AMDANet: Attention-Driven Multi-Perspective Discrepancy Alignment for RGB-Infrared Image Fusion and Segmentation

Supplementary Material

1. Supplementary Content

This supplementary material presents additional experimental results that are omitted from the main paper due to the space limit. In this supplementary material, we provide:

– visualization results of semantic segmentation on FMB, PST900, and MFNet datasets (Sec. 2);
– visualization results of image fusion on FMB, PST900, and MFNet datasets (Sec. 3);
– a discussion of the hyperparameters of the $\mathcal{L}_{total}$ loss function (Sec. 4);
– a discussion for the threshold τ of the SCT (Sec. 5);
– a discussion for the Local-Alignment and Global-Alignment of the FDAM (Sec. 6);
– performance analysis of Local-Alignment (Sec. 7)
– analysis of failure cases (Sec. 8)

2. Semantic Segmentation Results

To intuitively compare the performance of our method with advanced methods such as SegMiF [3] and MRFS [6] in semantic segmentation, Fig. 2, 3, and 4 present segmentation results on the FMB [3], MFNet [2], and PST900[4] datasets. From the first and second rows of Fig. 2, it can be observed that in nighttime environments, the differing imaging principles of infrared and visible light sensors result in significant feature discrepancies between the different modalities. These substantial inter-modal feature differences cause existing methods to retain some of the disparate features when constructing consistent fusion features, making it difficult to accurately segment fine and small objects. For instance, in the first row of Fig. 2, distant traffic signs are not segmented accurately by other methods. In contrast, our model utilizes anti-discrepancy causal inference to eliminate the discrepancies between multi-modal features, ensuring the accuracy of constructing fusion features. Benefiting from this comprehensive elimination of feature discrepancies, our method achieves more desirable semantic segmentation results. The same conclusion can be drawn from Fig. 3 and 4. For example, in the third row of Fig. 4, our method identifies and segments a person riding a bicycle in the distance more accurately.

3. Image Fusion Results

To intuitively compare the performance of our method with advanced methods such as SegMiF and MRFS in image fusion, Fig. 5, 6, and 7 present fusion results on the FMB, MFNet, and PST900 datasets. As shown in Fig. 5, compared to other methods, our approach demonstrates superior performance in enhancing fine texture details and improving realistic visual quality. For instance, in the fourth row of Fig. 5, under smoke-obscured conditions, the fusion results generated by other methods struggle to highlight the contours of the person behind the smoke. In contrast, our method effectively utilizes the features provided by the infrared image, clearly presenting the obscured person's characteristics. The reason for the above results is that we use the mutual feature mask learning strategy to promote the fusion of multimodal features, making the results directly reflect the complementary features between different modal features. Similar conclusions can be drawn from Fig. 6 and 7. For example, in the first row of Fig. 6, our method more clearly delineates the structural contours of pedestrians, and in the second row of Fig. 7, our method highlights the backpack with more prominent features independent of the background.

4. Discussion of the Hyperparameters (α_1, α_2, and α_3)

In the loss function $\mathcal{L}_{total}$ used by our ADMAnet, we balance the contributions of the semantic segmentation loss $\alpha_1\mathcal{L}_{seg}$, image fusion loss $\alpha_2\mathcal{L}_{fus}$, and mask consistency regularization loss $\alpha_3\mathcal{L}_{cr}$ during model training by adjusting the hyperparameters $\alpha_1 = 1$, $\alpha_2 = 0.5$, and $\alpha_3 = 0.5$. To validate the rationality of the selected hyperparameters, we discuss the effects of different α_1, α_2, and α_3 on the model's performance in Tab. 1, 2, and 3. The experiments evaluate the model's performance using the mean Intersection over Union (mIoU) metric.

Hyperparameter α_1. As shown in Tab. 1, the model's performance on all three datasets gradually decreases as α_1 decreases, with the worst performance observed at $\alpha_1 = 0.3$. This result indicates that a smaller α_1 limits the model's ability to learn how to construct features conducive to semantic segmentation from labeled data, leading to a disconnect between the constructed features and the segmentation task. In contrast, when $\alpha_1 = 1$, the model receives a stronger penalty related to the semantic segmentation task, enabling it to focus more on constructing fusion features suitable for segmentation.

