A Self-Controllable Adaptive Block for Multispectral Object Detection

Xiang Li[1,2], Jun Chen[1,2*]
[1]National Engineering Research Center for Multimedia Software, School of Computer Science,Wuhan University, Wuhan, China
[2]Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan, China

***Abstract*—Multispectral object detection has garnered significant attention due to its enhanced accuracy and robustness in challenging scenarios, such as low-light environments and severe weather conditions. However, a critical challenge persists: achieving effective fusion of visible (RGB) and infrared (IR) features is difficult when there are substantial discrepancies between the two modalities. For instance, visible features may be obstructed by various factors (e.g., smoke, fog or occlusions), while infrared features may lack distinguishable temperature differences from the background, so both scenarios hinder reliable feature utilization. To address these limitations, we propose a self-controllable adaptive block (SCAB). This block is designed to preserve the efficient feature representation of the original modality during model fine-tuning, while simultaneously integrating as much valid feature information as possible from the other modality through gradual iterative optimization. Experimental results on multiple RGB-IR datasets verify that our method achieves state-of-the-art performance.**

***Index Terms*—Multispectral object detection, Visible-infrared, Multimodal fusion**

I. Introduction

Multispectral object detection is an important task in computer vision, and in recent years, more and more scholars have devoted themselves to research in this direction. By integrating information from multiple modalities, detection accuracy and environmental adaptability can be improved, which has great potential for applications in multiple fields [1]–[3]. Visible cameras have shown notable limitations in delivering reliable imaging, and this is attributed to their narrow spectral bandwidth. This limitation becomes especially troublesome under low-illumination conditions and in adverse weather situations. For this reason, infrared (IR) modality boasting excellent adaptability to low-light environments has been increasingly used as supplementary information to boost the performance of the visible (RGB) modality. Subsequently, the combined application of RGB and IR images has been adopted in a growing number of semantic analysis tasks [4]–[7]. However, with the increasing application of multispectral object detection, the scenarios to be processed have become increasingly complex. Due to the influence of factors such as lighting, temperature, and weather, the characteristic information of visible modality and infrared modality will exhibit significantly different feature expressions, as shown in Fig. 1. In certain special scenarios, the model typically only benefits from the feature representation of a single modality to correctly recognize the target object—whereas the target in the other modality may be barely recognizable even to humans. A case in point is the visible modality, which struggles in low-light, smoky, or heavy fog environments; similarly, the infrared modality faces challenges in high-temperature scenarios. In this case, performing simple feature fusion is equivalent to introducing additional redundant feature information.

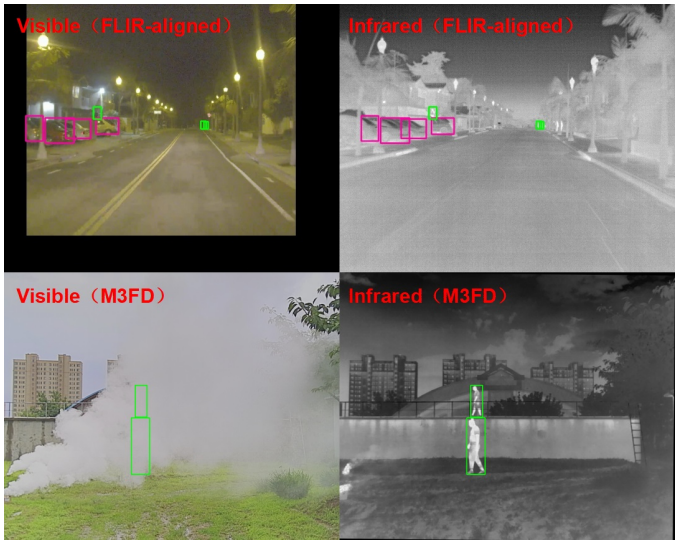


Fig. 1. The above group of images shows that small targets at night are extremely difficult to detect in the visible modality. The black borders in the images result from aligning cameras with different receptive fields. The group of images below demonstrates that in scenarios with smoke interference, the features extracted from the visible modality are completely ineffective.

