# Equivariant Multi-Modality Image Fusion

## SUPPLEMENTARY MATERIALS

Zixiang Zhao[1,2] Haowen Bai[1] Jiangshe Zhang[1] Yulun Zhang[3*] Kai Zhang[4]
Shuang Xu[5] Dongdong Chen[6*] Radu Timofte[2,7] Luc Van Gool[2,8]
[1]Xi'an Jiaotong University [2]ETH Zürich [3]Shanghai Jiao Tong University
[4]Nanjing University [5]Northwestern Polytechnical University [6]Heriot-Watt University
[7]University of Würzburg [8]INSAIT
zixiangzhao@stu.xjtu.edu.cn

## Abstract

*In this document, we provide additional supplementary information for the paper "Equivariant Multi-Modality Image Fusion". This file contains:*
*(I) The detail architecture for Restormer Block in U-Fuser module in Sec. 3.3.*
*(II) The specific network details for Pseudo sensing module in Sec. 3.3.*
*(III) Detailed illustration to the training&testing datasets in Sec. 4.1.*
*(IV) Detailed introduction for the selection of hyperparameters in EMMA.*
*(V) More qualitative comparison fusion results in Sec. 4.1 and Sec. 4.3.*
*(VI) Qualitative results for downstream infrared-visible applications in Sec. 4.2.*

## S-1. Detailed introduction for U-Fuser

In Sec. 3.3, the detailed architecture for the Restormer block [6] in *U-Fuser* $\mathcal{F}(\cdot,\cdot)$ is illustrated in Fig. S-1.

## S-2. Specific network details for Pseudo sensing module

In Sec. 3.3, the specific network details for U-Net-based [3] *Pseudo sensing module* $\mathcal{A}_i(\cdot)$ or $\mathcal{A}_v(\cdot)$ are in Tab. S-1. Downsampling and upsampling branches are represented by *DownConv* and *UpConv*, respectively.

## S-3. Detailed introduction to datasets

We adopt widely-used benchmarks MSRS [4], RoadScene [5], and M$^3$FD [2] for *Infrared-Visible image Fusion* (IVF), MSRS [4] and M$^3$FD [2] for *Multi-Modality Object Detection* (MMOD) and *Multi-Modality Semantic Segmentation* (MMSS), as well as Harvard Medical Image Dataset [1] for *Medical Image Fusion* (MIF), respectively.

- MSRS dataset[1]: 1083 pairs for IVF/MMSS training and 361 pairs for IVF/MMSS testing.
- RoadScene dataset[2]: 50 pairs for IVF validation and 50 pairs for IVF testing.
- M$^3$FD dataset[3]: 3360 pairs for MMOD training, 420 pairs for MMOD validation and 420 pairs for MMOD testing.
- Harvard Medical Image dataset[4]: 50 pairs for MIF testing.

*Corresponding authors.
[1]https://github.com/Linfeng-Tang/MSRS
[2]https://github.com/hanna-xu/RoadScene
[3]https://github.com/JinyuanLiu-CV/TarDAL
[4]http://www.med.harvard.edu/AANLIB/home.html