Hyperparameter α_2. As shown in Tab. 2, the model's performance significantly decreases when a smaller $\alpha_2 = 0.3$ is used. This is because the lack of strong supervision for the fusion process prevents the model from constructing accurate multimodal fusion features. Conversely, when a

Table 1. Evaluation of the α_1 ($\alpha_2 = 0.5, \alpha_3 = 0.5$).

Dataset	α_1=0.3	α_1=0.5	α_1=0.7	α_1=1
MFNet	58.7	60.2	61.6	**62.1**
FMB	59.5	62.6	63.5	**64.8**
PST900	80.9	86.5	88.1	**88.5**

Table 2. Evaluation of the α_2 ($\alpha_1 = 1, \alpha_3 = 0.5$).

Dataset	α_2=0.3	α_2=0.5	α_2=0.6	α_2=0.7
MFNet	60.5	**62.1**	61.8	61.3
FMB	61.8	**64.8**	63.6	63.3
PST900	85.4	**88.5**	88.2	87.6

Table 3. Evaluation of the α_3 ($\alpha_1 = 1, \alpha_2 = 0.5$).

Dataset	α_3=0.3	α_3=0.5	α_3=0.6	α_3=0.7
MFNet	59.4	**62.1**	61.2	60.7
FMB	61.2	**64.8**	64.2	61.3
PST900	86.7	**88.5**	88.3	84.6

larger $\alpha_2 = 0.7$ is used, performance also declines, as excessive focus on the fusion results causes the model to overlook the construction of fusion features suitable for semantic segmentation. In contrast, when $\alpha_2 = 0.5$, the model efficiently balances the fusion and segmentation processes, achieving optimal performance.

Hyperparameter α_3. As shown in Tab. 3, compared to $\alpha_3 = 0.3$, the model performs better when $\alpha_3 = 0.5$. This improvement is attributed to the fact that a larger α_3 allows the model to focus more on leveraging the inter-feature mask learning strategy to enhance multimodal feature fusion. However, the value of α_3 is not necessarily better when it is larger. When $\alpha_3 = 0.7$, the network's performance experiences a significant decline. This is because the inter-feature mask learning strategy is a complex contrastive learning approach. Over-reliance on this strategy increases the model's learning difficulty, making it challenging to construct fusion features for segmentation.

5. Discussion for the Threshold τ of the SCI

In the semantic consistency inference (SCI), we utilize semantic similarity as a bias indicator to measure the semantic consistency between modalities. Specifically, we introduce a threshold τ=0.4. When the semantic similarity between modalities is less than τ, the current multimodal features are considered to be significantly influenced by the bias of the encoder. To validate the rationality of τ=0.4, we compare the impact of different τ values on model performance. As shown in Tab. 4, when τ is set to larger values (e.g., 0.6 and 0.8), the model's performance declines. This is because a larger τ causes the model to overlook too many cross-modal semantic similarity features. Conversely, when τ is set to a smaller value (e.g., 0.2), the model's performance also decreases. This is because a smaller τ leads the model to excessively focus on inter-modal discrepancy features, increasing the risk of incorporating irrelevant discrepancies into feature fusion. Therefore, setting τ=0.4 allows the model to fully leverage inter-modal semantic similarity and facilitate the establishment of robust fusion features.

Table 4. Evaluation of the Threshold τ of the ACT.

Dataset	τ=0.2	τ=0.4	τ=0.6	τ=0.8
MFNet	60.3	**62.1**	60.7	60.4
FMB	61.3	**64.8**	63.5	62.6
PST900	85.4	**88.5**	87.2	86.3

Table 5. Ablations of the local-alignment and global-alignment.

