RegionFuse: Region-Adaptive Pixel Distribution Learning for Infrared and Visible Image Fusion

Supplementary Material

1. More Qualitative Comparison Results

More qualitative comparison results are shown in Fig. 1 and Fig. 2. Both our method and DDFM [5] are capable of effectively capturing information from both the visible and infrared modalities. However, our approach produces images with richer details and textures, while better preserving the overall illumination characteristics of the visible image.

Besides, we provide a visual example of objects spanning different illumination zones from the contrast-enhanced dataset (see Sec. 4.2, $k = 2.5$). In Fig. 3, our method injects more infrared information into the low-light regions while maintaining smooth transitions across regions, without introducing artifacts or abrupt pixel changes. This shows that our method can better capture cross-modal regional variations while preserving global semantic consistency.

2. More Exploration Studies

2.1. Sensitivity Analysis of Region Size

We conducted a sensitivity analysis on the region size, as presented in Table 1. The results show that the network achieves optimal performance with a region size of 48×48. Further reducing size to 16×16 leads to an out-of-memory error on a single RTX 4090 GPU under our training settings.

Table 1. Sensitivity analysis of region size on the MSRS test set.

Region Size	SF	AG	VIF	Qabf	Qcb
64×64	11.963	3.931	1.013	0.710	0.596
48×48	**12.115**	**3.995**	**1.063**	**0.723**	**0.601**
32×32	12.002	3.962	1.023	0.710	0.597
16×16	—	—	—	—	—

"—" indicates an out-of-memory error under this setting.

2.2. Sensitivity Analysis of Expert Number

As shown in Fig. 4, we performed a sensitivity analysis on the number of experts in MoRA. The results indicate that our method achieves the best AG metric with 4 experts and a top-3 selection, while it achieves the best SF metric under the E6-K1 configuration. All other metrics achieve their highest performance with the E4-K2 setting. Therefore, we employ 4 experts and select the top-2 experts for sparse routing in our experiments.

2.3. Sensitivity Analysis of λ in Loss Function

The hyperparameter λ is used to balance the contribution of the gradient loss to the total loss. Tab. 2 presents the sensitivity analysis of λ. The results indicate that the fusion performance remains stable across a wide range of λ values. We set $\lambda = 5.0$ as it achieves the best MR and RoR.

Table 2. Sensitivity analysis. MR: mean rank; RoR: rank of rank.

λ	SF	AG	VIF	Qabf	Qcb	MR↓	RoR↓
1.0	12.336	4.025	1.048	0.717	0.606	2.6	2
2.0	12.054	3.980	1.051	0.722	0.603	3.6	5
5.0	12.115	3.995	1.063	0.723	0.601	2.4	1
10	12.109	3.997	1.057	0.719	0.601	3.4	4
20	12.209	4.023	1.062	0.721	0.599	2.8	3

2.4. Sensitivity Analysis of μ in Loss Function

The hyperparameter μ is used to balance the contribution of the load loss to the total loss. To investigate its impact on the network's performance, we conducted a sensitivity analysis. We experimented with different μ values and evaluated the results on the MSRS dataset. As shown in Tab. 3, the model achieves optimal performance across all metrics when μ is set to 5.0. Therefore, we adopt $\mu = 5.0$ as the default setting in our experiments.

Table 3. Sensitivity analysis of μ on the MSRS test set.

μ	SF	AG	VIF	Qabf	Qcb
0.5	11.505	3.754	1.054	0.704	0.592
1	11.496	3.762	1.041	0.704	0.586
5.0	**12.115**	**3.995**	**1.063**	**0.723**	**0.601**
10	11.767	3.886	1.056	0.708	0.594
50	11.613	3.813	1.053	0.706	0.595

2.5. Expert Allocation Analysis

We incorporate a pixel distributionaware router into MoRA, which assigns regional features to specialized experts. In practice, the number of regions corresponding to specific pixel distributions is inherently imbalanced. For example, over- and under-exposed regions are rare in typical photographs. However, the standard load-balancing loss in sparse MoE [2] enforces a uniform allocation of samples across experts, which we argue impedes the learning of local pixel distributions. To address this, we adopted a two-stage training strategy. The load balancing loss was applied only for the first 5 epochs as an auxiliary loss to prevent

