Supplementary Materials of Text-DiFuse: An Interactive Multi-Modal Image Fusion Framework based on Text-modulated Diffusion Model

Hao Zhang,* **Lei Cao**,* **Jaiyi Ma**†
Electronic Information School
Wuhan University
Wuhan, China
{zhpersonalbox, jyma2010}@gmail.com, whu.caolei@whu.edu.cn

A   Data Configuration

A.1   Data for Training Diffusion Model

In our work, acquiring image restoration capability depends on pre-training a conditional diffusion model, which needs paired clean and degraded data. The clean data are used to build the loss function for supervision, while the degraded data act as conditioning inputs for the denoising network. Therefore, we use existing supervised datasets and additionally simulate a portion of the data to meet the requirements of mixed degradation. Our method primarily addresses three common types of degradation in the fusion scenario: improper lighting, color distortion, and noise. For improper lighting, we use $2,220$ image pairs from the MIT-Adobe FiveK Dataset [2], covering images with varying exposures and their corresponding ground truth manually adjusted by photography experts. For color distortion, we use $1,031$ image pairs from the Rendered WB dataset[1], including color-biased images under various light sources such as fluorescent, incandescent, and daylight, as well as corresponding reference images manually calibrated under the Adobe standard. For noise, we add Gaussian noise, pulse noise, Poisson noise, Rayleigh noise, and uniform noise to $2,220$ clean images from the MIT-Adobe FiveK Dataset and $2,220$ clean images from the MSRS dataset [4] to obtain noised images. All these image pairs constitute the complete dataset for training our diffusion model, driving our model's learning for compound degradation removal.

A.2   Clean Data for Training Fusion Module

Constructing Eqs. (9) and (10) actually involves very stringent data requirements. Specifically, they require a pair of degraded multi-modal images describing the same scene, along with their corresponding clean versions. Unfortunately, such a dataset is currently not available. To alleviate this challenge, we adopt a two-step strategy. Specifically, we first pre-train the diffusion model to learn the image restoration capability. In this step, we only need degraded-clean image pairs, without the need for paired multi-modal images that describes the same scene. Once the diffusion model is trained, it can be used to process existing degraded multi-modal image fusion datasets, to generate the required clean multi-modal image pairs. At this point, all the data required for constructing Eqs. (9) and (10) has been obtained.

---

*Equal Contribution

†Corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

B Limitation

Although our method shows advanced performance in multiple scenarios, it still has certain limitations. Specifically, the efficiency of our method is relatively low. To display it more intuitively, we implement an efficiency evaluation, including parameter number, and runtime, as reported in Table s1. It can be seen that our method has a relatively large number of parameters and a relatively long runtime. This is because the diffusion model requires multiple iterations of sampling. One piece of evidence is that another method based on the diffusion model, DDFM, also exhibits a large number of parameters and long runtime. In the future, we will study the acceleration strategy of the diffusion model and further improve its integration in multi-modal image fusion to increase operating efficiency.

Table s1: Statistical results of parameters and runtime.

	RFN-Nest	GANMcC	SDNet	U2Fusion	TarDAL	DeFusion	LRRNet	DDFM	MRFS	Ours
Parameter/M	30.10	2.28	0.07	0.66	0.30	7.87	0.20	552.66	134.96	157.35
Runtime/Second	1.28	4.83	1.01	3.46	0.58	0.29	0.27	131.74	4.39	31.63