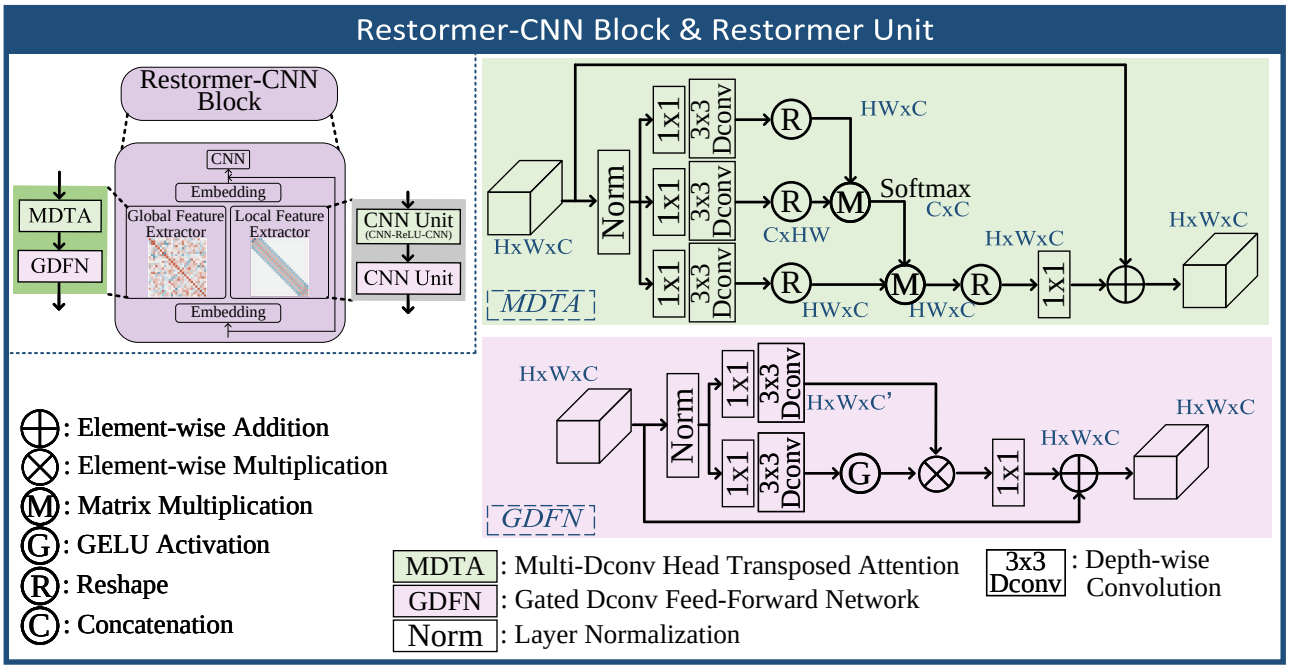


Figure S-1. Detail architecture for the Restormer-CNN block and the Restormer unit [6] in *U-Fuser* of EMMA.

| layer name | output size | detail architecture | layer name | output size | detail architecture |
|---|---|---|---|---|---|
| Input | 256×256 | $\begin{bmatrix} 3\times3, 32 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ | Output | 256×256 | $\begin{bmatrix} 3\times3, 32 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \begin{bmatrix} 3\times3, 1 \\ \text{BN} \\ \text{Sigmoid} \end{bmatrix}$ |
| DownConv1 | 256×256 | $\begin{bmatrix} 3\times3, 32 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$, Max-pooling | UpConv1 | 256×256 | Deconv, $\begin{bmatrix} 3\times3, 32 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ |
| DownConv2 | 128×128 | $\begin{bmatrix} 3\times3, 64 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$, Max-pooling | UpConv2 | 128×128 | Deconv, $\begin{bmatrix} 3\times3, 64 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ |
| DownConv3 | 64×64 | $\begin{bmatrix} 3\times3, 128 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$, Max-pooling | UpConv3 | 64×64 | Deconv, $\begin{bmatrix} 3\times3, 128 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ |
| DownConv4 | 32×32 | $\begin{bmatrix} 3\times3, 256 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$, Max-pooling | UpConv4 | 32×32 | Deconv, $\begin{bmatrix} 3\times3, 256 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ |
| DownConv5 | 16×16 | $\begin{bmatrix} 3\times3, 512 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$, Max-pooling | UpConv5 | 16×16 | Deconv, $\begin{bmatrix} 3\times3, 512 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ |
| Conv6 | 8×8 | $\begin{bmatrix} 3\times3, 1024 \\ \text{BN} \\ \text{LeakyReLU} \end{bmatrix} \times 2$ | | | |

Table S-1. The specific network details for *Pseudo sensing module* $\mathcal{A}_i(\cdot)$ or $\mathcal{A}_v(\cdot)$ of EMMA.

