Supplementary Material for “Towards Real-World HDR Video Reconstruction: A Large-Scale Benchmark Dataset and A Two-Stage Alignment Network ”

Yong Shu, Liquan Shen[*], Xiangyu Hu, Mengyao Li, Zihao Zhou
Shanghai University, Shanghai, China

Contents

[*] Corresponding author.

1. Evaluation of Our Proposed Dataset

1.1. Typical Scenes in Our Real-HDRV dataset

Our proposed Real-HDRV dataset contains 500 LDRs-HDRs video pairs covering a variety of scenes with high dynamic range. To ensure the diversity of our dataset, samples are collected from diverse light conditions (indoor, outdoor, daytime, and nighttime). To our best knowledge, our Real-HDRV is the largest real-world dataset for HDR video reconstruction.

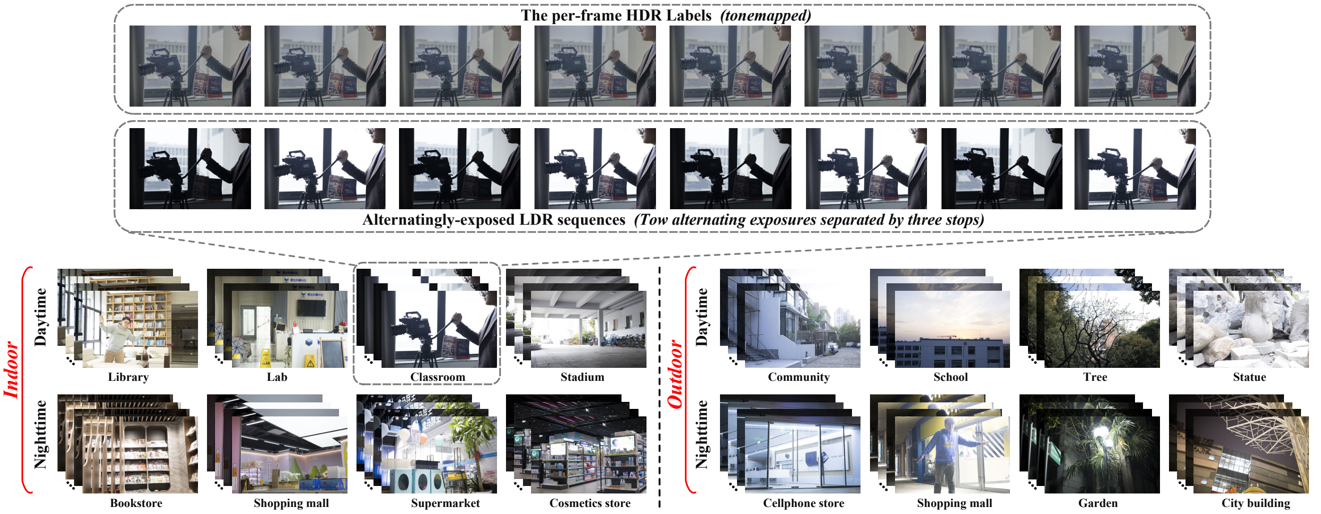

Figure S1. Some typical scenes in our Real-HDRV. We show the alternatingly-exposed LDR sequences and one specific LDRs-HDRs video pair. As seen, our dataset covers a wide variety of scenarios.

1.2. How different motions are created when capturing dataset?

Local motion: Motions between frames are created by people or moving objects, while the camera is static; *Camera motion:* The motion is created by moving the camera, while the scene is static; *Full motion:* Motions are created by people or moving objects and moving the camera.

1.3. More Qualitative Results on the Chen21 dataset

In this section, to demonstrate the effectiveness of our Real-HDRV, we show more visual comparisons on the real-world Chen21 dataset [1]. We train representative HDR reconstruction models (*i.e.*, AHDRNet [11], Kalantari19 [3], Chen21 [1], CA-ViT [7], and LAN-HDR [2]) on our Real-HDRV and the synthetic dataset [1], and evaluate the performance of trained models on the Chen21 dataset. All the following HDR results are tonemapped for visualization.

Figure S2 shows the visual results on the *static set* of the Chen21 dataset [1]. Obviously, the models trained with our Real-HDRV yield more excellent visual quality, while the models trained with the synthetic dataset [1] typically yield high-light blurs, color distortion or corrupted details. For example, in the areas of red rectangles in Figure S2 (the 1^{st} scene), the networks trained on the synthetic dataset typically generate severe high-light blurs in the reconstructed results, while the models trained with our Real-HDRV are able to produce the more appealing results. This is because our Real-HDRV is collected from real-world scenes, which can provide the real degradation distribution.