C More Visual Comparisons

As the length of the main text is limited, we provide more visual comparisons here to demonstrate the advantages of our Text-DiFuse, involving infrared and visible image fusion, and medical image fusion. We first provide the visual results of infrared and visible image fusion in Fig. s1. Clearly, our method can recover scene information from degraded environments and remove composite degradations including color deviation, improper lighting, noise, *etc*. On the contrary, competitors cannot do these, which will inevitably lead to the loss of useful features in the subsequent multi-modal information fusion. Besides, the visual results of medical image fusion are presented in Fig. s2. It can be seen that our method is the best at highlighting the body tissue structure. At the same time, it is also the best in maintaining the functional distribution characterized by colors. In comparison, the comparative methods either weaken the tissue structure or lead to inaccurate functional distribution. Overall, these visual results show that our Text-DiFuse achieves state-of-the-art performance for the task of multi-modal image fusion.
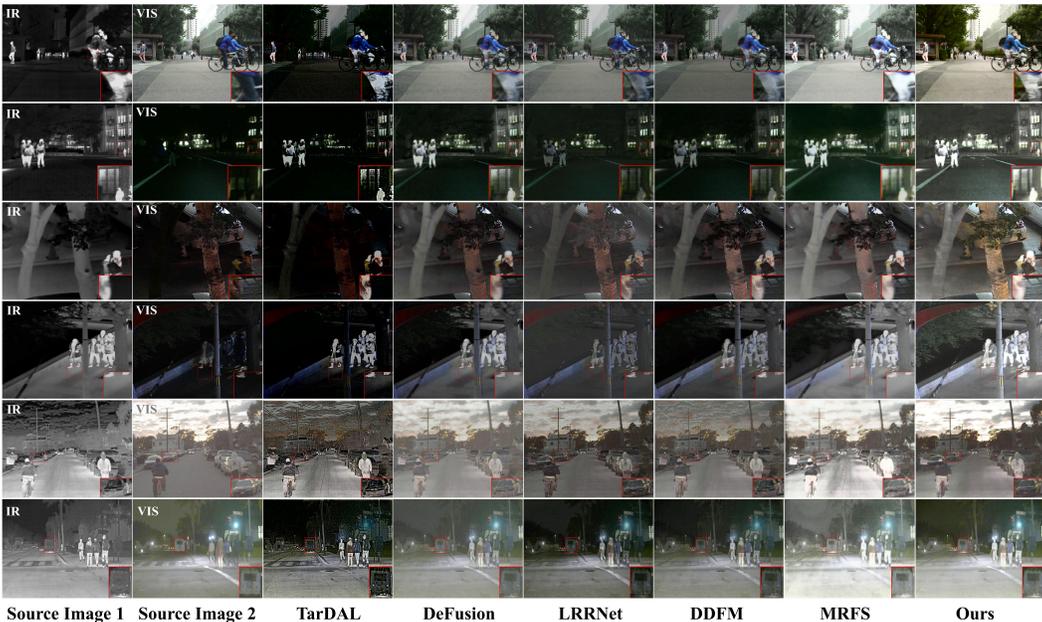
Figure s1: Visual results of infrared and visible image fusion.

2

| Source Image 1 | Source Image 2 | TarDAL | DeFusion | LRRNet | DDFM | MRFS | Ours |
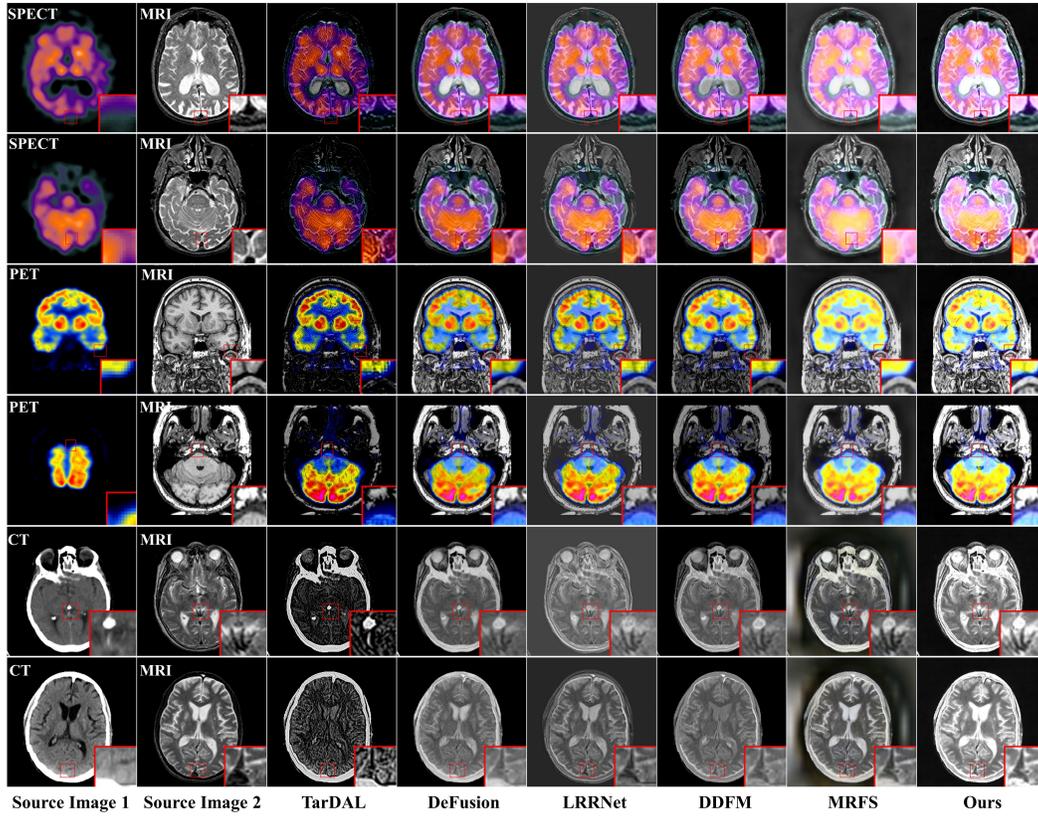
Figure s2: Visual results of medical image fusion.

