STYLE-COHERENT MULTI-MODALITY IMAGE FUSION (SUPPLEMENTAL MATERIAL)

Anonymous authors
Paper under double-blind review

1 FOURIER PRIOR EMBEDDED BLOCK

The Fourier Prior Embedded (FPE) block is designed to capture frequency information of multi-modal features through two main processes: Fourier Spatial Interaction (FSI) and Fourier Channel Interaction (FCI), as illustrated in Fig. 1. Specifically, for a given feature $\boldsymbol{X}$, the fast Fourier transform is first applied, resulting in real and imaginary components denoted as $\text{Real}(\boldsymbol{X})$ and $\text{Im}(\boldsymbol{X})$, respectively. The Fourier Spatial Interaction (FSI) process then operates on these components independently to maintain fidelity in frequency manipulation. This process can be formulated as:

$$\text{Real}(\boldsymbol{X}'_S) = \text{ReLU}(\text{DConv}(\text{Real}(\boldsymbol{X}))), \text{Im}(\boldsymbol{X}'_S) = \text{ReLU}(\text{DConv}(\text{Im}(X))), \tag{1}$$

where $\text{ReLU}(\cdot)$ is the ReLU function, and $\text{DConv}(\cdot)$ indicates the depth-wise convolution. Following FSI, the spatially enhanced feature is merged with the original spatial feature through concatenation and convolution, denoted as $\boldsymbol{X}_S$.

Following the Fourier Spatial Interaction (FSI) process, the spatially enhanced feature $\boldsymbol{X}_S$ undergoes further refinement through FCI. FCI enhances the channel-wise details of the feature frequencies using point-wise convolution, which is formulated as follows:

$$\text{Real}(\boldsymbol{X}'_C) = \text{ReLU}(\text{Conv}_1(\text{Real}(\boldsymbol{X}_S))), \text{Im}(\boldsymbol{X}'_C) = \text{ReLU}(\text{Conv}_1(\text{Im}(\boldsymbol{X}_S))), \tag{2}$$

where Conv_1 indicates the 1×1 convolution. Finally, a similar merging process occurs after the FPI, yielding the output of the FPE module, which achieves global modeling for both spatial and channel dimensions. Overall, FPE effectively extracts frequency information from the modality features of each branch. Subsequently, these features are enhanced across modalities using SNF. The enhanced features then serve as input for the next FPE block.

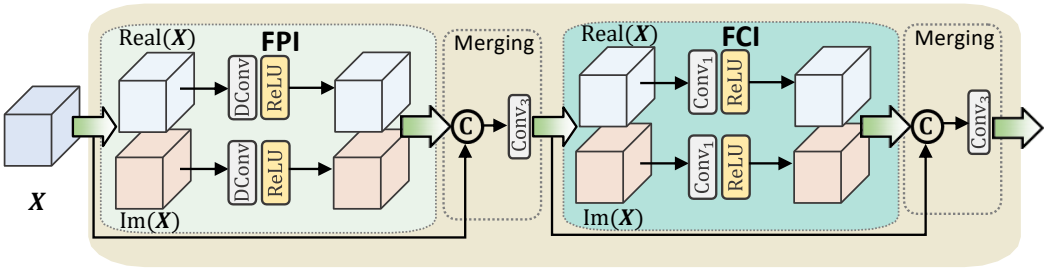

Figure 1: The architecture of FPE.

2 MORE TRAINING DETAILS

For the IVF task, we directly quote results on the MSRS [1] and RoadScene[2] datasets from (Zhao et al., 2023a; 2024). By training the available codes, we obtain results on the TNO[3] dataset.

[1] https://github.com/Linfeng-Tang/MSRS
[2] https://github.com/hanna-xu/RoadScene
[3] https://figshare.com/articles/dataset/TNOImageFusionDataset/1008029

For the MIF task, we obtain results on the Harvard medical [4] dataset by applying the trained IVF models to the MIF task without fine-tuning.

For the BIF task, we use the released code to retrain models with the same data settings as ours, obtaining results on the ATC [5] dataset. The training data is cropped to patches of size 256×256.

For downstream applications, we follow (Zhao et al., 2023a; 2024) for implementing semantic segmentation and object detection. For semantic segmentation, we directly utilize trained IVF models to obtain fusion images on the MSRS dataset. Following the dataset splits in (Tang et al., 2022b), we retrain the DeeplabV3+ segmentation model with cross-entropy loss using suggested hyperparameter settings from the released code[6]. For object detection, We utilize trained IVF models and obtain fusion images on the M^3FD[7] dataset, where 3,360 images are used for training, 420 images for validation, and 420 images for testing. We retrain the YoLo+ detection model using suggested hyperparameter settings from the released code[8].

