Multi-Grained Text-Guided Image Fusion for Multi-Exposure and Multi-Focus Scenarios

Supplementary Material

A1. More Qualitative Comparison for Multi-Exposure Image Fusion

We provide additional qualitative results for multi-exposure image fusion (MEF). As shown in Fig. A1, most prior methods fail to maintain appropriate luminance. AGAL [23], HoLoCo [25], and FILM [80] produce results suffering from considerable detail loss in high-intensity regions. In contrast, our method consistently preserves optimal brightness and contrast, while effectively retaining structural and textural information, such as the details of grass and trees outside the window, thereby delivering perceptually superior fusion outputs. This demonstrates the effectiveness and robustness of our method, where the multi-grained textual guidance modulates visual features at corresponding levels, and the hierarchical supervision together with the saliency-driven visual enrichment module, further improve cross-modal feature alignment and auxiliary text utilization.

A2. More Qualitative Comparison for Multi-Focus Image Fusion

We present additional qualitative results for multi-focus image fusion (MFF). As shown in Fig. A2, most compared methods suffer from insufficient detail preservation and noticeable artifacts. U2Fusion [56] and FILM [80] suffer from overexposure, leading to unnatural appearances in facial regions. In contrast, our method consistently generates visually balanced outputs with enhanced detail, natural luminance, and artifact-free quality, as demonstrated by the clear text and distinct lines on the book page. These results demonstrate the effectiveness and robustness of our method in preserving focused details and avoiding artifacts in MFF.

A3. Infrared and Visible Image Fusion

To further validate the generalizability and robustness of our proposed **M**ulti-grained **T**ext-guided **I**mage **F**usion (MTIF) framework, we extend its application from multi-exposure image fusion (MEF) and multi-focus image fusion (MFF) tasks to the infrared and visible image fusion (IVF). IVF aims to synthesize a single image that integrates thermal radiation cues from infrared sensors and fine-grained texture and color details from RGB cameras, enabling more comprehensive scene perception under adverse conditions such as low illumination, smoke, or haze [46]. Despite its benefits, IVF remains a challenging task due to the significant modality gap between infrared and visible images, including disparities in spectral characteristics, spatial structures, and semantic content [74].

Dataset and Setup. Following Zhao et al.[80], we use a subset of the MSRS[45] dataset for training, with the remainder reserved for testing. This dataset includes a wide variety of image pairs from different scenes. To further assess the generalizability of MTIF, we conduct experiments on additional datasets, including M3FD [22], which covers four major scenarios with various environments, illumination, seasons, and weather, and RoadScene [57], which contains road scene images captured in various outdoor settings. All training settings, including the number of epochs, optimization strategy, and model configurations, are consistent with those used in the MEF experiments. For the IVF task, we compare our method with nine state-of-the-art deep learning methods: TarDAL [22], DeFusion [20], MetaFusion [77], CDDFuse [78], LRRNet [12], MURF [60], DDFM [79], SegMIF [24], and FILM [80].

Results and Analysis. We present both qualitative and quantitative evaluations to assess the performance of our method. For the qualitative evaluation, Fig. A3 compares the input images with the fusion outputs generated by our method. The results clearly demonstrate that our method effectively preserves thermal radiation and fine details while minimizing artifacts through multi-grained textual guidance. For the quantitative results in Tab. A1, our method achieves superior performance in most metrics, demonstrating its effectiveness in preserving thermal and fine details and minimizing artifacts in diverse IVF tasks.

A4. Evaluation of Hyper-Parameters

We perform a hyper-parameter analysis of the loss function to assess the effects of different weighting strategies. As shown in Tab. A2, excessive weighting of α_1 or α_2 degrades fine detail preservation. Increasing β_1 slightly decreases contrast, while decreasing it reduces global information preservation. For β_2, overweighting compromises structural perception, whereas insufficient weighting underutilizes saliency, impairing overall visual quality. The proposed default configuration achieves an optimal balance and consistently achieves optimal performance on both multi-exposure image fusion and multi-focus image fusion tasks.