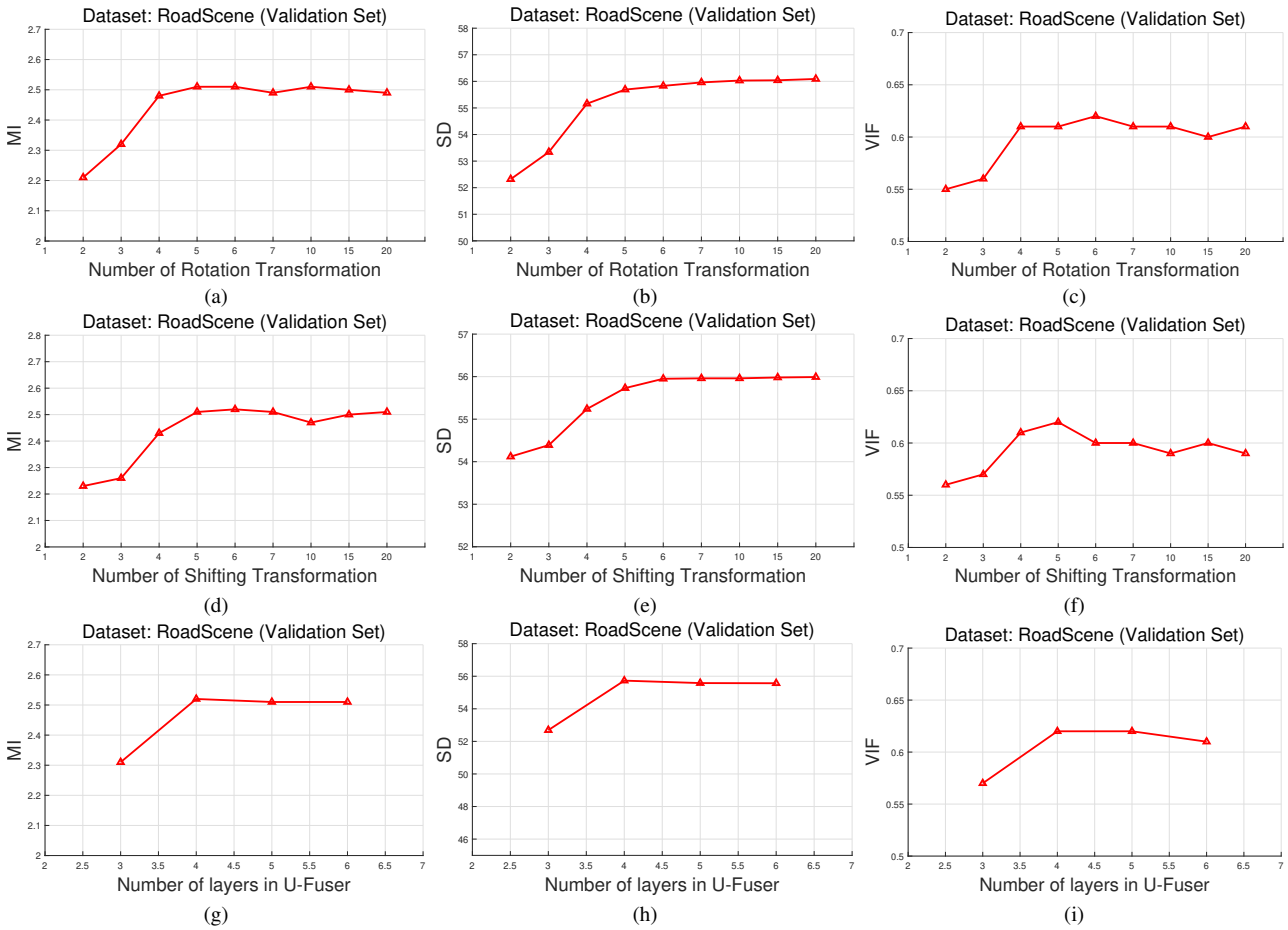


Figure S-2. Fusion results for different configurations of EMMA on the validation set. (a)-(c): The quality of fused images corresponding to different numbers of rotation transformations $G_R = 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 10,\ 15,\ 20$. (d)-(f): The quality of fused images corresponding to different numbers of shift transformations $G_S = 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 10,\ 15,\ 20$. (g)-(i): Results for different layer numbers in U-Fuser: $P = 3,\ 4,\ 5,\ 6$.

