AFUNet: Cross-Iterative Alignment-Fusion Synergy for HDR Reconstruction via Deep Unfolding Paradigm Supplementary Material

Xinyue Li[1], Zhangkai Ni[1*], Wenhan Yang[2]
[1]Tongji University, [2]Pengcheng Laboratory
2252065@tongji.edu.cn, zkni@tongji.edu.cn, yangwh@pcl.ac.cn

This supplementary material serves as an appendix to our main paper. In Section 1, we present the details of the Half Quadratic Splitting (HQS) Algorithm utilized in our formulation. Section 2 presents the algorithm flow of our proposed method AFUNet. In Section 3, we provide more experiment details. In Section 4, we provide the comparison of computational costs of AFUNet. Section 5 presents our additional quantitative comparison of some classical and state-of-the-art methods using perceptual metrics. Section 6 presents the visual analysis of our proposed unfolding framework, focusing on the feature maps of variables in our formulation. In Section 7, we further conduct ablation studies, giving more comprehensive visual results comparison and analysis on the effectiveness of tailored components, the number of iterative stages, and different paradigms. Section 8 showcases more visual reconstruction results of AFUNet.

1. The Formulation of Half Quadratic Splitting

Half Quadratic Splitting (HQS) [2] technique has been proven effective in solving maximum a posterior (MAP) problems, involving the data fidelity term and regularization terms, which is:

$$\hat{x} = \arg\min_{x} \frac{1}{2}\|y - Ax\|_2^2 + \lambda_1 \Psi_1(x) + \lambda_3 \Psi_3(x), \quad (1)$$

where λ_1 and λ_3 are weighting parameters that control the penalty strength of the regularizers.

The optimization problem Eq. (1) can be solved by introducing an auxiliary variable u and v to match the regularization terms $\Psi_1(\cdot)$ and $\Psi_3(\cdot)$, respectively.

$$\begin{aligned} \arg\min_{x,u,v} \frac{1}{2}\|y - Ax\|_2^2 + \lambda_1 \Psi_1(u) + \lambda_3 \Psi_3(v). \\ \text{s.t.} \quad u = x, v = x \end{aligned} \quad (2)$$

The HQS method solves the following problem, reformulated from Eq. (2):

$$\begin{aligned} \arg\min_{x,u,v} \frac{1}{2}\|y - Ax\|_2^2 + \lambda_1 \Psi_1(u) + \lambda_3 \Psi_3(v) \\ + \frac{\beta_1}{2}\|u - x\|_2^2 + \frac{\beta_3}{2}\|v - x\|_2^2, \end{aligned} \quad (3)$$

where β_1 and β_3 are penalty parameters of the Lagrangian term. Eq. (3) can be solved in an iterative strategy, HQS optimizes u, v, x in an alternating fashion by solving the following three subproblems separately:

$$u^t = \arg\min_{u} \frac{\beta_1}{2}\|u - x^{t-1}\|_2^2 + \lambda_1 \Psi_1(u), \quad (4a)$$

$$v^t = \arg\min_{v} \frac{\beta_3}{2}\|v - x^{t-1}\|_2^2 + \lambda_3 \Psi_3(v), \quad (4b)$$

$$x^t = \arg\min_{x} \frac{1}{2}\|y_2 - D_2 x\|_2^2 + \frac{\beta_1}{2}\|u^t - x\|_2^2 + \frac{\beta_3}{2}\|v^t - x\|_2^2. \quad (4c)$$

The update processes of u^t and v^t in Eq. (4a) and Eq. (4b) are non-trivial and need to adopt deep neural networks as regularizers. Note that Eq. (4c) is a least-squares problem with a quadratic penalty term with a closed-form solution.

HQS method can decouple the data fidelity terms and the regularization terms by introducing auxiliary variables, splitting the complex problem into several subproblems, and then tackling them in an iterative and alternating manner. Each variable, u, v and x, has its distinct physical meaning in the optimization problem. The optimization of these three variables in Eq. (4a), Eq. (4b) and Eq. (4c) can complement each other, approaching the ultimate optimal solution from different starting points. This design increases the likelihood of finding the global optimum and avoids the risk of getting trapped in local optima.

*Corresponding author.

Algorithm 1 Proposed AFUNet

Input: LDR images y_1, y_2, y_3, the number of reconstruction stages T.

