Supplementary Materials for "UniFuse: A Unified All-in-One Framework for Multi-Modal Medical Image Fusion Under Diverse Degradations and Misalignments"

Dayong Su[1], Yafei Zhang[1], Huafeng Li[1]*, Jinxing Li[2], Yu Liu[3]
[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[3]Department of Biomedical Engineering, Hefei University of Technology
dayongsu@outlook.com, {zyfeimail, hfchina99}@163.com,
lijinxing158@hit.edu.cn, yuliu@hfut.edu.cn

1. Further Experimental Comparisons

To comprehensively evaluate the performance of our proposed method, we provide additional visual comparisons of fusion results. Figure 1 presents source images from different datasets with various types of degradation. After processing these images using different fusion methods, the corresponding fusion results are shown in Figure 2. As observed in Figure 2, our method exhibits significant advantages in artifact suppression, noise elimination, feature alignment, and the restoration of contrast and detail. These results further demonstrate the superiority of our method in terms of visual quality.

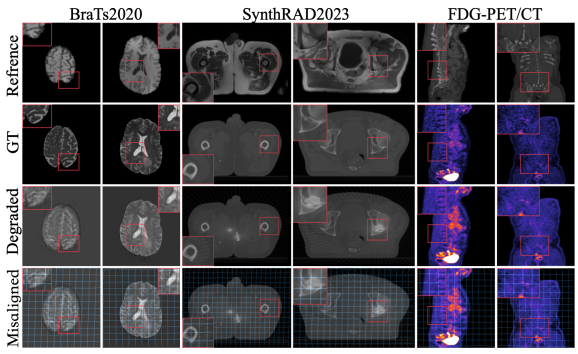

Figure 1. Source images to be fused and corresponding GTs of degraded images. The first row shows the reference images, while the second row presents the GTs of degraded images. The third and fourth rows display misaligned degraded images and their distortions.

*Corresponding author: Huafeng Li (hfchina99@163.com).

2. Evaluation of Alignment Performance

To evaluate the alignment performance of our method, we use the deformation field generated by the UniFuse network for image alignment and compare it with other image alignment methods, including VoxelM [1], TransM [2], and CorrMLP [3]. To quantify the alignment results, we adopt commonly used alignment metrics: mean squared error (Q_{mse}), correlation coefficient (Q_{cc}), and structural similarity (Q_{ssim}). To ensure a fair comparison, all unaligned degraded images are first processed by AMIR for image quality restoration before feeding them into the comparison methods. As evident from Table 1, our method demonstrates significant overall advantages across different datasets. This is primarily attributed to the specially designed degradation-aware prompt module, which ensures robust feature alignment performance when handling various types of data. To visually assess the quality of alignment, we generate alignment error maps by subtracting the aligned images from their corresponding label images. For better visualization, we render the error maps using the color spectrum shown below the figure, where perfectly aligned regions appear white. As clearly shown in the alignment error maps in Figure 3, our method exhibits a significant advantage in feature alignment compared to other methods.

3. Further Ablation Study

Effectiveness of DAPL. To visually analyze the effectiveness of DAPL, we visualize the results under the three settings in the DAPL ablation study, as shown in Figure 4. In Setting A, the fusion results of the network exhibit feature misalignment due to the lack of assistance from DAPL, which reduces the effectiveness of OUFR and subsequently affects the feature alignment performance of FA. In Set-

Table 1. Comparison of alignment performance across different methods and datasets.

