Supplementary Material to "MRFS: Mutually Reinforcing Image Fusion and Segmentation"

Hao Zhang[1], Xuhui Zuo[1], Jie Jiang[2], Chunchao Guo[2], Jiayi Ma[1*]
[1]Electronic Information School, Wuhan University, Wuhan, China
[2]Data Platform Department, Tencent, Shenzhen, China
{zhpersonalbox,jyma2010}@gmail.com, xuhuizuo2001@whu.edu.cn, {zeus, chunchaoguo}@tencent.com

1. Feature Visualization

We visualize the output features of the IGM-Att and PC-Att modules to intuitively verify the effectiveness of these modules, as shown in Fig. 1. Firstly, as the number of IGM-Att blocks increases, the multi-modal features gain deeper interactions. Therefore, the domain gap between visible and infrared features is effectively reduced, presenting increasingly similar appearances. Along with this, these features gradually tend to converge. As shown in Fig. 1 (a), high-response pedestrians in visible features are transmitted to infrared features, while high-response buildings in infrared features are also transmitted to infrared features. These observations indicate that our IGM-Att module achieves feature complementation and refinement as expected. Besides, the powerful feature aggregation capabilities of the PC-Att modules can be demonstrated by these features. It can be seen that the objects of interest including pedestrians and vehicles, are clearly given attention, which can promote their accurate segmentation. In addition, the mutually reinforcing relationship between vision and semantics can also be observed. On the one hand, the visual features that IGM-Att focuses on provide more effective contrast guidance for the semantic features that PC-Att is interested in. For instance, the third row of features in Figs. 1 (a) and (b) gradually transition from a global high-response distribution to a more contrasting distribution. On the other hand, the features of PC-Att facilitate the highlighting of objects in the features of IGM-Att. For example, in block 3 of Fig. 1 (b), the infrared and visible features outputted by the IGM-Att module are like those in PC-Att, which improves attention towards pedestrians. Finally, an interesting finding is that as the number of blocks continues to increase, the features output by the IGM-Att and PC-Att modules exhibit a gradually consistent appearance. Especially in the fourth block, the features almost look the same. In general, these results demonstrate the intrinsic consistency between vision and semantics, indicating that image fusion and semantic segmentation are mutually reinforcing.

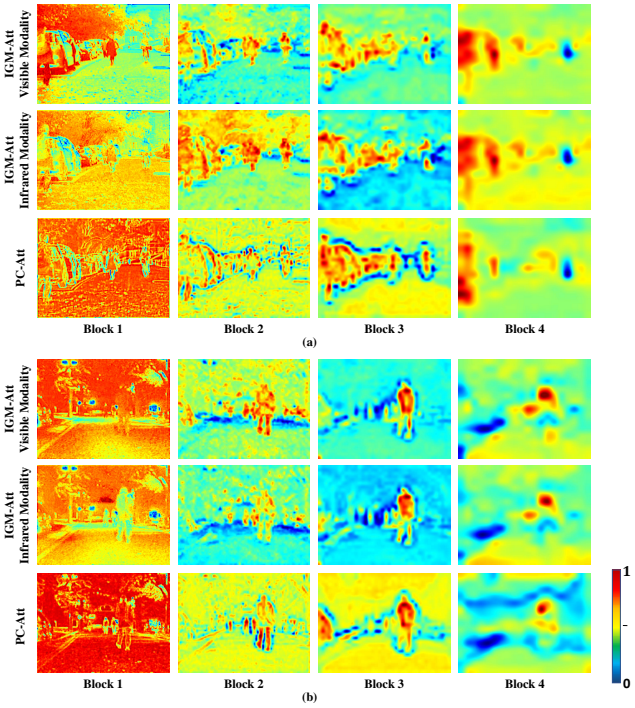

Figure 1. Visualization of the output features of the IGM-Att and PC-Att modules. (a) The 01463D image from the MFNet dataset; (b) The 01232N image from the MFNet dataset. Red denotes high-response features, while blue indicates low-response ones.

2. The Consistency Results of Ablation Studies

Here, we provide more qualitative fusion and segmentation results in ablation studies, as shown in Fig. 2. The configuration is the same as that in Section 4.4 of the main text. **Model I:** involves replacing salient information integration with a proportional strategy [12]; **Model II:** omits weaken information recovery; **Model III:** substitutes IGM-Att with conventional pooling-based attention [2]; **Model**