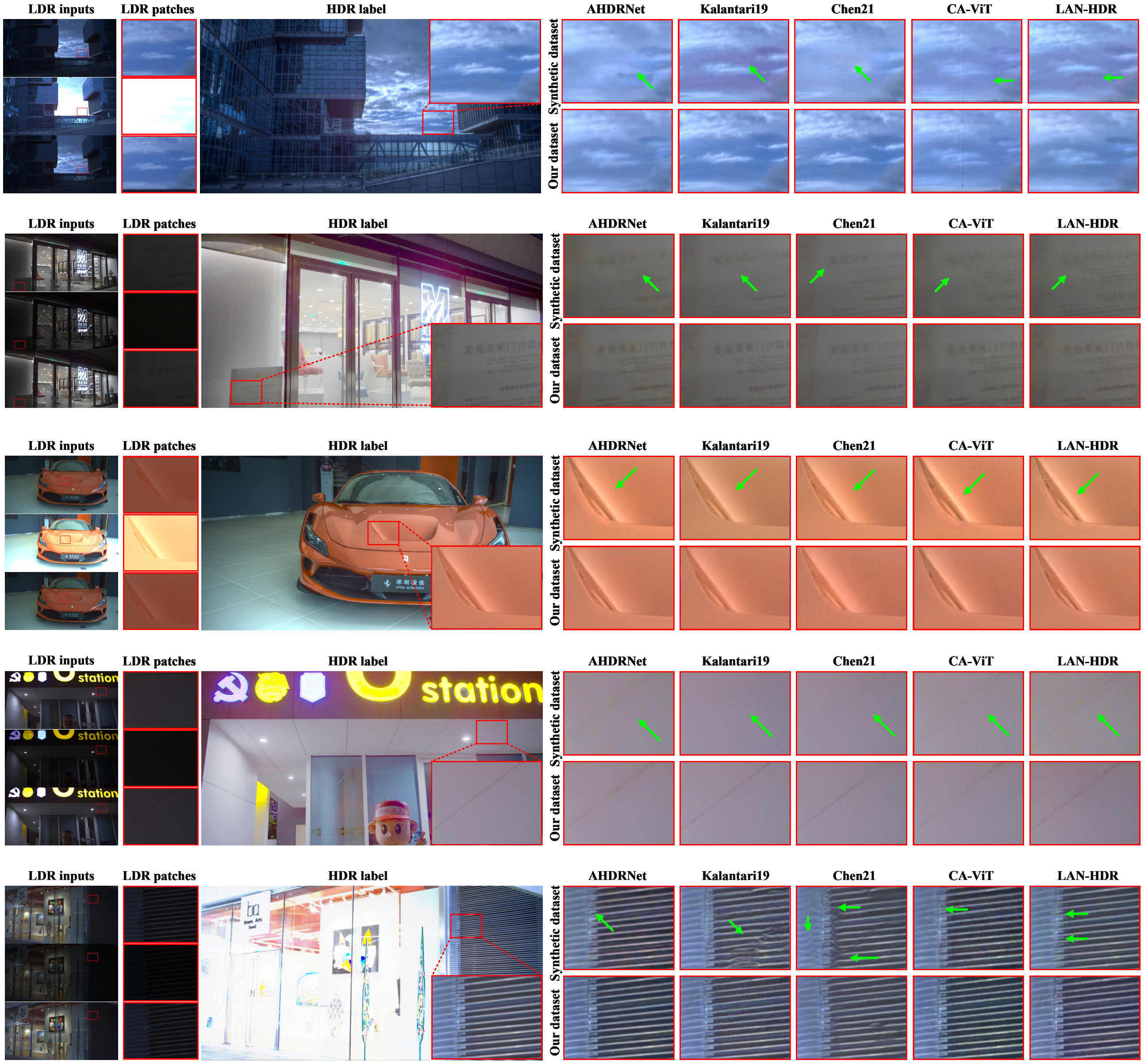

Figure S2. Visual comparison of different networks trained on our dataset and the synthetic dataset [1]. These test scenes are from the *static set* of the Chen21 dataset. Please zoom in for more details.

Figure S3 shows the visual results on the *dynamic set* of the Chen21 dataset [1]. Obviously, the models trained with our Real-HDRV yield better visual quality, while the models trained with the synthetic dataset [1] typically yield ghosting artifacts or corrupted details. For example, in the areas of red rectangles in Figure S3 (the 3^{rd} scene), the models trained on the synthetic dataset can not faithfully recover the details for the under-exposed regions and they are susceptible to ghosting artifacts, while the models trained on our Real-HDRV can generate more faithful results without introducing ghosting artifacts. These visual improvements reiterate the superiority of our Real-HDRV.