## S-4. Selection for the hyperparameters in EMMA

In Section 3.3, we claimed that by applying a series of transformations in the set $\mathcal{G}$, which includes shifts, rotations, and reflections, we can transform $\boldsymbol{f}$ into $\boldsymbol{f}_t$ to utilize imaging prior knowledge that the imaging responses are equivariant to the above-mentioned transformations. Here, we investigate the appropriate number of transformations for the EMMA paradigm. By varying the number of shift transformations $G_S$ and rotation transformations $G_R$ in $\mathcal{G}$ via grid search, we explore the impact of transformation numbers on the fusion result. Note that we determine the number of reflection transformations to be 3, as reflection operations only include horizontal flipping, vertical flipping, and horizontal & vertical flipping.

We record the fusion results corresponding to different $G_R$ and $G_S$ on the validation set in Figs. S-2a to S-2c and Figs. S-2d to S-2f, respectively. The metrics MI, SD and VIF are employed to determine the hyperparameters. The validation set contains 50 image pairs from RoadScene [5] datasets.

When $G_S < 5, G_R < 6$, the performance of the model capability is restricted. However, when $G_S > 5, G_R > 6$, further increasing the number of transformations does not significantly improve the fusion effect. Instead, it leads to more computational load and memory consumption. Therefore, we set $G_S = 5, G_R = 6$ for the number of equivariant transformations when training EMMA.

In addition, the number of layers $P$ in U-Fuser, which is the number of Restormer-CNN blocks in downsampling or upsampling branches, also affects the ability of feature extraction and information fusion. Therefore, we have also shown the results for $P = 3, 4, 5, 6$ in Figs. S-2g to S-2i. It is apparent that U-Fuser with too shallow or too deep architecture will reduce

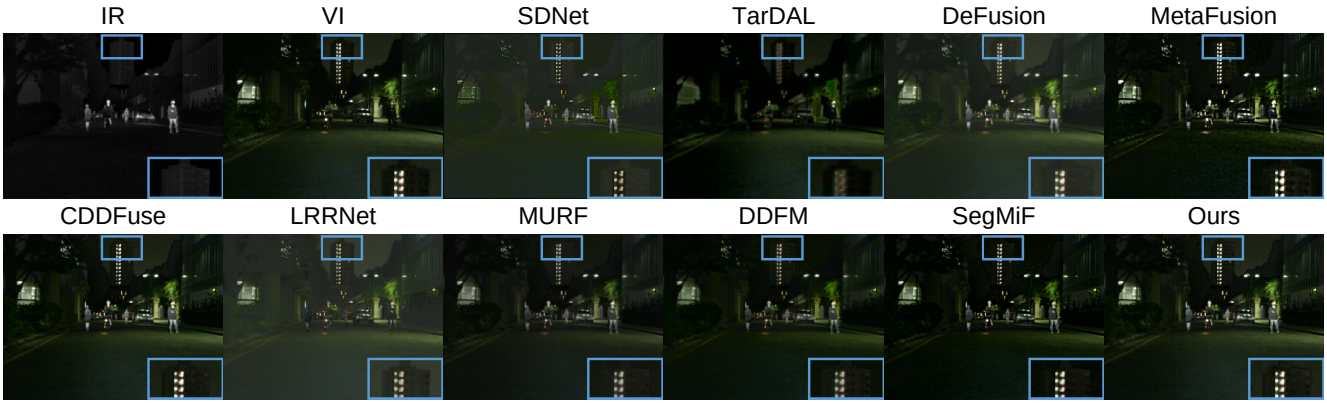


Figure S-3. Visual comparison for *Infrared-Visible image Fusion*.


Figure S-4. Visual comparison for *Infrared-Visible image Fusion*.

the effectiveness of the model, and $P = 4$ is a relatively optimal choice.

Finally, to have a good balance of model performance and computational cost, we set $\{G_R = 6, G_S = 5, P = 4\}$ for the following experiments.

## S-5. More qualitative comparison fusion results

More qualitative comparisons for *Infrared-Visible image Fusion* results are displayed in Figs. S-3 to S-6. Our method better integrates thermal radiation information in infrared images and detailed textures in visible images. Objects in dark regions are clearly highlighted, so that foreground targets can be easily distinguished from the background. Additionally, background details that are difficult to identify due to the low illumination have clear edges and abundant contour information, which help us understand the scene better.

More qualitative comparisons for *Medical Image Fusion* results are shown in Figs. S-7 and S-8. Our EMMA can better preserve the detailed texture and highlight the structure information than other methods.