Dataset	Local-Alignment	Global-Alignment	SFE	mIoU	mAP
MFNet	✗	✔	✔	59.2	73.9
	✔	✗	✔	58.5	71.7
	✔	✔	✗	61.2	75.8
	✔	✔	✔	**62.1**	**77.1**

6. Discussion for the Local-Alignment and Global-Alignment of the FDAM

In the Feature Discrepancy Alignment Module (FDAM), we employ two sub-modules, local-alignment and global-alignment, to align cross-modal feature representations from both local and global perspectives. To demonstrate the effectiveness of these modules, we compare their impact on model performance. As shown in Tab. 5, removing either module leads to a decline in model performance. Furthermore, compared to removing the local alignment module, the performance drop is more significant when the global alignment module is removed. This is because, during feature fusion and semantic segmentation, long-range contextual information plays a more critical role in modeling continuous semantic features, thereby improving the completeness of segmentation masks. Therefore, these experiments validate the importance of aligning multimodal features from both local and global perspectives.

In the Global Alignment sub-module, for the input feature streams, we first pass them through the Salient Feature Enhancement (SFE) module to enhance key features across different modalities. The purpose of SFE is to highlight semantically reliable regions prior to cross-modal interaction, suppress inconsistent responses, and guide the cross-attention mechanism to focus more accurately on relevant features across modalities. To evaluate the effectiveness of SFE, we conduct an ablation study. As shown in Tab. 5, removing SFE leads to a performance drop, which confirms its effectiveness in facilitating the alignment of key features

Table 6. Quantitative comparison with CBAM and DANet.

Method	FMB dataset		MFNet dataset	
	mIoU	mAP	mIoU	mAP
with CBAM [5]	63.3	74.8	60.8	73.1
with DANet [1]	63.6	75.2	60.5	74.3
***with* Ours**	**64.8**	**77.1**	**62.1**	**77.4**

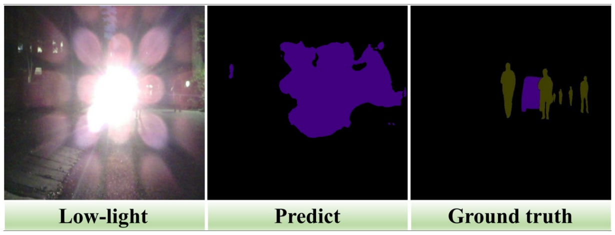

Figure 1. Failure cases analysis of AMDANet.

across modalities. This also indicates that relying solely on standard cross-attention is insufficient to capture the dependencies between modalities effectively.

7. Performance analysis of Local-Alignment

To further validate the performance of Local-Alignment, we compare it with similar attention-based methods such as CBAM [5] and DANet [1]. The experimental results are presented in Tab. 6. As shown in Table 7, compared to CBAM and DANet, which both couple spatial and channel attention, our Local-Alignment achieves a significant improvement in model performance. For example, compared to CBAM, our method improves mIoU by 1.5% and 1.3% on the FMB and MFNet datasets, respectively. This improvement is attributed to the fact that, unlike CBAM and DANet, our Local-Alignment extracts differential clues by minimizing the feature residual between the input and output, thereby dynamically guiding spatial and channel attention while adaptively suppressing misleading features. This approach allows for a deeper optimization of attention, rather than simply enhancing the features.

8. Failure Cases

The core of AMDANet lies in establishing complementary features by locating cross-modal semantic anchors. Although AMDANet performs strongly in establishing cross-modal semantic consistency features, its performance may degrade when the features of one modality are severely missing. As shown in Fig. 1, overexposure causes semantic degradation, leading to severe segmentation errors in AMDANet. This is because the loss of RGB modality features makes it difficult to locate reliable semantic anchors, thus hindering the establishment of semantic consistency fused features.

To address this issue, in future work, we consider introducing an adaptive weighting mechanism. This mechanism can automatically adjust the contribution weights of different modalities when one modality's features are severely lost, reducing the impact of information loss on the model's performance. Additionally, by incorporating deep feature reconstruction techniques, we can further supplement the missing semantic information in one modality and restore the localization of key semantic anchors, ensuring the effective maintenance of cross-modal semantic consistency.