Output: Final reconstructed HDR image $\hat{x}$.

Initialization:

Initialize the stage number $t = 1$ and ceiling T.

Initialize the reconstructed feature $f_x^0 = f_{y_2}$, alignment variable features $f_{\alpha_1}^0 = f_{y_1}$, $f_{\alpha_3}^0 = f_{y_3}$.

HDR Feature Reconstruction:

while $t \leq T$ **do**

  Feature Alignment:

  Update $f_{\alpha_1}^t$ by α_1-SAM ▷ Alignment Problem

  Update $f_{\alpha_3}^t$ by α_3-SAM ▷ Alignment Problem

  Feature Fusion:

  Update $f_{u_s}^t$, $f_{v_s}^t$ by SFM ▷ Fusion Problem

  Update f_u^t by U-CFM ▷ Fusion Problem

  Update f_v^t by V-CFM ▷ Fusion Problem

  Update $f_{x_p}^t$ by DCM ▷ Fusion Problem

  Update f_x^t from $\{f_u^t, f_{x_p}^t, f_v^t\}$ ▷ Fusion Problem

  $t = t + 1$ ▷ Next Stage

end while

HDR Image Reconstruction:

$\hat{x} = \mathrm{Sigmoid}(\mathrm{Conv}(f_x^{\mathrm{T}} + \mathrm{Conv}(f_{y_2})))$

Output reconstructed HDR image $\hat{x}$.

2. Algorithm

Our AFUNet consists of three processes: Initialization, HDR Feature Reconstruction, and HDR Image Reconstruction. HDR Feature Reconstruction can be further divided into two subprocesses, *i.e.,* Feature Alignment and Feature Fusion. In the Feature Alignment subprocess, we update $f_{\alpha_1}^t$ and $f_{\alpha_3}^t$. In the Feature Fusion subprocess, we update f_u^t, f_v^t and then f_x^t. The algorithm of our proposed method AFUNet is summarized in Algorithm 1.

3. Experiment Details

In this section, we present additional experiment details that are not included in the main paper due to space limitations.

Dataset Details. All methods are trained using three publicly available datasets, employing identical training settings: Kalantari's dataset [5] consists of 74 samples for training and 15 for testing, all captured under authentic environmental conditions. Each sample comprises three LDR images with exposure variations of { -2, 0, 2 } or { -3, 0, 3 }. Tel's dataset [9] consists of 108 samples for training and 36 for testing, similar to Kalantari's dataset, all captured under real-world conditions. Each sample comprises three LDR images with exposure variations of { -2, 0, 2 }. Different from Kalantari's dataset and Tel's dataset, Hu's dataset [4] is a synthetic dataset designed to emulate sensor realism, generated through a game engine. This dataset contains images captured at three distinct exposure levels { -2, 0, 2 }. We use the initial 85 samples for training and the remaining 15 samples for testing following [4].

Training Details. We harness the Adam optimization strategy [6] along with a cosine annealing scheme. The channel C is set to 72. The number of attention heads for cross-attention and self-attention in the Spatial Alignment Module (SAM) and Spatial Fusion Module (SFM) are both set to 4 in all iterative stages.

4. Computational Costs

The computational costs of our proposed method are presented in Tab. 1, which compares the inference time and parameter quantities across different methods. Among them, AFUNet demonstrates the fastest inference speed while maintaining a relatively small number of parameters. This indicates that AFUNet effectively balances computational complexity and performance.

Method	CA-ViT	SCTNet	DiffHDR	AFUNet
Time (s)	<u>2.445s</u>	3.399s	10.222s	**1.940s**
Param. (M)	1.22M	**0.99M**	74.99M	<u>1.16M</u>

Table 1. The inference time and parameters of different methods. For each item, the best result is blodfaced, and the second best is underlined.

5. Perceptual Metrics

As illustrated in Tab. 2, we additionally compute various common perceptual metrics, including Fréchet Inception Distance (FID) [3], Learned Perceptual Image Patch Similarity (LPIPS) [15], Visual Saliency-based Index (VSI) [14], and Deep Image Structure and Texture Similarity (DISTS) [1]. Tonemaping is applied for computing these perceptual metrics due to the domain difference between HDR images and natural images. According to the results shown in Tab. 2, our method AFUNet still performs well on perceptual metrics, with the strong capability to reconstruct high-quality HDR images that are visually satisfactory and aligned with human perception.

