Supplementary Material for "Revisiting Spatial-Frequency Information Integration with Hierarchical Perspective for Panchromatic and Multi-Spectral Image Fusion"

Jiangtong Tan[1], Jie Huang[1], Naishan Zheng[1], Man Zhou[1], Keyu Yan[1], Danfeng Hong[2], Feng Zhao[1]*
[1]University of Science and Technology of China, [2]Chinese Academy of Sciences
{jttan hj0117 nszheng manman keyu}@mail.ustc.edu.cn, hongdf@aircas.ac.cn, fzhao956@ustc.edu.cn

1. More result of ablation

In this section, we will provide more ablation experimental results. Firstly, we replaced the Global Fourier block with Local Fourier block to further validate the effectiveness of hierarchical information and the importance of global information. Secondly, we removed the branch exchanging operation to verify the effectiveness of this operation. As shown in Tab. 1, when we replaced the Global Fourier block with Local Fourier block, the metrics declined, indicating that the global Fourier information cannot be absent in hierarchical information. We can also observe that the model performance declined when removing the branch exchanging operation. This indicates that the interaction between the Local Fourier branch and the Global Fourier branch is important for the fusion process.

We also conduct ablation experiments on different numbers of SGLI modules. As shown in Tab. 2, the performance gradually improved with an increase in the number of SGLI modules, but it also led to an increase in the number of parameters. Therefore, we chose five SGLI modules as a compromise solution.

Config	Global frequency	Branch exchanging	PSNR↑	SSIM↑	SAM↓	ERGAS↓
(I)	✗	✓	42.0545	0.9707	0.0219	0.8984
(II)	✓	✗	42.1687	0.9712	0.0216	0.8874
Ours	✓	✓	**42.2319**	**0.9714**	**0.0215**	**0.8807**

Table 1. Ablation studies comparison on the WorldView-II datasets. The best and the second best values are highlighted in **bold**.

2. More result of experiment for generalization

In this section, we will provide more result of other fusion tasks including visible and infrared image fusion, and depth image SR.

Number of SGLI blocks	PSNR↑	SSIM↑	SAM↓	ERGAS↓
1	40.9116	0.9641	0.0251	1.0396
2	41.6499	0.9685	0.0231	0.9469
3	41.8385	0.9696	0.0225	0.9204
4	42.0210	0.9705	0.0219	0.9001
5	**42.2319**	**0.9714**	**0.0215**	**0.8807**

Table 2. Ablation studies comparison on the WorldView-II datasets. The best and the second best values are highlighted in **bold**.

2.1. Datasets and Benchmarks

Visible and infrared image fusion. We perform extensive experiments on three publicly available datasets: M3FD [11], RoadScene [21], and TNO [18]. We compare our proposed model with nine state-of-the-art visible and infrared image fusion methods: DDcGAN [13], DenseFuse [8], AUIF [24], DIDFuse [23], ReCoNet [5], SDNet [22], TarDAL [11], U2Fusion [21], and UMFusion [19]. **Depth image SR.** We utilize three depth image SR datasets: NYU v2 [16], Middlebury [15], and Lu [12]. We compare our proposed model with eight state-of-the-art depth image SR methods: GF [4], DGF [20], DJF [9], DMSG [6], DJFR [10], DSRNet [1], PacNet [17], and FDKN [7].

2.2. Implementation details

We implement our method with PyTorch on NVIDIA GTX 3090 GPU. In visible and infrared image fusion, we use the Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ to update our model with a batch size of 8 and learning rate is set to 1×10^{-4}. The patch size is set to 128×128. To comprehensively evaluate the fusion results for visible and infrared image fusion, we utilize metrics such as mutual information (MI) [14], visual information fidelity (VIF) [3], and feature mutual information (FMI) [2], where higher values indicate better performance. In depth image SR, we use Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$ with batch size of 1 and

*Corresponding author.