3 Biological Image Fusion

From Tab. 1, it can be observed that our proposed method outperformed the other MMIF methods in most of the evaluation metrics. This comprehensively demonstrates the superior performance of our proposed style-coherent fusion model (SCFNet). Fig. 2 shows that our SCFNet effectively captures cellular structural features from PC while suppressing noise from GFP.

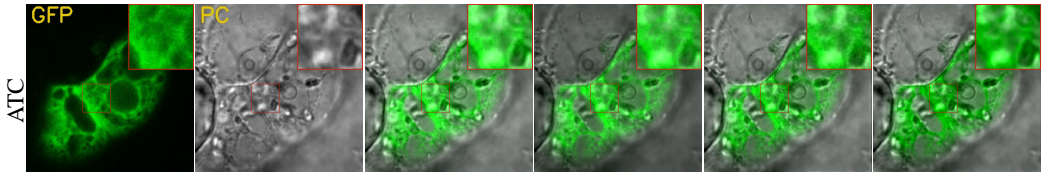

Figure 2: More visualizations of BIF on ATC (Koroleva et al., 2005) dataset.

Table 1: Quantitative results of BIF on ATC dataset.

BIF on ATC dataset							
Methods	EN↑	SD↑	SF↑	AG↑	Qbaf↑	VIF↑	SSIM↑
TarD	6.08	40.22	9.48	2.31	0.52	0.86	0.85
DeF	6.46	42.63	9.60	2.80	0.61	0.92	0.92
MURF	6.15	41.82	9.91	2.67	0.60	0.96	0.81
CDDFuse	6.70	48.38	12.56	3.73	<u>0.63</u>	**1.05**	1.00
DDFM	6.53	47.04	11.44	2.51	0.59	0.95	0.96
EMMA	<u>6.71</u>	<u>49.13</u>	<u>12.58</u>	<u>3.76</u>	0.58	0.97	<u>1.04</u>
SCFNet	**6.82**	**51.34**	**14.01**	**4.04**	**0.64**	<u>1.01</u>	**1.15**

4 More Analyses

4.1 Analyses of style-alignment fusion

We further analyze that SAF enables the selection of alignment across different modalities, including visible or infrared domains. It is important to note that SAF uses a well-defined source distribution to guide the diverse features into a unified domain, rather than directly defining the target domain.

[4] http://www.med.harvard.edu/AANLIB/home.html
[5] http://data.jic.bbsrc.ac.uk/gfp
[6] https://github.com/VainF/DeepLabV3Plus-Pytorch
[7] https://github.com/JinyuanLiu-CV/TarDAL
[8] https://github.com/ultralytics/yolov5

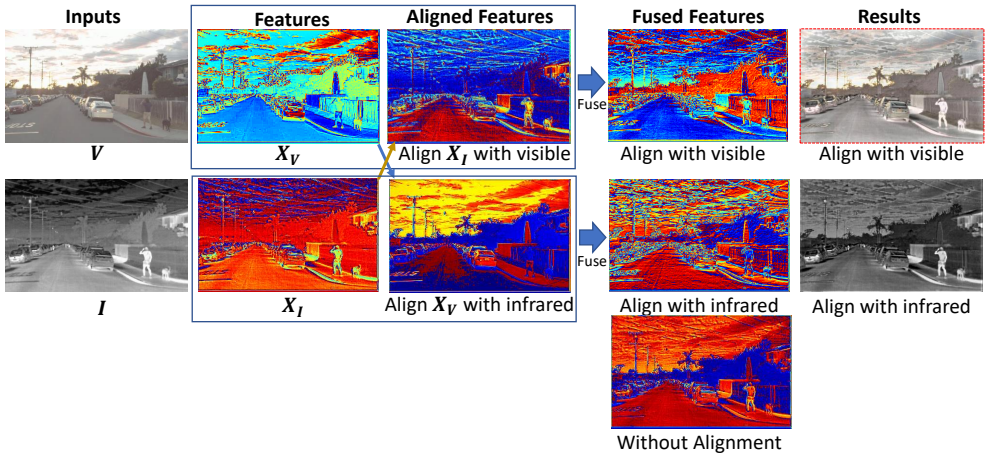

Figure 3: The visualization results of the feature alignment strategies based on style-alignment fusion module (SAF). It includes the alignment of features with the visible and infrared domains, as well as the fused results. The fused features with alignment retain more scene details. The result aligned with the visible domain exhibit superior visual performance.

Table 2: Ablation studies of the selection of alignment across different domains in SAF on RoadScene dataset.