6. Visual Analysis of Feature maps

In this section, we give a further analysis of our proposed cross-iterative Alignment and Fusion deep Unfolding Network by extracting the feature maps of main variables from each iterative stage in our formulation. Fig. 1 illustrates the feature reconstruction process within the feature domain, with annotations highlighting the intermediate variables of the feature maps we will present in the following. Specif-

Datasets	Models	GT	DHDR[11]	AHDR[12]	HDR-GAN[8]	CA-ViT[7]	DiffHDR[13]	SCTNet[9]	Ours
Kalantari [5]	FID ↓	0	12.97	12.26	10.71	8.39	7.83	8.53	7.62
	LIPIS ↓	0	0.0094	0.0075	0.0070	0.0054	0.0058	0.0060	0.0053
	DISTS ↓	0	0.0129	0.0107	0.0090	0.0067	0.0065	0.0060	0.0061
	VSI ↑	100	99.77	99.71	99.75	99.80	99.80	99.78	99.81
Hu [4]	FID ↓	0	11.91	12.08	10.27	4.56	3.98	4.04	3.74
	LIPIS ↓	0	0.0561	0.1049	0.0547	0.0506	0.0435	0.0037	0.0035
	DISTS ↓	0	0.0374	0.0835	0.0362	0.0289	0.0231	0.0029	0.0027
	VSI ↑	100	0.9256	0.8605	0.9600	0.9712	0.9745	0.9985	0.9985
Tel [9]	FID ↓	0	12.20	9.95	15.48	8.60	9.28	5.68	6.52
	LIPIS ↓	0	0.012	0.085	0.011	0.0074	0.0075	0.0074	0.0073
	DISTS ↓	0	0.0225	0.0155	0.0208	0.0148	0.0156	0.0107	0.0147
	VSI ↑	100	0.9971	0.9978	0.9979	0.9984	0.9985	0.9981	0.9985

Table 2. Quantitative comparison of proposed network with several state-of-the-art methods on Kalantari's dataset [5], Hu's dataset [4], and Tel's dataset [9].

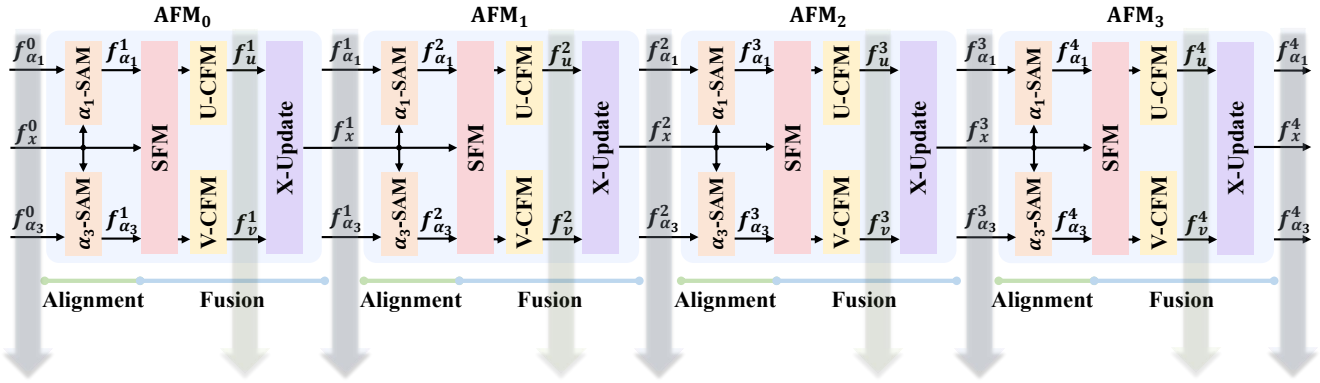

Figure 1. We extract intermediate variables' feature maps from the feature reconstruction process of AFUNet. The variables are highlighted in the figure, involving the initial three variables—$f^0_{\alpha_1}$, f^0_x, $f^0_{\alpha_3}$ and $f^t_{\alpha_1}$, f^t_x, $f^t_{\alpha_3}$, f^t_u, f^t_v in each stage, $t = 1, 2, 3, 4$.

