Customized Fusion: A Closed-Loop Dynamic Network for Adaptive Multi-Task-Aware Infrared-Visible Image Fusion

Supplementary Material

1. More Details of VFN

In the first stage, we train the VFN to focus on generating visually guided fused images. In the second stage, the VFN is frozen, while the RSC module assists in adapting the VFN to various downstream task requirements. As shown in Figure 1 (a), the VFN consists of $2L$ Feature Extraction Blocks (FEB) and one Fusion Feature Reconstruction Block (FRB). Each FEB mainly extracts features from infrared or visible images. As illustrated in Figure 1 (b), when $L = 1$, the FEB contains a single convolutional block designed to enhance the feature channels. When $L > 1$, each FEB is composed of three convolutional layers. The FRB is primarily responsible for fusing the extracted infrared and visible features to generate the final fused image. As depicted in Figure 1 (c), the FRB consists of Global Average Pooling (GAP), Global Max Pooling (GMP), Mean Pooling (MeanP), Max Pooling (MaxP), and L FEBs.

Specifically, the infrared and visible images $\boldsymbol{I}_{ir/vi}$ are separately fed into L FEBs to extract features $\{\boldsymbol{F}^l_{ir/vi}\}^{L-1}_{l=1}$ from each modality, where $\boldsymbol{F}^l_{ir/vi}$ denotes the feature maps obtained from the l-th FEB. To reconstruct the fused image, the output feature $\boldsymbol{F}^L_{ir/vi}$ from the final FEB is forwarded to the FRB. Within the FRB, we apply the Sobel operator to compute the gradient maps, which are concatenated along the channel dimension to form a combined gradient map $\boldsymbol{G} = [\nabla \boldsymbol{F}^L_{ir}, \nabla \boldsymbol{F}^L_{vi}]$. Here, $[\cdot]$ denotes channel-wise concatenation, and ∇ represents the Sobel gradient operator. To enhance the representation of multimodal features, we apply GAP, GMP, MeanP, and MaxP operations on $\boldsymbol{G}$, and then combine their outputs through element-wise multiplication. The resulting feature map is normalized by a Sigmoid function to produce the attention matrix $\boldsymbol{A}$:

$$\begin{aligned} &\boldsymbol{G}_f = p_f(\boldsymbol{G}), f \in \{\mathrm{GAP}, \mathrm{GMP}, \mathrm{MeanP}, \mathrm{MaxP}\}, \\ &\boldsymbol{A} = \mathrm{Sigmoid}(\boldsymbol{G}_{GAP} \odot \boldsymbol{G}_{GMP} \odot \boldsymbol{G}_{MeanP} \odot \boldsymbol{G}_{MaxP}), \end{aligned} \tag{1}$$

where $p_f(\cdot)$ denotes different types of pooling operations, and $\odot$ represents element-wise multiplication. The attention map $\boldsymbol{A}$ is then multiplied element-wise with $[\boldsymbol{F}^L_{ir}, \boldsymbol{F}^L_{vi}]$, and the result is fed into L FEBs and three convolutional blocks to generate the final fused image $\boldsymbol{I}_f$.

2. More Comparisons on Fusion Performance

We further conduct a qualitative comparison with several state-of-the-art fusion methods, including CoCo [4], TDAL [2], IRFS [7], SMiF [3], MRFS [9], TIMF [5], and SAGE [8]. As shown in Figure 2, the qualitative results demonstrate that our method preserves fine details of pedestrians and vehicles while effectively highlighting thermal radiation cues. It achieves a good balance between infrared and visible modalities, producing visually pleasing fusion results. In contrast, existing methods often exhibit modality imbalance, overly emphasizing either infrared or visible information.

Table 1. Quantitative evaluation results on the YOLOv5 and YOLOv11 detectors. The best performance for each metric is marked with a Red background.

Methods	YOLOv5	YOLOv11
MRFS [9]	0.6189	0.6206
Ours	0.6304	0.6430

Table 2. Quantitative comparison between the full model and its ablated variants across multiple downstream tasks. The best performance for each metric is marked with a Red background.

