One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion

Supplementary Material

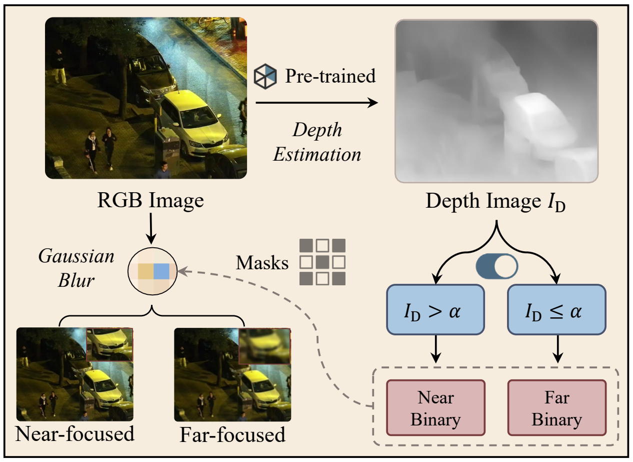

Figure 1. An illustration of the proposed data augmentation process.

1. Anonymous Code Repository

The anonymised code repository for this work is publicly available at https://anonymous.4open.science/r/GIFNet-821F.

2. Experimental Environment

During training, the batch size is set to 8, and the GIFNet model is optimised using the Adam optimiser [3] with a learning rate of 10^{-3}. All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. The results for all competitors are derived from their official implementations to ensure consistency and reproducibility.

3. Data Augmentation

This section details the process of constructing a joint dataset, based on the IVIF task, for cross-task interaction between IVIF-MFIF and IVIF-MEIF tasks.

3.1. Multi-Focus Image Fusion (MFIF)

We propose a data augmentation strategy to generate Multi-Focus Image Fusion (MFIF) data from the LLVIP Infrared and Visible Image Fusion (IVIF) dataset [2] (see Fig. 1).

MFIF images typically feature regions that appear blurred or clear depending on their depth of field. Since existing fusion datasets do not include depth maps, we utilise a pre-trained single-view depth estimation model [4] to derive a depth map I_D from the corresponding RGB image I_{vis}. For each pixel (x, y), the depth value $I_D(x, y)$ is normalised to the range $[0, 1]$. Next, a random threshold $\alpha \in (0, 1)$ is applied to segment the depth image into near-focused and far-focused regions. These regions are represented by binary masks B_n and B_f, respectively, defined as:

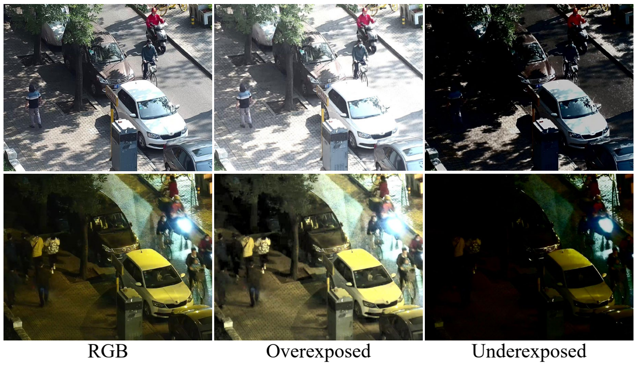

Figure 2. Examples of artificially generated data for the joint training of multi-exposure image fusion.

$$B_n(x, y) = \begin{cases} 1, & I_d(x, y) > \alpha \\ 0, & I_d(x, y) \leq \alpha \end{cases}, \tag{1}$$

$$B_f(x, y) = 1 - B_n(x, y). \tag{2}$$

Using these binary masks, we simulate focus variation in RGB images. For the near-focused image I_n, we introduce blurring as follows:

$$I_n(x, y) = \begin{cases} \text{gBlur}(I_{vis(x,y)}), & B_n(x, y) = 0 \\ I_{vis}(x, y) \quad , & B_n(x, y) = 1 \end{cases}, \tag{3}$$

where gBlur represents Gaussian Blur. Similarly, we can generate the far-focused image I_f based on the far-binary map B_f.