SAF	EN ↑	SF ↑	Qbaf↑	VIF ↑	SSIM↑
Align w/ infrared domain	**7.63**	**19.06**	0.53	0.68	1.03
Align w/ visible domain (Original)	7.55	18.32	**0.56**	**0.72**	**1.21**

Compared to the channel-wise fusion of features without alignment, the features fused using SAF exhibit clearer details in the sky and vehicles. This improvement highlights the effectiveness of the SAF approach in preserving and enhancing critical scene elements. When SAF aligns the infrared domain, as shown in Fig. 3, the explicitly aligned infrared domain $\boldsymbol{X}_V$ exhibits higher contrast on vehicles and more prominent landmarks. The resulting fusion preserves fine details, but some edges appear overly accentuated. When SAF aligns with the visible domain, the explicitly aligned $\boldsymbol{X}_I$ retains complete information on thermal tasks and power lines. As illustrated in the third row of Fig. 3, the aligned fusion features preserve the complete scene details of the source modalities, differing only in the visual effects of the reconstructed images. The fusion result with visible domain alignment is more visually satisfactory, striking a better balance between preserving details and maintaining natural appearance.

From the results in Tab. 2, it is evident that although aligning with the infrared domain increase the information entropy of the fused image, the visual quality assessments including metrics such as Qbaf, VIF, and SSIM tend to significantly decrease. Therefore, we prioritize alignment with the domain that contains more information to avoid excessive adjustments that could deteriorate the visual quality of the fused image.

4.2 Analyses of adaptive reconstruction loss

Based on the ablation study in Sec. 4.5 of the main paper that compares different training strategies, we provide further visual comparisons to validate the effectiveness of our proposed adaptive reconstruction loss function.

First, we compare the image-level supervision signals ($\text{Max}(\boldsymbol{V}, \boldsymbol{I})$) generated by traditional loss functions (Zhao et al., 2023a; Tang et al., 2022a) and our proposed loss function with supervision signals ($\text{Max}(\mathcal{R}(\boldsymbol{V}), \mathcal{R}(\boldsymbol{I}))$). It is evident that the learnable rescaled function $\mathcal{R}$, defined in Eq.9 in the main paper, significantly enhances the utilization of multi-modality images to supervise the model, resulting in more comprehensive information, particularly for people obscured by smoke

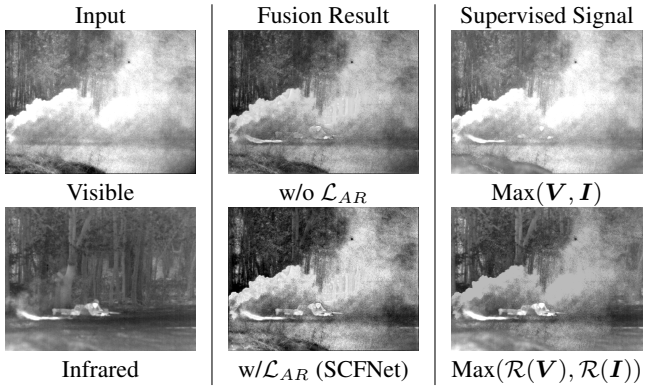

Figure 4: Visualizations of ablation studies for $\mathcal{L}_{AR}$. The supervision signals in $\mathcal{L}_{AR}$, compared to existing loss functions, guide our SCFNet in generating complete scene details with higher contrast.

Table 3: Ablation studies of β in Eq.9 and operations in Eq.11 on RoadScene dataset.

	Methods	EN↑	SD↑	SF↑	Qbaf↑	VIF↑
β in Eq.9	Mean($\boldsymbol{I}$)	7.31	53.72	17.61	0.52	0.70
	Learnable	7.41	53.90	18.04	0.54	0.71
	Mean($\boldsymbol{V}$) (Original)	**7.55**	**55.29**	**18.32**	**0.56**	**0.72**
Operations in Eq.11	Mean($\cdot$)	6.67	48.5	15.13	0.53	0.66
	Separation	7.42	53.14	17.99	0.54	0.71
	Max($\cdot$) (Original)	**7.55**	**55.29**	**18.32**	**0.56**	**0.72**

and details of tress, while enhancing contrast and suppressing noise to ensure that the blurry, noisy visible image content is effectively presented. Additionally, the visualization of "w/o $\mathcal{L}_{AR}$" further demonstrates that fusion models trained without the $\mathcal{L}_{AR}$ loss are prone to degradation, resulting in low image contrast, making the person in smoke less noticeable. Therefore, the proposed adaptive reconstruction loss function addresses the absence of GT and effectively guides the model to generate high-quality images with complete information, facilitating further improvements in downstream applications such as detection.

