Pseudo Visible Feature Fine-Grained Fusion for Thermal Object Detection appendix

Ting Li[1], Mao Ye[1*], Tianwen Wu[1], Nianxin Li[1], Shuaifeng Li[1], Song Tang[2], and Luping Ji[1]
[1] University of Electronic Science and Technology of China
[2] University of Shanghai for Science and Technology
lting1103@gmail.com, cvlab.uestc@gmail.com

A. Preliminary

A.1. Mamba

State Space Models (SSMs) [3] have emerged as a robust foundation in deep learning, drawing from traditional control theory and providing linear scalability with sequence length for long-range dependency modeling. Structured State Space Sequence Models (S4) and Mamba both employ a classical continuous system, which maps a 1D function or sequence, denoted as $x(t) \in \mathbb{R}$, through a hidden state $h(t) \in \mathbb{R}^N$ to an output $y(t) \in \mathbb{R}$. The SSMs are formulated as the following linear Ordinary Differential Equations (ODEs):

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h'(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ represents the state matrix, while $\mathbf{B} \in \mathbb{R}^{N\times 1}$ and $\mathbf{C} \in \mathbb{R}^{1\times N}$ are the projection parameters. These continuous parameters $\mathbf{A}$ and $\mathbf{B}$ are then discretized using a timescale parameter $\mathbf{\Delta}$. The Zero-Order Hold (ZOH) method is typically employed for this discretization, defined as follows:

$$\begin{aligned} \bar{\mathbf{A}} &= \exp(\mathbf{\Delta A}), \\ \bar{\mathbf{B}} &= (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I} \cdot \mathbf{\Delta B}. \end{aligned} \tag{2}$$

Upon discretization, Eq.1 can be transformed into the following RNN form with a step size $\mathbf{\Delta}$:

$$\begin{aligned} h_t &= \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t. \end{aligned} \tag{3}$$

Moreover, Eq.3 can be equivalently converted into the following CNN form:

$$\begin{aligned} \bar{\mathbf{K}} &= (\mathbf{C}\bar{\mathbf{B}}, \mathbf{C}\bar{\mathbf{A}}\bar{\mathbf{B}}, \ldots, \mathbf{C}\bar{\mathbf{A}}^{M-1}\bar{\mathbf{B}}), \\ \mathbf{y} &= \mathbf{x} \circledast \bar{\mathbf{K}}. \end{aligned} \tag{4}$$

*The corresponding author.

where $\circledast$ represents the convolution operation, $\bar{\mathbf{K}} \in \mathbb{R}^M$ is a structured convolution kernel, and M denotes the length of the input sequence $\mathbf{x}$.

B. More Details on Our Method

B.1. VMamba-based Multi-scale Feature Enhancement block

From our perspective, the current cross-modal fusion method [10] lacks effective feature extraction, as it relies solely on existing single-modality feature extractors for the concatenated features from two modalities. Inspired by the VMamba [8] block's capability to extract feature and model long-range dependencies, we incorporate the VMamba block to further process the multi-scale feature $f^i (i \in \{3, 4, 5\})$ extracted by CSPDarknet [1].

Initially, the feature f^i is flattened to form the input token sequence $T^i \in \mathbb{R}^{B\times N\times C}$, where B, N, C denote batch size, sequence length, and the number of channel, respectively. The sequence is first normalized by a normalization layer and then projected into $\mathbf{x} \in \mathbb{R}^{B\times N\times P}$ and $r \in \mathbb{R}^{B\times N\times P}$ through a linear layer. Subsequently, a 1D convolution layer with SiLU activation is applied to $\mathbf{x}$, yielding $\mathbf{x}'$. This intermediate representation $\mathbf{x}'$ is further linearly projected into $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$. A timescale parameter $\mathbf{\Delta}$ is then employed to discretize $\mathbf{A}$ and $\mathbf{B}$ into $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. The output $\mathbf{y}$ is computed using the SSM, as described in Eq.4. Following this, $\mathbf{y}$ is gated by r and combined with the input T^i to produce the output sequence that preserves the original shape of the input. Finally, a reshape operation is applied to this output sequence, resulting in the final refined feature f^i.

B.2. Graph node updating based on GRU.