*Corresponding author

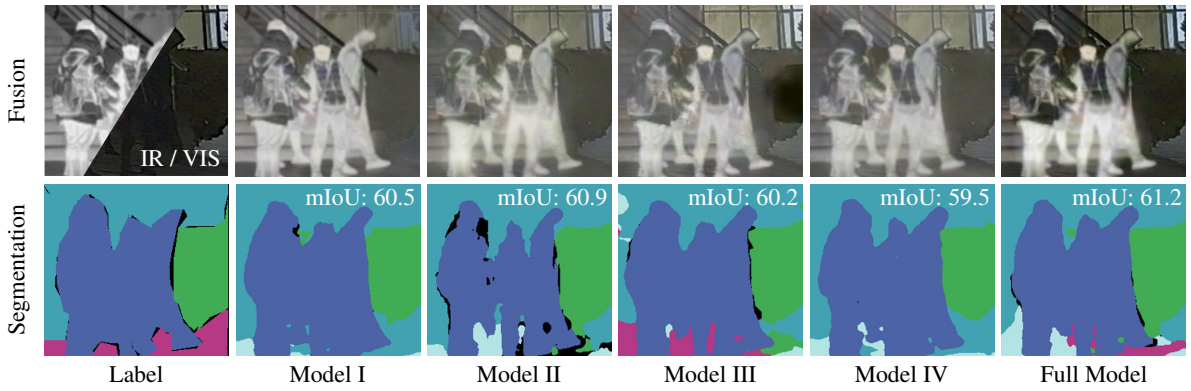

Figure 2. Qualitative fusion and segmentation results of ablation studies.

IV: replaces PC-Att with cross-attention-based feature integration [14]. From these results, we can see that better visual perception areas in the fused results correspond to better segmentation results. Specifically, our full model presents high-contrast results, clearly distinguishing background and foreground. Correspondingly, the full model can better separate foreground pedestrians and give more reasonable outer contours when performing segmentation. In contrast, removing any of these designs all diminishes the performance of image fusion and semantic segmentation. This means that each designs plays a important role in our method.

3. The Impact of Network Paradigm on IGM-Att and PC-Att

Image fusion is a low-level vision task dedicated to providing better visual perception, so it prioritizes local features for perception. Semantic segmentation is a high-level vision task that requires a full understanding of the scene, so it emphasizes a fusion of both local and global features. For the above considerations, we use CNN-based IGM-Att for image fusion and collaborate it with transformer-based PC-Att for segmentation. To verify the rationality of the network paradigm settings of IGM-Att and PC-Att, we provide the versions of transformer-based (IGM2Trans) and CNN-based (PC2CNN). As shown in Figs. 3 (I) (a)-(b), both IGM2Trans and PC2CNN yield degraded segmentation. Besides, Figs. 3 (II) (a)-(b) demonstrating that IGM2Trans negatively impacts fusion, while PC2CNN has minimal effect. These results validate our network paradigm choices for IGM-Att and PC-Att.

4. How PC-Att Affect Image Fusion

Although PC-Att is not directly connected to the image fusion head, it still influences image fusion by sharing IGM-Att. As joint optimization continues, their gradients can interact through chain conduction, thereby building a mutually promoting mechanism for image fusion and segmentation. To verify the above conclusion, We remove IGM-Att (rmIGM) and PC-Att (rmPC) for ablations, respectively. Figs. 3 (I) (c)-(d) and Figs. 3 (II) (c)-(d) indicate that both rmIGM and rmPC lead to reduced performance of image fusion and segmentation. These findings show that cascading IGM-Att and PC-Att improves both image fusion and segmentation.

5. Generalization Experiment

To verify the generalization performance of our MRFS, we test the model pre-trained on the MFNet dataset [3] on the LLVIP dataset [10]. The comparative methods include SDNet [11], U2Fusion [10], SeAFusion [8], DetFusion [7], DATFuse [9], CDDFuse [13], TGFuse [5], and SegMiF [4]. The visual results for objective evaluation are presented in Fig. 4. From the four scenes we present, it can be seen that our MRFS effectively recovers weak details and enhances overall visual quality. For instance, in Fig. 4 (a), our fused image improves the visibility, especially in the corners, and better highlights faint thermal objects compared to other methods. In Fig. 4 (b), our fused image enhances the visibility of the tree canopy while sufficiently retaining the digital color fidelity of the car. This means that our method preserves the original color of the car and does not make it appear as blue as other methods. In Fig. 4 (c), our fused image enhances the details of the car and provides another example of the faint thermal objects highlighting and the visibility of the tree canopy enhancing. In Fig. 4 (d), the overall background visibility and color fidelity in our fused image have been effectively enhanced. Simultaneously, the faint thermal objects have been better highlighted compared to other methods. We also provide the evaluation results of the information content and contrast of the fused image employing non-reference metrics, specifically entropy (EN). [6] and standard deviation (SD) [1]. In Fig. 5, our MRFS still retains its advanced performance.

