *2020 IEEE International Conference on Image Processing (ICIP)*, p. 276, 2020.

[22] Y. Hu, L. Shi, L. Yao, and L. Weng, "Dual attention feature fusion for visible-infrared object detection," *International Conference on Artificial Neural Networks*, p. 53, 2023.

[23] Y. Zeng, T. Liang, Y. Jin, and Y. Li, "Mmi-det: Exploring multi-modal integration for visible and infrared object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, p. 11198, 2024.

[24] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan, and W. Yang, "Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection," *Pattern Recognition*, vol. 145, p. 109913, 2024.

[25] T. Zhao, B. Liu, Y. Gao, Y. Sun, M. Yuan, and X. Wei, "Rethinking multi-modal object detection from the perspective of mono-modality feature learning," *arXiv preprint arXiv:2503.11780*, 2025.