Supplementary Material for LUT-Fuse: Towards Extremely Fast Infrared and Visible Image Fusion via Distillation to Learnable Look-Up Tables

Xunpeng Yi[1,*], Yibing Zhang[1,*], Xinyu Xiang[1], Qinglong Yan[1], Han Xu[2], Jiayi Ma[1,†]
[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Automation, Southeast University, Nanjing 210096, China
{yixunpeng, zhangyibing, xiangxinyu, qinglong_yan}@whu.edu.cn,
xu_han@seu.edu.cn, jyma2010@gmail.com

1. MM-Net

Inspired by the Text-IF [1], we introduce a Multi-Modal Network (MM-Net) for infrared and visible image fusion. The network comprises two parallel encoders with self-attention, a cross attention module for cross-modal feature interaction, and a decoder to reconstruct the fused image. In the following, we provide a detailed description of the architecture.

Parallel Encoders with Self-Attention. To capture salient features from both modalities, we first extract feature maps from the infrared image I_{ir} and the visible image I_{vis} using two parallel encoders:

$$F_{ir} = E_{ir}(I_{ir}), \tag{1}$$

$$F_{vis} = E_{vis}(I_{vis}), \tag{2}$$

where E_{ir} and E_{vis} denote the encoder networks for the infrared and visible images, respectively. Then, a self-attention mechanism is employed to capture long-range dependencies within the feature maps. For an input feature map F, the self-attention operation is defined as:

$$\tilde{F} = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

where Q, K, and V are the query, key, and value matrices obtained from F through linear projections, and d_k is the dimensionality of the key vectors. This process enriches the extracted features by emphasizing important regions and suppressing irrelevant information.

Cross Attention for Cross-Modal Feature Interaction Following the self-attention mechanism, the refined feature maps $\tilde{F}_{ir}$ and $\tilde{F}_{vis}$ are further processed using a cross attention mechanism to enable effective cross-modal interaction. The cross attention is computed as:

$$F'_{ir} = softmax\left(\frac{Q_{vis}K_{ir}^T}{\sqrt{d_k}}\right)V_{ir}, \tag{4}$$

$$F'_{vis} = softmax\left(\frac{Q_{ir}K_{vis}^T}{\sqrt{d_k}}\right)V_{vis}. \tag{5}$$

Feature Decoding. The features F'_{ir} and F'_{vis} are concatenated along the channel dimension, and form a comprehensive feature representation:

$$F_{fuse} = Concat(F'_{ir}, F'_{vis}). \tag{6}$$

The fused feature map F_{fuse} is then passed through a decoder network $\mathcal{D}$ to reconstruct the final fused image:

$$I_{fuse} = \mathcal{D}(F_{fuse}). \tag{7}$$

The decoder is designed to progressively integrate multi-scale features and ultimately output a fused image with a collection of rich textures and salient targets.

2. More Experiments

As the length of the main paper is limited, we present more visual comparisons here to show advantages of LUT-Fuse.

The results on the LLVIP are presented in Figs. r1-r3. Our LUT-Fuse demonstrates three significant advantages, attributed to the LUT architecture specifically designed for multi-modal fusion tasks. First, our method effectively highlights thermal targets. As shown in the three sets of results, the pixel intensity of thermal targets in our outputs is the highest, indicating that they are the most prominent. Second, our results show better brightness and finer detail, in the second and third sets, revealing more scene content with greater clarity. Last, our approach produces more vivid and natural colors, leading to perceptually superior results.