To address these challenges, extensive research efforts have been devoted to developing diverse multispectral fusion approaches, which can be broadly categorized into three main types: pixel-level fusion [8], feature-level fusion [9] and decision-level fusion [10]. At the pixel level, fusion techniques first integrate multi-modal input images into a single composite image, which is then fed into a unified detection model. The core focus of such methods [11]–[13] lies in synthesizing a fused image that encapsulates comprehensive information from all input modalities, followed by leveraging this composite image to perform target de-

* Corresponding author: chenj.whu@gmail.com.
This research was supported by the National Nature Science Foundation of China (62071338, 62072347, U1903214, 61876135).

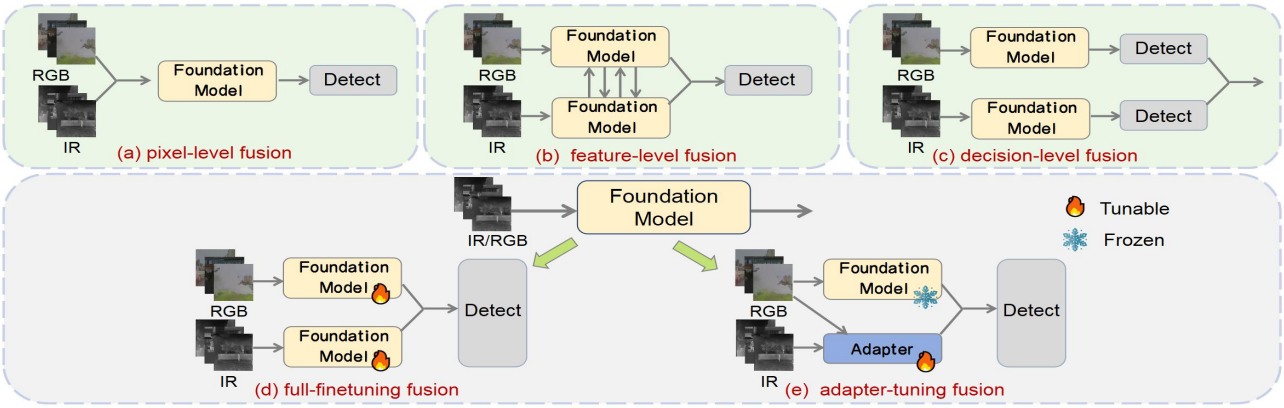


Fig. 2. Structural diagrams of five multispectral fusion approaches.

tection tasks. Feature-level fusion has emerged as the preferred choice among the majority of researchers [14]–[17]; this paradigm embeds fusion mechanisms directly into the detection framework, forming an integrated end-to-end system based on either convolutional neural networks (CNNs) [18], [19] or Transformer-based architectures [20]–[22]. In contrast, decision-level fusion methodologies [23] employ probabilistic distribution-based techniques to combine the outputs of multiple independent detection models, where the final detection result is derived by aggregating the predictions generated by each individual detector. These methods primarily enhance the network's performance on multispectral datasets by increasing the number and complexity of network parameters, though they have not yet explored specific measures to mitigate the issue of inconsistent feature expressions across different channels within the same feature space. As illustrated in Fig. 2, the structural differences of five typical multispectral fusion approaches are systematically presented, covering the three traditional types mentioned above and two RGB-IR specific fusion strategies.

In practice, RGB-IR object detectors [24], [25] adopt RGB-based models as strong baselines and fine-tune them to extract features from both RGB and IR modalities. To a certain extent, this full fine-tuning strategy can effectively integrate key feature information of the two modalities; however, it suffers from two major drawbacks: the impairment of prior knowledge encoded in the baseline model and the excessive storage overhead caused by redundant parameters. Inspired by the Adapter framework [26], we propose a self-controllable adaptive block (SCAB), which can preserve the efficient feature representation capability of the original modality during model fine-tuning, while maximizing the integration of effective feature information from the other modality through gradual iterative learning. Specifically, we first train a high-performance pre-trained model on a single modality using any mainstream object detector. Subsequently, a self-controllable adaptive block (SCAB) is deployed across two branches that share the same network architecture as the pre-trained model, and the SCAB separately shares the weights of the pre-trained network with these two branches. The design of the control mechanism is critical to the performance of the proposed block; therefore, we adopt zero-convolution as the control gate. For a specific layer in the network, the weights of the main branch are first frozen, followed by the application of a zero-convolution operation on the features of the secondary branch. The output of this operation is then fused with the features of the main branch, which is further fed as the input to the next layer of the main branch. If the modality differences are significant, ordinary fine-tuning methods may lead to overfitting, modality collapse, and catastrophic forgetting. So we draw on the experience of the zero convolution used in the text-to-image field. Overall, our contributions are summarized as follows:

- The self-controllable adaptive block (SCAB) is very simple and can be added to various neural network layers, such as resnet block, conv-bn-relu block, multi-head attention block, transformer block, etc.
- This architecture ensures that harmful noise is not added to the deep features of the large backbone model at the beginning of training.
- Zero convolution layers can help control parameters in a network and generate control signals to adjust network behavior. They do not involve highly variable weights typically associated with convolution, so there is no need for complex feature extraction for the task. This makes zero convolution layers advantageous in control tasks and typically more effective than traditional convolution layers in control applications.
- We validated the effectiveness of SCAB through ablation study and achieved state-of-the-arts results on mainstream multispectral datasets.

II. Related Work

A. Universal Object Detection Algorithm

Object detection models play a vital role in multispectral detection, facilitating the automatic identification and localization of objects within multispectral images. In recent years,

deep learning—especially models based on convolutional neural networks (CNNs)—has brought about significant advancements in detection efficiency and accuracy by leveraging specialized network architectures and loss functions. These models fall into two main categories: single-stage models (such as the YOLO [27] series, SSD [28], and RetinaNet [29]) and two-stage models (including Faster R-CNN [30] and Cascade R-CNN [31]). Single-stage models are celebrated for their speed, making them well-suited for real-time applications, while two-stage models are distinguished by their high precision, rendering them ideal for scenarios where accurate object localization is critical.

B. RGB-IR Object Detection Algorithm

Researchers have proposed numerous multispectral detection algorithms based on excellent object detection algorithms. Zhang et al. [32] investigate a two-stream SSD framework to extract contextually enhanced features for RGB-IR object detection. Additionally, AR-CNN [33] is developed based on Faster R-CNN to align RGB and IR features. With the rise of transformers, Yuan et al. put forward a complementary fusion transformer (CFT) [34] module to attain superior detection outcomes. Moreover, C^{2}Former [35] is a new transformer block that can be integrated into existing pre-trained models to enhance intra-modality and inter-modality feature representations.

The aforementioned approaches strive to develop task-specific structures with the aim of enhancing performance in relevant downstream tasks. They either train the designed models entirely from the ground up or employ a full finetuning strategy on pre-trained models. Unlike these methods, we have designed a self-controllable adaptive block (SCAB) that can gradually learn control signals through training, and can reduce the introduction of noise while increasing modal interaction.

III. Method

A. Overall Architecture

By adding self-controllable adaptive block (SCAB) to the base network, we obtain a new finetune network called self-controllable adaptive network (SCAN). The overall architecture of SCAN is illustrated in Fig. 3. Firstly, we select any mainstream object detection network for pre-training on a single modality, and choose the single modality with better performance based on different multispectral datasets. Then we use mainstream object detection networks with the same structure and divide them into two branches. Choose the one with better performence on single modality as the main branch. Then share the pre-trained network weights to the corresponding positions in the two branches, and freeze the weights of the main branch. Finally, we add SCABs on each layer corresponding to the two branches, which can provide excellent support to the main branch by gradually integrating the effective features of the other branch.