To enhance the nodes' ability to capture complex feature interactions and retain relevant information from neighboring nodes, we employ GRU-based [2] updates for the graph nodes over L iterations. The latter section shows that setting $L = 3$ yields the best detection performance. The node

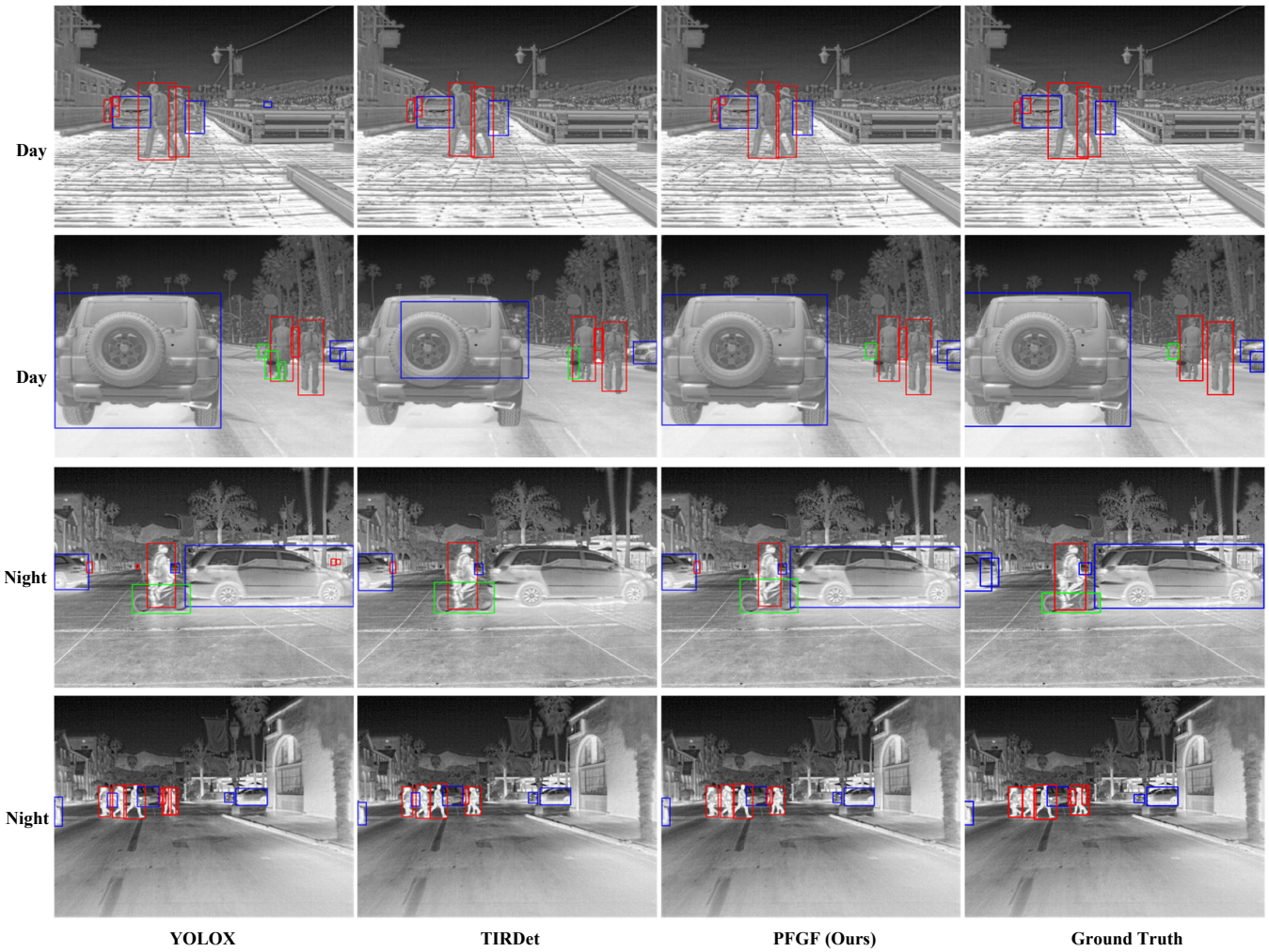

Figure 1. Detection visualization on FLIR dataset. The results of baseline (YOLOX), TIRDet, PFGF (ours), and ground truth are represented in the first, second, third, and fourth columns, respectively. Correspondingly, the first two rows present samples during the day, and the last two rows present samples at night. Red, blue, and green boxes denote the detected objects "person", "car", and "bicycle", respectively.

