
Appendix for Efficient Rectified Flow for Image Fusion

Anonymous Author(s)

Affiliation

Address

email

1 A Detailed derivation of formulas

2 In this section, we provide a detailed proof regarding posterior sampling in diffusion-based image
3 fusion methods. The noise predicted by the diffusion model at time step t is often related to the score
4 at the current time step. According to [4], the specific formulation can be expressed as:

$$\epsilon_\phi(x_t, t) = -\sqrt{1 - \alpha_t} \nabla_{x_t} \log p(x_t), \quad (1)$$

5 In posterior sampling, we also need to take into account the guidance from image fusion, denoted
6 as i, v . Therefore, what we need to solve is $\nabla_{f_t} \log p(f_t | i, v)$, which can be expressed using Bayes'
7 theorem as:

$$\begin{aligned} p_t(f_t | i, v) &= \frac{p_t(i, v | f_t) \cdot p_t(f_t)}{p_t(i, v)}, \\ \Rightarrow \log p_t(f_t | i, v) &= \log p_t(i, v | f_t) + \log p_t(f_t) - \log p_t(i, v), \\ \Rightarrow \nabla_{f_t} \log p_t(f_t | i, v) &= \nabla_{f_t} \log p_t(f_t) + \nabla_{f_t} \log p_t(i, v | f_t). \end{aligned} \quad (2)$$

8 Among them, $\nabla_{f_t} \log p(f_t | i, v)$ and $\nabla_{f_t} \log p(f_t)$ can be expressed by Equation 1 as follows:

$$\begin{aligned} \epsilon_\phi(f_t, t) &= -\sqrt{1 - \alpha_t} \nabla_{f_t} \log p_t(f_t), \\ \epsilon'_\phi(f_t, t | i, v) &= -\sqrt{1 - \alpha_t} \nabla_{f_t} \log p_t(f_t | i, v). \end{aligned} \quad (3)$$

9 Therefore, the final equation can be expressed as:

$$\begin{aligned} \epsilon'_\phi &= \epsilon_\phi(f_t, t) - \sqrt{1 - \alpha_t} \nabla_{f_t} \log p_t(i, v | f_t), \\ &\approx \epsilon_\phi(f_t, t) - \sqrt{1 - \alpha_t} \nabla_{f_t} \log p_t(i, v | \tilde{f}_{0|t}), \\ &\approx \epsilon_\phi(f_t, t) - \rho \sqrt{1 - \alpha_t} \nabla_{f_t} \|i, v - \mathcal{M}(\hat{f}_0(f_t))\|_2^2. \end{aligned} \quad (4)$$

10 Therefore, we inject the image fusion prior by correcting the predicted noise during the sampling
11 process, thereby achieving high-quality image fusion based on the diffusion model.

12 B Details about the training loss

13 In the main text, the loss $\mathcal{L}_{\text{fusion}}$ used in the training stage II is defined as follows:

$$\mathcal{L}_{\text{fusion}} = \lambda_{\text{int}} \mathcal{L}_{\text{int}} + \lambda_{\text{SSIM}} \mathcal{L}_{\text{SSIM}} + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}} + \lambda_{\text{color}} \mathcal{L}_{\text{color}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}. \quad (5)$$

14 Followed by [6], we use the Intensity Loss to encourage the model to focus on salient features in the
15 fused image, which is specifically defined as:

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} \|I_f - \max(I_{\text{vis}}^g, I_{\text{ir}}^g)\|_1. \quad (6)$$

16 Here, I_{vis}^g and I_{ir}^g are the ground truth corresponding to the fused image. We also use $\mathcal{L}_{\text{SSIM}}$ to train
17 the model so that the fused image is structurally as similar as possible to the two input images. It is
18 defined as:

$$\mathcal{L}_{\text{SSIM}} = (1 - \text{SSIM}(I_f, I_{\text{vis}}^g)) + \mu(1 - \text{SSIM}(I_f, I_{\text{ir}}^g)). \quad (7)$$

19 We also compute the gradient loss $\mathcal{L}_{\text{grad}}$ to ensure the similarity between the fused image and the
 20 input images in terms of edge features:

$$\mathcal{L}_{\text{grad}} = \frac{1}{HW} \|\nabla I_f - \max(\nabla I_{\text{vis}}^g, \nabla I_{\text{ir}}^g)\|_1, \quad (8)$$

21 and use $\mathcal{L}_{\text{color}}$ to keep consistent color with input images:

$$\mathcal{L}_{\text{color}} = \frac{1}{HW} \|\mathcal{F}_{CbCr}(I_f) - \mathcal{F}_{CbCr}(I_{\text{vis}}^g)\|_1. \quad (9)$$