ically, we display the feature maps of the initial three variables $f^0_{\alpha_1}$, f^0_x, $f^0_{\alpha_3}$, as well as the $f^t_{\alpha_1}$, f^t_x, $f^t_{\alpha_3}$, f^t_v, and f^t_u at each stage within the process. As shown in Fig. 2 and Fig. 3, the top three rows demonstrate the initial and each stage variables $f^t_{\alpha_1}$, f^t_x and $f^t_{\alpha_3}$ $(t = 0, 1, \ldots, 4)$. The bottom two rows demonstrate f^t_u and f^t_v $(t = 1, 2, 3, 4)$. Examining Fig. 2 and Fig. 3, we can observe clear misalignment issues, which are caused by large motion in dynamic scenes, in the 0th and 1st stages. While going through the iterative stages, these artifacts are incrementally reduced, and the intermediate feature map quality in the reconstruction process is progressively improved. The two alignment auxiliary variables $f^t_{\alpha_1}$ and $f^t_{\alpha_3}$ progressively align to the intermediate reconstructed feature in the top three rows. The bottom three rows display the two introduced auxiliary variables, f^t_u and f^t_v. These variables serve as additional reconstruction target features, complementing f^t_x. Together, they approach the ultimate optimum, providing essential information for HDR feature reconstruction.

7. Ablation Study

7.1. The Effectiveness of Components

The visual comparison of AFUNet and its variants M1-M4 trained on Kalantari's dataset [5] is shown in Fig. 4. The effectiveness of each meticulously designed component within AFUNet is evident. Visual comparisons between {M2, M3, M4} and M1, highlight a pronounced enhancement in reconstruction clarity and robustness, validating the necessity of these components within our framework for improved reconstruction and occlusion handling capabilities in dynamic scenarios.

7.2. Different Iterative Stages

The quantitative results of AFUNet with different iterative stages we provided in the main text are presented in a more intuitive graphical manner in Fig. 5. With the increment of the stage number, an overall upward trend is observed, accompanied by slight fluctuations at stages 4, 5, and 6 cases.