feature f_j^i is updated using $\mathcal{U}_{GRU}(f_j^i, m_j^{i,l-1})$ following l-1 message passing. The operation $\mathcal{U}_{GRU}$ is defined in detail as follows:

$$v_j^{i,l} = \sigma\left(W_v m_j^{i,l-1} + U_v f_j^{i,l-1} + b_v\right),$$
$$r_j^{i,l} = \sigma\left(W_r m_j^{i,l-1} + U_r f_j^{i,l-1} + b_r\right),$$
$$\tilde{h}_j^{i,l} = \tanh\left(W_h m_j^{i,l-1} + r_j^{i,l} \odot (U_h f_j^{i,l-1}) + b_h\right),$$
$$f_j^{i,l} = v_j^{i,l} \odot f_j^{i,l-1} + (1 - v_j^{i,l}) \odot \tilde{h}_j^{i,l},$$

where $v_j^{i,l}$ and $r^{i,l}j$ are the update and reset gates, respectively, controlling the balance between preserving past features and incorporating new information from its neighbors. $\tilde{h}_j^{i,l}$ represents the candidate feature. W and U denote the weight matrix, b is bias term. $f_j^{i,l}$ is the updated feature, allowing each node to dynamically adjust its representation and effectively capture complex interactions within the graph.

C. More Experimental Results

Visualization comparison. We conduct a visual comparison between the proposed PFGF method, the base detector, and the publicly available mono-modality method TIRDet [10] on the FLIR, LLVIP, and Autonomous Vehicle datasets. The results are obtained at a confidence score [9] threshold of 0.5. The detection results are shown in Fig. 1, 2, and 3. The proposed PFGF method enhances the performance by increasing the True Positive (TP) detections and the confidence level of object detection in complex scenes. Additionally, in daytime scenarios, our method