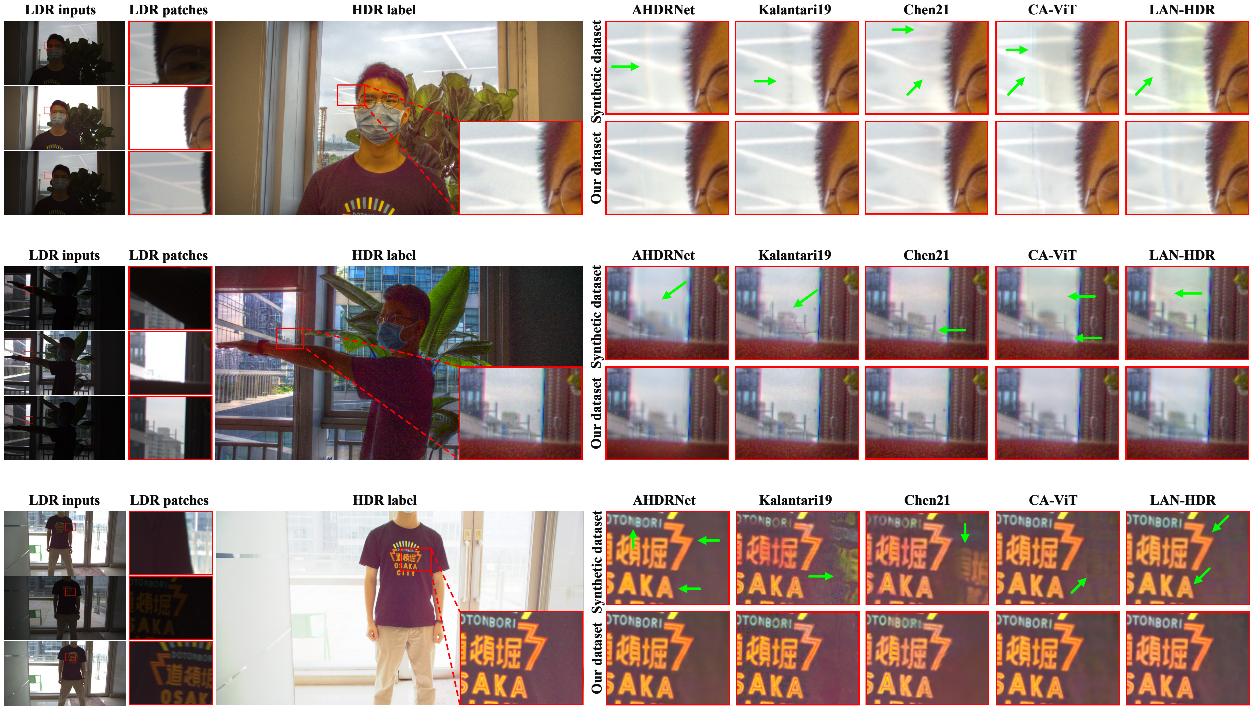

Figure S3. Visual comparison of different models trained on our dataset and the synthetic dataset [1]. These test scenes are from the *dynamic set* of the Chen21 dataset [1]. Please zoom in for more details.

Figure S4 shows the visual results on one of the *unlabeled sequences* of the Chen21 dataset [1]. As seen, when the reference frame is high-exposure, the models trained on our dataset can recover more clear and faithful details for the over-exposed regions, while the models trained on the synthetic dataset generate ghosting artifacts or color distortions.

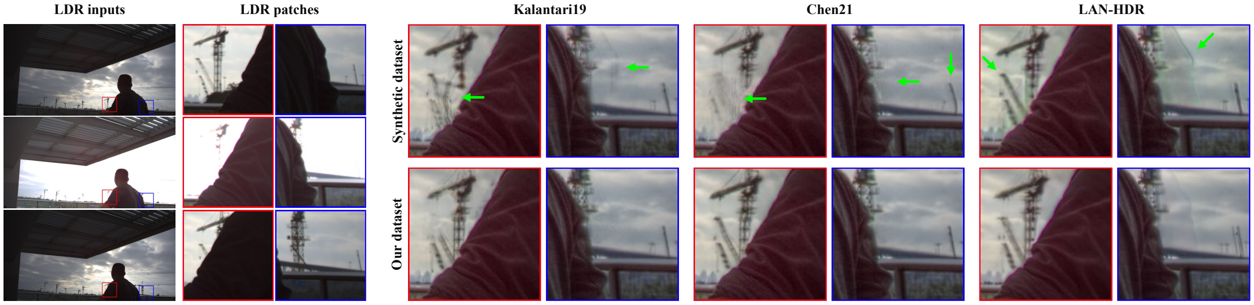

Figure S4. Visual comparison of different models trained on our dataset and the synthetic dataset [1]. The scene is one of the *unlabeled sequences* from the Chen21 dataset [1]. Please zoom in for more details.