| Source Image 1 | Source Image 2 | TarDAL | DeFusion | LRRNet | DDFM | MRFS | Ours |
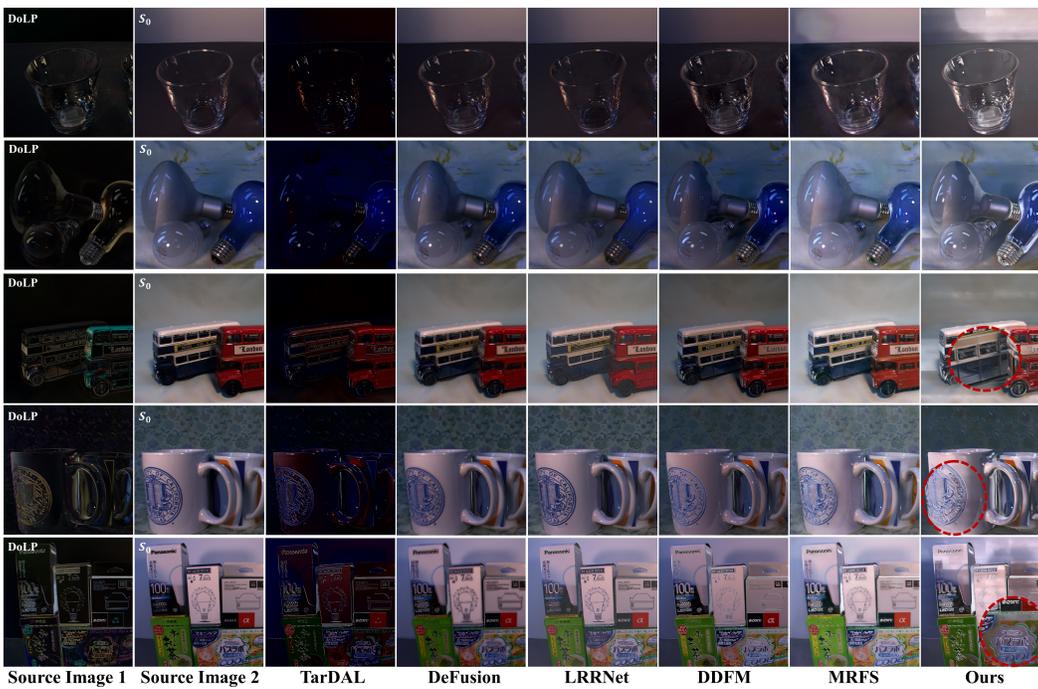
Figure s3: Visual results on extended polarization image fusion.