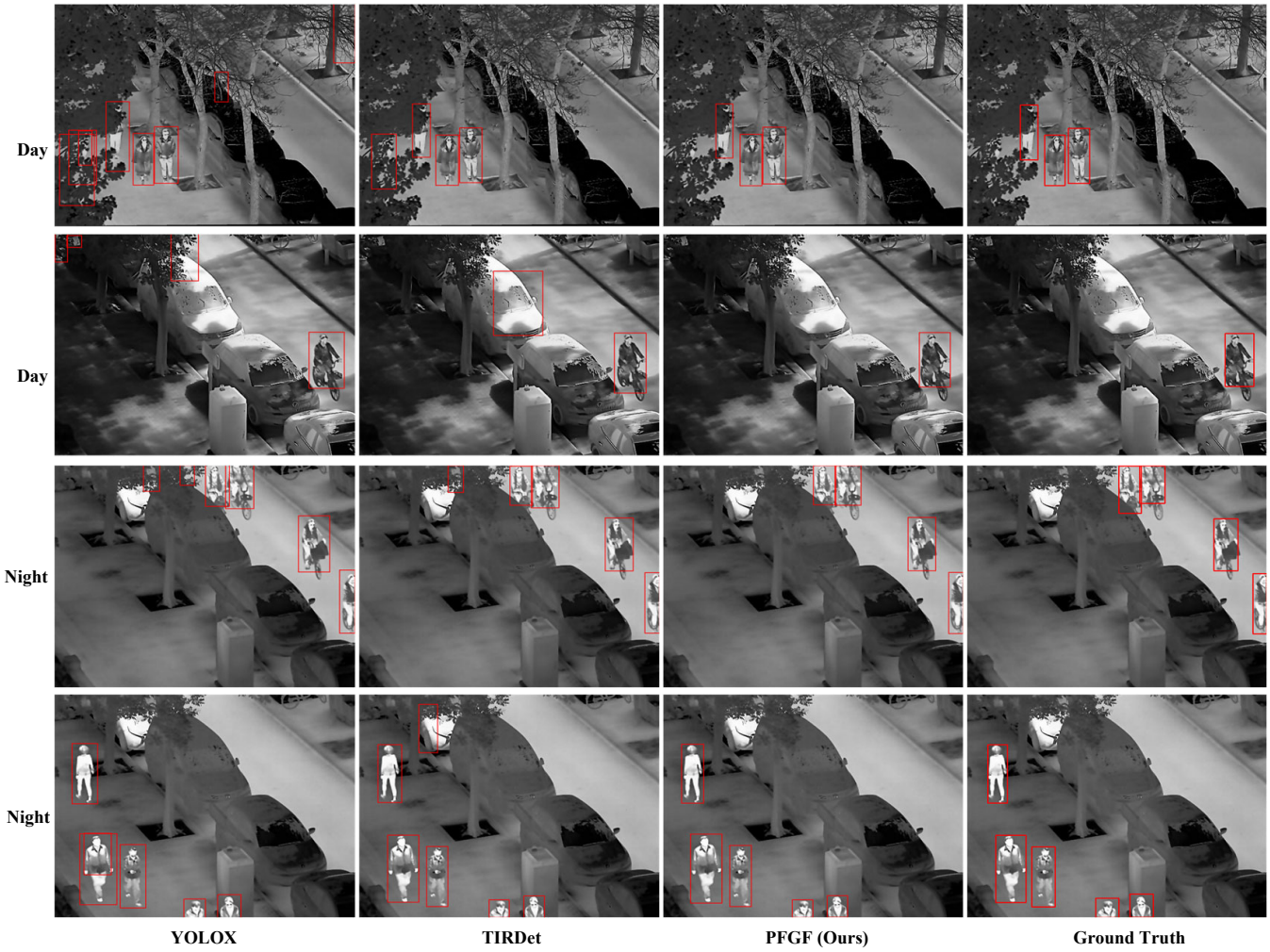

Figure 2. Detection visualization on LLVIP dataset. The results of baseline (YOLOX), TIRDet, PFGF (ours), and ground truth are represented in the first, second, third, and fourth columns, respectively. Correspondingly, the first two rows present samples during the day, and the last two rows present samples at night. Red boxes indicate detected "person" objects.

shows a reduced False Positive (FP) detections in thermal images. This demonstrates the effective integration of visible information, enabling accurate detection of easily ambiguous objects. In contrast, the TIRDet method exhibits unreliable detections and False Negative (FN) detections, particularly with the car category.

Sensitivity analysis. To assess the impact of the number of nodes n and message passing iterations L within the graph on both FLIR and LLVIP datasets, we present the detection results for various node quantities while keeping all other parameters constant, as depicted in Fig. 4 and Fig. 5. The detection performance is suboptimal with a single node in the graph, but there is a significant improvement when the number of nodes is increased to three. Further increasing the number of nodes to 5 yields only marginal performance enhancements, with results closely matching those obtained using $n = 3$. However, this increase in nodes also leads to a reduction in the network's operational efficiency. Moreover, an excessive number of nodes leads to performance deterioration due to the presence of redundant information, a trend that is particularly evident in the LLVIP dataset. As a result, utilizing 3 nodes strikes an optimal balance between speed and accuracy. The impact of the message passing iterations L on detection performance is consistent with the findings related to the number of nodes. To ensure an optimal trade-off between speed and performance, we set $L = 3$.

Effective analysis on the numbers of using Mamba. In our ablation studies, we have validated the effectiveness of VMamba-based Multi-scale Feature Enhancement (MEM) block, Inter-Mamba block, and CKI strategy. Here, we further evaluate the impact of the number using Mamba blocks in different places, as shown in Fig. 6 and Fig. 7. The num-