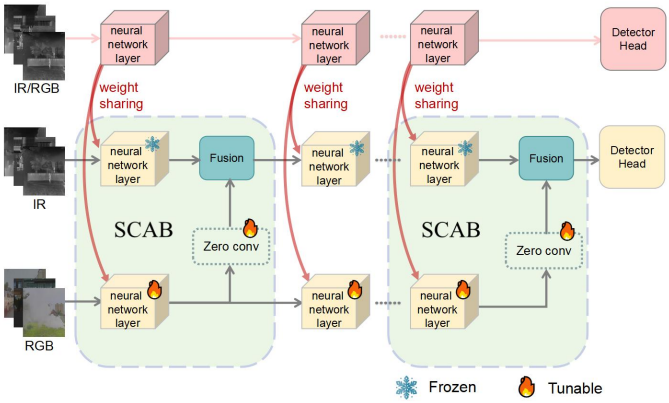


Fig. 3. The overall architecture of SCAN. Share the weights of the pre-trained model with two branches and freeze the weights of the main branch. Then design SCAB based on zero convolution for iterative feature fusion between the two branches. According to the different backbone architectures, the neural network layer can be any structure. But at the same stage, it is necessary to maintain complete consistency in all branches.

B. How Self-Controllable Adaptive Block Works

A self-controllable adaptive block injects additional conditions into a neural network layer. Suppose $\mathcal{N}(\cdot;\Theta)$ is such a trained neural block, with parameters Θ, that transforms an input feature map $\mathcal{X}$, into another feature map $\mathcal{O}$ as:

$$\mathcal{O} = \mathcal{N}(\mathcal{X}, \Theta) \tag{1}$$

The features map of the tunable other branch $\mathcal{X}_o$ is connected to the frozen main branch with zero convolution layer, denoted $\mathcal{Z}(\cdot;\cdot)$. Specifically, $\mathcal{Z}(\cdot;\cdot)$ is a 1×1 convolution layer with both weight and bias initialized to zeros. And $\mathcal{F}(\cdot)$ represents a feature fusion module which can be Concat, Add, weighted addition, or other complex fusion methods. The complete output of the SCAB then computes:

$$\mathcal{O}_m = \mathcal{F}(\mathcal{N}(\mathcal{X}_m;\Theta_{frozen}),\ \mathcal{Z}\left(\mathcal{N}(\mathcal{X}_o;\Theta_{tunable});\Theta_z\right)) \tag{2}$$

In the initial training step, the weight and bias parameters of a zero convolution layer are both initialized to zero. As a result, each of the $\mathcal{Z}(\cdot;\cdot)$ terms in Equation (2) computes to zero, and

$$\mathcal{O}_m = \mathcal{O} \tag{3}$$

In this manner, when training starts, harmful noise of the other branch is unable to affect the hidden states of the neural network layers. Furthermore, given that $\mathcal{Z}\left(\mathcal{N}(\mathcal{X}_o);\Theta_z\right)$ equals 0 and the other branch also takes in the input image $\mathcal{X}$, maintains full functionality and preserves the capabilities of the pretrained model—enabling it to act as a robust backbone for subsequent learning. Zero convolutions safeguard this backbone by removing random noise (which manifests as gradients) during the initial training stages.

C. Self-Controllable Adaptive Net for Multispectral Detection

The above introduces the principle of how SCAB works. Other fields that utilize zero convolution often have a significant difference in scale between the source and target domains. But in multispectral object detection, the size of the

source domain and the target domain are the same. Therefore, training convergence is much easier and less prone to gradient vanishing problems. So we added SCABs at every layer of the network to verify the effectiveness of the blocks. Furthermore, to better verify the efficiency and plug-and-play capability of SCAB, two dual-branch methods, namely YOLOv11 [36] and CFT, are selected as the backbones of our detection framework in the subsequent experiments.

IV. Experiments

A. Datasets

To verify the effectiveness, feasibility, and generalization ability of our detection algorithm in complex scenarios, we utilized three open-source multispectral object detection datasets. A brief overview of each dataset is presented below:

M3FD [37]: Gathered via dual-optical cameras and infrared sensors, this dataset comprises 4,200 image pairs, with annotations covering six object categories (including humans, cars, and trucks). It is extensively used in both image fusion and object detection tasks. For experimental purposes, 3,360 of these images are selected as the training set, while the remaining 840 serve as the validation set.