BraTs2020 Dataset				SynthRAD2023 Dataset				FDG PET/CT Dataset			
Methods	$Q_{mse}\downarrow$	$Q_{cc}\uparrow$	$Q_{ssim}\uparrow$	Methods	$Q_{mse}\downarrow$	$Q_{cc}\uparrow$	$Q_{ssim}\uparrow$	Methods	$Q_{mse}\downarrow$	$Q_{cc}\uparrow$	$Q_{ssim}\uparrow$
VoxelM	1.49E-3	0.456	0.885	VoxelM	7.02E-4	0.485	0.902	VoxelM	3.01E-4	0.477	0.927
TransM	5.07E-4	0.485	0.961	TransM	7.24E-4	0.484	0.906	TransM	2.92E-4	0.478	0.919
CorrMLP	4.89E-4	0.485	0.962	CorrMLP	5.12E-4	0.490	0.931	CorrMLP	2.42E-4	0.482	**0.968**
Ours	**4.21E-4**	**0.488**	**0.966**	Ours	**3.48E-4**	**0.492**	**0.951**	Ours	**9.52E-5**	**0.493**	0.935

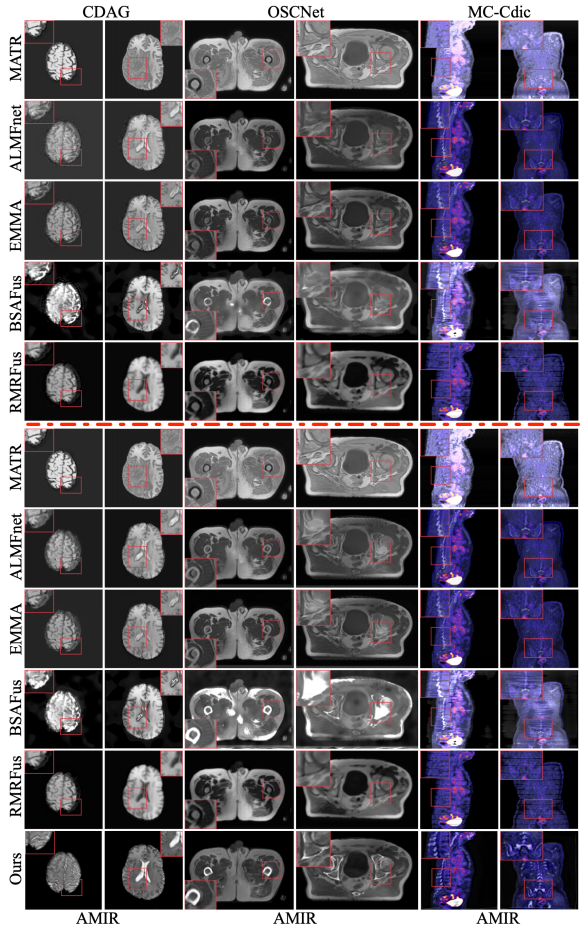

Figure 2. Comparison of fusion results using different methods. Rows 1–11 show results from various fusion approaches. Top and bottom text indicate the image restoration methods used above and below the red line, while left text specifies the fusion method for each row.

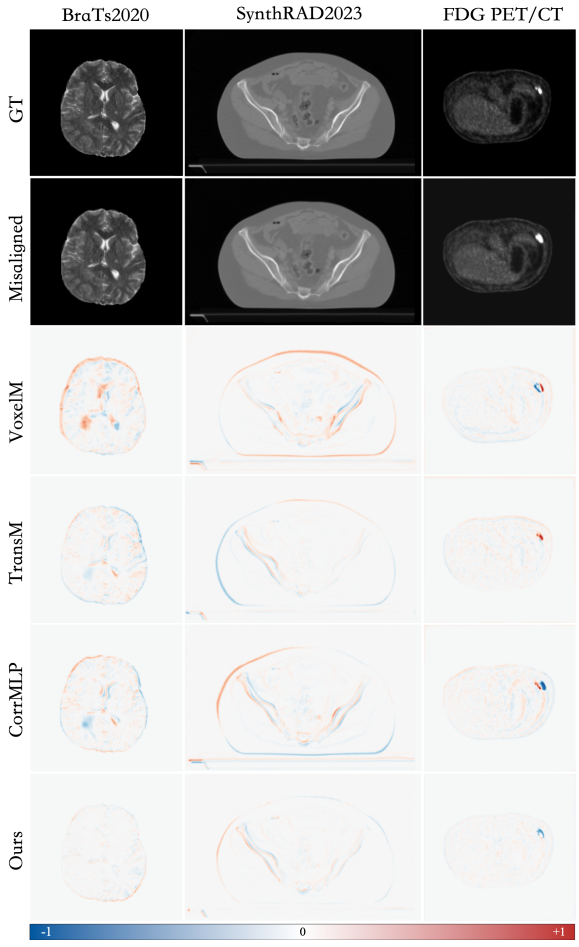

Figure 3. Visual comparison of alignment performance. The first row shows the aligned image labels, the second row presents the misaligned images, and the third to sixth rows display the alignment error maps for different methods. The closer the image information is to white, the better the alignment performance.