## S-6. Qualitative results for Downstream Infrared-Visible applications

The qualitative results for infrared-visible object detection and semantic segmentation are exhibited in Figs. S-9 and S-10 as well as Figs. S-11 and S-12, respectively. In object detection, EMMA can improve detection accuracy by fusing thermal radiation information and highlighting the difficult-to-observe objects. Therefore, small objects can be better detected.

For the segmentation task, EMMA better integrates the edge and contour information in the source images, which enhances the ability of our model to perceive the object boundary, and makes the segmentation more accurate. Therefore, EMMA can capture segmentation details in certain small regions that cannot be obtained by other methods.

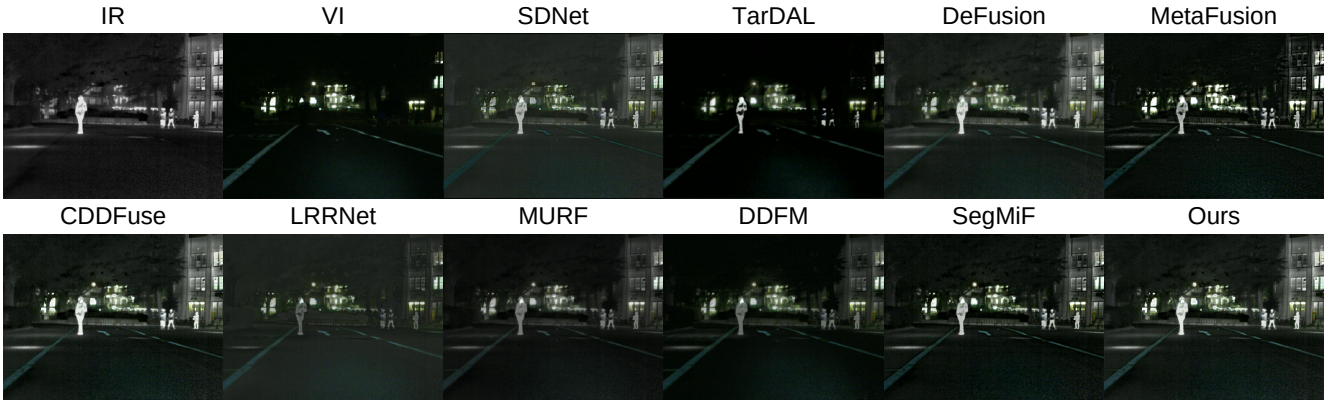


Figure S-5. Visual comparison for *Infrared-Visible image Fusion*.

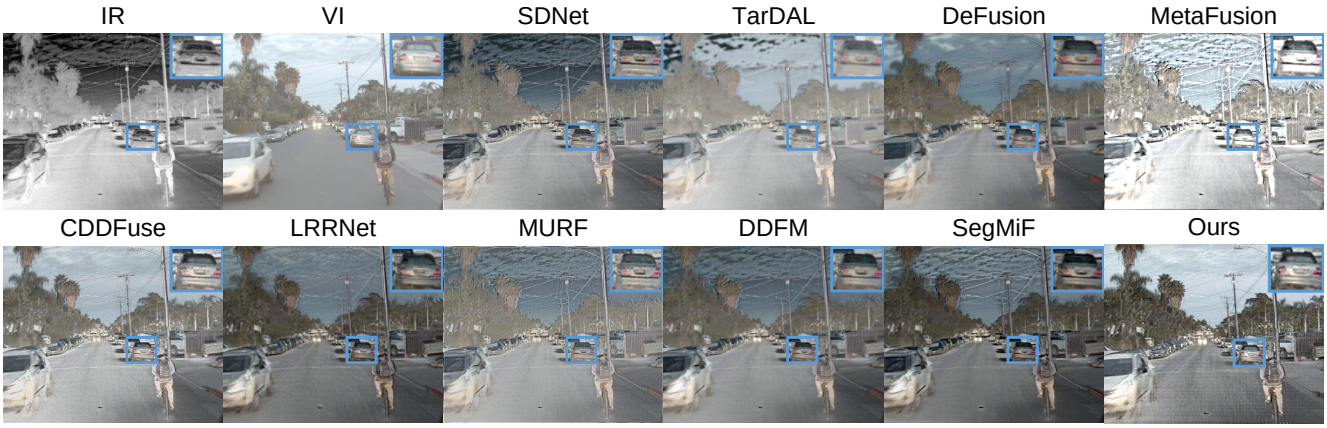


Figure S-6. Visual comparison for *Infrared-Visible image Fusion*.