Methods	OD	Seg	SOD	
	$\mathrm{mAP}_{50\to95}\uparrow$	mIoU $\uparrow$	$\mathrm{mF}_\beta\uparrow$	$\mathrm{E}_m\uparrow$
Model VI	0.6289	60.13	0.8121	0.9085
Model V	0.6277	60.06	0.8128	0.9086
Ours	0.6304	60.34	0.8129	0.9087

3. More Comparisons on an Advanced Detector

To further evaluate the adaptability of the proposed method to more advanced detectors, we deploy our method on YOLOv11 [1] for quantitative evaluation. As shown in Table 1, the proposed method achieves significantly higher $\mathrm{mAP}_{50\to95}$ than the "task network retraining" approach (retraining based on the representative fusion method MRFS [9]) on both YOLOv5 and YOLOv11. These quantitative results demonstrate that our method maintains clear advantages on more advanced detectors.

4. More Ablation Studies

Effectiveness of BVB. The Basis Vector Bank (BVB) stores a set of basis vectors for generating task-specific convolutional kernel parameters. To evaluate its effectiveness, we design **Model IV**, which removes the BVB and directly predicts convolutional kernel parameters using only semantic

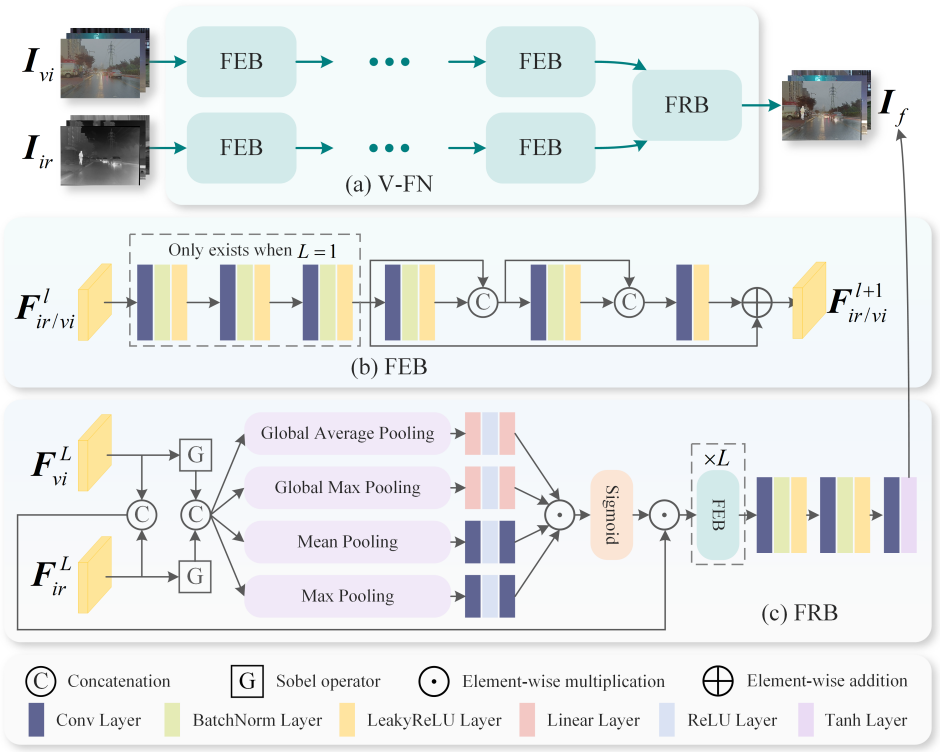

Figure 1. Network architecture of VFN. The VFN (a) consists of a Feature Extraction Blocks (FEB) (b) and a Fusion Feature Reconstruction Block (FRB) (c).