References

[1] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[2] 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 *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115, 2017. 1

[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 8115–8124, 2023. 1

[4] Shreyas S. Shivakumar, Neil Rodrigues, Alex Zhou, Ian D. Miller, Vijay Kumar, and Camillo J. Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9441–9447, 2020. 1

[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[6] 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 26974–26983, 2024. 1

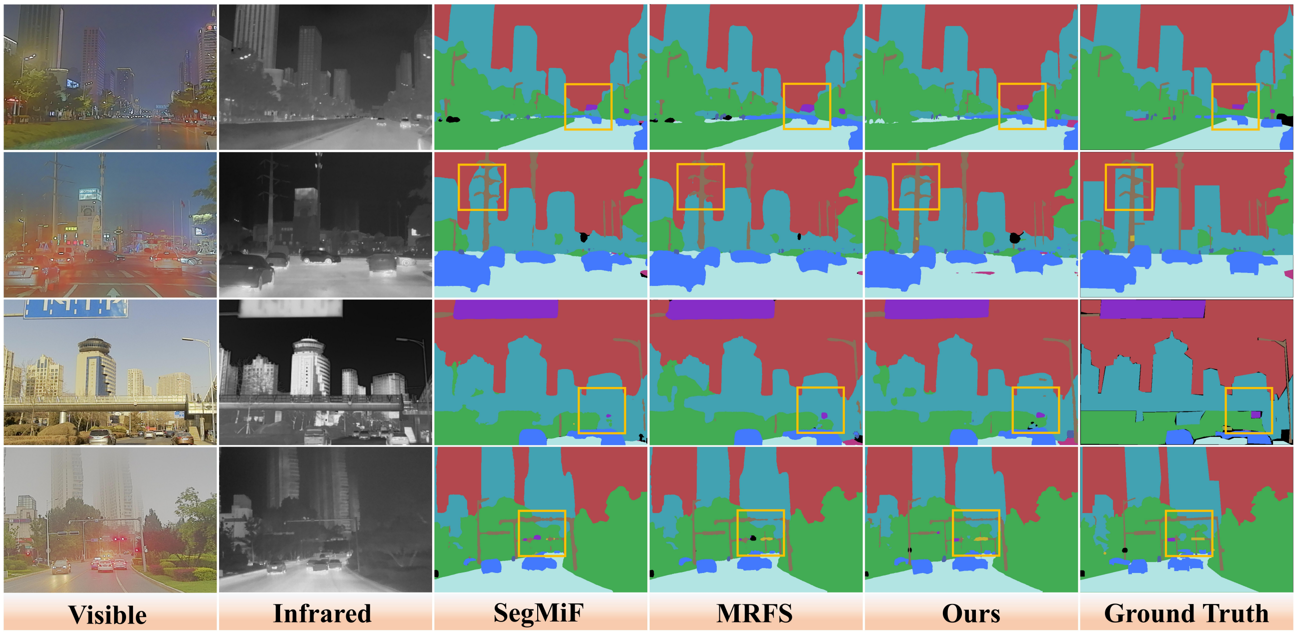

Figure 2. Visualization results of semantic segmentation on FMB dataset. The yellow box shows areas with obvious differences.