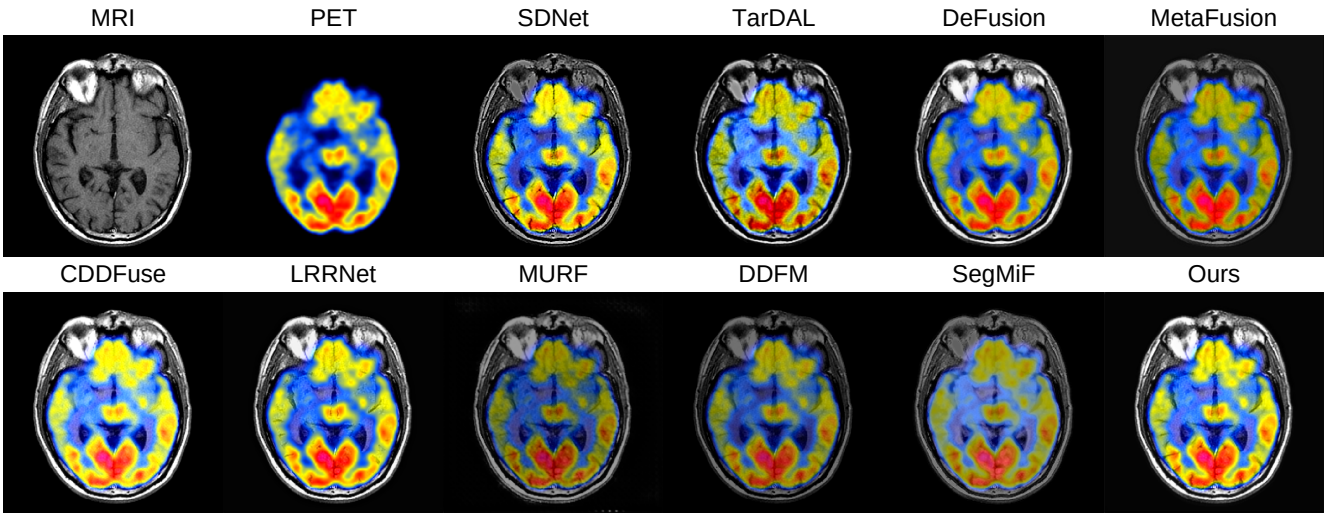


Figure S-7. Visual comparison for *Medical Image Fusion*.

