Supplementary Materials for iHDR: Iterative HDR Imaging with Arbitrary Number of Exposures

Anonymous IEEE ICIP Submission

V. Physics-based Tonemapping

The goal of the physics-based tonemapping network is to enable the iterative fusion scheme by converting the HDR output of DiHDR from the linear domain back to the nonlinear domain of the LDR input images. Consequently, the tone-mapped results can serve as inputs for subsequent fusion steps. Traditional tone-mapping methods, such as μ-law, can map HDR images to nonlinear scales for display purposes. However, these tone-mapped images often exhibit significant variations from the captured images, particularly in terms of brightness and contrast. Moreover, learning-based tonemapping approaches are prone to color biases and visual artifacts. Such biases, even if minor, can accumulate and be exacerbated across fusion iterations, ultimately compromising the quality of the results.

We address this issue by modeling the LDR imaging process. A realistic imaging model can be formulated as follows, similar to [1], [5]. Consider an LDR image, $\mathbf{L}$, captured at an exposure time of t where the underlying HDR scene irradiance map is represented by $\mathbf{H}$.

$$\mathbf{L} = \texttt{ADC}\Big\{\xi \times \texttt{Clip}\Big\{\text{Poisson}\big(t \times \text{QE} \times (\mathbf{H} + \mu_{\text{dark}})\big)\Big\} + \mathcal{N}(0, \sigma^2_{\text{read}})\Big\}^{1/\gamma} \tag{9}$$

where ξ is the conversion gain, QE is the quantum efficiency, μ_{dark} is the dark current, and σ_{read} is the read noise standard deviation. Here, Poisson represents the Poisson distribution characterizing the photon arriving process and the dark current effect, and $\mathcal{N}$ represents the Gaussian distribution characterizing the sensor noise. $\texttt{ADC}\{\cdot\}$ is the analog-to-digital conversion and $\texttt{Clip}\{\cdot\}$ is the full well capacity induced saturation effect. We assume a linear camera response function for CMOS sensors and that the imperfections in the pixel array, ADC, and color filter array have been mitigated.

Since our goal is to convert the estimated $\widehat{\mathbf{H}}$ to the domain of $\mathbf{L}$ while preserving essential HDR information, we can remove the random perturbations and lossy processes in the imaging model. For HDR datasets, we simplify the parameters ξ, t, and QE by absorbing them into one exposure-related scalar $c = 4.5$ (Equation. (5)), thus providing a physically motivated initial estimate.

VI. Datasets

Our 9-input LDR Dataset. This paper collects 10 exposure brackets with labels, each containing 9 frames of dynamic scenes with EV values of ±4, ±3, ±2, ±1, and 0. We capture the data using a Sony α6400 camera mounted on a tripod.

