SAFNet: Selective Alignment Fusion Network for Efficient HDR Imaging - Supplementary Material

Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen✉

vivo Mobile Communication Co., Ltd, China
{ltkong,libra,cokexiong,haozhang,guhong,jinwei.chen}@vivo.com

1 Challenge123 Dataset

The existing labeled multi-exposure HDR datasets [3–5, 9] have facilitated research in related fields. Nevertheless, results of recent methods [1, 6, 9, 11] tend to be saturated due to their limited evaluative ability [4, 9]. We attribute this phenomenon to their relatively small motion magnitude between LDR inputs and relatively small saturation ratio of the reference LDR image. To widen the performance gap between different algorithms, we propose a new challenging multi-exposure HDR dataset with enhanced motion range and saturated regions, whose statistics comparison with existing datasets [4,9] is listed in Table 1 of our main paper. In the supplementary material, we elaborate construction details of the proposed Challenge123 HDR dataset.

To capture LDR raw image, we use a vivo X90 Pro+ phone equipped with high-end Sony IMX 989 sensor under different lighting conditions, containing indoor, outdoor, daytime and nighttime scenarios. To obtain LDR images with different exposures in a controlled dynamic scene, we first make the camera to automatically adjust exposure, white balance and focus parameters according to the integrated algorithms of the smart phone for better adaptation. Secondly, we fix all camera parameters except for the shutter speed to capture LDR sequences with three different exposures, *i.e.*, under-, middle- and over-exposure. Our mobile phone is fixed on a tripod to keep steady, and we take 10 to 100 successive frames per exposure for subsequent denoising. Generally, the number of shots grows when exposure time decreases, varying based on different noise levels. Finally, we can obtain the noise-free LDR raw image for each exposure by averaging all successive raw frames under the same camera parameter.

To acquire HDR ground truth, we first copy above noise-free LDR raw images with their corresponding camera parameter files from the mobile phone to a desktop computer. Then, we perform a relatively comprehensive ISP simulation pipeline to generate high quality LDR images in the linear domain, whose implementation details are depicted in Figure 1. Specifically, we use the parameter parser and simulation tool provided by Qualcomm, which can parse the camera parameter '.bin' file into '.xml' file, and simulate a relatively complete Image Front End (IFE) pipeline based on '.raw' and '.xml' files, respectively. We dump the intermediate result before Global Tone Mapping (GTM) in RGB color space as our simulated LDR image in linear domain. Finally, we merge above linear

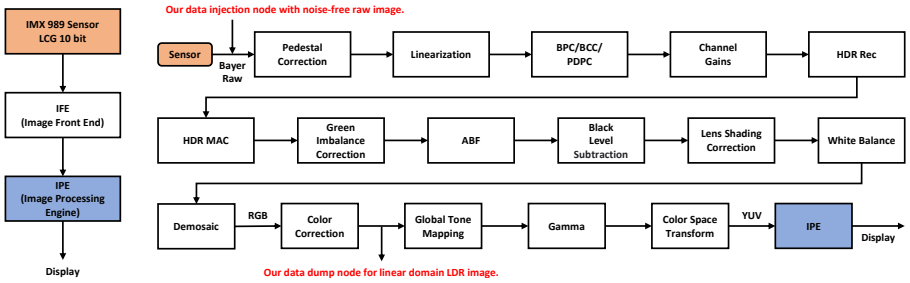

Fig. 1: Details of our ISP simulation pipeline. The left part shows overall framework of the ISP pipeline for Qualcomm platform. The right part presents details of the Image Front End (IFE) in Qualcomm platform. We inject Bayer raw data before 'Pedestal Correction' and dump simulated LDR image before 'Global Tone Mapping'.

domain LDR images of three different exposures by using the weighting function in [2, 4], generating high quality HDR ground truth and reference LDR image. As for the non-reference LDR images, we follow the same acquisition process as before, but move the mobile phone with a relatively large camera pose to create relatively large inter-frame motion. Also, exposure time of the reference LDR image is set to a larger or a smaller value than the normal one for generating more saturated regions. The above two approaches can make our paired LDR-HDR dataset more challenging than the existing ones [4, 9], which has been analyzed in Table 1 and Table 3 of our main paper.

Based on above data collection and processing strategy, we develop a labeled multi-exposure HDR dataset, called Challenge123 dataset, including 96 training samples and 27 test samples, covering diverse lighting conditions, shooting time, motion modes and scene structures. To enhance the applicability of our dataset and promote future research, for each of three content-related moving scenes, we further create under-, middle- and over-exposure LDR images and corresponding HDR image. It means that for each of our 96 training scenes, we have $3\times2\times1 = 6$ exposure combination for training theoretically, while all experiments on our Challenge123 dataset in this paper adopt under-, middle- and over-exposure LDR images by the time order like previous methods.