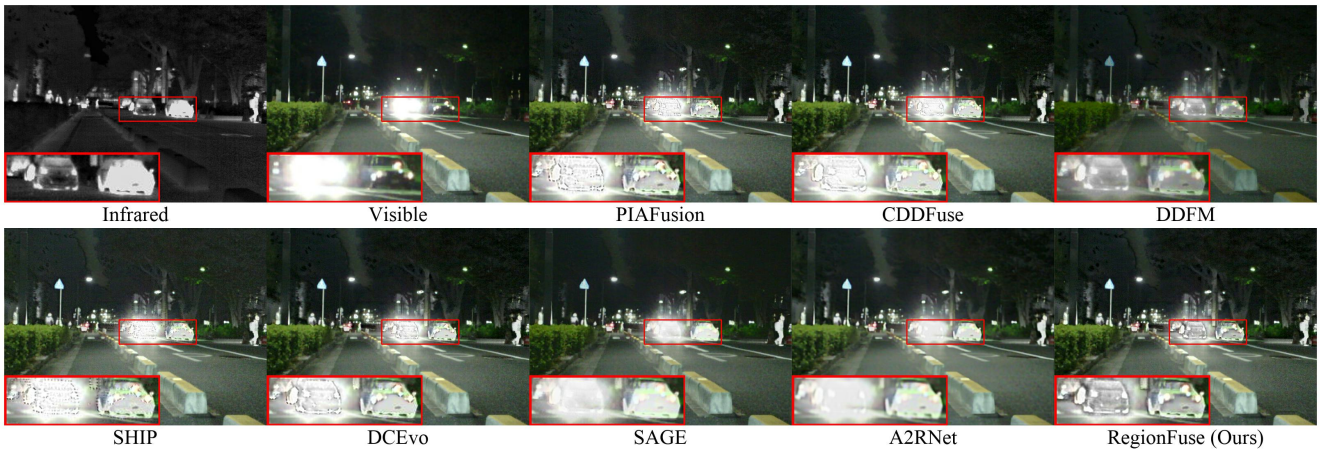


Figure 1. Qualitative comparison on the MSRS dataset.

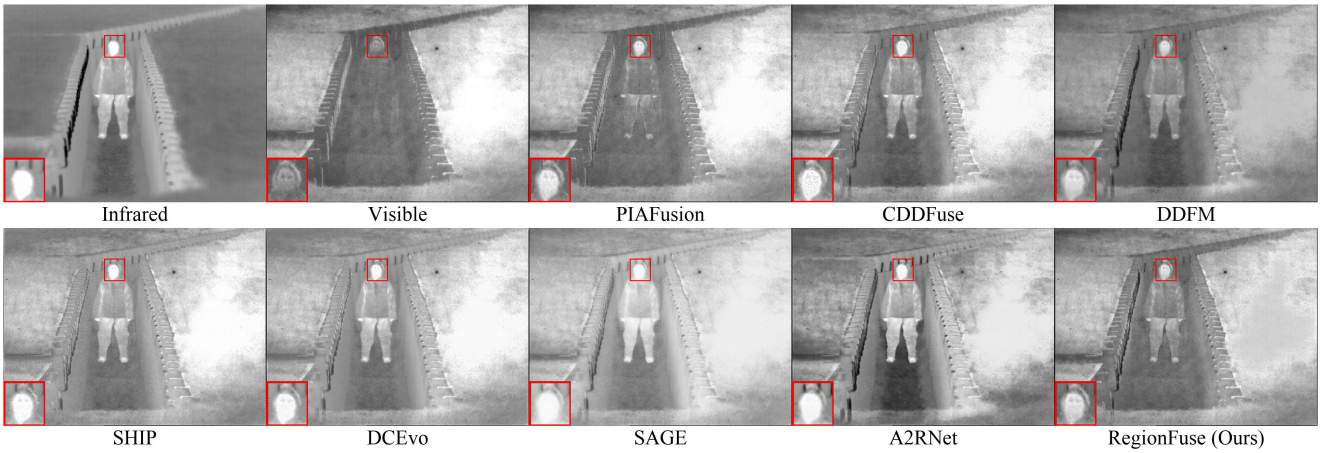


Figure 2. Qualitative comparison on the TNO dataset.

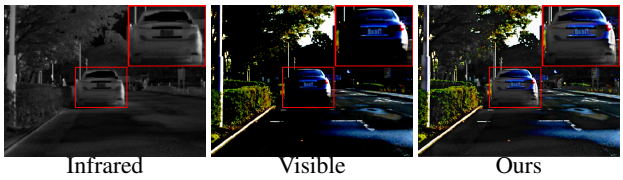


Figure 3. Visual example on enhanced dataset ($k = 2.5$).

the router from favoring a small subset of experts, and was discarded thereafter.

As shown in Fig. 5, after the first five epochs, the allocation of regions across experts changes dynamically, with more regions being routed to Experts 1 and 2. However, the initial phase of load-balanced training still guarantees that every expert receives a sufficient number of regions, ensuring all experts are engaged in the learning process.