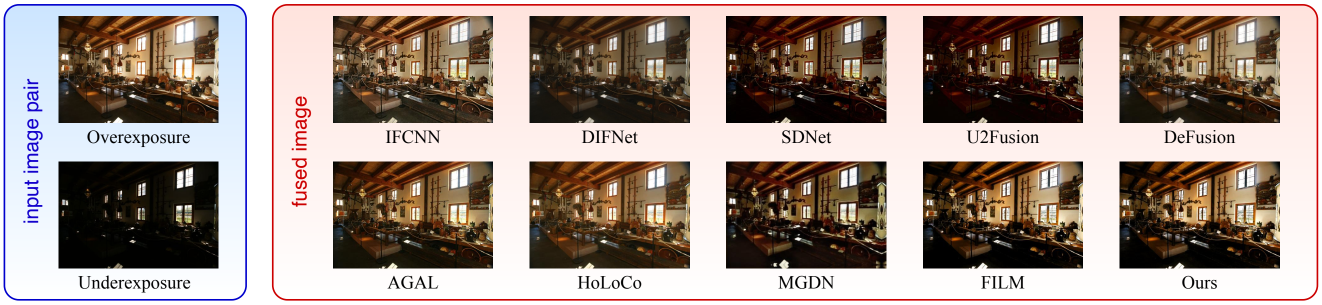

Figure A1. Visualization comparison of fusion results on the MEFB [72] dataset for the multi-exposure image fusion task.

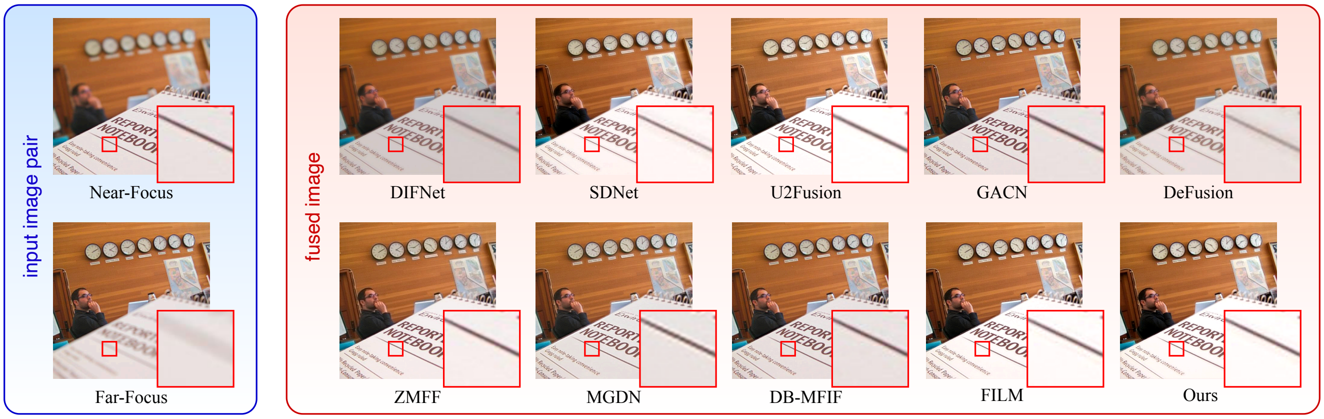

Figure A2. Visualization comparison of fusion outputs on the Lytro [35] dataset for the multi-focus image fusion task.

Table A1. Quantitative results of infrared and visible image fusion. The **bold** markers represent the best values.