2 More Results and Analysis

Results on Kalantari 17 Dataset. In Figure 2, we present one more visual comparison on Kalantari 17 test dataset [4], which compares on non-rigid foreground motion and rigid background motion areas. It is obvious that our SAFNet can not only deal with complex motion and occlusion cases like attention-based methods [6, 7, 12, 13], but also generate faithful scene structures as alignment-based approaches [4, 10].

To fairly compare recent SOTA methods [6, 9, 12, 13] with proposed SAFNet on the well-known Kalantari 17 dataset [4], we further train all these algorithms

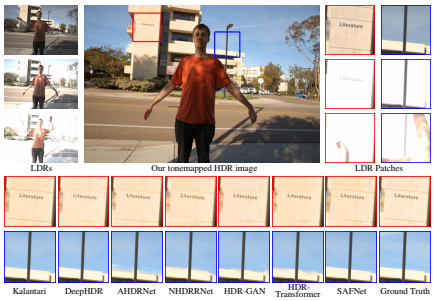

Fig. 2: Visual Comparison on Kalantari 17 test dataset [4]. Zoom in for best view.

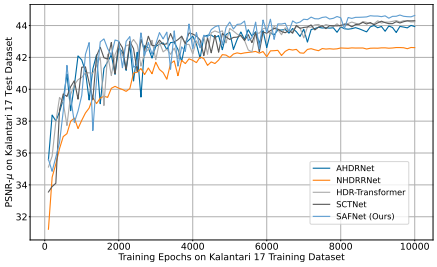

Fig. 3: PSNR-μ on Kalantari 17 benchmark [4] during the whole training process. All algorithms are compared fairly.

under the same data augmentation approach and learning schedule, whose results are summarized in Figure 3. As can be seen, proposed SAFNet exceeds AHDRNet [12], NHDRRNet [13], HDR-Transformer [6] and SCTNet [9] on PSNR-μ consistently in the convergence stage. We attribute the reason to that attention-based multi-exposure HDR methods show relatively poor generalization ability on new dynamic scenes, which tend to overfit on existing training samples. Differently, our SAFNet based on explicit flow alignment generalize well on unobserved dynamic scenes. Besides, our joint refinement decoder can adaptively adjust fusion weights according to flow uncertainty in current location, that can merge high quality HDR image with much fewer ghosting artifacts.

Results on Challenge123 Dataset. In Figure 4, we show six more visual comparisons on our developed Challenge123 test dataset, including both daytime and nighttime scenarios. In the top left figure, there are less artifacts on the murals of our SAFNet prediction. In the top right figure, logos and texts reconstructed by our method are more distinct and realistic. In the middle left figure, the building edge and the sky are more faithful regarding to proposed algorithm. In the middle right figure, texture details synthesized by proposed approach are more natural and coherent. In the bottom left figure, clouds generated by our SAFNet are more consistent and true-colored. In the bottom right figure, signs and texts predicted by our network look more comfortable even if the reference frame is ill-exposed. Summarily, proposed SAFNet can generate more favorable HDR images in challenging motion and exposure scenes.

Note that our proposed Challenge123 dataset aims to widen the performance gap between different algorithms for ease of analysis. The similar result can also be observed on Kalantari 17 Dataset [4]. For example, the first result of HDR-Transformer [6] in Figure 6 of our main paper contains block artifacts at the door handle. With the increasing popularity of high-resolution photography, offsets of several hundred pixels are more common in multi-frame HDR imaging, especially in bracket exposure or night scene modes with long time-lapses.