LLVIP [38]: Acquired using visible and infrared cameras, this dataset is composed of 15,488 image pairs, with its annotation work focusing specifically on the pedestrian category. Each image—measuring 640×512 pixels in size—has undergone pre-registration. The dataset is divided into two subsets: 12,025 image pairs for training and 3,463 image pairs for testing. It is primarily utilized in low-light vision tasks.

FLIR-aligned [39]: Captured via infrared thermographic cameras, this dataset primarily includes annotations for three object categories: pedestrians, cars, and bicycles. Each image, with a resolution of 640×512 pixels, has undergone pre-registration. The dataset consists of 4,129 training image pairs and 1,013 testing image pairs, and it is widely used in object detection tasks—especially in complex scenarios like nighttime and low-light environments.

B. Implementation details

Training was carried out on a single NVIDIA A100 GPU. The total number of pre-training epochs was set to 300, while the number of fine-tuning epochs was set to 200. The initial batch size for training was 16. The Adam optimizer was used for the training process, with its parameters configured as follows: an initial learning rate of 1.0×10^{-3} and a momentum of 0.937. To ensure the consistency of input data distribution and computational efficiency, the image size is fixed to 640 × 640 during both training and inference phases. To comprehensively assess the proposed model, standard evaluation metrics are adopted, including mAP@[0.5:0.95], model parameters, and inference time. Specifically, mAP@0.5 evaluates detection accuracy, which is the mean of category-wise Average Precision (AP) at an intersection over union (IoU) threshold of 0.5, while mAP@[0.5:0.95] reflects the model's robustness across different IoU thresholds (ranging from 0.5 to 0.95 with a step of 0.05). Model parameters, measured in millions of parameters (MParams), indicate the model's complexity and storage cost, and inference time (in milliseconds per image) measures its inference efficiency, with all tests conducted on a unified hardware platform.

TABLE I
Performance Comparison on the M^3FD Datasets

Method	↑mAP@.5	↑mAP@[.5:.95]	↓Params(M)	↓Time(ms)
U2Fusion[TPAMI '20] [11]	79.2	53.8	0.66	43
CDDFuse[CVPR '23] [12]	80.3	54.9	1.19	96
PIAFusion[IF '22] [13]	80.6	54.9	-	55
TarDAL[CVPR '22] [37]	81.0	54.9	0.3	30
MetaFusion[CVPR '23] [8]	-	56.5	-	15
Fusion-Mamba[TMM '25] [4]†	88.0	61.9	287.6	78
CFT[arXiv '21] [34]†	81.8	55.4	206	68
SCAN-CFT(ours)†	86.2	60.4	334.7	68.3
SCAN-YOLOv11-nano(ours)†	82.7	56.5	6.97	3.95
SCAN-YOLOv11(ours)†	**88.4**	**62.8**	56.4	18.7

† denotes the adoption of the YOLO series as backbone (same for subsequent tables).

TABLE II
Performance Comparison on the LLVIP Datasets

Method	↑mAP@.5	↑mAP@[.5:.95]	↓Params(M)	↓Time(ms)
U2Fusion[TPAMI '20] [11]	87.1	52.6	0.66	43
TarDAL[CVPR '22] [37]	94.5	63.3	0.3	30
MetaFusion[CVPR '23] [8]	91.0	56.9	-	15
PoolFuser[AAAI '23] [1]	80.3	38.4	-	49
ProbEn[ECCV '22] [10]	93.4	51.5	945.9	846.9
CSSA[CVPR '23] [14]	94.3	59.2	-	31
TFDet[TNNLS '25] [2]	95.7	56.1	-	130
UniRGB-IR[MM '25] [5]	96.1	63.2	147	-
Fusion-Mamba[TMM '25] [4]†	97.0	64.3	287.6	78
CFT[arXiv '21] [34]†	96.9	63.6	206	68
SCAN-CFT(ours)†	**97.5**	65.9	334.7	68.3
SCAN-YOLOv11-nano(ours)†	96.4	65.8	6.97	3.95
SCAN-YOLOv11(ours)†	97.0	**68.5**	56.4	18.7