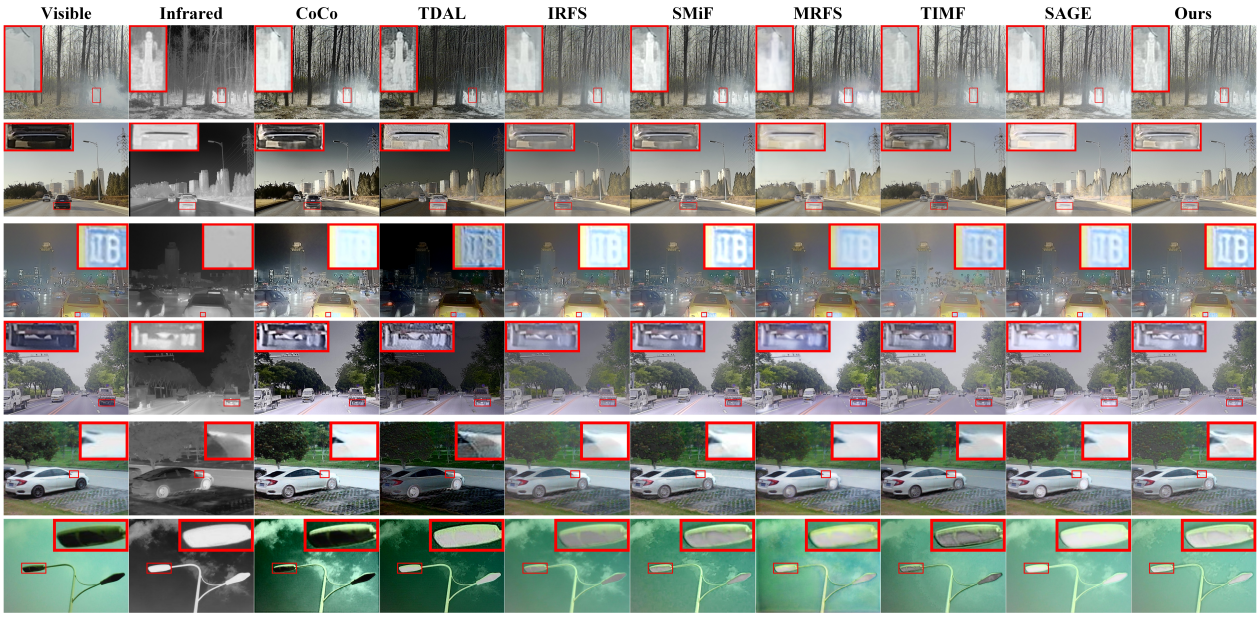

Figure 2. Qualitative comparison between the proposed method and existing state-of-the-art approaches. The first and second columns show the visible and infrared source images, respectively, while the third to tenth columns display the fused results of different comparison methods. The top two rows, middle two rows, and bottom two rows correspond to samples from the M^3FD, FMB, and VT5000 datasets, respectively.

Table 3. Quantitative analysis of the hyperparameter L across multiple downstream tasks. The hyperparameter settings of our proposed method are highlighted in **bold**, while the best performance for each metric is marked with a Red background.

L	OD	Seg	SOD		Params (M)	FLOPs (G)
	$mAP_{50\to95}$ ↑	mIoU ↑	mF_β ↑	E_m ↑		
2	0.6304	60.34	0.8129	0.9087	0.46	174.06
3	0.6262	60.38	0.8118	0.9095	0.89	348.12
4	0.6300	60.55	0.8119	0.9091	1.32	522.16

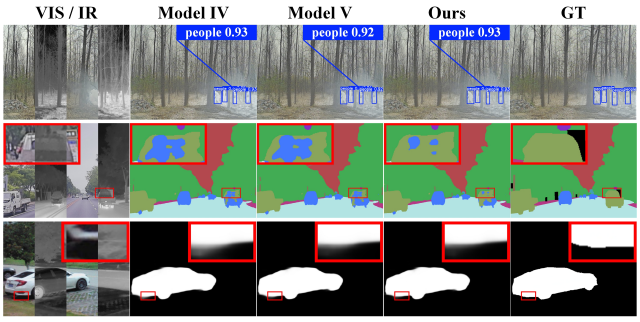

Figure 3. Qualitative comparison between the full model and the ablation models (Model IV and Model V).

features from downstream tasks and multimodal features. As shown in Table 2 and Figure 3, Model IV exhibits a clear performance degradation across all downstream tasks, with both quantitative and qualitative results significantly inferior to those of the full model. In contrast, the full model equipped with the BVB achieves superior and more balanced performance across multiple tasks. These results clearly demonstrate the pivotal role of the BVB in enhancing multi-task adaptability.

Effectiveness of A2SI. The Architecture-Adaptive Semantic Injection (A2SI) block adopts a multi-branch architecture for task-specific semantic extraction and adaptively adjusts the structure of each branch based on semantic features from different tasks. To evaluate its effectiveness, we design **Model V**, where the multi-branch structure is fixed and no longer adaptively adjusted based on task-specific semantic features. As shown in Table 2 and Figure 3, Model V exhibits a clear bias toward the object detection task and exhibits a notable decline in multi-task adaptability compared with the full model. In summary, the results confirm the essential role of the A2SI block in achieving effective multi-task adaptation.