ting B, the fusion results show significant contrast deviations and a loss of source image details due to the absence of degradation-aware prompts in the fusion and restoration network, making it difficult for the network to maintain consistent fusion performance across different degradation task scenarios. In Setting C, the fusion results display a decline in both feature alignment and detail restoration performance. Only when DAPL is fully retained do the fusion results achieve optimal visual quality.

Effectiveness of OUFR. To visually validate the effectiveness of OUFR, we replace the Spatial Mamba component with a Transformer and the standard Mamba, respec-

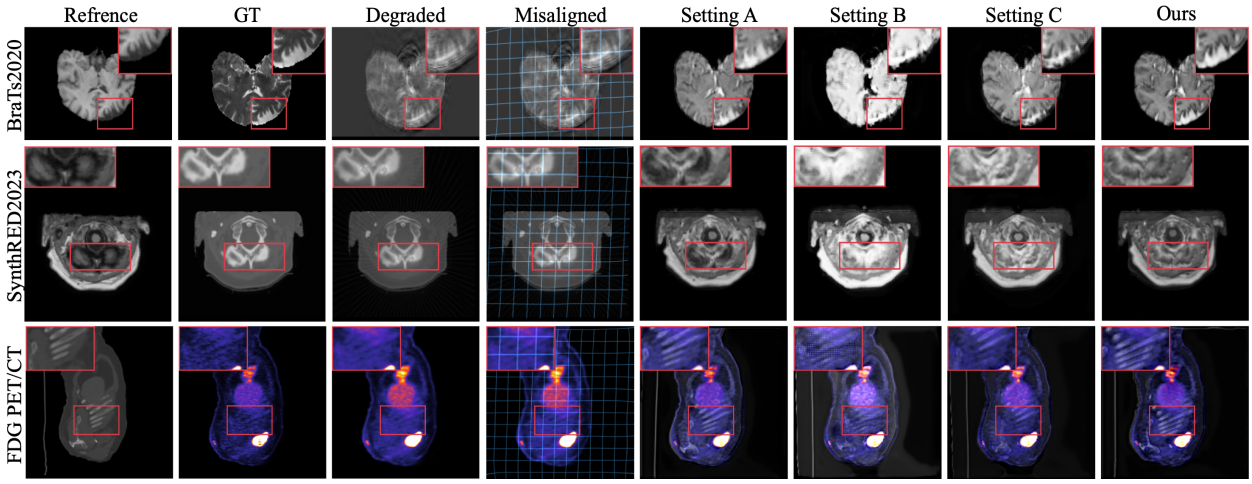

Figure 4. Visual comparison of DAPL's effectiveness. The first column shows the reference image without degradation, the second column presents the ground truth (GT) of the degraded image, the third column displays the misaligned degraded image, the fourth column highlights the distortions in the misaligned degraded image, and the fifth to eighth columns show the fusion results under different experimental settings.