TABLE III
Performance Comparison on the FLIR Datasets

Method	↑mAP@.5	↑mAP@[.5:.95]	↓Params(M)	↓Time(ms)
TarDAL[CVPR '22] [37]	42.9	-	0.3	30
GAFF[WACV '21] [3]	74.6	37.4	31.4	61
CFT[arXiv '21] [34]†	78.7	40.2	206	68
ProbEn[ECCV '22] [10]	75.5	37.9	945.9	846.9
CSSA[CVPR '23] [14]	79.2	41.3	-	31
ICAFusion[PR '24] [20]†	79.2	41.4	120.21	26
UniRGB-IR[MM '25] [5]	**81.4**	44.1	147	-
CFT[arXiv '21] [34]†	78.7	40.2	206	68
SCAN-CFT(ours)†	81.0	41.9	334.7	68.3
SCAN-YOLOv11-nano(ours)†	77.6	42.5	6.97	3.95
SCAN-YOLOv11(ours)†	79.8	**44.2**	56.4	18.7

C. Comparison with State-of-the-Arts models

Pixel-level fusion methods form the first category in Table I, and are tabulated separately in Tables II and III for consistency. The rationale for this separate tabulation lies in the fact that the reported computational overhead of these methods is confined solely to the fusion module, and does not account for the inference latency of a full detector—a critical component for real-world detection deployments. SCAN-YOLOv11-nano is a lightweight variant of our proposed model, which realizes further parameter reduction and inference speedup with only a marginal sacrifice in detection accuracy. Leveraging the inherent efficiency of YOLO-series architectures, we additionally

(a) Ground Truth (b) CFT (c) Ours

Fig. 4. Comparison visualization results with CFT.

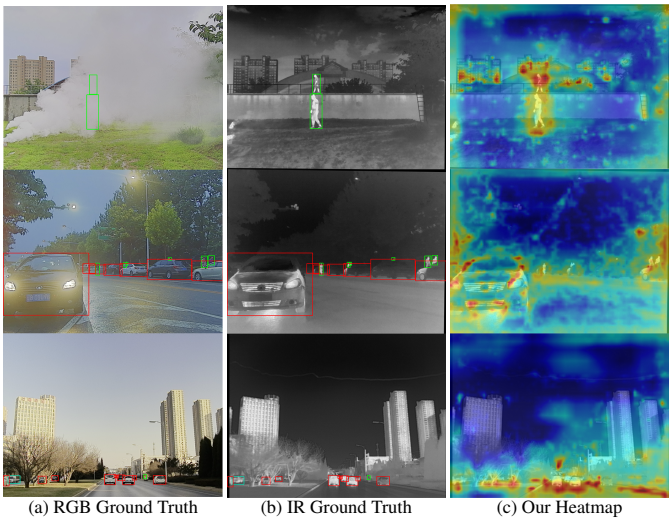
conduct comparative experiments against YOLO-backbone-based methods to conduct a more impartial evaluation of the significant efficiency gains achieved by our approach. Visual comparative results between our method and CFT—a representative algorithm within the YOLO-backbone-based paradigm—are illustrated in Fig. 4. As can be observed from the visualization, CFT is more prone to generating false detections compared to our method, and the confidence scores of its successfully detected bounding boxes are consistently lower. Through extensive evaluations across multiple diverse datasets, we observe that several state-of-the-art competing methods suffer from performance degradation on specific datasets due to intrinsic disparities in data characteristics. In stark contrast, our proposed method maintains consistently high detection accuracy across all evaluated datasets, while simultaneously delivering drastically enhanced inference speed and a streamlined parameter footprint.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT MODALITIES ON FLIR DATASET

Method	Modality	↑mAP@.5	↑mAP@[.5:.95]
YOLOv11-nano	visible	60.4	29.2
	infrared	70.5	37.3
	cross-modality	72.1	38.5
SCAN-YOLOv11-nano(ours)	cross-modality	**78.1**	**42.4**