3
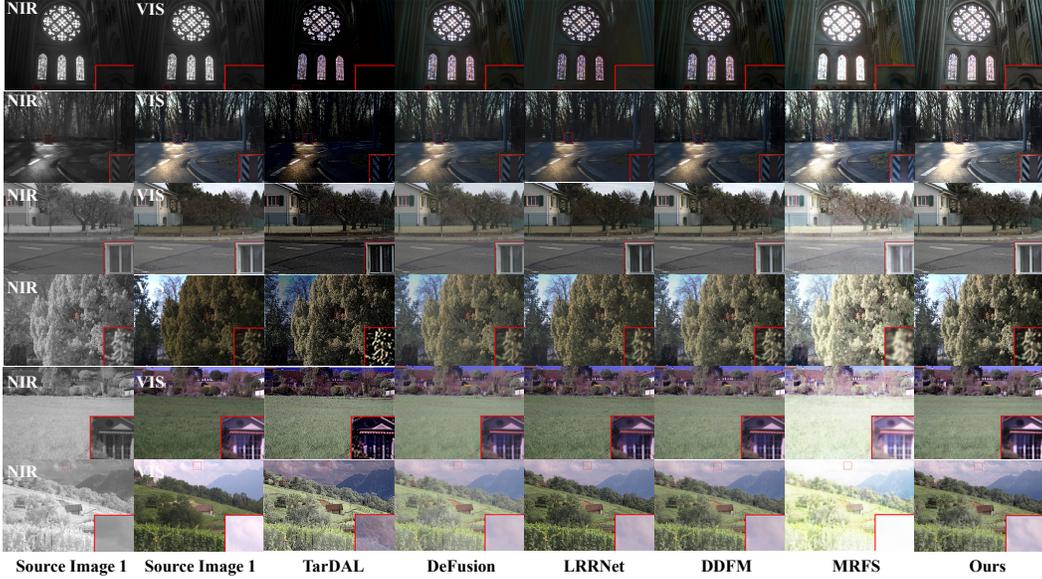
Figure s4: Visual results on extended polarization image fusion.

Table s2: Quantitative results on extended application scenarios.

Sce.	Polarization Fusion					RGB-NIR Fusion				
	EN	AG	SD	SCD	VIF	EN	AG	SD	SCD	VIF
RFN.	6.54	2.45	33.05	1.70	0.55	6.99	3.47	43.93	0.56	0.96
GANM.	6.23	2.84	28.63	1.54	0.51	6.88	4.65	39.73	0.46	0.94
SDN.	5.44	3.37	18.89	1.04	0.49	7.00	6.55	48.06	0.91	1.11
U2F.	5.97	3.54	28.00	1.52	0.50	6.77	6.49	39.92	0.40	0.92
Tar.	4.19	4.45	21.74	0.66	0.24	5.42	**10.19**	49.08	0.42	0.51
DeF.	6.53	2.87	35.87	1.36	0.63	6.99	4.45	43.66	0.45	1.16
LRR.	6.30	2.79	33.30	0.86	0.45	6.53	4.20	39.11	0.30	0.81
DDFM	6.55	3.61	32.87	1.76	**0.65**	7.02	4.80	44.94	0.79	**1.26**
MRFS	7.08	3.33	47.58	1.59	0.64	7.50	5.05	57.37	0.95	0.96
Ours	**7.19**	**5.39**	**48.85**	**1.78**	0.55	**7.52**	6.61	**63.04**	**1.42**	1.16

D Extended Application

We extend our Text-DiFuse to the polarization image fusion scenario and the near-infrared and visible image fusion scenario. Firstly, the purpose of polarization image fusion is to fuse polarization information and intensity images, to produce images with more clearly visible textures and a more comprehensive description of objects in the scene. The visual results of the polarization image fusion are presented in Fig. s3. It can be observed that our method still exhibits good performance. It effectively integrates the structure contained in the polarization information into the intensity image, demonstrating enhanced visualization that far exceeds that of other competitors. Secondly, the visual results of near-infrared and visible image fusion are shown in Fig. s4, where our method effectively integrates texture details from the near-infrared band with those from the visible image, while preserving natural color attributes of the visible image. Notably, the inherent image restoration capability of our method allows it to produce vivid fused images in underexposed scenes without causing overexposure like MRFS, as seen in the results of the first row. In addition, we conduct quantitative experiments to evaluate the objective application performance, as presented in Table s2. The proposed Test-DiFuse demonstrates good quantitative performance in these two application scenarios, ranking first in most metrics. Overall, our method can be generalized to the polarization image fusion and the near-infrared and visible image fusion scenarios with promising performance.