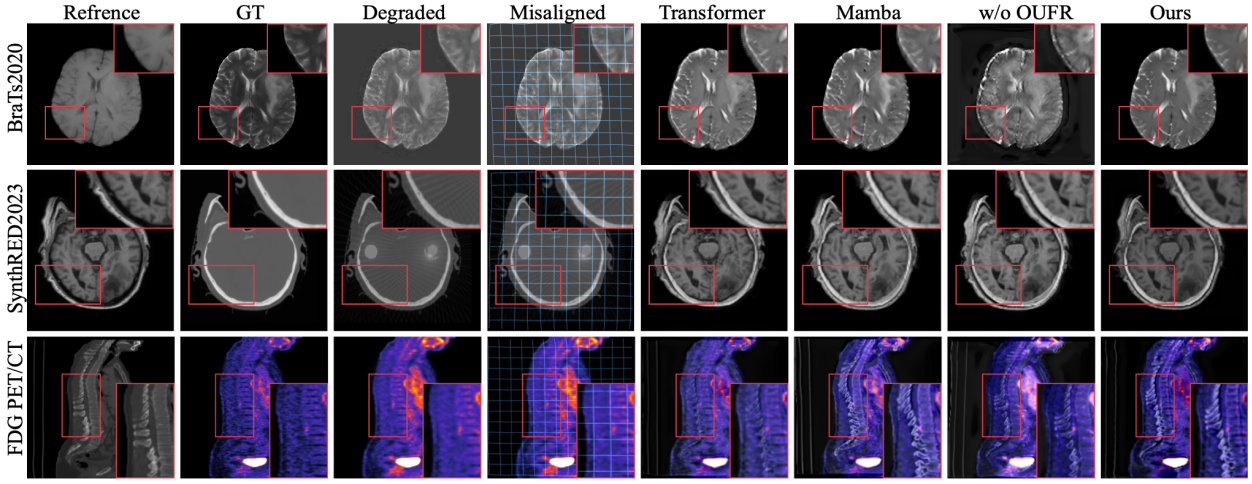

Figure 5. Visual comparison of OUFR's effectiveness. The first column shows the reference image without degradation, the second column presents the ground truth (GT) of the degraded image, the third column displays the misaligned degraded image, the fourth column highlights the distortions in the misaligned degraded image, and the fifth to eighth columns show the fusion results under different experimental settings.

tively, and also conduct an experiment where the entire OUFR is removed for comparison. As shown in Figure 5, when Spatial Mamba is replaced by either the Transformer or the standard Mamba, feature misalignment occurs in the fusion results. This phenomenon becomes even more pronounced when the entire OUFR is removed. This is because the introduction of Spatial Mamba effectively eliminates modality differences between source image features, allowing the feature alignment process to proceed without being affected by these differences, thus producing higher-quality fused images. Only when OUFR is fully present does the network achieve optimal fusion quality, further demonstrating the effectiveness of OUFR.

Effectiveness of FA. To visually analyze the impact of using multiple RegBLKs jointly in FA on the fusion results, we conduct an ablation study by adjusting the number of RegBLKs and observing their effects. As shown in Figure 6, when no RegBLKs are used in FA, significant feature misalignment occurs in the fused image. When the number of RegBLKs (J) is set to 4, feature alignment achieves optimal results.