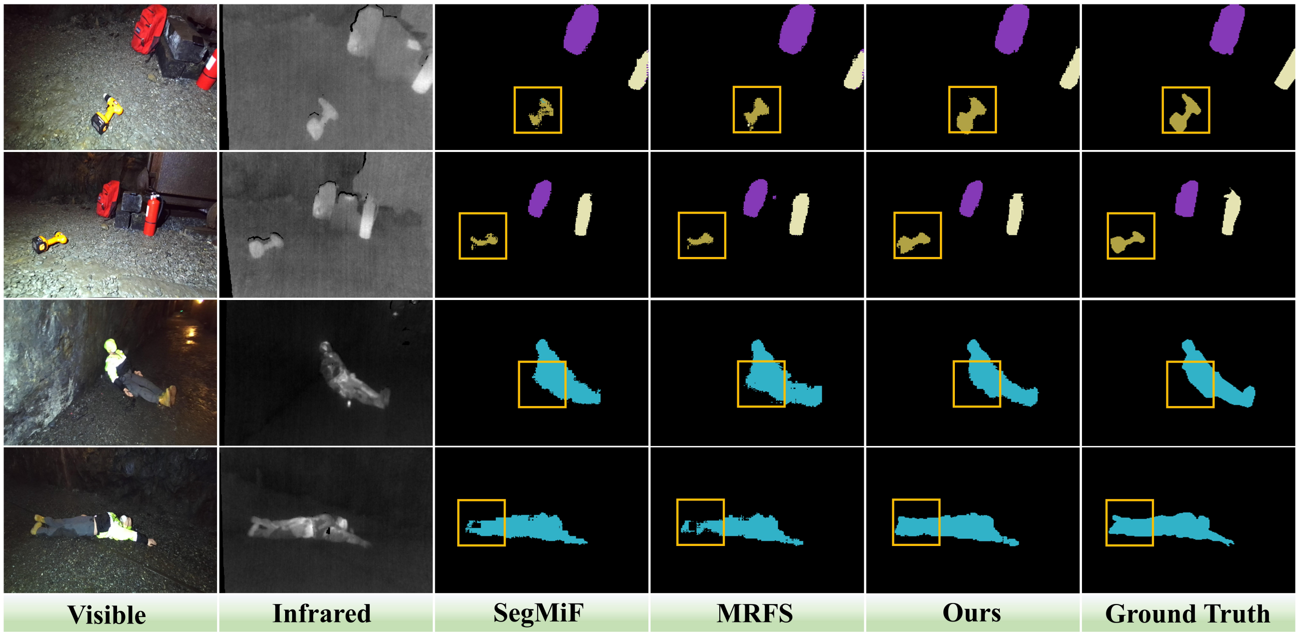

Figure 3. Visualization results of semantic segmentation on PST900 dataset. The yellow box shows areas with obvious differences.

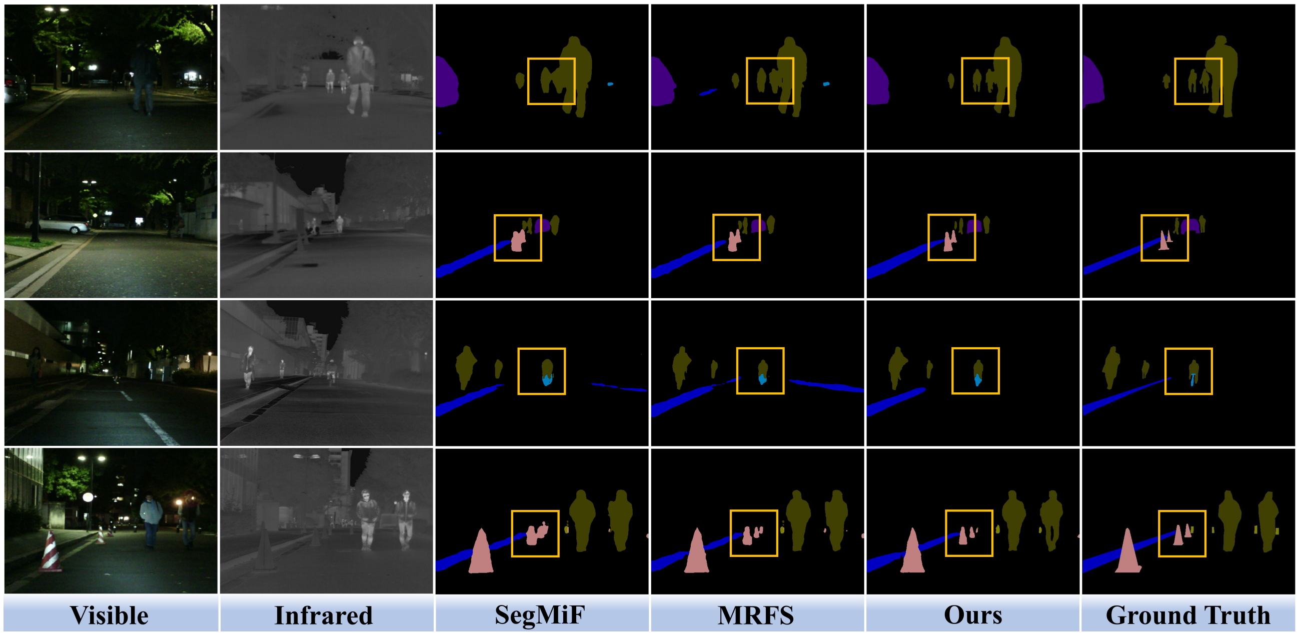

Figure 4. Visualization results of semantic segmentation on MFNet dataset. The yellow box shows areas with obvious differences.