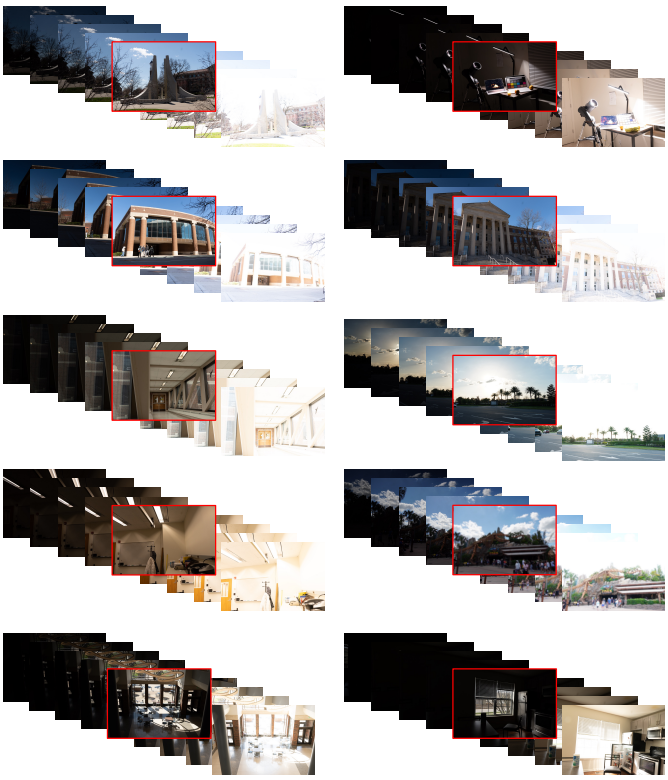

Fig. 9: Test set generated by this paper. Our dataset highlights full motions and 9-frames from EV -4 to EV 4. The reference frame (EV 0) is marked by a red box in each subset.

The resolution of all images is downsampled to 1500 $\times$ 1000 pixels, as shown in Fig. 9.

VII. Computational Cost

Table. IV presents a comparison of computational cost for various methods. The penultimate column lists the parameters of DiHDR, along with the wall time for processing two inputs. The last column provides the total time for DiHDR and ToneNet to process three input frames (2 DiHDR passes and 1 ToneNet pass) Additionally, we compare GMACs for different methods. We observe that the computational time and complexity of iHDR is lower than or comparable to HDR-Transformer and SCTNet on 3 inputs of 1500×1000 images.

VIII. Additional Visual Results

A. Experiments on HDR Deghosting.

Results on 2-input SIG17 Dataset. Fig. 10 - Fig. 12 show the results of HDR deghosting experiments. Our method

TABLE IV: Computational cost comparison of the proposed solution against other SOTA methods. The input size is set to 1000 × 1500 pixels. The speed is measured on an NVIDIA A100 GPU.

Method	DeepHDR [7]	AHDR [8]	HDR-GAN [4]	HDR-Transf. [3]	SCTNet [6]	iHDR
GMACs	1453.70	2166.69	778.81	981.81	293.77	374.76
Time (s)	0.29	0.35	4.85	6.86	7.12	6.93

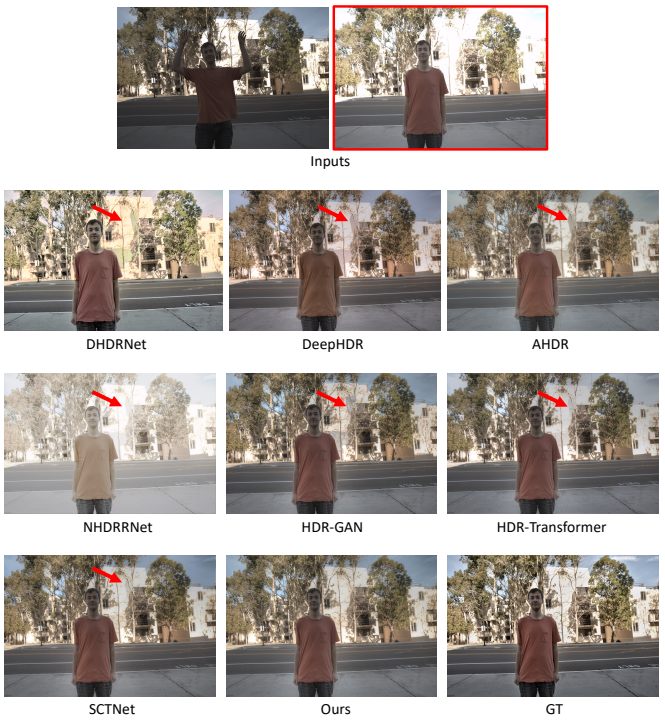

Fig. 10: Qualitative comparison on *Scene. 1* from the SIG17 [2] dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

outperforms other methods in suppressing ghosting artifacts.

Results of Generalization Performance. Since all (2-input) HDR deghosting methods are trained on the 2-input SIG17[2] dataset, we explore their generalization performance on the untrained SCTNet dataset[6]. Fig. 13 - Fig. 18 demonstrate the results of these methods tested on out-of-domain data. Our method outperforms others in ghosting suppression, color reproduction, and detail preservation.