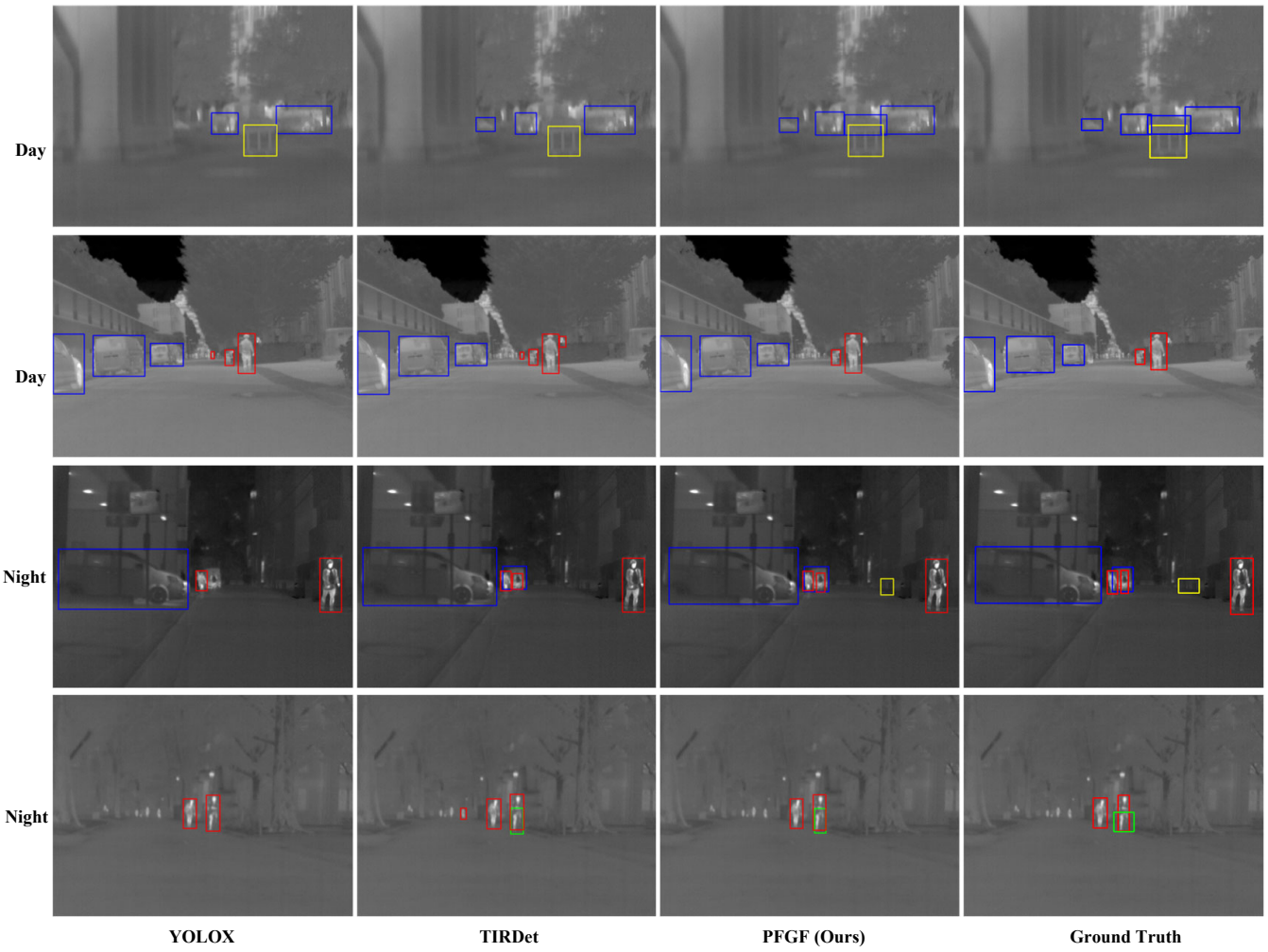

Figure 3. Detection visualization on Autonomous Vehicles dataset. The results of baseline (YOLOX), TIRDet, PFGF (ours), and ground truth are represented in the first, second, third, and fourth columns, respectively. Correspondingly, the first two rows present samples during the day, and the last two rows present samples at night. Red, blue, green, yellow, and cyan boxes represent the detected objects "person", "car", "bike", "color_cone", and "car_stop", respectively.

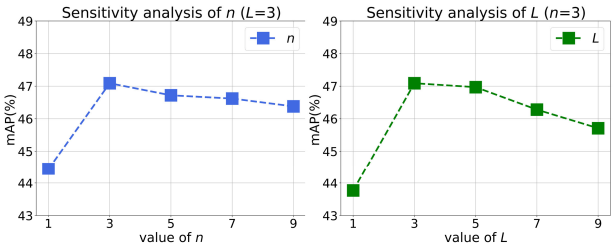

Figure 4. Hyperparameter analysis with respect to n and L on the FLIR dataset.

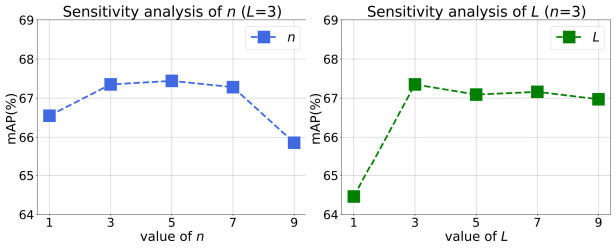

Figure 5. Hyperparameter analysis with respect to n and L on the LLVIP dataset.

ber of VMamba blocks used in MEM is denoted by K, the number of Inter-Mamba blocks in cross-modality fusion by O, and the number of VMamba blocks in CKI by D. We varied the number of each type of Mamba block from 1 to 5 while keeping other parameters constant.