To comprehensively investigate the performance of dif-

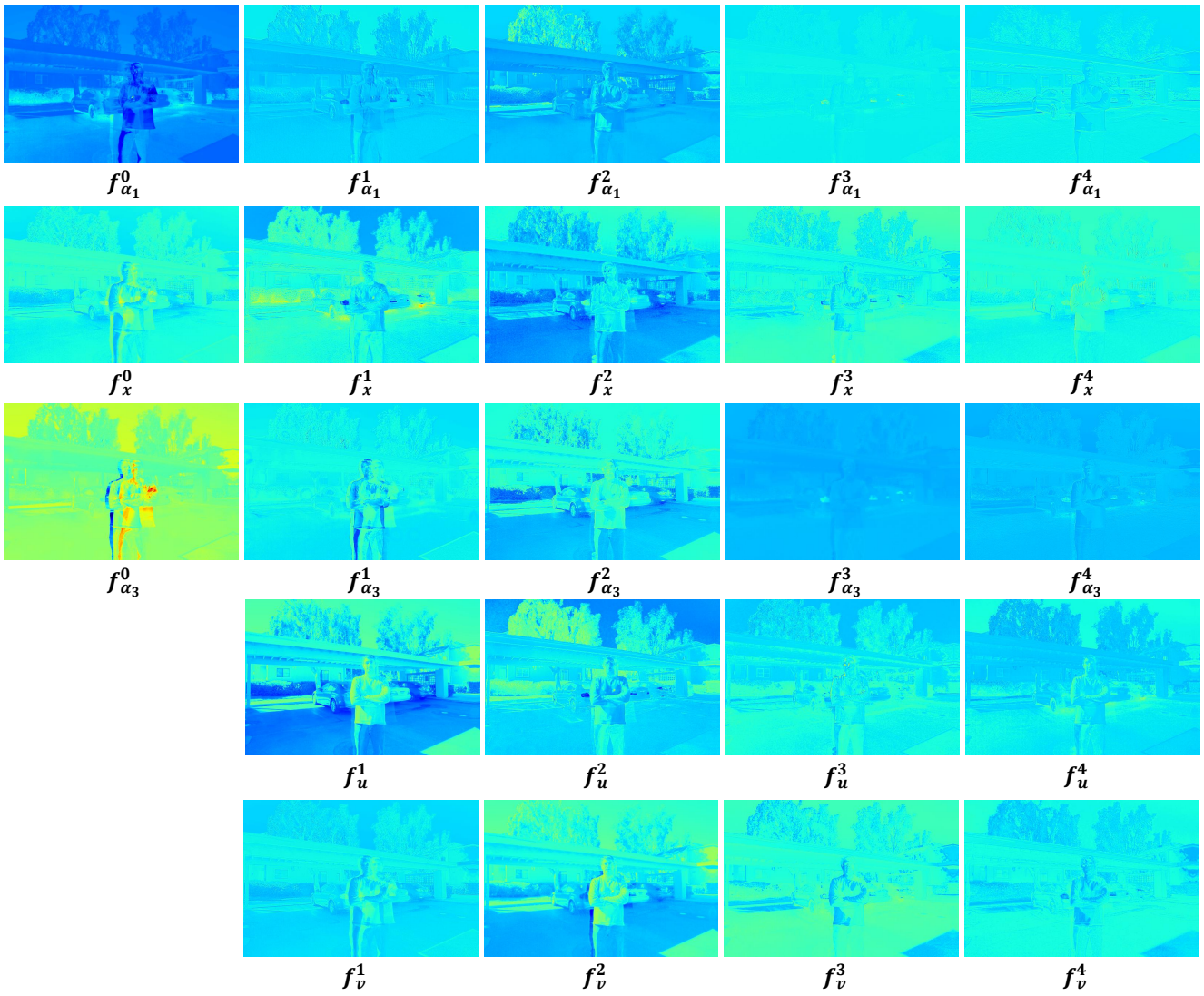

Figure 2. Feature maps of variables in our formulation, *i.e.,* $f_{\alpha_1}^t$, f_x^t, $f_{\alpha_3}^t$, f_u^t and f_v^t. These feature maps are extracted from each stage of AFUNet trained on Kalantari's dataset [5].

ferent stages, we give more visual comparisons of various stage numbers. We denote the variations with distinct numbers of iterative stages as AFUNet-{2,3,4,5,6}, with AFUNet-4 being the default configuration, while the others adhere to the same training specifications as described in the main text. We summarize the visual comparisons in Fig. 6. In line with the quantitative outcomes, the quality of HDR visual effects is enhanced with an increasing number of stages.

7.3. Different Paradigms

In the main text, we discuss a novel paradigm "FA" and compare it with the "AF" paradigm, *i.e.,* AFUNet. The visual comparison of the two paradigms is shown in Fig. 7, consistent with the quantitative results. Additionally, the visual comparison of different feature maps on two paradigms is shown in Fig. 8. At the same stage, the "AF" paradigm outperforms the "FA" paradigm, yielding superior results and fewer artifacts. The preliminary alignment facilitates subsequent fusion operations, demonstrating the necessity and effectiveness of our carefully tailored alignment modules. This also highlights the synergy between alignment and fusion.

8. Additional Qualitative Results

In this section, we present additional qualitative results. Fig. 9, Fig. 10 and Fig. 11 show visual results for various motion and poor exposure condition cases which are challenging in Kalantari's dataset [5], Tel's dataset [9], and Hu's