4

Figure s5: Visual results of three-channel direct processing.

Table s3: Quantitative results of three-channel direct processing.

MSRS	EN↑	AG↑	SD↑	SCD↑	VIF↑	CIECAM16↓
3-channel	6.84	**3.66**	39.18	1.39	0.72	3.15
Ours	**7.08**	3.31	**47.44**	**1.44**	**0.76**	**1.59**

Figure s6: Visual comparison of InstructIR plus Fusion.

E   Brightness-Chrominance Separation

Image fusion requires a high level of color fidelity to the scene. Taking the infrared and visible image fusion as an example, the colors in the fused image are required to be as consistent as possible with those in the visible image. Therefore, by independently purifying and preserving the chrominance components in the visible image, our method can effectively and conveniently achieve color fidelity. Next, we discuss why our method does not directly process three-channel images. Firstly, from the perspective of image restoration alone, directly processing color images is entirely feasible. However, our method requires embedding information fusion into the latent layers of the diffusion model used for image restoration. This means that features from the gray infrared image could potentially interfere with the color distribution of features from the visible image. In particular, this interference occurs in the highly nonlinear latent space, where some small changes can be amplified by the decoder to produce large color distortions. In this case, ensuring the expected color fidelity is very difficult. Second, the interference is directly related to the way multi-modal features are fused. In our method, we use a nonlinear neural network called the Fusion Control Module to perform information aggregation, which is guided to retain significant thermal radiation objects while preserving rich background textures. These two goals correspond to the similarity loss functions (see Eqs. (9) and (10)) based on the indicators of pixel intensity and gradient. Under such optimization guidance, it is difficult to avoid disrupting the color distribution in the features from the visible image. For verifying, we adapt our proposed method to directly process three-channel images without separating brightness and chrominance components, and the results are presented in Fig. s5. Clearly, color distortion occurs. Furthermore, we implement quantitative evaluation in Table s3. The direct processing strategy decreases the color score CIECAM16 and also negatively affects other metrics to varying degrees

F   All-in-One Restoration Plus Fusion

We conduct comparisons using the all-in-one restoration method InstructIR [3] followed by several advanced image fusion methods. First, we input different text prompts into InstructIR to address improper lighting, noise, and color distortion. The restored images are then fused with advanced fusion methods. Visual results are shown in Fig. s6. Our method which deeply couples image restoration and fusion, shows better performance than these methods following a concatenation

5

Table s4: Quantitative comparison of InstructIR plus Fusion.