3.2. Multi-exposure Image Fusion

We further describe the creation of a joint IVIF-MEIF dataset to support the training of our GIFNet framework. Following the approach used for MFIF, we process RGB images from the LLVIP dataset by altering their exposure levels to produce overexposed and underexposed images.

In FusionBooster [1], the authors provide a concise information probe model for decomposing different fusion results, with the subsequent booster layer being used to enhance the separate components. Thus, we take advantage of this information probe in this approach to produce the overexposed and underexposed components of the RGB images.

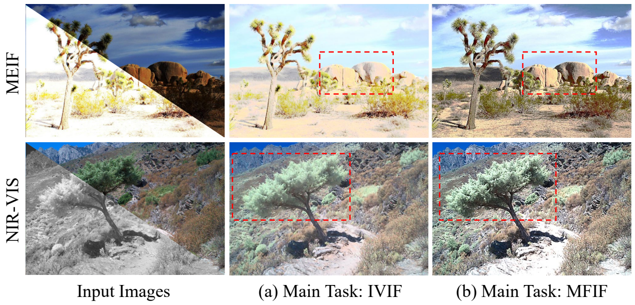

Figure 3. Impact of the main task selection in the inference phase.

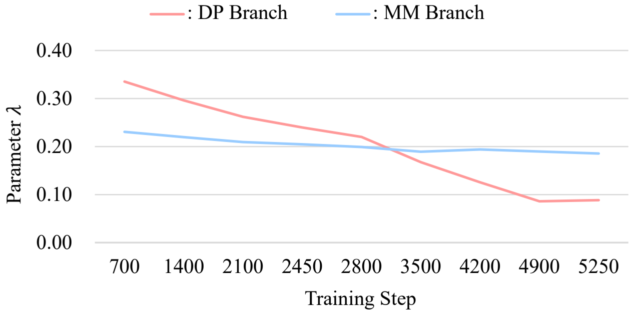

Figure 4. Visualisation of the controllable factor λ in the cross-fusion gating mechanism (DP: Digital Photography, MM: Multi-Modal).

As illustrated in Fig. 2, the pre-trained model successfully adjusts the exposure levels of the input images, producing high-quality artificial datasets that align with the anticipated exposure variations. These datasets are crucial for harmonising training across IVIF-MFIF and IVIF-MEIF tasks, reducing domain gaps and enabling consistent feature extraction.

4. Main Task Selection for the Inference Phase

In our method, we alternatively select the main task and auxiliary task during training, but this configuration must be fixed during the inference phase. We conduct experiments regarding this selection on two unseen fusion tasks: MEIF and NIR-VIS. As shown in Fig. 3a, when using IVIF as the main task, GIFNet struggles to control exposure settings in poorly exposed conditions. Although the task-specific features in this branch facilitate the preservation of significant information, it fails to maintain a natural visual effect (as in the NIR-VIS task). In contrast, when MFIF is designated as the main task (Fig. 3b), GIFNet produces visually robust fused images with appropriate exposure settings and effective utilisation of both modalities, exhibiting superior generalisation ability. Therefore, in the test stage, we choose to consider MFIF as the main task,