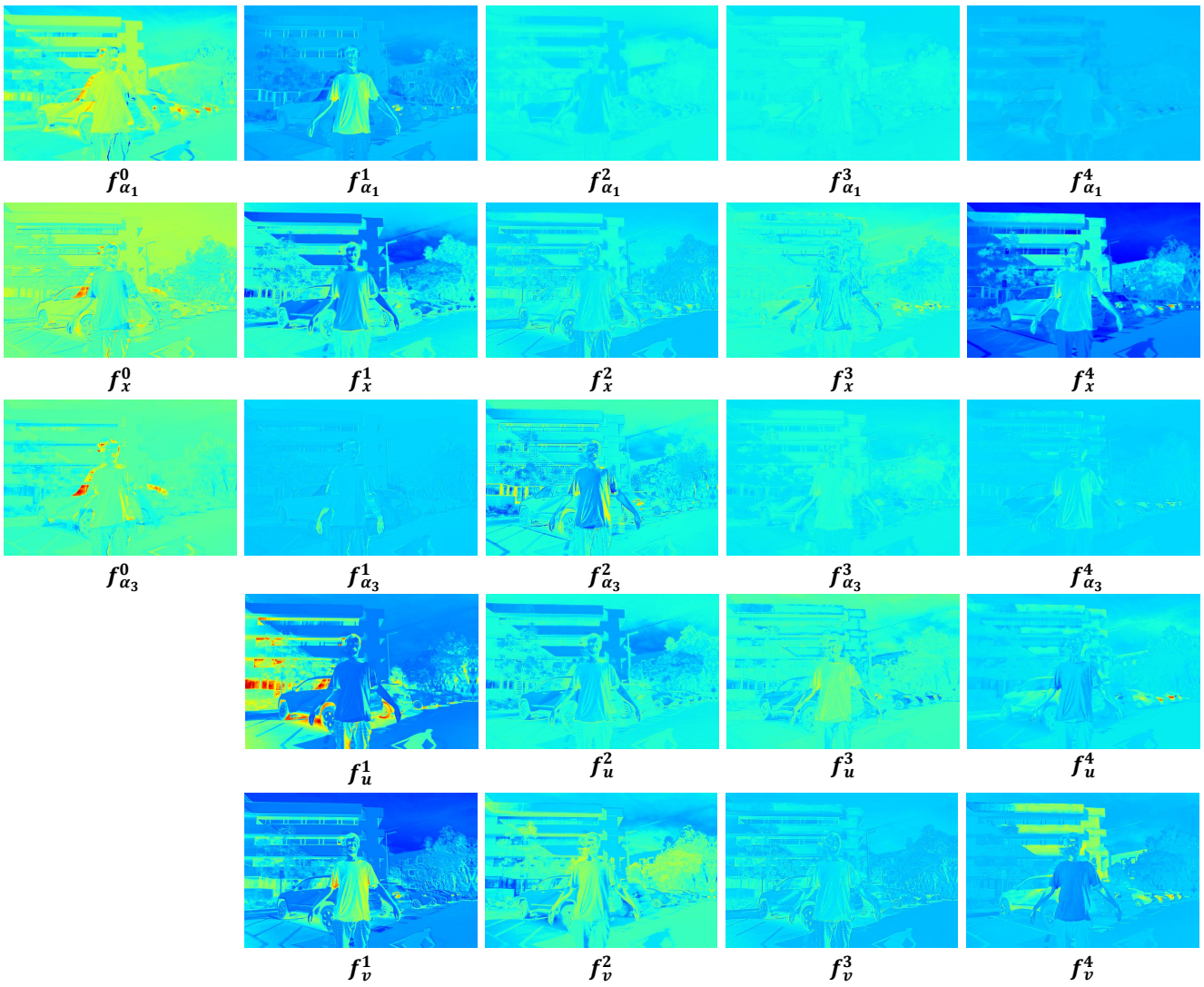

Figure 3. Feature maps of variables in our formulation, *i.e.*, $f_{\alpha_1}^t$, f_x^t, $f_{\alpha_3}^t$, f_u^t and f_v^t. These feature maps are extracted from each stage of AFUNet trained on Kalantari's dataset [5].

dataset [4], respectively. Fig. 12 provides additional qualitative results without ground truth in Tursen's dataset [10].

References

[1] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

[2] Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30, 2017.

[4] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1163–1170, 2013.

[5] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, 2017.

[6] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proceedings of the European Conference on Computer Vision*, pages 344–360, 2022.