Method	RoadScene			TNO			M3FD		
	MI↑	VIF↑	FMI↑	MI↑	VIF↑	FMI↑	MI↑	VIF↑	FMI↑
DDcGAN	2.6178	0.5946	0.859	1.8470	0.6737	0.858	2.5397	0.7684	0.836
DenseFuse	3.1276	0.8025	0.868	2.4018	0.7997	0.890	2.9297	0.7621	0.863
AUIF	3.1110	0.8466	0.856	2.2714	0.8146	0.879	3.0490	0.8192	0.845
DIDFuse	3.1840	0.8274	0.853	2.4422	0.8286	0.863	3.0476	0.8770	0.831
ReCoNet	3.1594	0.7956	0.858	2.4263	0.8266	0.878	3.0495	0.8184	0.845
SDNet	3.4225	0.8207	0.863	2.1860	0.7624	0.883	3.2315	0.6784	0.846
TarDAL	3.4640	0.7872	0.852	2.6480	0.8601	0.881	3.1624	0.8100	0.825
U2Fusion	2.8110	0.7402	0.861	1.9225	0.6878	0.879	2.7590	0.7091	0.850
UMFusion	3.2019	0.7913	0.866	2.2474	0.7169	0.888	3.0871	0.7089	0.855
Ours	**4.8114**	**0.8671**	**0.878**	**4.2646**	**0.9012**	**0.898**	**5.8933**	**0.9261**	**0.908**

Table 3. Quantitative comparison of our method with other state-of-the art methods on M3FD,RoadScne, and TNO datasets. The best values are highlighted in **bold**.

Method	Middlebury			Lu			NYU v2			Average		
	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$	$\times 4$	$\times 8$	$\times 16$
Bicubic	2.47	4.65	7.49	2.63	5.23	8.77	4.71	8.29	13.17	3.27	6.06	9.81
GF	3.24	4.36	6.79	4.18	5.34	8.02	5.84	7.86	12.41	4.42	5.85	9.07
DGF	1.94	3.36	5.81	2.45	4.42	7.26	3.21	5.92	10.45	2.53	4.57	7.84
DJF	1.68	3.24	5.62	1.65	3.96	6.75	2.80	5.33	9.46	2.04	4.18	7.28
DMSG	1.88	3.45	6.28	2.30	4.17	7.22	3.02	5.38	9.17	2.40	4.33	7.17
DJFR	1.32	3.19	5.57	1.15	3.57	6.77	2.38	4.94	9.18	1.62	3.90	7.17
DSRNet	1.77	3.05	4.96	1.77	3.10	5.11	3.00	5.16	8.41	2.18	3.77	6.16
PacNet	1.32	2.62	4.58	1.20	2.33	5.19	1.89	3.33	6.78	1.47	2.76	5.53
FDKN	1.08	2.17	4.50	0.82	**2.10**	**5.05**	1.86	3.58	6.96	1.25	2.62	5.50
Ours	**1.07**	**2.04**	**4.02**	**0.81**	2.19	5.19	**1.53**	**3.19**	**6.44**	**1.13**	**2.47**	**5.21**

Table 4. Average RMSE performance comparison for scale factors $\times 4$, $\times 8$ and $\times 16$ with bicubic down-sampling. The best values are highlighted in **bold**.

learning rate is set to 1×10^{-4}. The model's performance is evaluated using the root mean squared error (RMSE) as the default metric.

For the model implementation, PAN image is set to be infrared image and LRMS is set to be visible image in visible and infrared image fusion, while in depth image SR, PAN image is set to be natural image and LRMS is set to be depth image.

2.3. Comparison

From the Tab. 3 and Tab. 4, it can be observed that our method achieved metrics that are almost superior to the SOTA methods on all datasets. Although it did not reach the optimal performance on the Lu dataset in Tab. 4, it achieved the highest average performance metric. This further demonstrates the generalization ability of our method. It is worth noting that in this experiment, we only tested its generalization ability and did not specifically focus on whether it is the state-of-the-art.

3. More Detailed Description about Dataset

In this section, we will delve into the details of the dataset of pan-sharpening , visible and infrared image fusion, and depth image SR .

3.1. Dataset of pan-sharpening

In our experiments, we use three pan-sharpening datasets including WorldViewII, WorldViewIII, and GaoFen2. WorldViewII dataset consists of 760 image pairs for training, and 80 image pairs for testing. WorldViewIII dataset consists of 2150 image pairs for training, and 200 image pairs for testing. GaoFen2 dataset consists of 2712 image pairs for training, and 200 image pairs for testing.

3.2. Dataset of visible and infrared image fusion

In our experiments, we perform extensive experiments on three publicly available datasets: M3FD, RoadScene, and TNO. The M3FD dataset consists of 4200 paired infrared and visible images, with 3900 images designated for training and 300 images for testing. In order to assess the generalizability of our method, we train our model on the M3FD dataset, and evaluate it on the RoadScene and TNO datasets. Since these two datasets do not have a predefined split, we randomly select 25 image pairs from each dataset for comparison purposes.