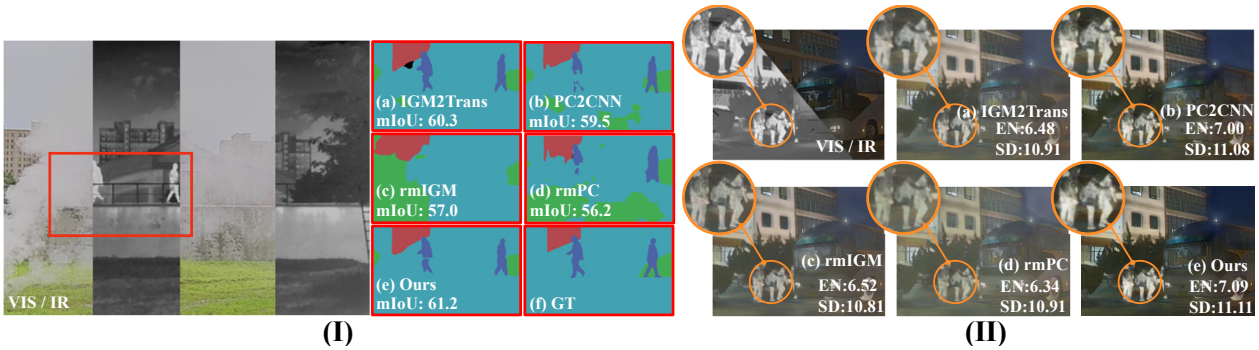

Figure 3. Results of the Impact of network paradigm on IGM-Att and PC-Att.

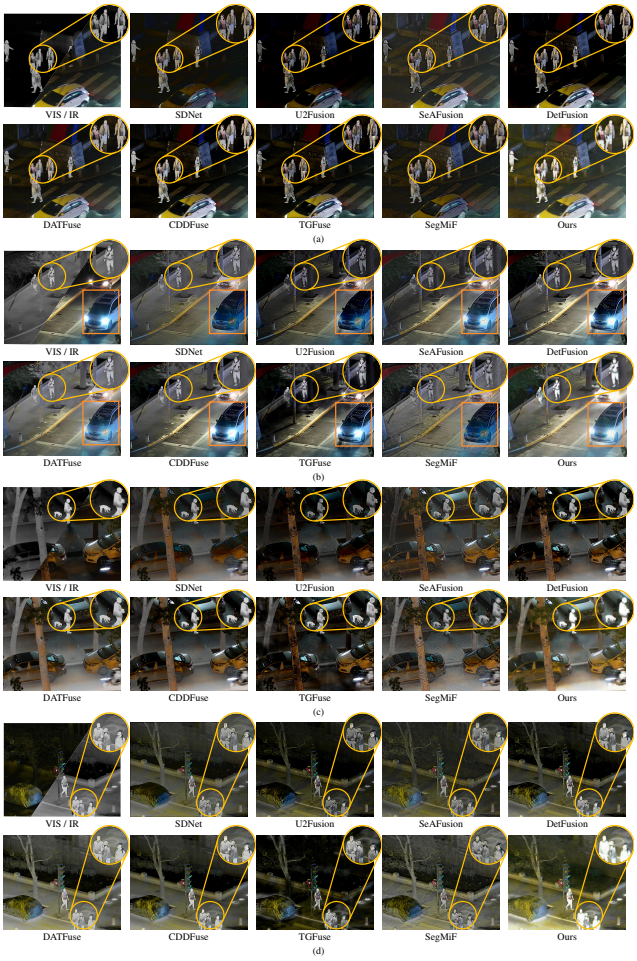

Figure 4. Qualitative generalization on the LLVIP dataset.

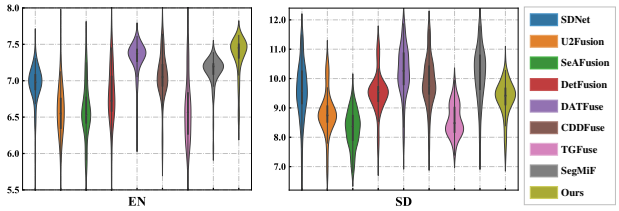

Figure 5. Quantitative generalization on the LLVIP dataset.

References

[1] V Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12):1890–1896, 2015. 2

[2] Fuqin Deng, Hua Feng, Mingjian Liang, Hongmin Wang, Yong Yang, Yuan Gao, Junfeng Chen, Junjie Hu, Xiyue Guo, and Tin Lun Lam. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4467–4473, 2021. 1

[3] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5108–5115, 2017. 2

[4] 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*, pages 8115–8124, 2023. 2

[5] Dongyu Rao, Tianyang Xu, and Xiao-Jun Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023. 2

[6] J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 2

[7] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the ACM International Conference on Multimedia*, pages 4003–4011, 2022. 2

[8] Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-

time infrared and visible image fusion network. *Information Fusion*, 82:28–42, 2022. 2

[9] Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023. 2

[10] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2022. 2

[11] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129(10):2761–2785, 2021. 2

[12] Hao Zhang, Han Xu, Yang Xiao, Xiaojie Guo, and Jiayi Ma. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12797–12804, 2020. 1

[13] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5906–5916, 2023. 2

[14] Heng Zhou, Chunna Tian, Zhenxi Zhang, Qizheng Huo, Yongqiang Xie, and Zhongbo Li. Multispectral fusion transformer network for rgb-thermal urban scene semantic segmentation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2