[8] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE*

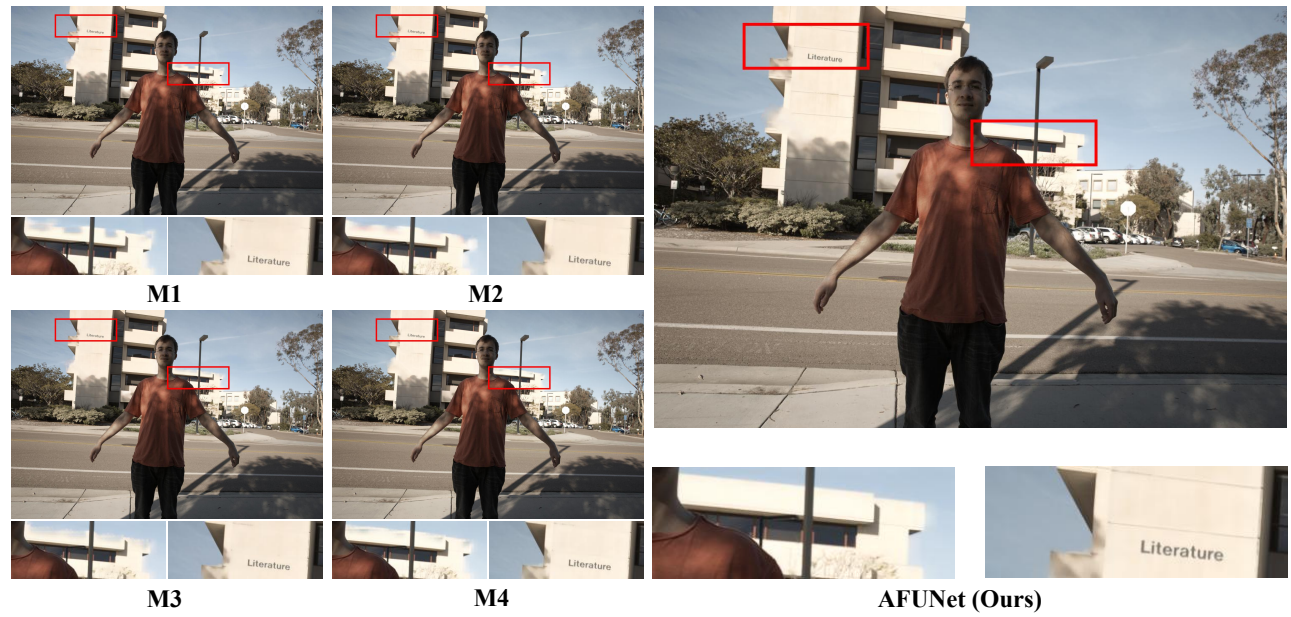

Figure 4. The visual comparison of AFUNet and its variants of different components M1-M4 in Kalantari's dataset [5].

Transactions on Image Processing, 30:3885–3896, 2021.

[9] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free HDR deghosting with semantics consistent transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12836–12845, 2023.

[10] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for HDR images. In *Computer Graphics Forum*, pages 139–152, 2016.

[11] Shangzhe Wu, Xu Jiarui, Tai Yu-Wing, and Tang. Chi-

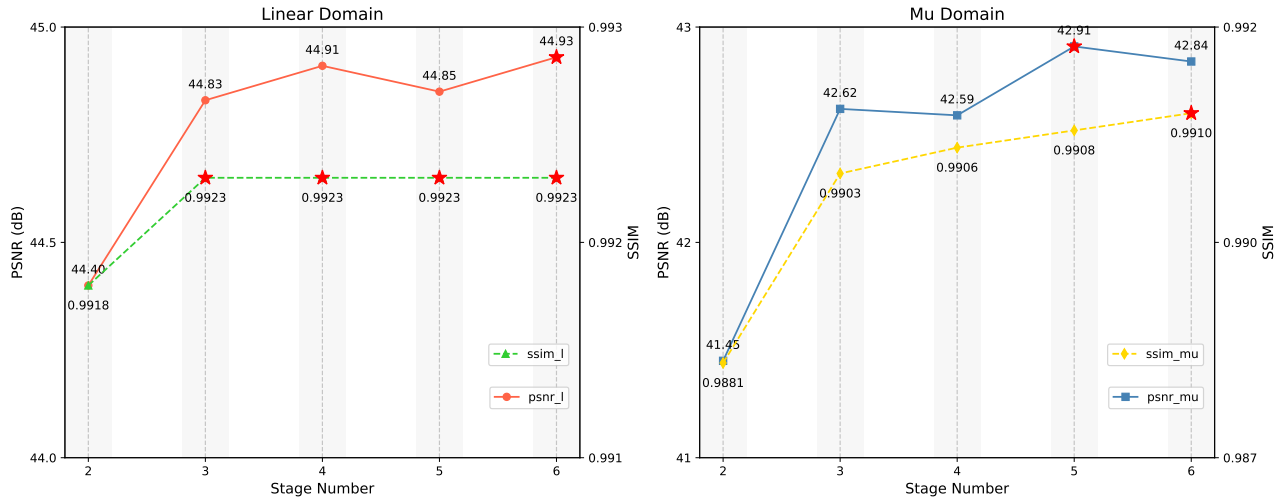

Figure 5. Quantitative comparison in Kalantari's dataset [5] between AFUNet and its variants of different iterative reconstruction stages. Red star markers indicate the highest values of metrics in the linear domain (left) and the tone-mapped domain (right).