B. Flex Imaging

We validate the capability of the proposed iHDR to handle an arbitrary number of inputs on our collected dataset and compare it with other 3-input frameworks. Fig. 19 - Fig. 21 demonstrate the visual results.

IX. Q & A

In this section, we list a few questions and answers that might interest readers.

Fig. 11: Qualitative comparison on *Scene. 2* from the SIG17 [2] dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

Q1. What is the major advantage of ToneNet over other tonemapping methods?

Answer: Since ToneNet is inspired by the sensor's physics model, it does not introduce systematic bias into the mapping process. Other tonemapping methods, as shown in Fig. **??**, tend to render images based on oversimplified models and hence cause an inevitable domain shift. After iterative fusion steps, the outputs tonemapped by other methods will easily accumulate, making the system unstable.

Q2. How does your method perform on the (3-input) SIG17 dataset?

Answer: Beating the SOTA 3-input HDR deghosting methods is not our priority in this paper. Our method introduces a flexible fusion approach for HDR imaging. Moreover, it is noteworthy that our method was only trained on the 2-input dataset and tested on the 3-input dataset, while other methods were trained on the 3-input manner. From this perspective, our results are not bad.

Q3. Why are all PSNR/SSIM scores so low on your own test dataset?

Answer: This is caused by our collected dataset, which has large full motions (both foreground and background) and many exposure-uneven scenes. Generating high-quality HDR images on our dataset is more challenging than other datasets.

Fig. 12: Qualitative comparison on *Scene. 3* from the SIG17 [2] dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

Fig. 13: Qualitative comparison on *Scene. A* from the SCTNet dataset[6]. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

Q4. You propose iHDR to process HDR imaging with a flexible number of inputs, so can I say that the more frames you use, the better the results?

Answer: Not exactly. The optimal number of frames depends on the scene. For example, for cases with uniform and moderate exposure, two frames are enough for HDR; more frames will not produce better results and may even introduce more harmful artifacts.

Q5. Can you talk about significant challenges for (your or others') HDR imaging approaches?

Answer: Dataset. The primary drawback of HDR imaging lies in the scarcity of comprehensive datasets. Existing HDR imaging datasets are often limited in size and diversity, lacking a wide range of real-world scenes encompassing various lighting conditions (daytime/nighttime, low-light, indoor/outdoor) and camera settings (ISO, aperture, focal length, exposure time). Consequently, methods trained on such datasets may exhibit poor generalization capability when applied to real-world scenarios.

References

[1] Y. Chi, X. Zhang, and S. H. Chan. HDR imaging with spatially varying signal-to-noise ratios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5724–5734, 2023.

[2] N. K. Kalantari and R. Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4), 2017.

[3] Z. Liu, Y. Wang, B. Zeng, and S. Liu. Ghost-free high dynamic range imaging with context-aware Transformer. In *European Conference on Computer Vision*, pages 344–360, 2022.

[4] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. H. Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

[5] X. Qu, Y. Chi, and S. H. Chan. Spatially varying exposure with 2-by-2 multiplexing: Optimality and universality. *IEEE Transactions on Computational Imaging*, 10:261–276, 2024.

[6] S. Tel, Z. Wu, Y. Zhang, B. Heyrman, C. Demonceaux, R. Timofte, and D. Ginhac. Alignment-free HDR deghosting with semantics consistent Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[7] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang. Deep high dynamic range imaging with large foreground motions. In *European Conference on Computer Vision*, 2018.

[8] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. 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.

[9] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang. Deep HDR imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020.