3.3. Dataset of depth image SR

In our experiments, we utilize three depth image SR datasets: NYU v2, Middlebury, and Lu. The NYU v2 dataset comprises 1449 RGB-D image pairs, while the Middlebury dataset consists of 30 RGB-D image pairs and the Lu dataset contains 6 RGB-D image pairs. For training our proposed network, we utilize the first 1000 RGB-D image pairs from the NYU v2 dataset, and then we evaluate the trained model on the remaining 449 RGB-D image pairs. To generate the low-resolution depth map, we follow [7] experimental protocol, which involves applying bicubic operation at different ratios (×4, ×8, and ×16). We directly test the trained model on the NYU v2 dataset, as well as on the additional Middlebury and Lu datasets.

References

[1] Chunle Guo, Chongyi Li, Jichang Guo, Runmin Cong, Huazhu Fu, and Ping Han. Hierarchical features driven residual learning for depth map super-resolution. *IEEE Transactions on Image Processing*, 28(5):2545–2557, 2018. 1

[2] Mohammad Bagher Akbari Haghighat, Ali Aghagolzadeh, and Hadi Seyedarabi. A non-reference image fusion metric based on mutual information of image features. *Computers & Electrical Engineering*, 37(5):744–756, 2011. 1

[3] Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information fusion*, 14(2):127–135, 2013. 1

[4] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2012. 1

[5] Zhanbo Huang, Jinyuan Liu, Xin Fan, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In *European Conference on Computer Vision*, pages 539–555. Springer, 2022. 1

[6] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 353–369. Springer, 2016. 1

[7] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision*, 129(2):579–600, 2021. 1, 3

[8] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018. 1

[9] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 154–169. Springer, 2016. 1

[10] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Joint image filtering with deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1909–1923, 2019. 1

[11] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022. 1

[12] Si Lu, Xiaofeng Ren, and Feng Liu. Depth enhancement via low-rank matrix completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3390–3397, 2014. 1

[13] Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020. 1

[14] Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):1, 2002. 1

[15] Daniel Scharstein and Chris Pal. Learning conditional random fields for stereo. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 1

[16] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1

[17] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11166–11175, 2019. 1

[18] Alexander Toet. The tno multiband image data collection. *Data in brief*, 15:249–251, 2017. 1

[19] D Wang, J Liu, X Fan, and R Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. arxiv 2022. *arXiv preprint arXiv:2205.11876*. 1

[20] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1838–1847, 2018. 1

[21] Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020. 1

[22] Hao Zhang and Jiayi Ma. Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion. *International Journal of Computer Vision*, 129:2761–2785, 2021. 1