References

[1] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024. 1

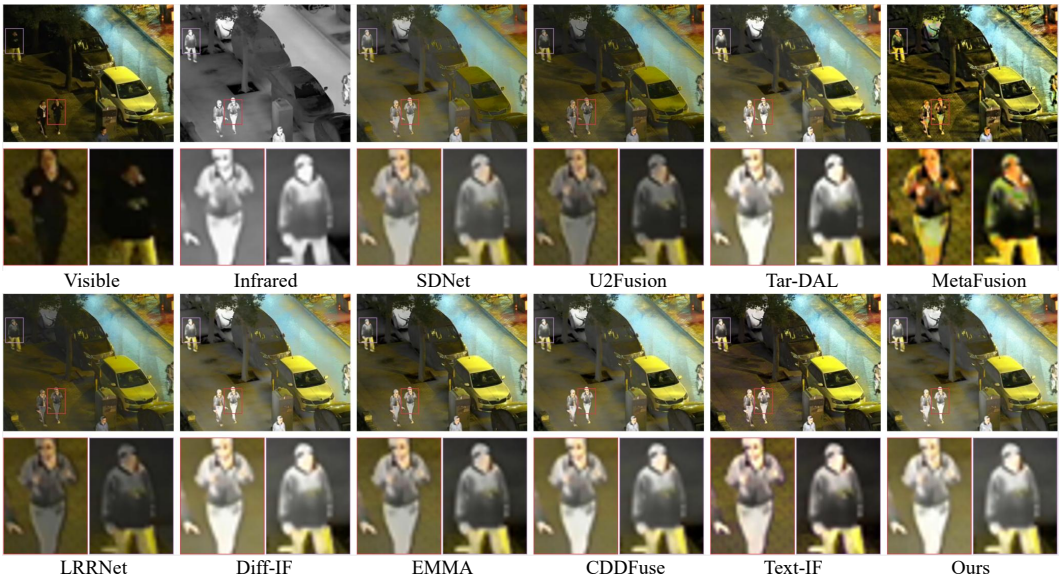

Figure r1. Qualitative comparison of our proposed LUT-Fuse with the state-of-the-art methods

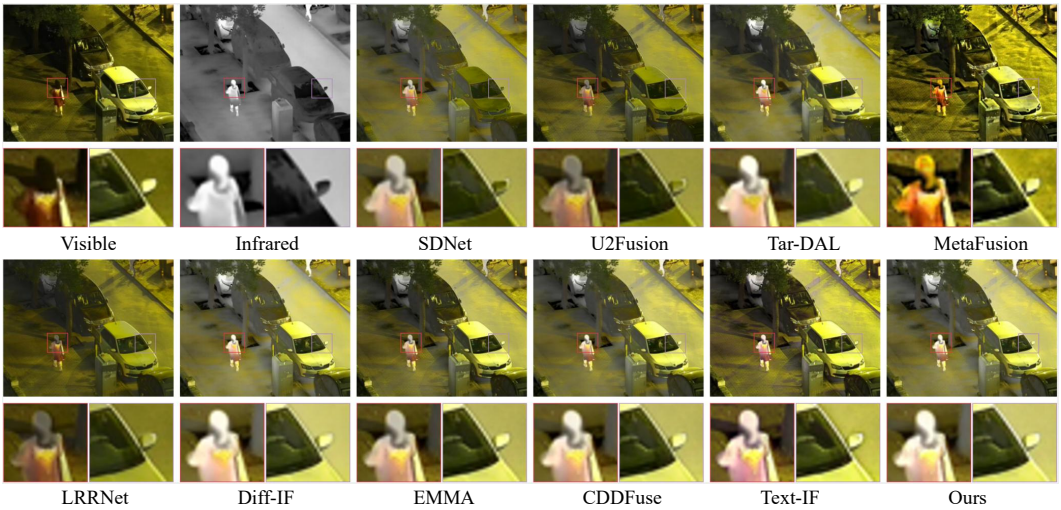

Figure r2. Qualitative comparison of our proposed LUT-Fuse with the state-of-the-art methods

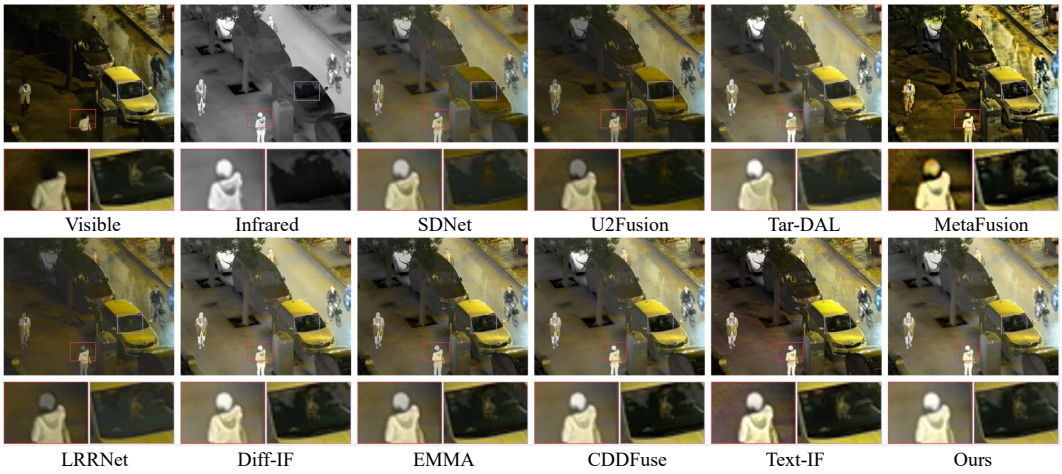

Figure r3. Qualitative comparison of our proposed LUT-Fuse with the state-of-the-art methods