22 As mentioned in the main text, we use $\mathcal{L}_{\text{mask}}$ for saliency-guided regional fusion, which can be
 23 represented as:

$$\mathcal{L}_{\text{mask}} = \|\mathcal{W}_v \cdot I_v + \mathcal{W}_{ir} \cdot I_{ir} - I_f\|_1. \quad (10)$$

24 Here, \mathcal{W}_v and \mathcal{W}_{ir} denote the saliency-based weight maps computed from the corresponding input
 25 images. \mathcal{W}_v and \mathcal{W}_{ir} are computed based on pixel-level visual saliency maps. Specifically, they are
 26 estimated by measuring the sparsity of the image pixel histograms: the sparser the pixel distribution,
 27 the higher the corresponding saliency. The computation can be formulated as:

$$\text{Saliency}(i) = \sum_{j=0}^{255} |i - j| \cdot \text{Hist}(j). \quad (11)$$

28 Here, i represents the grayscale value of the current pixel, and j denotes the grayscale value of the
 29 traversed pixels. By computing the saliency values $S_{ir}(i, j)$ and $S_v(i, j)$ for each pixel, we can obtain
 30 the corresponding \mathcal{W}_v and \mathcal{W}_{ir} . The specific formulas are given as:

$$\mathcal{W}_v(i, j) = \mu_v + S_v(i, j) - \mu_v \cdot S_{ir}(i, j) \quad (12)$$

$$\mathcal{W}_{ir}(i, j) = 1 - \mathcal{W}_v(i, j) \quad (13)$$

32 C Additional experiments of our method

Table 1: Quantitative comparison on Lytro, MFFW and MFI-WHU datasets. The best and second best results are highlighted in **bold** and underline.

Dataset	Lytro				MFFW				MFI-WHU			
Method	MI	CC	Qcb	PSNR	MI	CC	Qcb	PSNR	MI	CC	Qcb	PSNR
DeFusion [2]	6.27	0.97	0.59	77.2	5.59	0.97	0.55	74.4	6.01	0.97	0.69	77.2
TC-MoA [9]	7.45	0.97	0.76	74.8	5.34	0.96	0.63	72.8	6.76	0.97	0.75	75.6
Text-IF [6]	5.63	0.97	<u>0.65</u>	71.9	5.26	0.96	<u>0.61</u>	70.2	5.31	0.97	0.62	72.3
DDFM [8]	3.53	0.85	0.41	67.2	3.33	0.73	0.38	64.5	2.99	0.74	0.43	66.4
CCF [1]	5.15	0.96	0.49	66.8	4.47	0.95	0.47	66.6	4.95	0.96	0.47	67.1
Ours	<u>6.58</u>	0.98	0.61	<u>76.1</u>	5.80	0.97	0.55	71.7	<u>6.38</u>	0.98	<u>0.71</u>	<u>75.7</u>

33 **More Quantitative Comparison of Multi-Focus Fusion** As shown in Table 1, we conducted
 34 comparisons on three datasets in MFIF: Lytro [3], MFFW [5], and MFI-WHU [7]. Our method
 35 outperforms other approaches on most metrics. Specifically, it achieves either the first or second best
 36 performance¹ on 9 different metrics, surpassing other comparison methods and demonstrating the
 37 superiority of our approach in the Multi-Focus Fusion task.

38 References

- 39 [1] Bing Cao, Xingxin Xu, Pengfei Zhu, Qilong Wang, and Qinghua Hu. Conditional controllable
 40 image fusion. *arXiv preprint arXiv:2411.01573*, 2024.
- 41 [2] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A
 42 self-supervised decomposition approach for image fusion. In *European Conference on Computer*
 43 *Vision*, pages 719–735. Springer, 2022.

¹Compare the percentiles of the same value.

- 44 [3] Mansour Nejati, Shadrokh Samavi, and Shahram Shirani. Multi-focus image fusion using
45 dictionary-based sparse representation. *Information fusion*, 25:72–84, 2015.
- 46 [4] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and
47 Ben Poole. Score-based generative modeling through stochastic differential equations. In
48 *International Conference on Learning Representations*, 2021.
- 49 [5] Shuang Xu, Xiaoli Wei, Chunxia Zhang, Junmin Liu, and Jianshe Zhang. Mffw: A new dataset
50 for multi-focus image fusion. *arXiv preprint arXiv:2002.04780*, 2020.
- 51 [6] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text
52 guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF*
53 *Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024.
- 54 [7] Hao Zhang, Zhuliang Le, Zhenfeng Shao, Han Xu, and Jiayi Ma. Mff-gan: An unsupervised
55 generative adversarial network with adaptive and gradient joint constraints for multi-focus image
56 fusion. *Information Fusion*, 66:40–53, 2021.
- 57 [8] Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jianshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang,
58 Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: denoising diffusion model for multi-
59 modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer*
60 *Vision*, pages 8082–8093, 2023.
- 61 [9] Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for
62 general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and*
63 *pattern recognition*, pages 7099–7108, 2024.