Methods		MSRS Dataset				
		EN ↑	AG ↑	SD ↑	SCD ↑	VIF ↑
InstructIR	RFN-Nest	6.62	2.20	31.63	1.39	0.74
	GANMcC	6.27	1.94	26.52	1.20	0.63
	SDNet	5.44	2.62	18.24	1.09	0.56
	U2Fusion	6.43	3.19	33.79	1.38	0.76
	TarDAL	4.53	2.99	24.87	0.96	0.29
	DeFusion	7.03	2.86	43.76	1.03	0.84
	LRRNet	6.86	2.73	37.91	0.76	0.74
	DDFM	6.67	2.39	31.55	1.32	0.81
	MRFS	**7.40**	3.28	45.98	0.96	**0.89**
Ours		7.08	**3.31**	**47.44**	**1.44**	0.76
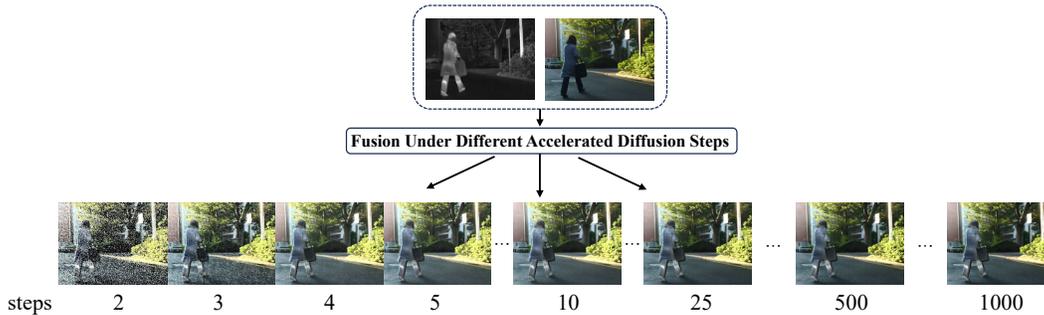
Figure s7: Visual results with different sampling steps.
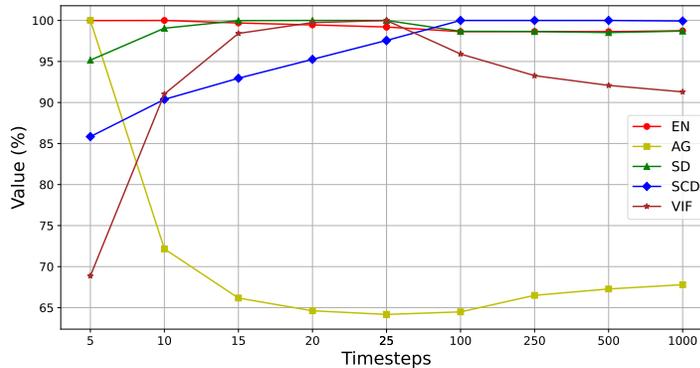
Figure s8: Metric changes with different sampling steps.

strategy. In particular, our method can balance thermally salient object retaining and degradation removal, while competitors cannot. Furthermore, the quantitative results in Table s4 also prove the advantages of our method.

G Analysis of Sampling Steps

In our method, image restoration and information integration are mutually coupled. This is reflected in the physical connection, where a fusion control module is embedded within the internal structure of the diffusion model. Once all the networks are trained, we can follow the standard diffusion model testing procedure, which involves performing T steps of continuous sampling. It is worth noting that information fusion needs to be performed at each sampling step. In this case, the only factor affecting the final fusion result is the number of sampling steps. More sampling steps mean better

6

performance, but they also result in significant time consumption. Therefore, setting an appropriate number of sampling steps is a matter worth discussing. In tasks where the ground truth is available, the number of sampling steps can be well determined by checking whether the generated results are sufficiently close to the ground truth. However, for the image fusion task, where ground-truth data do not exist, we rely on visual perception and multiple no-reference metrics to make the assessment. Specifically, we set the number of sampling steps to 2, 3, 4, 5, 10, 25, 500, and 1000, with qualitative and quantitative results shown in Figs. s7 and s8. Notably, each metric is normalized along the step dimension for easier presentation. It can be observed that as the number of steps increases, noise is gradually removed and the scene texture becomes increasingly refined. Corresponding to the quantitative results, 25 steps achieve good performance saturation, with subsequent increases in the number of steps resulting in only slight fluctuations in scores. Note that the only exception is AG, as it is affected by noise during the diffusion process. Therefore, in our experimental section, the number of sampling steps is set to 25.

H Broader Impacts

This paper is devoted to solving the problem of multi-modal image fusion under degraded scenes to provide high-quality fused results suitable for human and machine perception. Therefore, it can be expected that this work will demonstrate positive social impacts in many fields. For example, it can help drivers better perceive the road conditions ahead in environments with poor visibility through information fusion, such as at night, to improve driving safety. For another example, it can help poor areas that only have low-quality medical imaging equipment to enhance the perception of the body's condition through information recovery and fusion, thereby assisting in disease diagnosis and treatment. As far as we know, this work does not appear to have any negative social impacts and the risks are extremely low.

References

[1] Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1535–1544, 2019.

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 97–104, 2011.

[3] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2(4), 2024.

[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. *Information Fusion*, 83:79–92, 2022.

7