Effectiveness of UFR&F. To visually demonstrate the

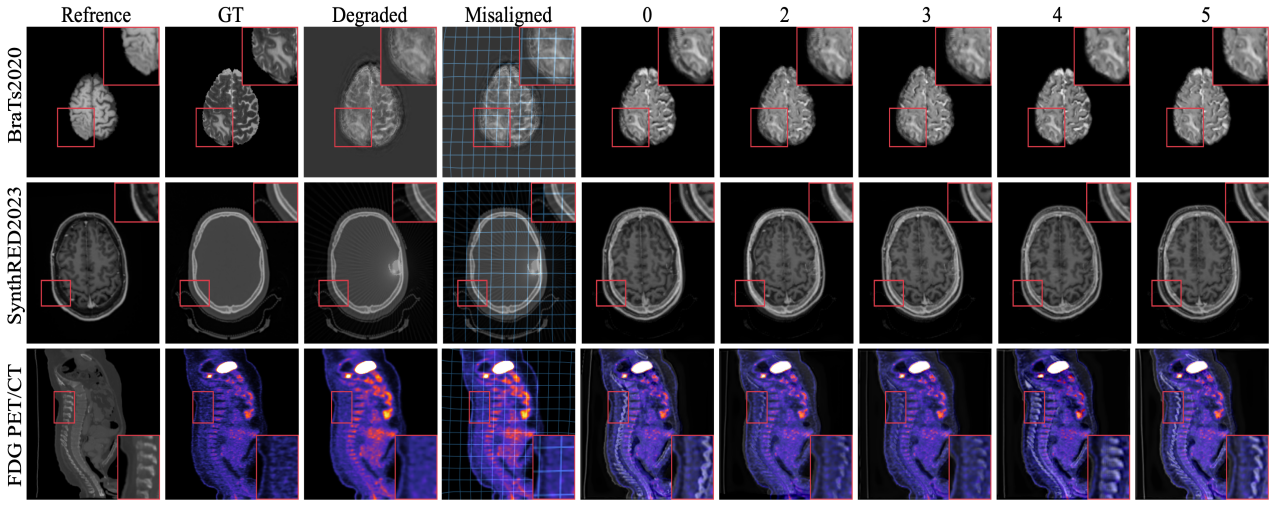

Figure 6. Visual comparison of FA's effectiveness. The first column shows the reference image without degradation, the second column presents the ground truth (GT) of the degraded image, the third column displays the misaligned degraded image, the fourth column highlights the distortions in the misaligned degraded image, and the fifth to ninth columns show the fusion results under different experimental settings.

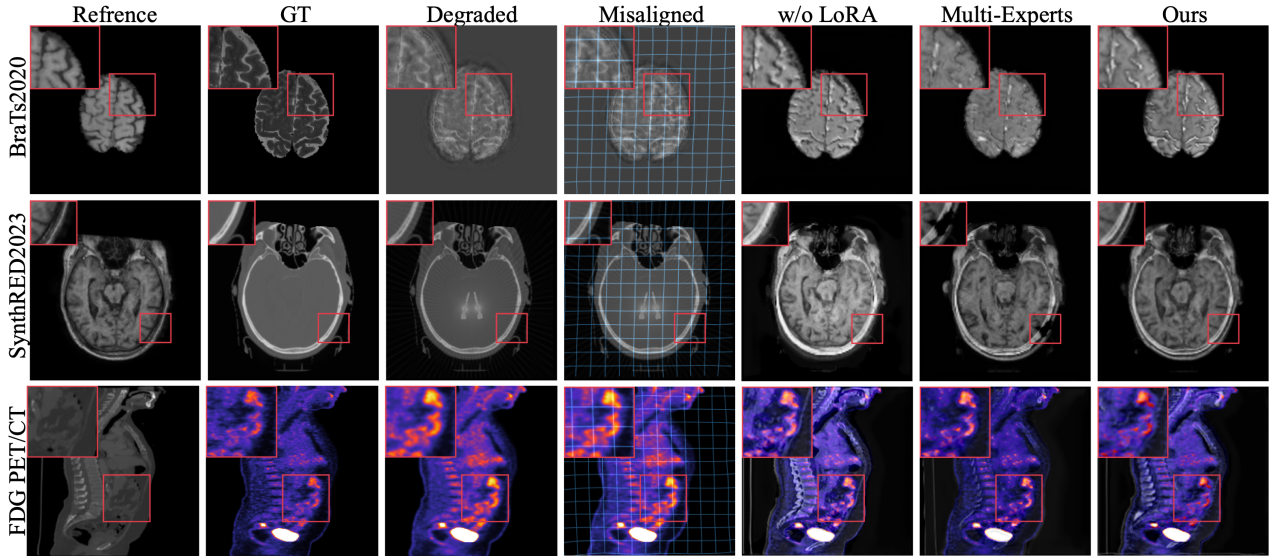

Figure 7. Visual comparison of UFR&F's effectiveness. The first column shows the reference image without degradation, the second column presents the ground truth (GT) of the degraded image, the third column displays the misaligned degraded image, the fourth column highlights the distortions in the misaligned degraded image, and the fifth to ninth columns show the fusion results under different experimental settings.

effectiveness of UFR&F, we design two sets of experiments. In the first set, we remove the LoRA branch from ALSN. In the second set, we replace the ALSN branch with a standard multi-expert architecture. As shown in Figure 7, after removing the LoRA branch, the network is unable to accurately restore the contrast information in the source images, and the degradation removal effect is compromised due to the loss of the network's adaptive capability to different types of data. When using the multi-expert architecture, feature loss occurs in the fusion results due to the training convergence issues mentioned in the main text. Only when using the complete UFR&F does the network's fusion results exhibit the best visual quality.