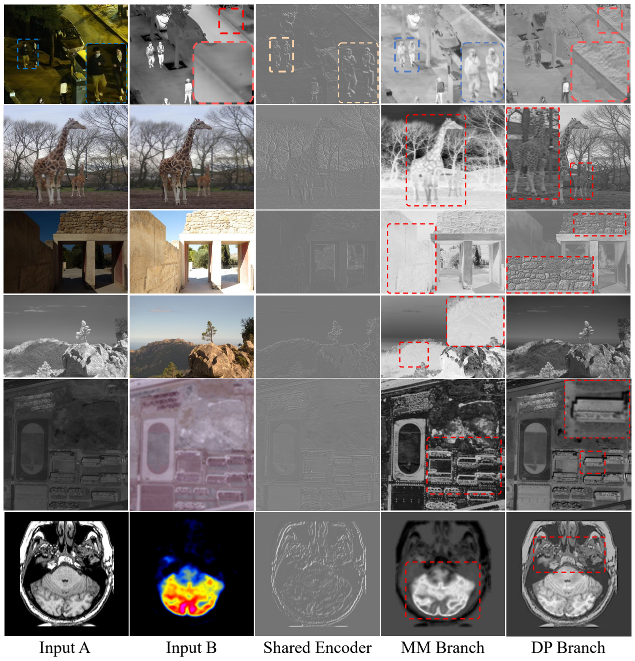

Figure 5. Visualisations of the feature maps from the shared-encoder and the two branches on various image fusion tasks.

5. An Analysis of the Cross-Task Interaction

During the training process, a cross-fusion gating mechanism is used to iteratively optimise the multi-modal and digital photography branches. In this section, we visualise the learnable parameter λ of these two branches to provide an intuitive understanding of the cross-task interaction process. As shown in Fig. 4, the controllable factors from these branches converge to stable values: approximately 0.08 for the DP branch and 0.18 for the MM branch. This imbalance can be attributed to the fact that the DP task has ground truth images for supervised training, thus requiring fewer auxiliary features from the other unsupervised trained branch. Additionally, the curve with a decreasing trend indicates that the dependency among different branches lessens as the training process proceeds.

6. Feature Visualisation

We present visualisations of the feature maps from different components: the shared encoder (S-Enc), the MM branch, and the DP branch, as shown in Fig. 5. The S-Enc, driven by the image reconstruction objective, captures foundational image features, such as target contours and structural details, which are essential for high-quality image fusion.

The MM and DP branch visualisations reveal the distinct contributions of each branch to the fusion process. For instance, in the first case, MM features focus on preserving salient information from the source inputs, such as ther-

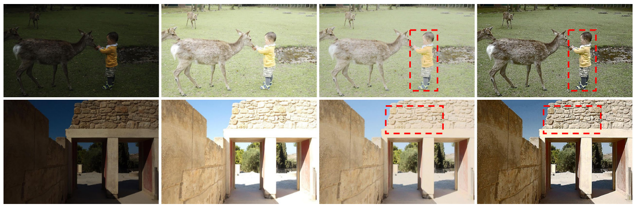
Underexposed Overexposed w/o DA w/ DA

Figure 6. Qualitative results of the ablation study regarding the Data Augmentation (DA) technique on the Multi-Exposure Image Fusion (MEIF) task.

Case	Training Data	EI	VIF	SCD	AG
w/o DA	LLVIP+MFI-WHU	58.64	1.17	0.51	6.22
GIFNet (w/ DA)	LLVIP*	111.27	2.52	1.04	12.00

Table 1. Quantitative results from the ablation study of the data augmentation technique on the unseen MEIF task.

mal targets. Meanwhile, DP features enhance finer details, capturing sharper edges and more defined textures, as well as clearer shadows on the ground. Similar patterns are observed across other seen and unseen fusion tasks. Notably, the additional learning of digital photography features consistently benefits various fusion tasks by producing the necessary features for visually robust outputs, as seen in the third example (MEIF task) where enhanced texture details are prominent.

7. Ablation Experiments about the Data Augmentation

This section presents ablation experiments to assess the efficacy of the proposed data augmentation strategy. Specifically, we examine the performance of the GIFNet model when trained without augmented data. Instead, we utilise the widely adopted MFI-WHU dataset [5] to construct the training set for the MFIF task. Since the shared RGB modality between the MM and DP tasks is absent, we adapt the public loss function of the DP branch by incorporating the ground-truth images from MFI-WHU. The modified loss function is defined as follows:

$$\mathcal{L}_{\text{pub}} = \mathcal{L}_{\text{ssim}}(I_r, I_{\text{gt}}) + \mathcal{L}_{\text{mse}}(I_r, I_{\text{gt}}), \tag{4}$$

where I_{gt} denotes the fully-focused image, I_r is the output of the reconstruction (REC) branch.