5. Hyperparameter Analysis

The proposed method involves four key hyperparameters, namely the number of A2SI blocks L, the balance coefficient δ between the reward loss ℓ_r^n and the penalty loss ℓ_p^n, the number of semantic extraction branches M in the A2SI block, and the embedding dimension of the basis vectors e_2. In this work, we set $L = 2$, $\delta = 5$, $M = 4$, and $e_2 = 256$. To verify the effectiveness of these hyperparameter settings on Object Detection (OD), Semantic Segmentation (Seg), and Salient Object Detection (SOD) tasks, we conduct hyperparameter analysis on the M^3FD [2], FMB [3], and VT5000 [6] datasets.

Impact of L on Performance. To verify the effectiveness of the hyperparameter L , we conduct a search within the range of 2 to 4. Specifically, we set three groups of hyperparameters with $L = 2$, $L = 3$, and $L = 4$, respectively. As shown in Table 3, with the increase of L, the performance of the fused images on the semantic segmentation task gradually improves. However, the performance on the object detection task decreases to varying degrees, while that on the salient object detection task remains relatively stable. In addition, as L increases, both the number of parameters and computational cost grow significantly. Therefore, considering the overall performance and efficiency, we set $L = 2$ as the default value.

Impact of δ on Performance. To verify the effectiveness of the hyperparameter δ, we conduct a search within the range of 1 to 10. Specifically, three groups of hyperparameters are set to $\delta = 1$, $\delta = 5$, and $\delta = 10$, respectively. As shown in Table 4, when $\delta = 1$, it is difficult for CLDyN to achieve balanced performance across multiple downstream tasks. Conversely, when $\delta = 10$, it becomes challenging to maintain a balance between the penalty loss ℓ_p^n and the reward loss ℓ_r^n, which negatively affects the optimization of all tasks. Only when $\delta = 5$ does the VFN achieves relatively stable and superior performance across all downstream tasks. Therefore, we set the hyperparameter $\delta = 5$ as the default value.

Impact of M on Performance. To verify the effectiveness of the hyperparameter M, we perform a search within the range of 1 to 8. Specifically, five groups of hyperparameters are set as $M = 1$, $M = 2$, $M = 4$, $M = 6$, and $M = 8$, respectively. As shown in Table 5, setting too few or too many semantic extraction branches weakens the adaptability of the method across multiple downstream tasks. When $M = 4$, the fused images achieve the best performance on both object detection and semantic segmenta-

Table 4. Quantitative analysis of the hyperparameter δ across multiple downstream tasks. The hyperparameter settings of our proposed method are highlighted in **bold**, while the best performance for each metric is marked with a Red background.

δ	OD	Seg	SOD	
	$mAP_{50\to95}$ ↑	mIoU ↑	mF_β ↑	E_m ↑
1	0.6278	60.17	0.8123	0.9088
5	0.6304	60.34	0.8129	0.9087
10	0.6293	60.15	0.8125	0.9085

Table 5. Quantitative analysis of the hyperparameter M across multiple downstream tasks. The hyperparameter settings of our proposed method are highlighted in **bold**, while the best performance for each metric is marked with a Red background.

M	OD	Seg	SOD	
	$mAP_{50\to95}$ ↑	mIoU ↑	mF_β ↑	E_m ↑
1	0.6289	60.07	0.8119	0.9086
2	0.6294	60.15	0.8123	0.9086
4	0.6304	60.34	0.8129	0.9087
6	0.6293	60.29	0.8132	0.9091
8	0.6299	60.23	0.8118	0.9084

tion tasks, and perform comparably well on salient object detection. Therefore, we set the hyperparameter $M = 4$ as the default value.

Impact of e_2 on Performance. To verify the effectiveness of the embedding dimension e_2, we perform a search within the range of 128 to 512. Specifically, the embedding dimensions of the basis vectors are set to 128, 256, and 512, respectively. As shown in Table 6, when $e_2 = 256$, the proposed method achieves the strongest adaptability across multiple downstream tasks, maintaining consistently superior performance in all tasks. Therefore, we set $e_2 = 256$ as the default configuration.