D. Ablation Studies

Table IV quantifies the performance gains attained by integrating SCAB into the baseline model. The cross-modality implementation of YOLOv11 adopted in this ablation study leverages a fundamental pixel-level fusion strategy, and the experimental results herein validate the effectiveness of the proposed SCAB module. Table V presents an in-depth component-wise validation of the internal modules of SCAB, which not only verifies the efficacy of the Frozen and Zero Conv modules but also substantiates the proposed self-controllable design principle of our framework. Furthermore, even when both the Frozen and Zero Conv modules are ablated, the model still achieves a significantly higher accuracy than the baseline in Table IV. This performance retention is attributed to the fact that the core architecture of SCAB remains a valid adapter module in the absence of these specific components, thereby corroborating the inherent adaptive characteristic of the proposed SCAB design. It should be noted that the calculation of model weights in Table V incorporates two components: the pre-trained weights and the adapter weights. All values reported in the preceding tables have included both components in their computation, while the inference time is evaluated by considering only the adapter component.



(a) RGB Ground Truth (b) IR Ground Truth (c) Our Heatmap

Fig. 5. Visualization of RGB and IR ground truth annotations, and the corresponding attention heatmap of our method.

TABLE V
ABLATION STUDY ON FLIR DATASET

Frozen	Zero Conv	↑mAP@.5	↑mAP@[.5:.95]	↓Params(M)	↓Time(ms)
		76.3	40.9	3.53 + 2.51	3.9
✓		77.1	41.3	3.54 + 2.51	3.9
✓	✓	**78.1**	**42.4**	4.46 + 2.51	3.95

E. Visualization

To further verify the effectiveness of the proposed method, in addition to the object detection visualization results presented in Fig. 4, this section conducts an in-depth analysis of the model's decision-making process by visualizing attention heatmaps, as shown in Fig. 5. By comparing the ground truth annotations of RGB and IR modalities with the attention distributions generated by the model, the degree of attention paid by the network to target regions during inference can

be intuitively observed. Experimental results demonstrate that the proposed method can effectively focus on the key regions where targets are located, suppress the interference from background noise and irrelevant regions, thereby achieving accurate perception and effective fusion of multi-modal information. This verifies the rationality and superiority of the designed Self-Controllable Adaptive Block.

V. Conclusion

In this paper, we propose a self-controllable adaptive block (SCAB) designed to preserve the efficient feature representation of the original modality during model fine-tuning. Simultaneously, it maximizes the integration of valid feature information from the other modality through a process of gradual iteration.The experiment in this article found that SCAB has a greater improvement on small models, of course, this is also because we have added SCAB at every layer of the network, which increases the difficulty of the large model converging to the global optimum. In the future, we will explore better ways to add SCAB in network architecture.

References

[1] Y. Cao, Y. Fan, J. Bin, and Z. Liu, "Lightweight transformer for multi-modal object detection (student abstract)," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 16 172–16 173.

[2] X. Zhang, X. Zhang, J. Wang, J. Ying, Z. Sheng, H. Yu, C. Li, and H.-L. Shen, "Tfdet: Target-aware fusion for rgb-t pedestrian detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 13 276–13 290, 2025.

[3] H. Zhang, É. Fromont, S. Lefèvre, and B. Avignon, "Guided attentive feature fusion for multispectral pedestrian detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 72–80.

[4] W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, G. Guo, and B. Zhang, "Fusion-mamba for cross-modality object detection," *IEEE Trans. Multimedia*, vol. 27, pp. 7392–7406, 2025.

[5] M. Yuan, B. Cui, T. Zhao, J. Wang, S. Fu, X. Yang, and X. Wei, "Unirgb-ir: A unified framework for visible-infrared semantic tasks via adapter tuning," in *Proc. 33rd ACM Int. Conf. Multimedia*, 2025, pp. 2409–2418.

[6] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Detfusion: A detection-driven infrared and visible image fusion network," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 4003–4011.

[7] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma, "Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior," *Inf. Fusion*, vol. 110, p. 102450, 2024.

[8] W. Zhao, S. Xie, F. Zhao, Y. He, and H. Lu, "Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13 955–13 965.