To evaluate the impact of data augmentation, we employ the multi-exposure image fusion (MEIF) task, examining its role in enhancing GIFNet's generalisation capabilities. As illustrated in Fig. 6, the absence of the proposed data augmentation process led to GIFNet's inability to adequately regulate the brightness levels in fused images, resulting in a noticeable decline in visual quality.

The quantitative results in Table 1 align with these observations, confirming that GIFNet achieves superior performance when employing the augmentation technique. The substantial differences observed across experiments highlight the pivotal role of shared modalities in minimising domain gaps between fusion tasks. This demonstrates the significance of the cross-task interaction paradigm and its reliance on common modalities to achieve consistent performance improvements.

Table 2. Quantitative results of other methods with code and model available for the IVIF task. (†: Retraining required)

Method	Venue	Agnostic	Tasks	EI	AG	SCD	VIF	SF
DDFM	ICCV23	×	2	41.13	4.43	1.55	0.51	14.17
SegMIF	CVPR23	×	1	58.33	5.97	1.54	**0.90**	20.61
FBooster	IJCV24	×	3†	56.87	5.71	1.56	0.89	16.99
EMMA	CVPR24	×	2	58.32	5.81	1.55	0.85	18.42
FILM	ICML24	×	4†	58.06	5.96	1.54	0.88	21.69
GIFNet	Ours	✓	**6**	**62.46**	**6.70**	**1.61**	0.73	**25.67**

8. More Comparison with Advanced Methods

In Table 2, we now provide more results and an extra metric of the advanced methods with code and model available. The differences of advanced methods on these metrics are relatively minor. But our low-level task interaction paradigm, supporting more fusion tasks, effectively improves most of these image quality assessments. While the low-level and high-level tasks are complementary, as illustrated in the main page, the reliance on abstract semantics can degrade fusion quality, and the use of single, specific features may constrain generalisation ability.

9. More Qualitative Results on Different Fusion Tasks

Infrared and Visible Image Fusion Task: In this subsection, we present additional results on the IVIF task. As illustrated in Fig. 7, GIFNet demonstrates superior performance compared to other methods by effectively preserving infrared information (as seen in the first example) and retaining rich texture details from the input images (as shown in the second example).

Near Infrared and Visible Image Fusion Task: Fig. 8 showcases additional results of various methods on the NIR-VIS fusion task. Benefiting from the integration of the digital photography image fusion task, GIFNet effectively utilises the near-infrared modality, resulting in enhanced image quality with well-balanced illumination in the background. By contrast, other advanced approaches exhibit varying degrees of insufficient illumination issues, leading to suboptimal fused images.

Medical Image Fusion Task: More comparative results for the medical image fusion task are displayed in Fig. 9. As highlighted in specific regions, GIFNet distinctly outperforms other methods by better preserving enhanced edge