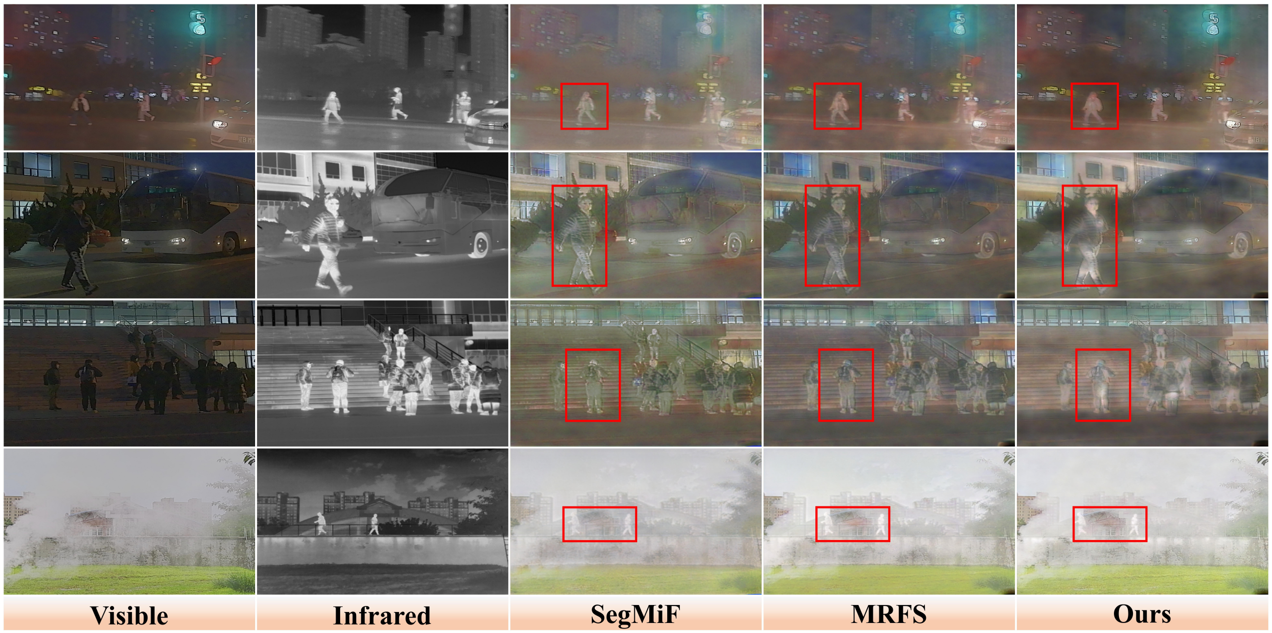

Figure 5. Visualization results of image fusion on FMB dataset. The red box shows areas with obvious differences.