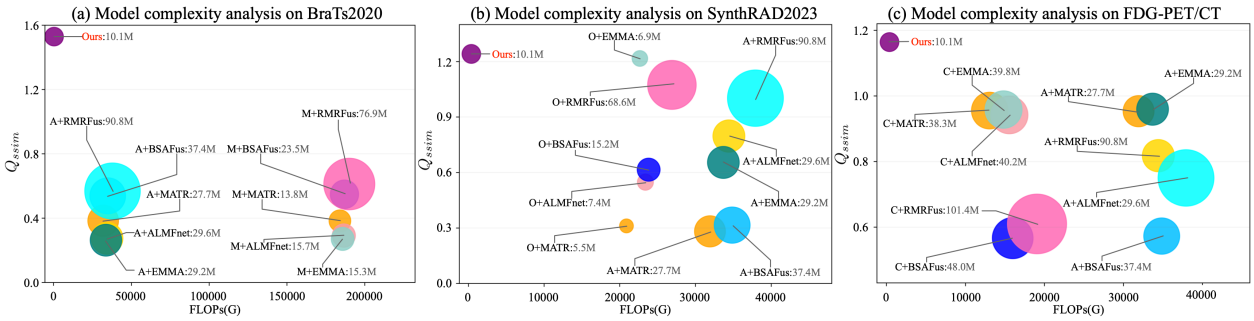

Figure 8. Model complexity analysis. The x-axis represents the FLOPs (in billions) for models processing $256 \times 256 \times 256$ image inputs, the y-axis denotes the average Q_{ssim} scores, and the bubble size indicates the number of parameters. Subplots (a), (b), and (c) correspond to the experimental results on different datasets.

4. Complexity Comparison

In practical applications, a model's parameter size and computational load directly determine its deployment difficulty and application cost. Therefore, we conducted a complexity analysis of our proposed method and compared it with existing approaches. Figure 8 presents the complexity comparison results of different methods. Our method achieves optimal performance while exhibiting significantly lower computational complexity than other methods, and its parameter size is also smaller than that of all the compared All-in-One restoration fusion frameworks. This advantage is attributed to our adoption of a single-stage design pattern, which substantially reduces both the number of model parameters and computational overhead. It is worth noting that although our method has more parameters than some single-degradation restoration fusion frameworks, this is due to the incorporation of 3D convolution in the network and the combination of multiple RegBLKs in the FA module.

5. Limitations of the Method

Despite the promising performance demonstrated by Unifuse, several limitations remain. First, the method assumes that the input images are degraded but not severely distorted, which may limit its effectiveness when dealing with highly noisy or corrupted images. Second, while Unifuse can handle alignment and fusion in a unified framework, its performance may still be sensitive to extreme misalignments or inconsistencies in image resolutions, as it relies on the assumption of relatively consistent input conditions. Third, the degradation-aware prompt learning module, while effective for a range of common degradation types, may not generalize well to all possible degradation scenarios or unseen image modalities. Finally, while the integration of ALSN allows for adaptive feature representation, its performance could be constrained by the complexity of the network, potentially leading to increased computational cost in resource-limited environments. Further research is needed to address these challenges, enhance the model's robustness, and improve its scalability across diverse real-world applications.

References

[1] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. 1

[2] Junyu Chen, Eric C Frey, Yufan He, William P Segars, Ye Li, and Yong Du. Transmorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022. 1

[3] Mingyuan Meng, Dagan Feng, Lei Bi, and Jinman Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9645–9654, 2024. 1