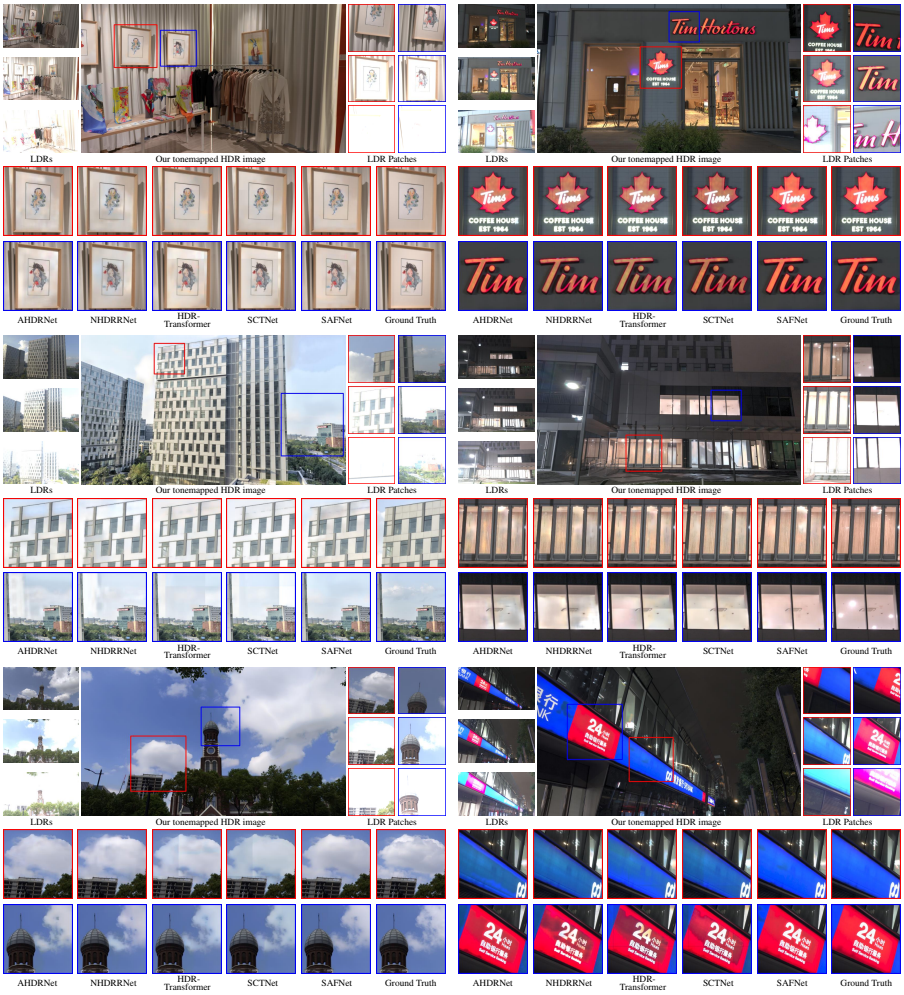

Fig. 4: Visual comparison of recent SOTA methods on our Challenge123 test dataset.

Results on Tel 23 Dataset. To verify the effectiveness of proposed approaches in more motion types and light conditions, we further train our SAFNet on Tel 23 [9] training set from scratch with the same learning schedule as on our Challenge123 dataset. Then, we evaluate our algorithm on Tel 23 test set, and compare the results in Table 1. It can be observed that our algorithm achieves best accuracy on PSNR-l, SSIM-μ and SSIM-l, but falls behind HDR-Transformer [6] and SCTNet [9] on PSNR-μ and HDR-VDP2. We attribute the reason to different motion and saturation characteristics between Tel 23 and Kalantari 17 datasets. Kalantari 17 and our developed Challenge123 datasets both contain regions that are both saturated and moving in the reference LDR image, which can evaluate not only the long-range texture aggregation ability but also the

Table 1: Quantitative comparison on Tel 23 test set [9]. The best result is in **bold**.

Method	PSNR-μ	PSNR-l	SSIM-μ	SSIM-l	HDR-VDP2
Sen [8]	39.97	44.21	0.9792	0.9932	67.20
Kalantari [4]	41.67	47.33	0.9838	0.9961	72.25
AHDRNet [12]	44.16	50.29	0.9896	0.9971	78.12
HDR-Transformer [6]	44.88	51.09	0.9904	0.9981	78.87
SCTNet [9]	**44.93**	51.73	0.9906	0.9981	**79.53**
SAFNet (Ours)	44.61	**51.97**	**0.9908**	**0.9982**	78.80

deghosting ability of multi-exposure HDR algorithms. Differently, backgrounds of Tel 23 dataset are almost static, while the moving people are nearly well-exposed in the reference LDR frame. Therefore, Tel 23 dataset can only evaluate the deghosting ability when merging multiple LDR images.

In conclusion, Transformer-based methods [6,9] are better to deal with multi-exposure image fusion for deghosting. Differently, our flow-based SAFNet are better to handle the moving texture aggregation in a region selective way.

References

1. Catley-Chandar, S., Tanay, T., Vandroux, L., Leonardis, A., Slabaugh, G., Pérez-Pellitero, E.: Flexhdr: Modeling alignment and exposure uncertainties for flexible hdr imaging. IEEE Transactions on Image Processing **31**, 5923–5935 (2022)
2. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques. p. 369–378. SIGGRAPH '97 (1997)
3. Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In: Digital Photography X. vol. 9023, p. 90230X (Mar 2014)
4. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4) (2017)
5. Liu, S., Zhang, X., Sun, L., Liang, Z., Zeng, H., Zhang, L.: Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
6. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: Computer Vision – ECCV 2022. pp. 344–360 (2022)
7. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.H.: Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. IEEE Transactions on Image Processing **30**, 3885–3896 (2021)
8. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based hdr reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6) (2012)
9. Tel, S., Wu, Z., Zhang, Y., Heyrman, B., Demonceaux, C., Timofte, R., Ginhac, D.: Alignment-free hdr deghosting with semantics consistent transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12836–12845 (2023)

10. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: Computer Vision – ECCV 2018. pp. 120–135 (2018)
11. Yan, Q., Chen, W., Zhang, S., Zhu, Y., Sun, J., Zhang, Y.: A unified hdr imaging method with pixel and patch level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22211–22220 (2023)
12. Yan, Q., Gong, D., Shi, Q., Hengel, A.v.d., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
13. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep hdr imaging via a non-local network. IEEE Transactions on Image Processing **29**, 4308–4322 (2020)