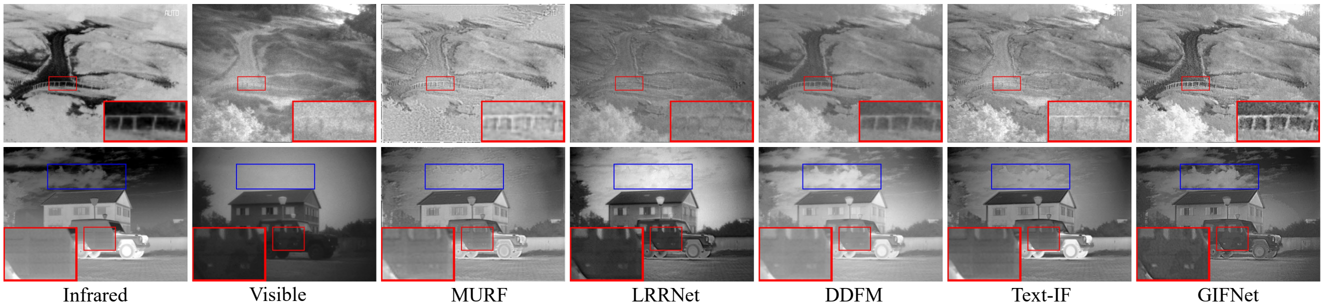

Figure 7. More qualitative results of different methods on the infrared and visible image fusion task. Compared with other methods, our GIFNet effectively preserves the infrared information and texture details from input images, achieving superior fusion quality in various scenarios.

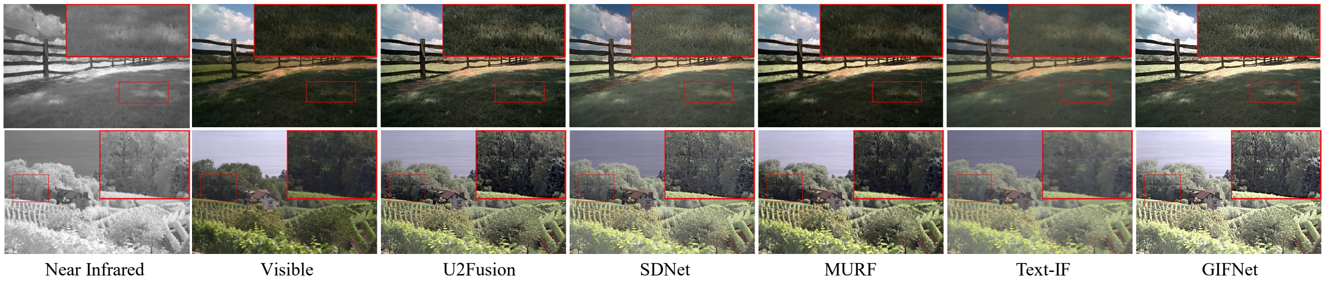

Figure 8. More qualitative results of different methods on the near infrared and visible image fusion task. GIFNet excels in maintaining appropriate illumination and clear background details, outperforming other methods prone to underexposure issues.

information from MRI images, retaining detailed textures of the brain structure. This demonstrates GIFNet's robustness and capability in handling medical imaging requirements.

Remote Task: Additional results on the pansharpening task are presented in Fig. 10. Similar to its performance in previous tasks, GIFNet consistently delivers clearer texture details from the Panchromatic modality (high spatial resolution) while accurately preserving the colour information from the Multispectral modality.

Multi-focus Image Fusion Task: Further results for the MFIF task are depicted in Fig. 11. As indicated by the highlighted regions, GIFNet's ability to achieve superior edge intensity, as reported in Section 4.4, is well reflected in these additional experiments. For instance, the method provides a clearer depiction of letters and sharper edge details in the background, illustrating its capacity for high-quality multi-focus fusion.

Multi-exposure Image Fusion Task: Additional results on the MEIF task are shown in Fig. 12. Compared to other advanced fusion methods, the proposed approach generates fused images with robust and natural exposure settings across diverse environments, including outdoor and indoor scenes. In contrast, other approaches struggle to produce images with clear details and balanced exposure, further highlighting GIFNet's effectiveness in this challenging task.