Figure 6. Visual comparison of AFUNet and its variants of different iterative stages on Kalantari's dataset [5].

Fusion-Alignment **Alignment-Fusion(AFUNet)**

Figure 7. Visual comparison between "AF" and "FA" paradigms in Kalantari's dataset [5].

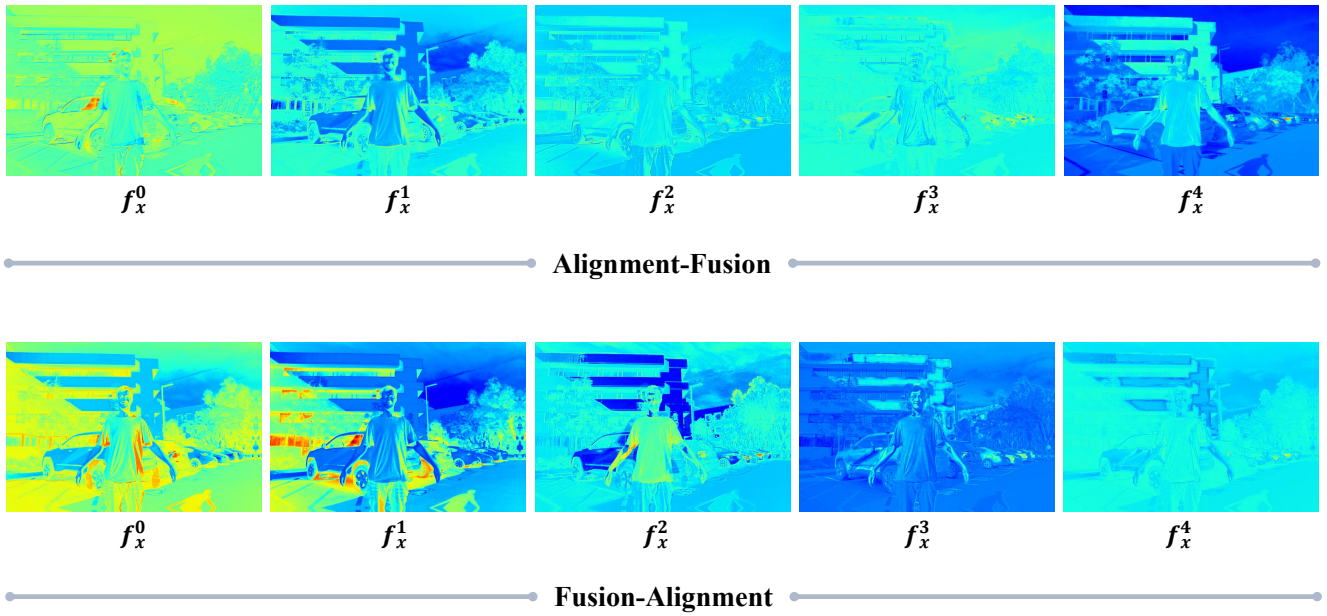

Figure 8. Visual comparison of feature maps extracted from AFUNet trained on Kalantari's dataset [5].

Keung. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision*, pages 117–132, 2018.

[12] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019.

[13] Qingsen Yan, Tao Hu, Yuan Sun, Hao Tang, Yu Zhu, Wei Dong, Luc Van Gool, and Yanning Zhang. Towards high-quality HDR deghosting with conditional diffusion models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[14] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10): 4270–4281, 2014.

[15] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

Figure 9. Qualitative results for challenging cases in Kalantari's dataset [5].

Figure 10. Qualitative results for challenging cases in Tel's dataset [9].

Figure 11. Qualitative results for challenging cases in Hu's dataset [4].

Figure 12. Qualitative results for challenging cases in Tursen's dataset without ground truth [10].