As illustrated in Fig. 6, for the VMamba block in MEM,

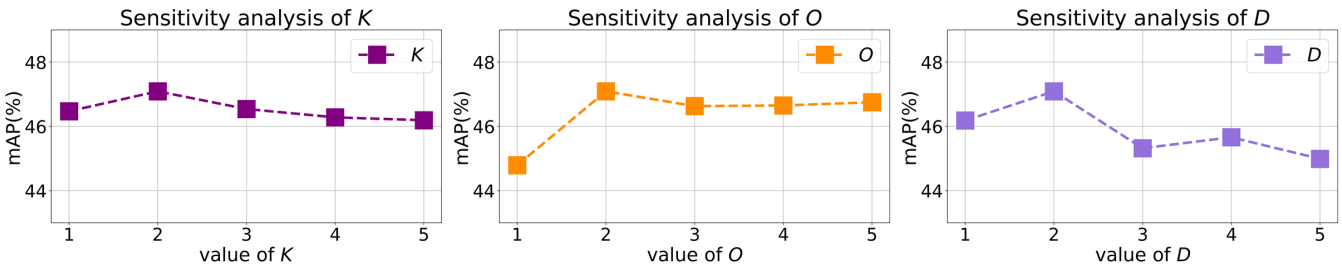

Figure 6. Hyperparameter analysis with respect to K, O and D on the FLIR dataset.

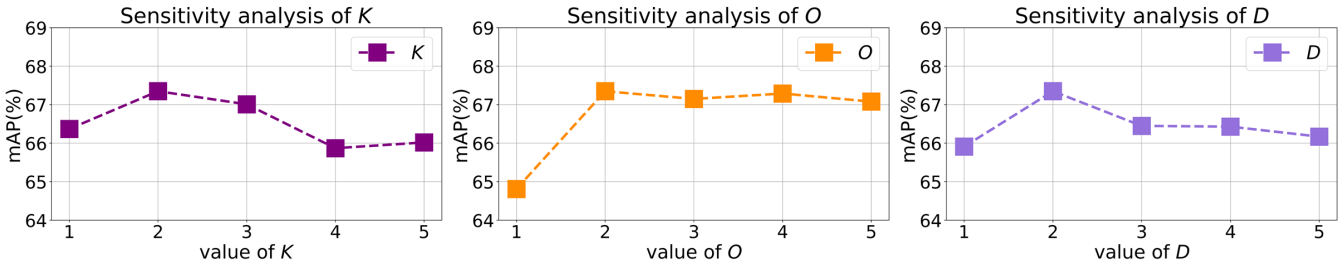

Figure 7. Hyperparameter analysis with respect to K, O and D on the LLVIP dataset.

Method	FLIR		LLVIP		AV	
	mAP	mAP50	mAP	mAP50	mAP	mAP50
FPN[7]	43.3	79.6	64.4	95.0	41.3	75.3
CKI*	44.4	81.8	66.3	96.1	43.7	76.6
CKI	**47.1**	**84.8**	**67.3**	**96.9**	**45.9**	**78.8**

Table 1. Comparisons of different cascade strategies on FLIR, LLVIP, and AV datasets. AV: Autonomous Vehicles dataset.

Method	mAP	mAP50	FLOPS(G)↓	Param(M)↓
IDA[6]	46.3	81.1	62.1	54.1
DIP[4]	-	77.3	156.0	59.1
IAH[11]	-	76.4	272.6†	114.5†
TIRDet[10]	44.3	81.4	405.1	55.8
PFGF	**47.1**	**84.8**	409.0	66.9

Table 2. Comparisons on FLIR dataset in terms of mAP, mAP50, mAP75, FLOPS, and Param.

detection performance remains relatively stable across different block numbers. The optimal performance is observed with 2 blocks, beyond which the performance gradually decreases. Increasing the number of Inter-Mamba blocks from 1 to 2 results in a substantial performance improvement. Beyond this point, the mAP value increased slowly, but at the cost of a higher computational burden. Consequently, we set the number to $O = 2$. For the VMamba blocks in the CKI, the optimal performance is also achieved with 2 blocks, with performance declining when the number of blocks is increased further.