## References

[1] Harvard Medical website. http://www.med.harvard.edu/AANLIB/home.html. 1

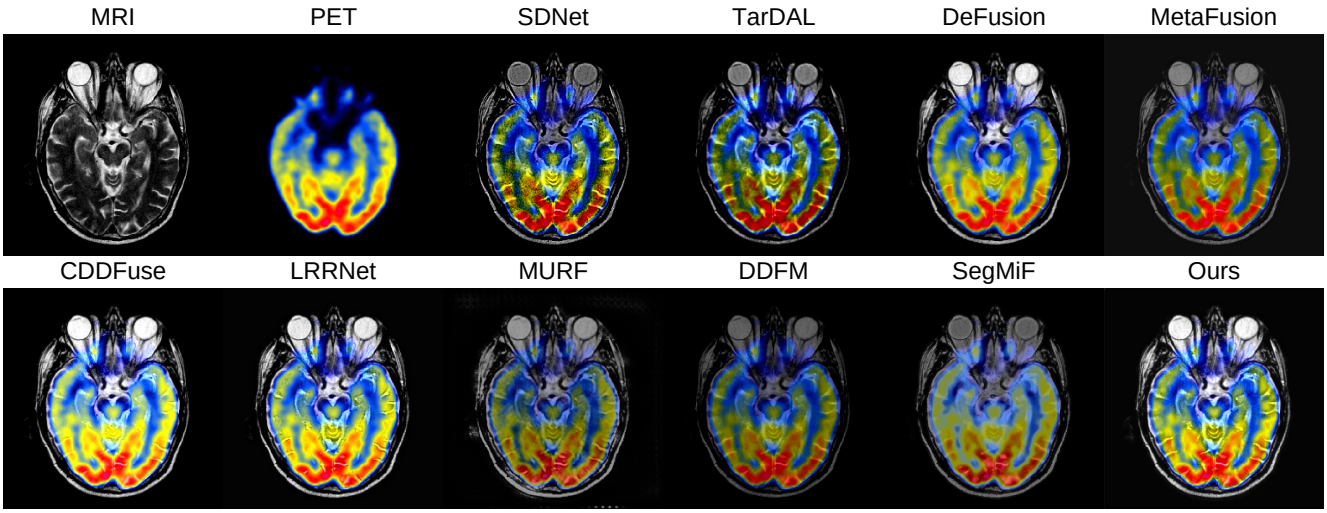


Figure S-8. Visual comparison for *Medical Image Fusion*.

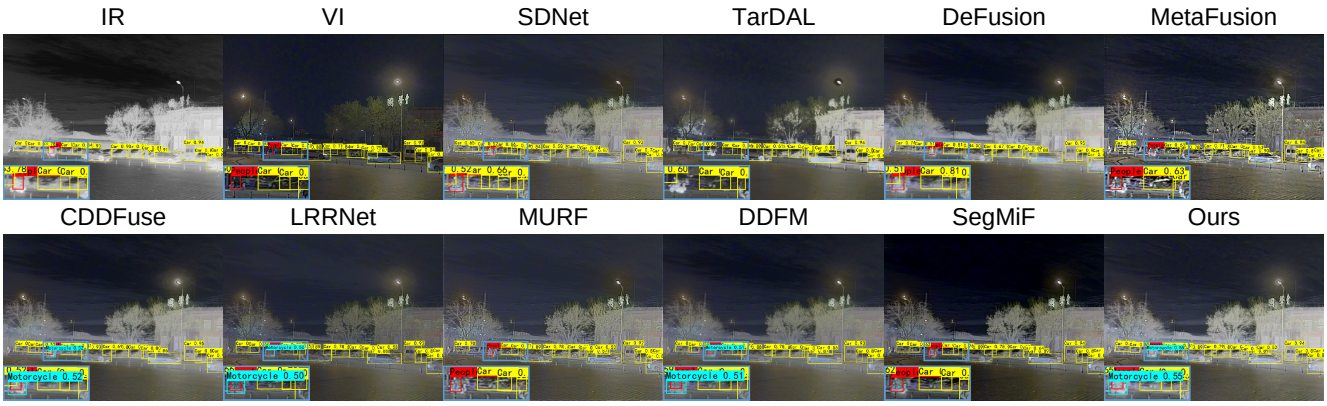


Figure S-9. Qualitative results for infrared-visible object detection on $M^3FD$ dataset.

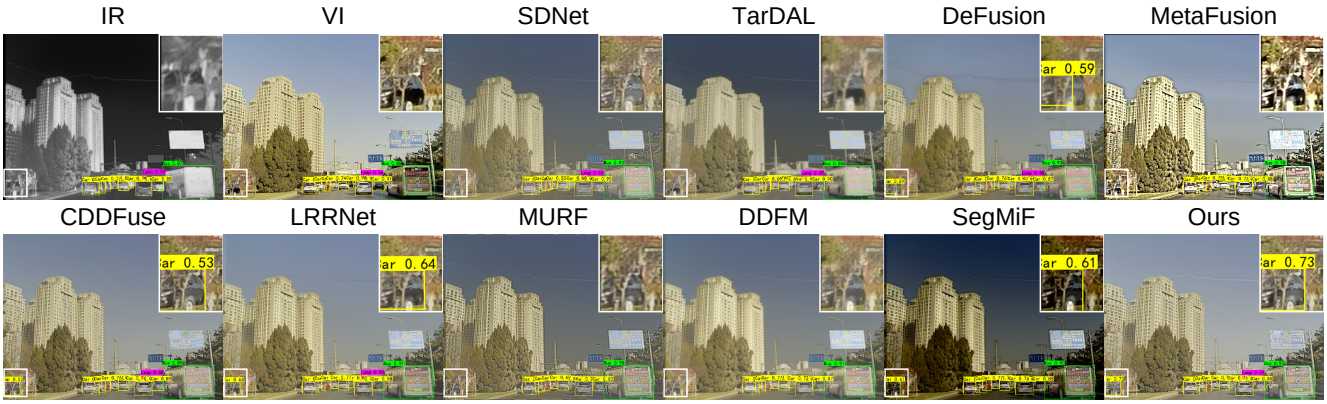


Figure S-10. Qualitative results for infrared-visible object detection on $M^3FD$ dataset.