2.6. Efficiency Study

Efficiency metrics are necessary to assess practical deployment. We therefore include Tab. 4 to report the parameters, FLOPs, and latency for all methods under the same setting (480×640 input, RTX 4090). The results show that our model has competitive efficiency compared with other Transformer-based baselines using standard MHA (e.g., CDDFuse/DCEvo).

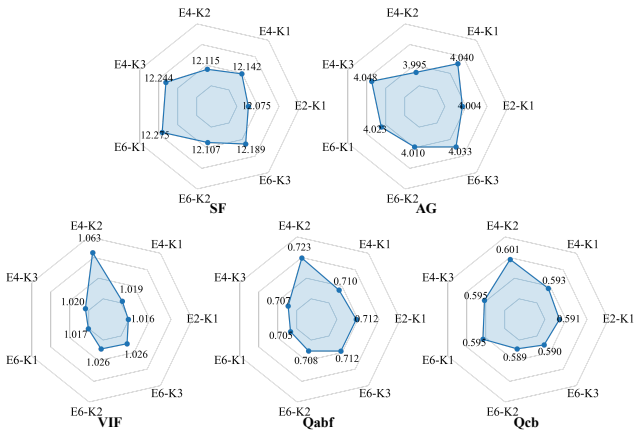


Figure 4. Sensitivity analysis of expert number on the MSRS test set. Ei-Kj indicates that each MoRA module contains i experts and sparsely selects the top-j experts.

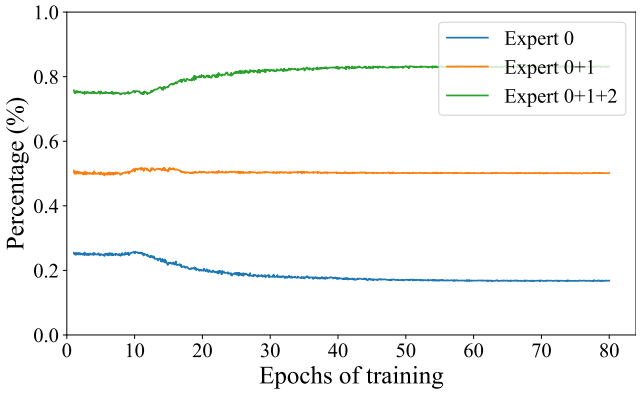


Figure 5. This figure illustrates the proportion of image regions assigned to each expert. Expert 0 denotes the standalone allocation percentage for the first expert, while Expert 0+1 and Expert 0+1+2 indicate the cumulative allocation percentages for the first two and first three experts, respectively.

Table 4. Efficiency comparison on MSRS dataset.

Methods	Params (M)	FLOPs (G)	Time (ms)
PIAF	1.18	361.06	16.50
CDDF	1.19	547.74	190.82
DDFM	552.81	5.36×10^5	3.31×10^4
SHIP	0.55	164.81	26.46
DCEvo	2.00	912.98	158.11
SAGE	0.14	20.21	2.06
A2R	10.61	171.19	36.34
Ours	2.82	455.82	168.87

2.7. Ablation Study on X-Restormer Block

While X-Restormer performs well in restoration, its direct use in fusion is non-trivial due to the need for balanced cross-modal integration. We conducted additional experiments to explore the impact of X-Restormer on different fusion networks. Replacing PIAFusion's encoder with X-Restormer leads to grainy artifacts (Fig. 6) despite some metric gains (Tab. 5), indicating unstable feature fusion. Thus, our performance gains cannot be attributed to X-Restormer alone, but to the fusion-oriented design of the MGT encoder and the overall framework.

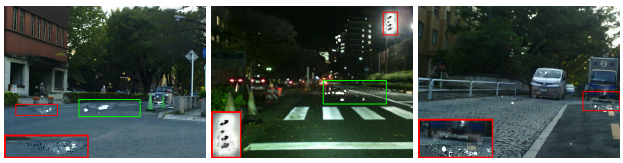


Figure 6. Visualization of grainy artifacts.