The Effectiveness of Cascade Knowledge Integration (CKI) strategy. A key design of our approach is the novel Cascade Knowledge Integration (CKI) strategy. To validate the effectiveness of the CKI strategy, we compare it with Feature Pyramid Networks (FPN) and CKI*. FPN is a common multi-level fusion strategy that merges high-level semantic features with lower-level spatial features. CKI* refers to the cascade propagation of knowledge from the high-level subgraph to the low-level subgraph. The data inputs for these strategies differ. For FPN, we first apply cross-modality fusion using Inter-Mamba to combine the latent pseudo-visible feature z with the lowest discriminative feature f^3, and then pass the fused result to FPN for further integration. In contrast, for CKI*, the latent pseudo-visible feature z is fused with the highest discriminative feature f^5 using Inter-Mamba, enabling information propagation from f^5 to f^3. As shown in Table 1, our CKI strategy consistently outperforms all other methods across all datasets, highlighting its superior ability to distill and leverage multi-level information compared to existing approaches.

Feature map visualization. To further validate the effectiveness of our proposed GMF module, we select two images each from the FLIR and LLVIP datasets to visualize and compare the feature maps of Cross-Modality Aggregation (CMA) [10], the Graph Interaction Module (GIM) [5], and our Graph-Mamba Fusion (GMF) at Stage-3. To highlight the activation regions within the feature maps, we compute the averages across the channel dimension and