Fig. 14: Qualitative comparison on *Scene. B* from the SCT-Net dataset[6]. Results obtained by DHDRNet [2], Deep-HDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

Fig. 15: Qualitative comparison on *Scene. C* from the SCT-Net dataset[6]. Results obtained by DHDRNet [2], Deep-HDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

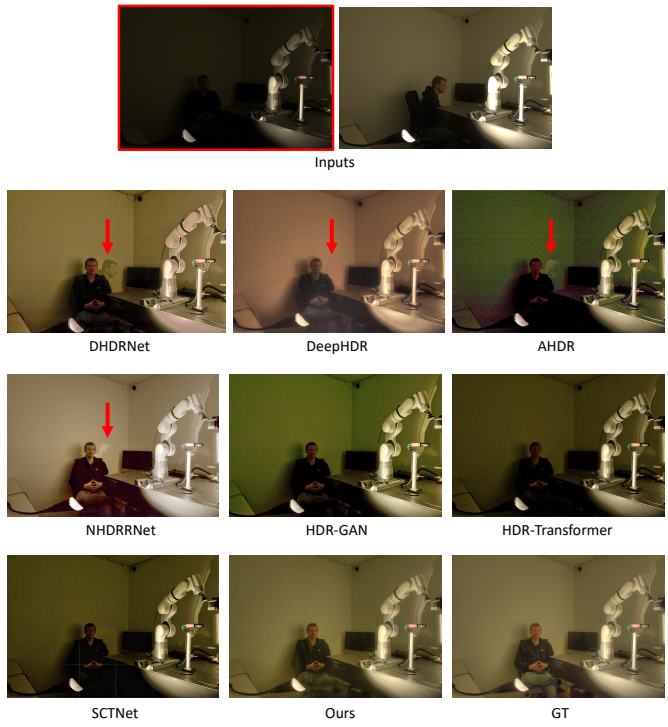

Fig. 16: Qualitative comparison on *Scene. D* from the SCT-Net dataset[6]. Results obtained by DHDRNet [2], Deep-HDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

Fig. 17: Qualitative comparison on *Scene. E* from the SCT-Net dataset[6]. Results obtained by DHDRNet [2], Deep-HDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

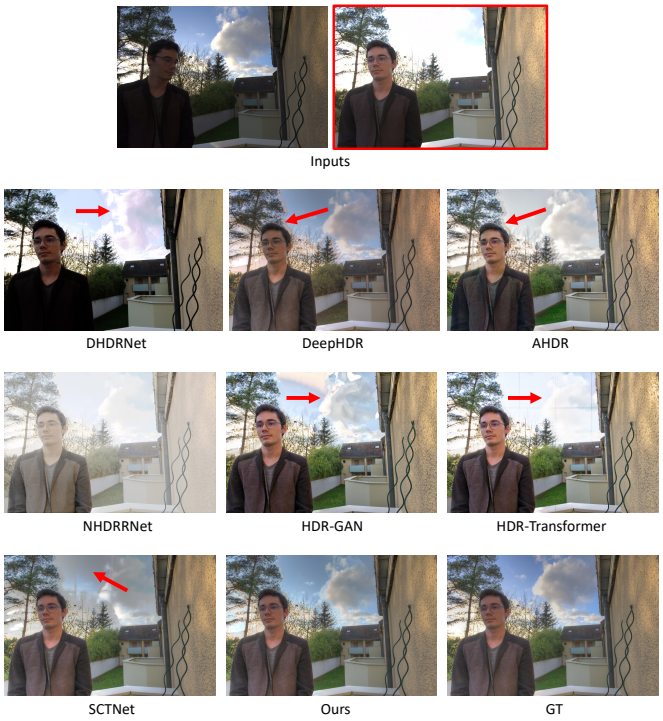

Fig. 18: Qualitative comparison on *Scene. F* from the SCTNet dataset[6]. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame.