[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 *CVPR*, pages 5792–5801. IEEE, 2022. 1

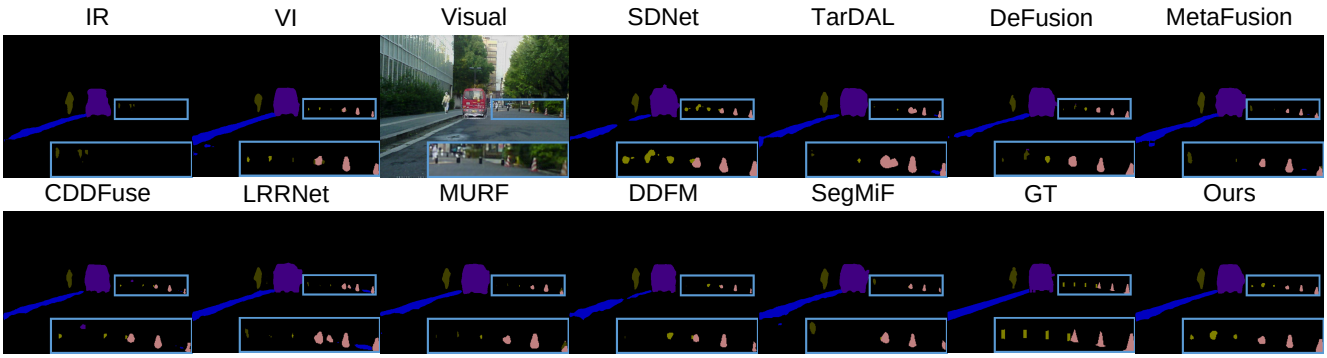


Figure S-11. Qualitative results for infrared-visible semantic segmentation on MSRS dataset.

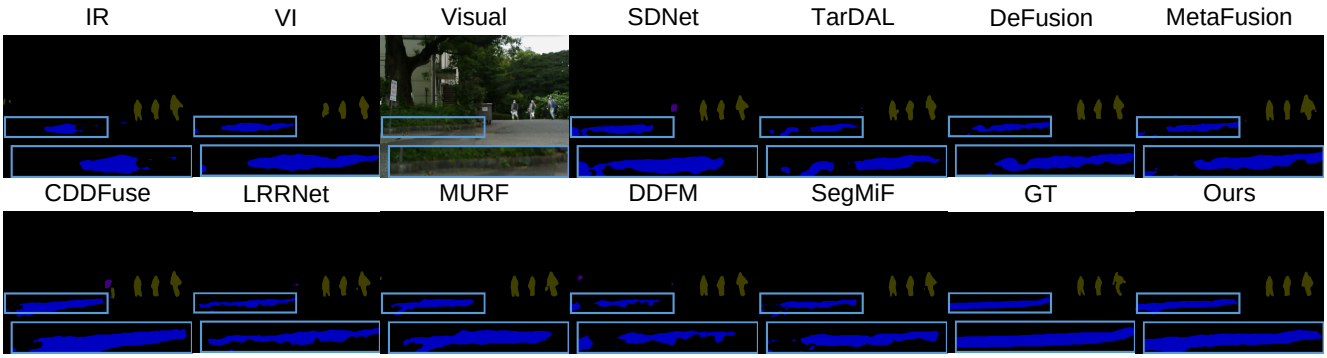


Figure S-12. Qualitative results for infrared-visible semantic segmentation on MSRS dataset.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[4] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. 1

[5] Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *AAAI Conference on Artificial Intelligence, AAAI*, pages 12484–12491, 2020. 1, 3

[6] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729. IEEE, 2022. 1, 2