We specifically analyze the β of the main paper Eq.9, which is set to Mean($\boldsymbol{X}$) to align $\hat{\boldsymbol{X}}_{\boldsymbol{F}}$ with the distribution of the visible image domain. The results in Tab. 3 indicate that setting $\beta =$ Mean(I), denoted as "Mean($\boldsymbol{I}$)", significantly reduces the effectiveness of fusion. This reduction in performance is primarily attributable to alignment conflicts with our proposed SAF which is oriented toward the visible domain. We also set β learnable and constrain it within [Mean($\boldsymbol{I}$), Mean($\boldsymbol{V}$)], denoted as "Learnable". Due to the negative impacts of the alignment conflict being mitigated, this results in a less pronounced reduction in performance. However, introducing too many learnable variables for the supervision signal leads to training instability. Consequently, fixing β at Mean(V) enables a more effective and stable adjustment of source images to align with the visible domain.

Additionally, in the main paper Eq.11, We follow Zhao et al. (2023a); Tang et al. (2022a) by utilizing the maximum pixel values as the supervision signal to enhance the overall clarity of the fused image. As shown in the following Tab. 3, replacing with the mean operation, denoted as "Mean($\cdot$)", leads to a significant decrease in performance and loss of a substantial amount of detail. Separate supervisions with $R(I)$ and $R(V)$ as Xu et al. (2022); Zhao et al. (2020); Xu et al. (2020a), denoted as "Separation". It also leads to decreased performance, as the fusion model tends to blend the two images with a smoothing effect.

5 LIMITATIONS

Similar to existing methods (Liu et al., 2023; Sun et al., 2022; Tang et al., 2022a; Zhao et al., 2024; 2023b; Yi et al., 2024), the proposed SCFNet primarily focuses on geometrically calibrated multi-

modality images. However, in some cases, capturing simultaneously from the same scene can be challenging due to differences in sensor perspectives and positions, often resulting in misalignments in the collected multi-modality images. Effectively training our SCFNet with misaligned multi-modality images will be the focus of our future work.

6 More Visualizations

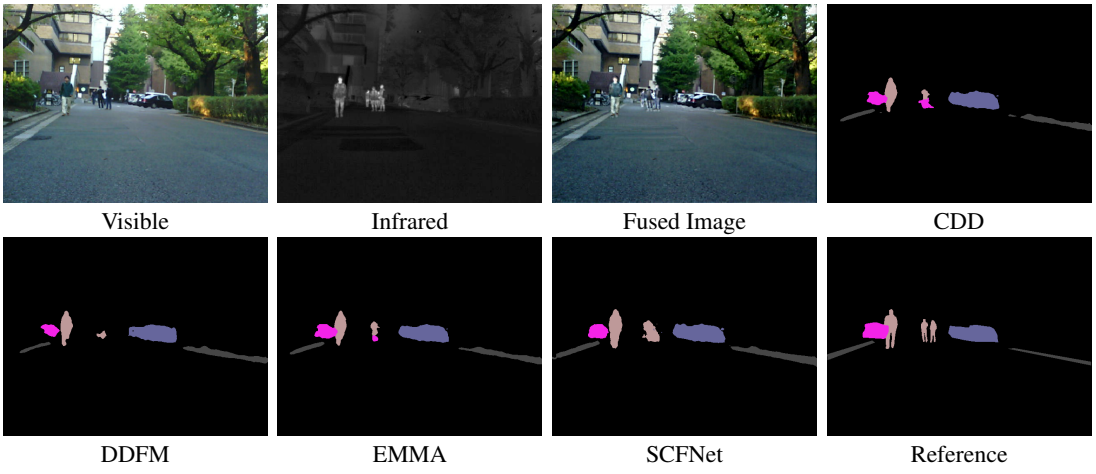

Figure 5: Visualizations of semantic segmentation on MSRS (Tang et al., 2022b) dataset.