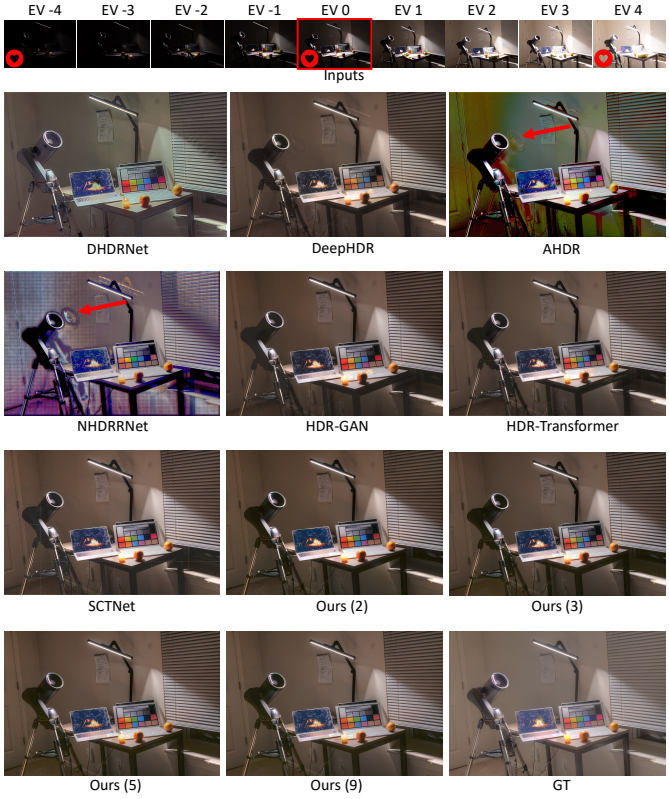

Fig. 19: Qualitative comparison on *Scene. I* from our dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame. All 3-input methods are fed with frames marked with a red heart symbol.

Fig. 20: Qualitative comparison on *Scene. II* from our dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame. All 3-input methods are fed with frames marked with a red heart symbol.

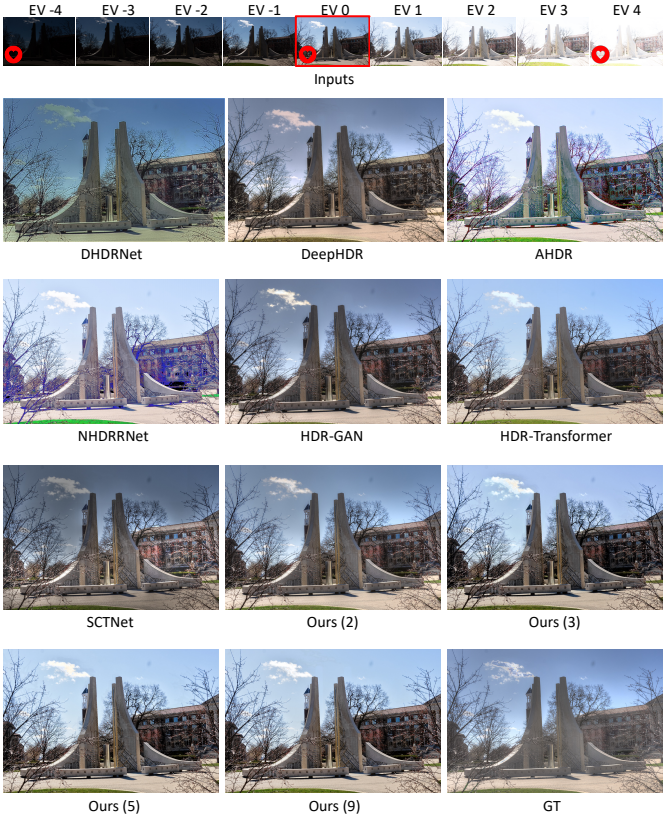

Fig. 21: Qualitative comparison on *Scene. III* from our dataset. Results obtained by DHDRNet [2], DeepHDR [7], AHDR [8], NHDRRNet [9], HDR-GAN [4], HDR-Transformer [3], SCTNet [6] and Ours. The red box represents the reference frame. All 3-input methods are fed with frames marked with a red heart symbol.