References

[1] Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, and Josef Kittler. Fusionbooster: A unified image fusion boosting paradigm. *International Journal of Computer Vision*, 2024. 1

[2] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 1

[5] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. *Information Fusion*, 66:40–53, 2021. 3

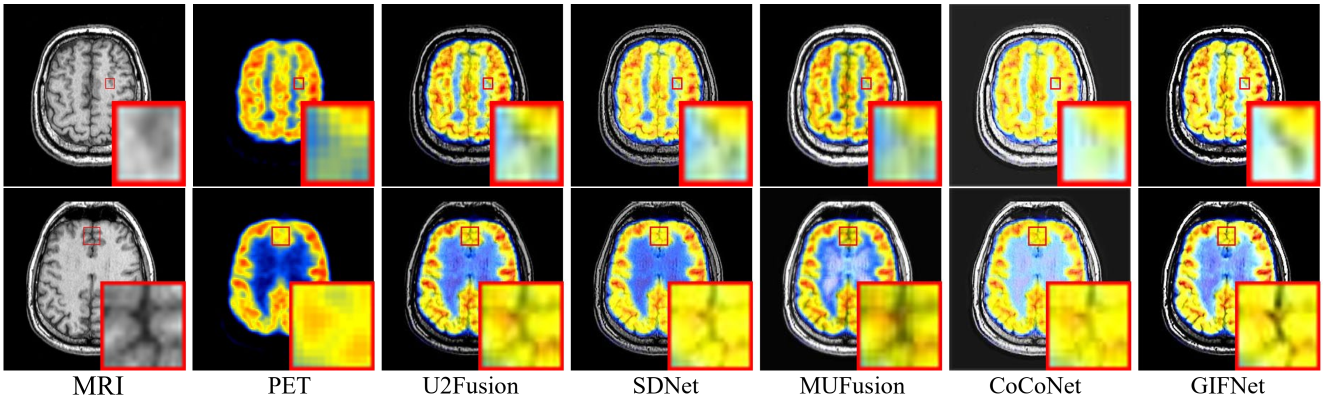

Figure 9. More qualitative results of different methods on the medical image fusion task. GIFNet demonstrates distinct advantages by better preserving enhanced brain structure details in the MRI images.

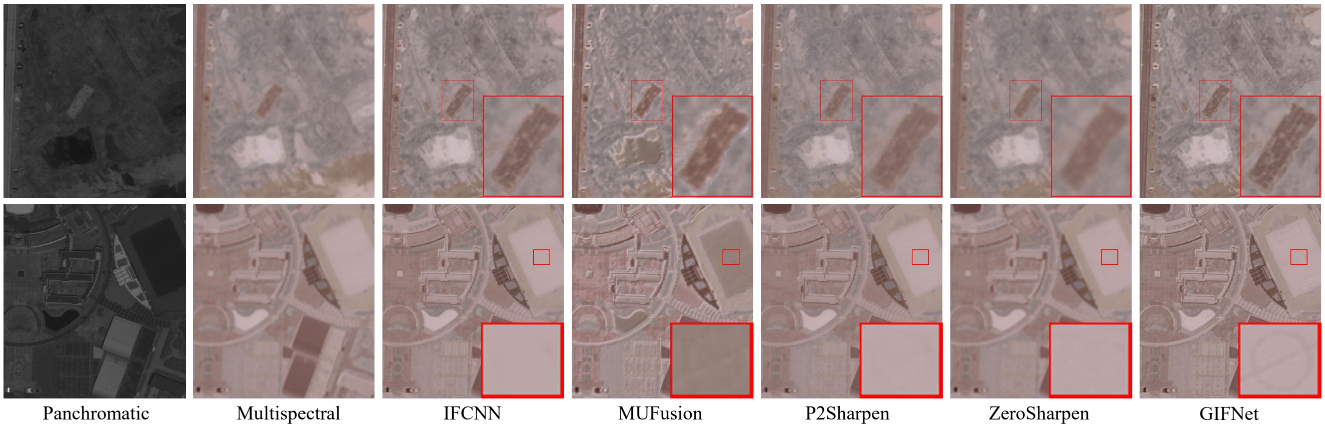

Figure 10. More qualitative results of different methods on the Pansharpening task. GIFNet consistently achieves clearer spatial details from the Panchromatic images and retains accurate colour information from the Multispectral images, ensuring high-quality fused results.