[23] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020. 1

[24] Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1186–1196, 2021. 1

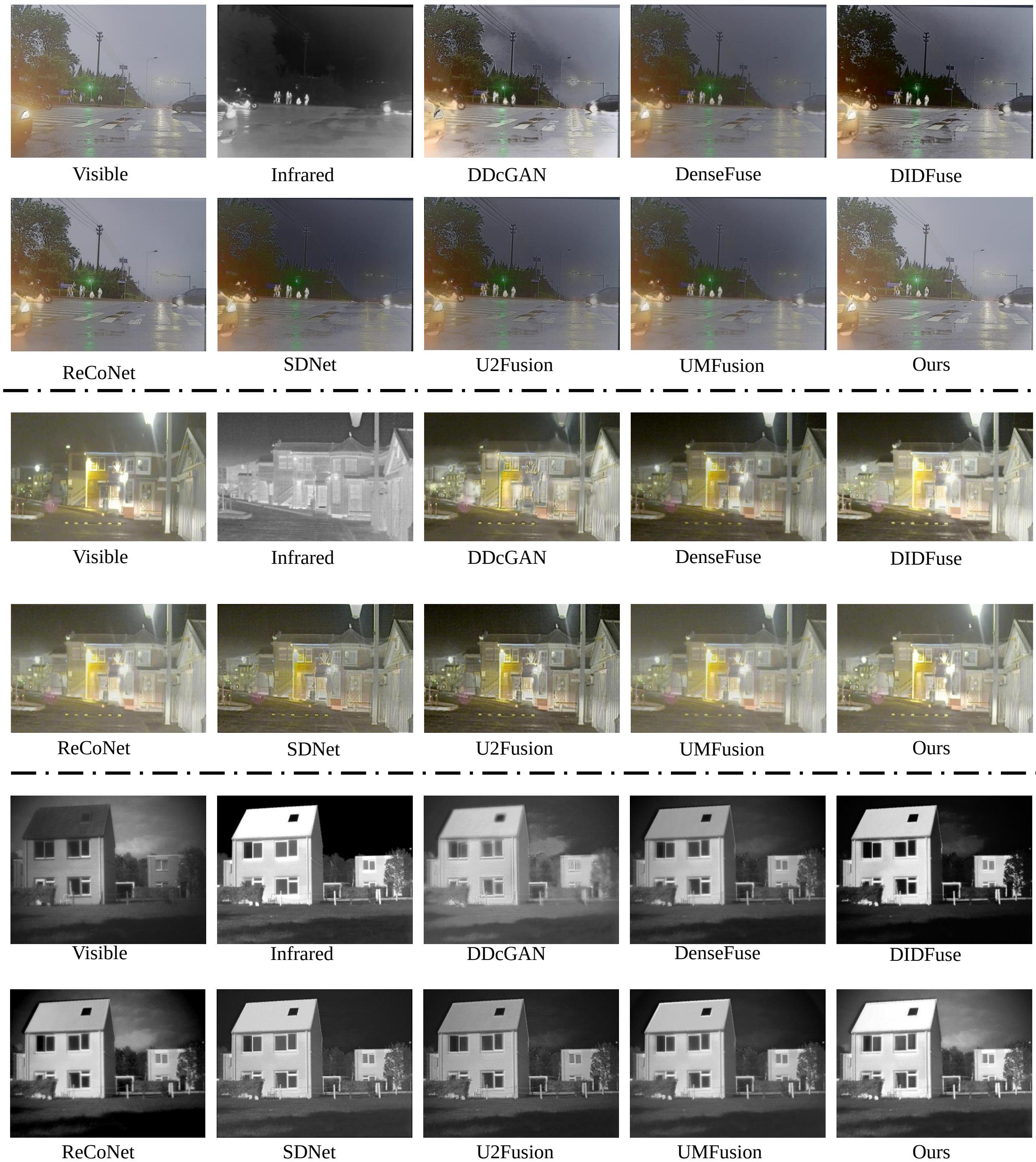

Figure 1. Qualitative results of different methods. From top to bottom: M3FD, RoadScene, and TNO datasets.

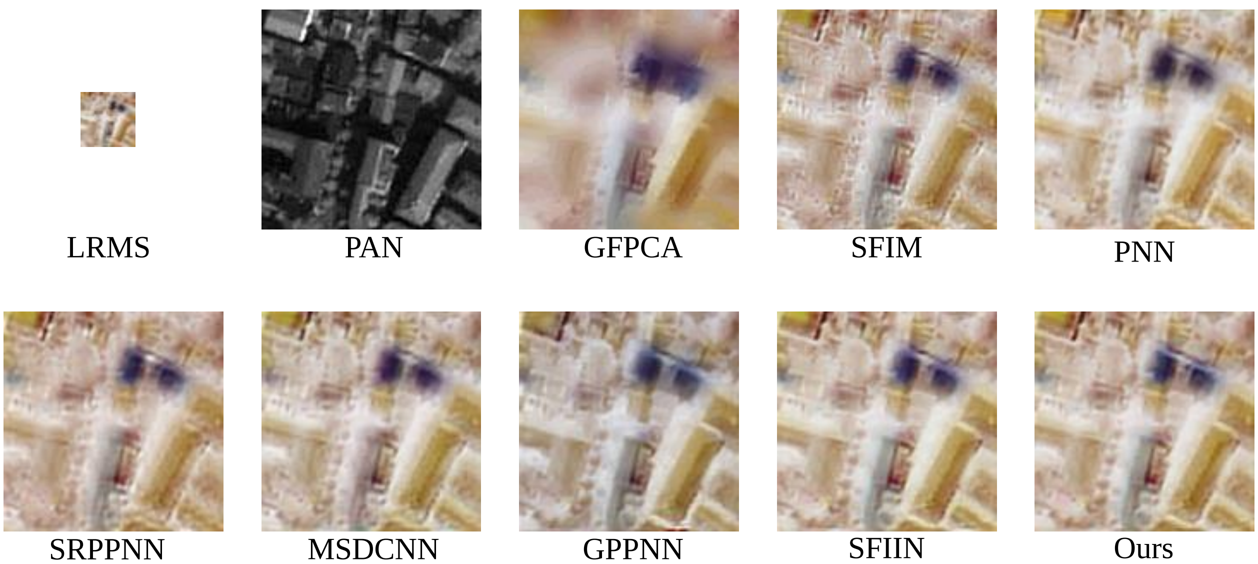

Figure 2. The result of our approach compared with other methods on real-world full-resolution scenes from the GaoFen2 dataset.