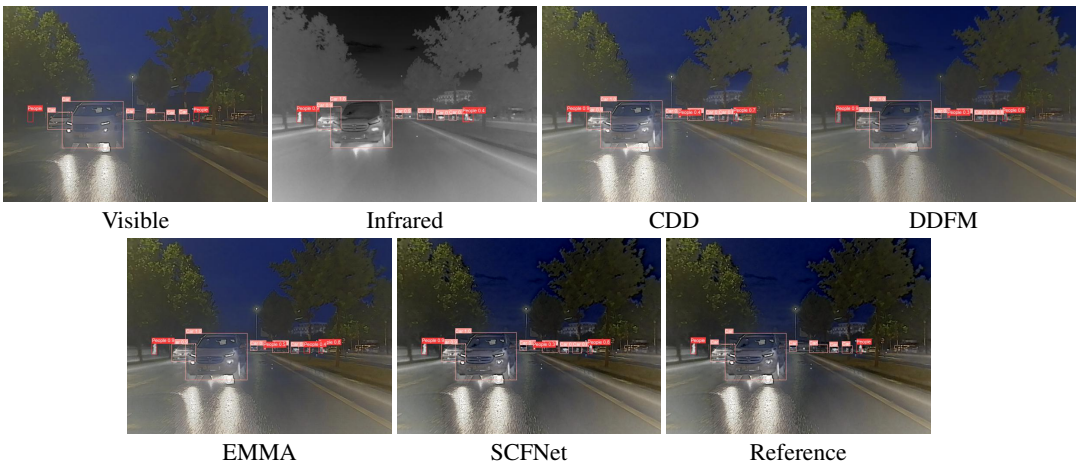

Figure 6: More visualizations of object detection on M^3FD (Liu et al., 2022) dataset.

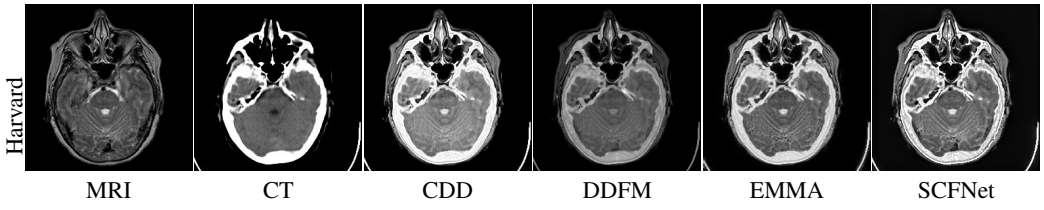

Figure 7: More visualizations of MIF on Harvard Medical (website) dataset.

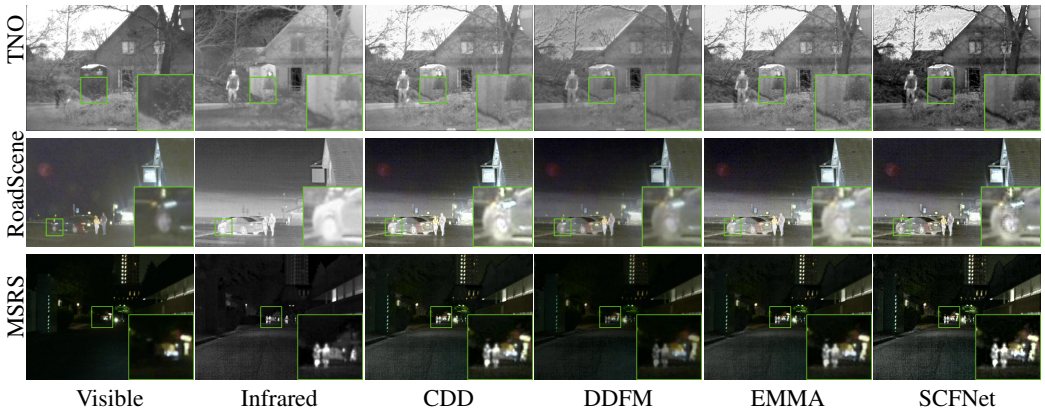

Figure 8: More visualizations of IVF on TNO (Toet & Hogervorst, 2012), RoadScene (Xu et al., 2020b), and MSRS (Tang et al., 2022b) datasets.

REFERENCES

Olga A Koroleva, Matthew L Tomlinson, David Leader, Peter Shaw, and John H Doonan. High-throughput protein localization in arabidopsis using agrobacterium-mediated transient expression of gfp-orf fusions. *The Plant Journal*, 41(1):162–174, 2005.

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*, pp. 5802–5811, 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8115–8124, 2023.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 4003–4011, 2022.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022a.

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, 2022b.

Alexander Toet and Maarten A Hogervorst. Progress in color night vision. *Optical Engineering*, 51 (1):010901–010901, 2012.

Harvard Medical website. http://www.med.harvard.edu/aanlib/home.html. 7, 8.

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, 2020a.

Han Xu, Jiayi Ma, Zhuliang Le, Junjun Jiang, and Xiaojie Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12484–12491, 2020b.

Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19679–19688, 2022.

Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, pp. 102450, 2024.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *Proceedings of IJCAI*, 2020.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5906–5916, 2023a.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8082–8093, 2023b.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.