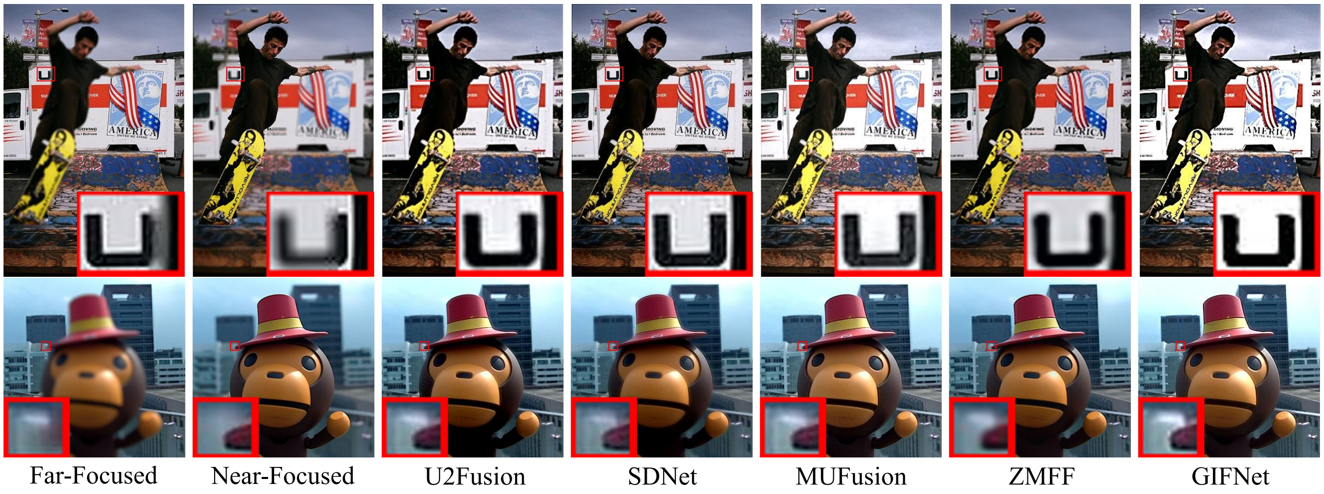

Figure 11. More qualitative results of different methods on the multi-focus image fusion task. As highlighted, GIFNet excels at providing sharper edges and clear visual details in the background regions, showcasing its superior edge intensity preservation.

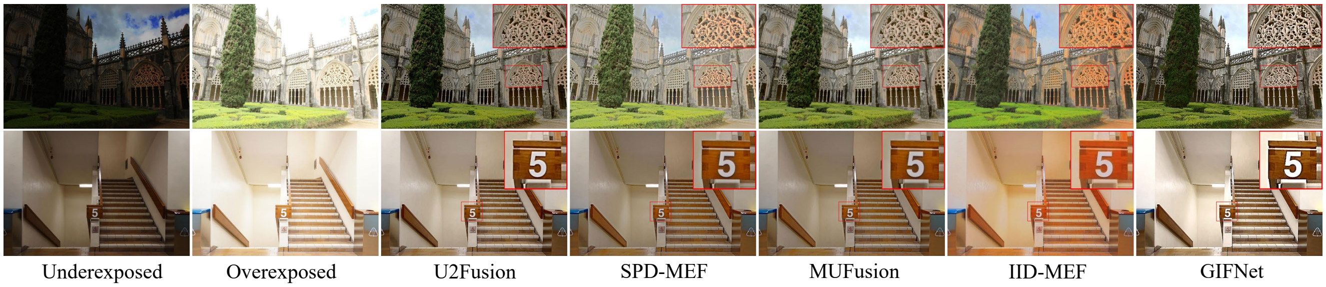

Figure 12. More qualitative results of different methods on the multi-exposure image fusion task. Our GIFNet produces fused images with balanced exposure settings and natural details in diverse environments, outperforming other approaches that struggle to maintain a consistent image quality.