Table 6. Quantitative analysis of the hyperparameter e_2 across multiple downstream tasks. The hyperparameter settings of our proposed method are highlighted in **bold**, while the best performance for each metric is marked with a Red background.

e_2	OD	Seg	SOD	
	$mAP_{50\to95}$ ↑	mIoU ↑	mF_β ↑	E_m ↑
128	0.6273	60.18	0.8118	0.9083
256	0.6304	60.34	0.8129	0.9087
512	0.6292	60.25	0.8130	0.9092

Table 7. Parameter counts and computational costs of the core components in the proposed method and their corresponding ablated variants.

Models	A2SI	BVB	Params (M)	FLOPs (G)
IV	✓	✗	0.11	1.88
V	✗	✓	0.35	173.04
Full	✓	✓	0.46	174.06

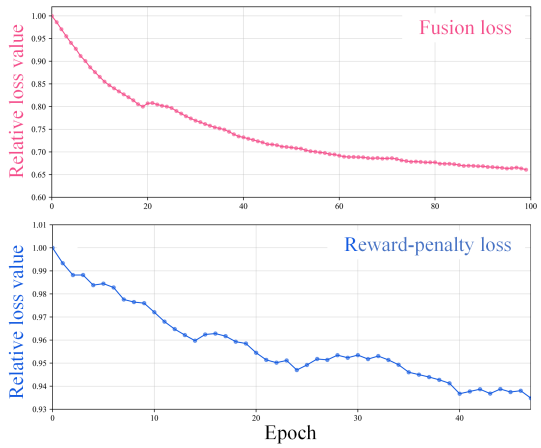

Figure 4. Training loss curves of the proposed method.

6. Convergence Analysis

To verify the stable optimization and convergence of the proposed method, we visualize the fusion loss and the reward-penalty loss during training. As shown in Figure 4, both losses decrease smoothly and gradually converge without catastrophic oscillations.

7. More Complexity Analysis

To better analyze the complexity of the proposed method, we report the parameter counts and computational costs of each core component, as shown in Table 7.

Table 8. Quantitative results under complex scenarios (exposure attenuation $\alpha \in [0.4, 0.8]$).

Complex scenarios	OD	Seg	SOD	
	$mAP_{50\to95}$ ↑	mIoU ↑	mF_β ↑	E_m ↑
✓	0.6284	59.73	0.8100	0.9074
✗	0.6304	60.34	0.8129	0.9087

8. Limitations and Future Work

The proposed method assumes that the input infrared and visible images are captured under normal weather condi-

tions. However, in open-world applications, the algorithm often encounters complex and dynamic environments such as heavy rain, low-light conditions, and severe noise, where its performance may degrade, as shown in Table 8. To address this challenge, future research will focus on developing a multi-task-aware infrared and visible image fusion framework that remains robust under extreme weather conditions, enabling reliable performance across diverse downstream tasks. In addition to external weather factors, the aging of internal sensor components can also deteriorate image quality, leading to increased noise, pixel defects, and intensified artifacts. Future work will further explore multi-task-aware infrared and visible image fusion under complex compound degradations, aiming to enhance the robustness and generalization of the algorithm in open-world scenarios.

References

[1] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 1

[2] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan 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 (CVPR)*, pages 5802–5811, 2022. 1, 3

[3] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8081–8090, 2023. 1, 3

[4] Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *International Journal of Computer Vision*, 132 (5):1748–1775, 2024. 1

[5] Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024. 1

[6] Zhengzheng Tu, Yan Ma, Zhun Li, Chenglong Li, Jieming Xu, and Yongtao Liu. Rgbt salient object detection: A large-scale dataset and benchmark. *IEEE Transactions on Multimedia*, 25:4163–4176, 2023. 3

[7] Di Wang, Jinyuan Liu, Risheng Liu, and Xin Fan. An interactively reinforced paradigm for joint infrared-visible image fusion and saliency object detection. *Information Fusion*, 98: 101828, 2023. 1

[8] Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17882–17891, 2025. 1

[9] Hao Zhang, Xuhui Zuo, Jie Jiang, Chunchao Guo, and Jiayi Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26964–26973, 2024. 1