	MSRS Infrared-Visible Fusion Dataset							**M^3FD Infrared-Visible Fusion Dataset**							**RoadScene Infrared-Visible Fusion Dataset**					
	EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf		EN	SD	SF	AG	VIF	Qabf
TarD [22]	5.28	25.22	5.98	1.83	0.42	0.18	TarD [22]	6.79	40.75	8.18	2.92	0.53	0.30	TarD [22]	7.25	47.57	11.46	4.23	0.56	0.43
DeF [20]	6.46	37.63	8.60	2.80	0.77	0.54	DeF [20]	6.84	35.09	9.65	3.37	0.59	0.42	DeF [20]	7.39	47.60	11.26	4.47	0.63	0.50
Meta [77]	5.65	24.97	9.99	3.40	0.63	0.48	Meta [77]	6.68	29.62	16.22	5.68	0.68	0.57	Meta [77]	6.87	31.95	14.40	5.55	0.55	0.46
CDDF [78]	6.70	**43.39**	11.56	3.74	1.05	0.69	CDDF [78]	7.08	41.29	16.49	5.42	0.78	0.63	CDDF [78]	7.41	**54.59**	17.04	6.07	0.63	0.51
LRR [12]	6.19	31.78	8.46	2.63	0.54	0.46	LRR [12]	6.60	30.19	11.69	3.95	0.57	0.51	LRR [12]	7.09	38.77	11.50	4.36	0.43	0.33
MURF [60]	5.04	20.63	10.49	3.38	0.44	0.36	MURF [60]	6.52	27.90	11.43	4.51	0.39	0.30	MURF [60]	6.91	33.46	13.74	5.31	0.53	0.47
DDFM [79]	6.19	29.26	7.44	2.51	0.73	0.48	DDFM [79]	6.72	31.15	9.84	3.42	0.63	0.47	DDFM [79]	7.27	42.94	10.89	4.20	0.63	0.50
SegM [24]	5.95	37.28	11.10	3.47	0.88	0.63	SegM [24]	6.89	35.64	16.11	5.52	0.78	0.65	SegM [24]	7.29	47.10	15.07	5.78	0.65	0.56
FILM [80]	6.72	43.17	11.70	3.84	1.06	0.73	FILM [80]	7.09	41.53	16.77	5.55	0.83	0.67	FILM [80]	7.43	49.25	17.34	6.60	**0.69**	0.62
Ours	**6.73**	43.31	**11.72**	**3.84**	**1.06**	**0.73**	Ours	**7.15**	**42.88**	**17.08**	**5.77**	**0.86**	**0.69**	Ours	**7.53**	54.13	**18.14**	**6.92**	0.68	**0.62**

Table A2. Ablation of hyper-parameters on multi-exposure and multi-focus image fusion tasks. The **bold** markers represent the best values.

Configuration	**SICE Multi-exposure Image Fusion Dataset**						**RealMFF Multi-focus Image Fusion Dataset**					
	EN	SD	SF	AG	VIF	Qabf	EN	SD	SF	AG	VIF	Qabf
$\alpha_1 = 20$	7.28	56.38	18.88	**5.52**	1.20	0.77	7.13	56.04	15.72	5.40	**1.59**	0.76
$\alpha_1 = 5$	7.26	59.52	18.81	5.43	1.43	0.76	7.13	55.30	15.45	5.28	1.58	0.76
$\alpha_2 = 5$	7.25	59.54	18.64	5.37	1.44	0.77	7.11	55.87	15.55	5.31	1.58	0.76
$\alpha_2 = 0.2$	7.28	60.06	18.72	5.43	1.46	0.77	7.13	**57.39**	15.59	5.39	1.59	0.75
$\beta_1 = 2$	7.13	58.30	**19.24**	5.50	1.42	0.75	7.13	56.01	15.34	5.26	1.58	0.76
$\beta_1 = 0.5$	7.26	59.69	18.81	5.41	1.46	0.77	7.13	56.73	15.69	5.36	1.58	0.76
$\beta_2 = 400$	7.27	59.11	18.82	5.43	1.44	0.76	7.12	55.31	15.49	5.29	1.58	0.76
$\beta_2 = 25$	7.25	59.49	18.84	5.47	1.45	0.76	7.11	54.83	15.56	5.31	1.57	0.76
Ours	**7.28**	**60.41**	18.78	5.40	**1.46**	**0.79**	**7.14**	56.03	**15.78**	**5.43**	1.56	**0.77**

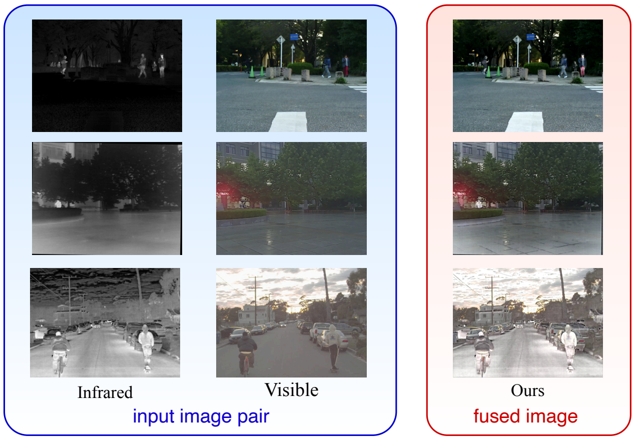

Figure A3. Visualization comparison of fusion results for the infrared-visible image fusion task.