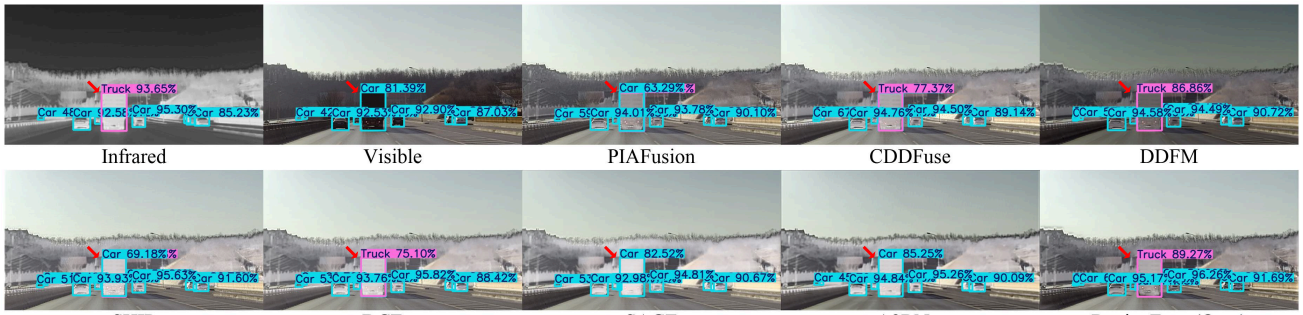
Table 5. Quantitative results on MSRS dataset (*: X-Restormer).

Methods	SF	AG	VIF	Qabf	Qcb
PIAF	11.49	3.75	0.99	0.69	0.58
PIAF*	**13.41**	**4.09**	0.97	0.69	0.58
Ours	12.12	3.99	**1.06**	**0.72**	**0.60**

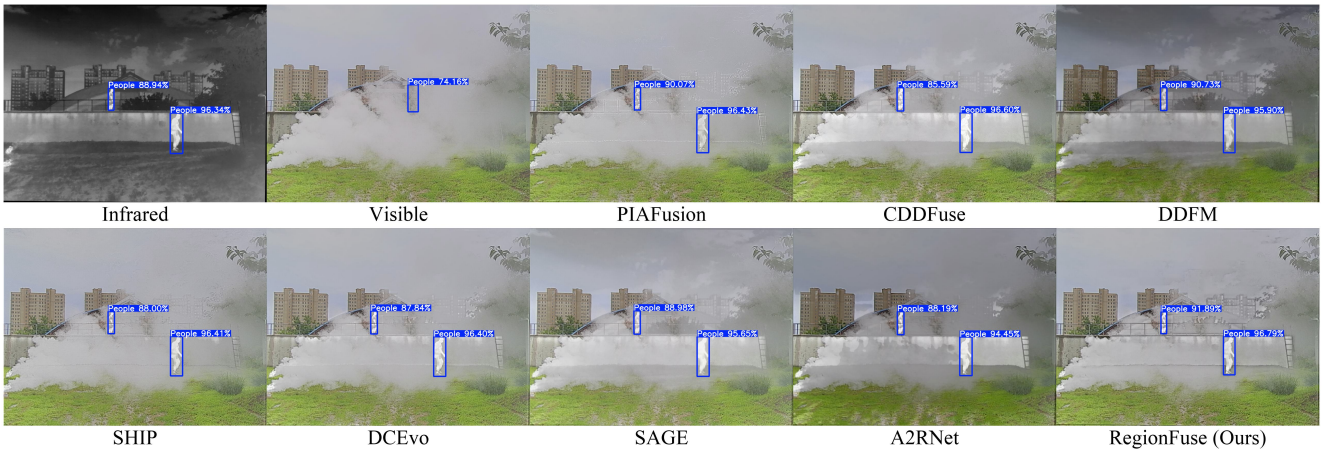
3. Visualization of Detection Results

As shown in Fig. 7, we visualize the detection results on the M3FD dataset. In the first example shown in Fig. 7a, the detector misclassifies a truck as a car when applied to the visible-only image, as well as the fused images generated by PIAFusion [3], SHIP [6], SAGE [4], and A2RNet [1]. In contrast, our RegionFuse correctly identifies the object and demonstrates the highest prediction confidence compared to the other fusion methods. The results indicate that our method can effectively enhance downstream detection performance.

In the second example shown in Fig. 7b, the human in the visible image is heavily obscured by smoke, causing the detector to fail to detect the person. However, the fused image integrates complementary information from both the visible and infrared modalities, enabling the detector to correctly recognize the presence of persons. In this case, our method produces fused images with the richest structural details and the highest visual fidelity, thereby enabling the detector to achieve the highest prediction confidence.



(a) Detection example in a road scene.



(b) Detection example in a smoke-occluded scene.

Figure 7. Visualization of detection results on the M3FD dataset.

References

[1] Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. A^2rnet: Adversarial attack resilient network for robust infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4770–4778, 2025. 3

[2] Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. 1

[3] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83:79–92, 2022. 3

[4] 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*, pages 17882–17891, 2025. 3

[5] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8082–8093, 2023. 1

[6] Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26384–26395, 2024. 3