[9] D. Rao, T. Xu, and X.-J. Wu, "Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network," *IEEE Trans. Image Process.*, 2023, early access, May 10, 2023.

[10] Y.-T. Chen, J. Shi, Z. Ye, C. Mertz, D. Ramanan, and S. Kong, "Multimodal object detection via probabilistic ensembling," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 139–158.

[11] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–581, 2020.

[12] Z. Zhao, H. Bai, and J. E. A. Zhang, "Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5906–5916.

[13] L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "Piafusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83, pp. 79–92, 2022.

[14] Y. Cao, J. Bin, J. Hamari, E. Blasch, and Z. Liu, "Multimodal object detection by channel switching and spatial attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 403–411.

[15] Q. Li, C. Zhang, Q. Hu, H. Fu, and P. Zhu, "Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection," *IEEE Trans. Multimedia*, vol. 25, pp. 3420–3431, 2023.

[16] R. Li, J. Xiang, F. Sun, Y. Yuan, L. Yuan, and S. Gou, "Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection," *IEEE Trans. Multimedia*, vol. 26, pp. 852–863, 2024.

[17] C. Tian, Z. Zhou, Y. Huang, G. Li, and Z. He, "Cross-modality proposal-guided feature mining for unregistered rgb-thermal pedestrian detection," *IEEE Trans. Multimedia*, vol. 26, pp. 6449–6461, 2024.

[18] X. Cheng, K. Geng, Z. Wang, J. Wang, Y. Sun, and P. Ding, "Slbaf-net: Super-lightweight bimodal adaptive fusion network for uav detection in low recognition environment," *Multimedia Tools Appl.*, vol. 82, no. 30, pp. 47 773–47 792, 2023.

[19] T. Zhao, M. Yuan, and X. Wei, "Removal and selection: Improving rgb-infrared object detection via coarse-to-fine fusion," *CoRR*, 2024, arXiv:2401.10731.

[20] 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 Recognit.*, vol. 145, p. 109913, 2024.

[21] Y. Zhu, X. Sun, M. Wang, and H. Huang, "Multi-modal feature pyramid transformer for rgb-infrared object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 9, pp. 9984–9995, Sep. 2023.

[22] S. Lee, J. Park, and J. Park, "Crossformer: Cross-guided attention for multimodal object detection," *Pattern Recognit. Lett.*, vol. 179, pp. 144–150, 2024.

[23] C. Li, D. Song, R. Tong, and M. Tang, "Illumination-aware faster r-cnn for robust multispectral pedestrian detection," *Pattern Recognit.*, vol. 85, pp. 161–171, 2019.

[24] M. Yuan, Y. Wang, and X. Wei, "Translation, scale and rotation: cross-modal alignment meets rgb-infrared vehicle detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 509–525.

[25] L. Zhang, Z. Liu, X. Zhu, Z. Song, X. Yang, Z. Lei, and H. Qiao, "Weakly aligned feature fusion for multimodal object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4145–4159, 2025.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. D. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2790–2799.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[31] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.

[32] L. Zhang, Z. Liu, S. Zhang, X. Yang, H. Qiao, K. Huang, and A. Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Inf. Fusion*, vol. 50, pp. 20–29, 2019.

[33] L. Zhang, X. Zhu, X. Chen, X. Yang, Z. Lei, and Z. Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5126–5136.

[34] Q. Fang, D. Han, and Z. Wang, "Cross-modality fusion transformer for multispectral object detection," *CoRR*, 2021, arXiv:2111.00273.

[35] M. Yuan and X. Wei, "C^2former: Calibrated and complementary transformer for rgb-infrared object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–12, 2024.

[36] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *CoRR*, 2024, arXiv:2410.17725.

[37] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, Z. Wei, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5802–5811.

[38] X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "Llvip: A visible-infrared paired dataset for low-light vision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3496–3504.

[39] H. Zhang, É. Fromont, S. Lefèvre, and B. Avignon, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *2020 IEEE Int. Conf. Image Process.*, 2020, pp. 276–280.