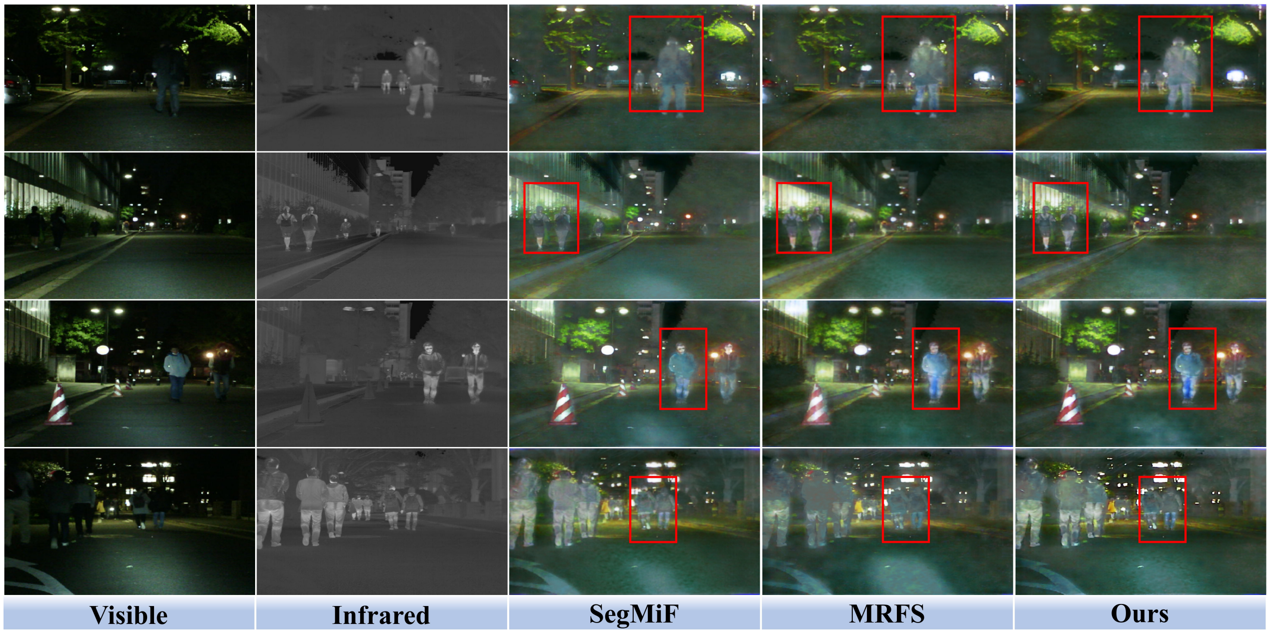

Figure 6. Visualization results of image fusion on MFNet dataset. The red box shows areas with obvious differences.

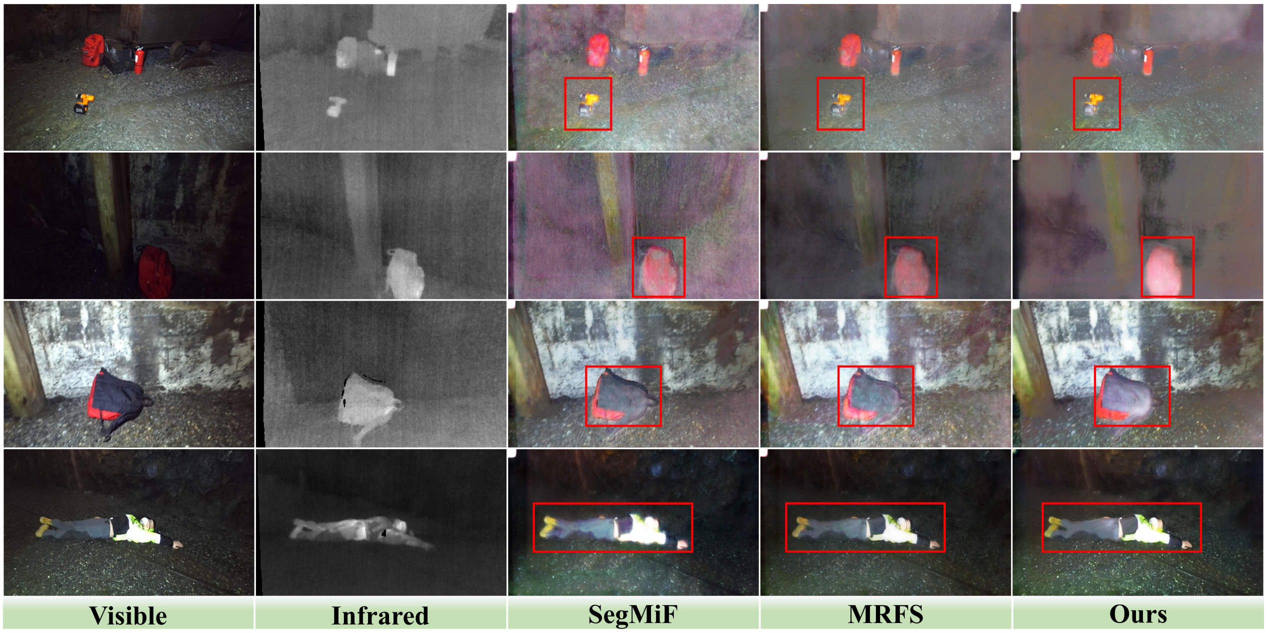

Figure 7. Visualization results of image fusion on PST900 dataset. The red box shows areas with obvious differences.