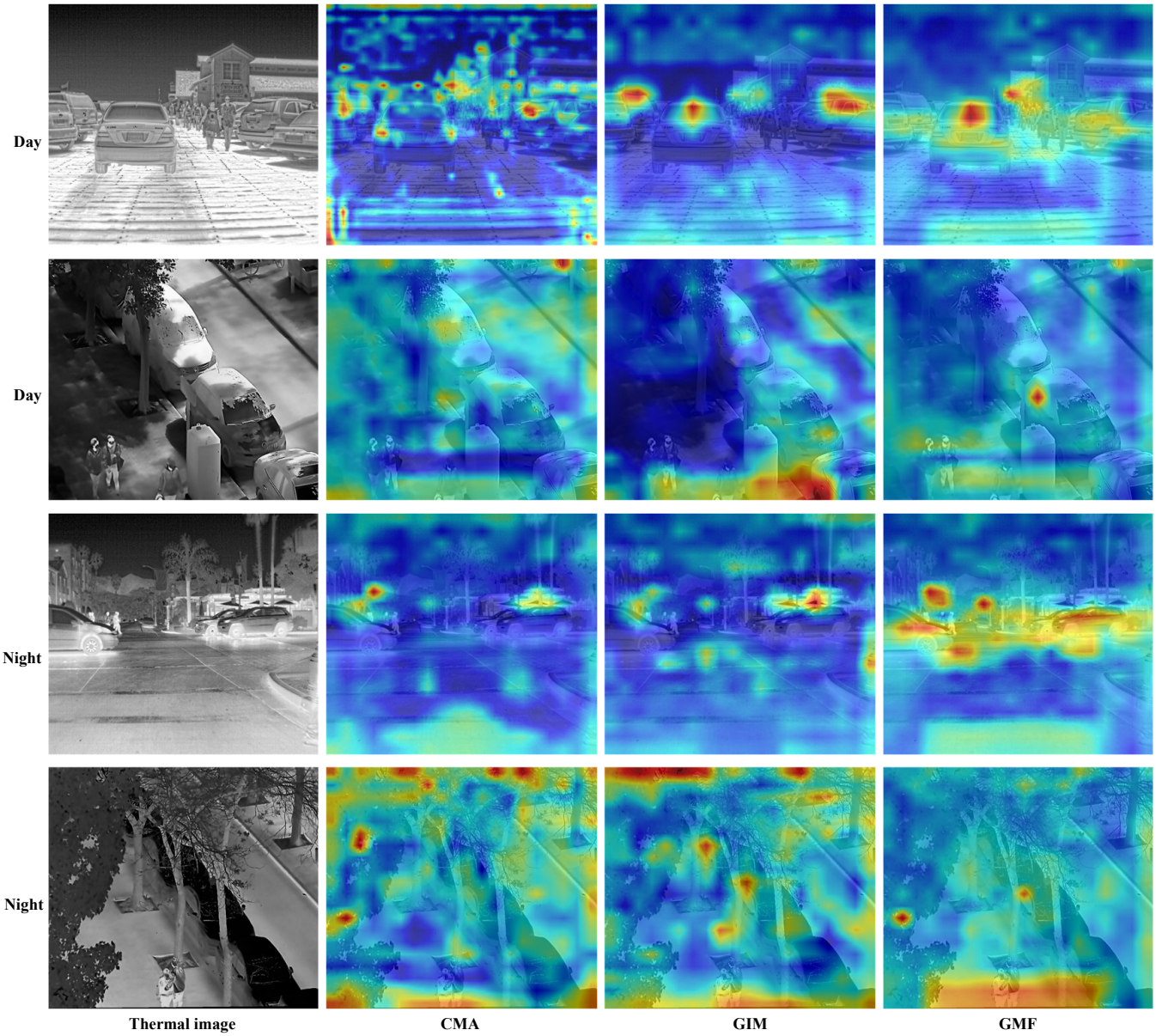

Figure 8. Visualization of feature maps in CMA, GIM, and GMF modules on FLIR and LLVIP datasets. The first and third rows are samples from FLIR dataset, while the second and fourth rows are from the LLVIP dataset. The first two rows represent samples taken under good lighting conditions during the day, whereas the last two rows captured under low-light conditions at night.

normalize the resulting feature maps. As shown in Fig. 8, the proposed GMF displays more activation regions compared to CMA and GIM. Specifically, the comparison between the visualizations from GIM and GMF reveals that while GIM emphasizes thermal information across the entire feature map, GMF focuses more on the target regions. This result demonstrates GMF's strong capability to effectively fuse thermal and generated visible information.

Comparisons for FLOPS and Param with others. Table 2 presents a comparison of the detection results and efficiency of different methods on the FLIR dataset. The symbol † indicates that the method is not open-source, and its FLOPS and Param are calculated based on our reproduced codes. For other open-source methods, the FLOPS and Param are computed using their official implementations. The results demonstrate that while the proposed method PFGF, significantly outperforms other approaches in detection accuracy, it exhibits certain limitations in terms of model size and efficiency. However, we consider this trade-off worthwhile. In the future, we aim to address these limi-

tations by developing a lightweight and robust T2V translation model and designing more efficient graph structures, such as dynamic graph networks, capable of optimizing connections and node relevance in real-time, to further enhance the robustness of thermal object detection under varying environmental conditions.

Analysis of Feature Map Subtraction Strategy. Feature map subtraction computes edge weights while preserving local details, as convolution ensures each position retains localized information. This method captures intrinsic differences with spatial consistency. Comparisons with patch-based distance weighting on the FLIR dataset show a 0.6% mAP drop, indicating that excessive reliance on neighborhood information can introduce bias and distort feature differences. These results validate the subtraction approach as an effective and efficient strategy for feature interaction.

D. Hardware specifications and software environment

We utilized an Nvidia-3090-24G GPU for our computations. On the software side, we employed the MMdetection [1] framework (version: 2.26.0) for object detection algorithms, based on the PyTorch [2] library (version 1.12.1). For further details, please refer to the official website provided in the footnotes.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection, 2020. 1

[2] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017. 1

[3] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in Neural Information Processing Systems*, 34:572–585, 2021. 1

[4] Yuxuan Hu, Ning Zhang, and Lubin Weng. Retrieve the visible feature to improve thermal pedestrian detection using discrepancy preserving memory network. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1125–1129. IEEE, 2023. 5

[5] Jiawei Li, Jiansheng Chen, Jinyuan Liu, and Huimin Ma. Learning a graph neural network with cross modality interaction for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4471–4479, 2023. 5

[6] Songtao Li, Mao Ye, Luping Ji, Song Tang, Yan Gan, and Xiatian Zhu. Illumination distribution-aware thermal pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 5

[7] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[8] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model, 2024. 1

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 2

[10] Zeyu Wang, Fabien Colonnier, Jinghong Zheng, Jyotibdha Acharya, Wenyu Jiang, and Kejie Huang. Tirdet: Mono-modality thermal infrared object detection based on prior thermal-to-visible translation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2663–2672, 2023. 1, 2, 5

[11] Qian Xie, Ta-Ying Cheng, Zhuangzhuang Dai, Vu Tran, Niki Trigoni, and Andrew Markham. Illumination-aware hallucination-based domain adaptation for thermal pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 5

[1] https://github.com/open-mmlab/mmdetection

[2] https://pytorch.org/