2. The architecture of global alignment module (GAM)

Our CNN-based GAM adaptively learns to compensate the global motion for the **alternating-exposed** inputs through end-to-end training, performing alignment in an exposure-invariant manner. Figure S5 shows the detailed architecture of our GAM, given the input $\{X_j|j = i-1, i, i+1\}$, the GAM firstly uses a shared encoding layer to extract feature maps G_j with 16 channels from inputs. Then, the features $\{G_j|j = i-1, i+1\}$ of neighboring frames are fed into the weights estimation module $E\left(.\right)$ along with the feature map G_i of the reference image to obtain the corresponding weights $\{\alpha_{1k}, \alpha_{2k}\}$, generating the global offsets $O_{i-1,i}$ and $O_{i+1,i}$.

Note that our method works for images with different sizes. We use the global average pooling at the end of GAM to get 8-channel weights, weighting 8 pre-defined fixed offsets [12] (size: 8×H×W×2) to generate the final global offsets (size: H×W×2). The global offsets are then used for spatially warping the neighboring frames.

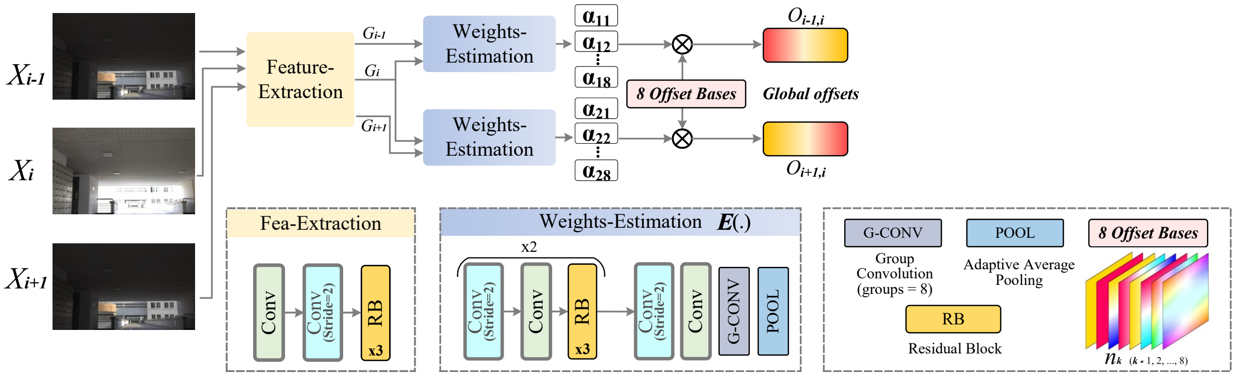

Figure S5. The detailed architecture of our proposed GAM.

3. Evaluation of Our Proposed Method

In this section, we further evaluate our method on the Chen21 dataset [1] to demonstrate the generalization of our proposed method. Then, we provide more visual results on our Real-HDRV. We compare our method with prevalent state-of-the-art HDR video reconstruction methods [1–3, 5, 13] and state-of-the-art HDR deghosting methods [7, 11] on the Chen21 dataset for a comprehensive evaluation. Note that all methods are trained on our Real-HDRV.

3.1. Quantitative Results

Table S1 shows the quantitative comparison on the Chen21 dataset [1]. Our method achieves superior or comparable performances to state-of-the-art methods. Although some methods [1, 7, 13] shows slightly better scores in some evaluation metrics (*e.g.*, PSNR-μ) for dynamic set, they suffer from the ghosting artifacts for under-exposed regions (see Figure S6).

Table S1. Quantitative comparison between our method and other methods on the *dynamic set* and *static set* of the Chen21 dataset [1].

	Evaluation on the *dynamic set*				Evaluation on the *static set*			
Method	HDR-VDP-2	PSNR-μ	SSIM-μ	HDR-VQM	HDR-VDP-2	PSNR-μ	SSIM-μ	HDR-VQM
AHDRNet [11]	62.51	45.02	0.9741	89.60	59.53	40.16	0.9589	77.84
Kalantari19 [1]	61.39	45.31	0.9689	86.95	59.69	41.19	0.9336	81.83
Chen21 [1]	61.39	45.65	0.9716	90.33	59.46	41.37	0.9419	81.43
CA-ViT [7]	62.43	45.19	0.9744	90.41	59.05	39.91	0.9570	77.69
Yue23 [13]	62.73	45.31	0.9693	90.53	60.22	40.81	0.9572	82.50
LAN-HDR [2]	62.83	45.38	0.9743	88.96	59.78	40.09	0.9565	79.29
Ours	63.57	45.53	0.9739	90.22	62.01	41.65	0.9621	87.34

3.2. Qualitative Results

Figure S6 shows the visual results on the *dynamic set* of the Chen21 dataset [1]. As seen, our method can recover the details for the under-exposed areas without introducing ghosting artifacts, while other competing methods typically generate severe ghosting artifacts. This is because these methods either lack an alignment module or can not effectively perform alignment for the input frames, and hence they are prone to introducing unaligned contents from the neighboring frames.

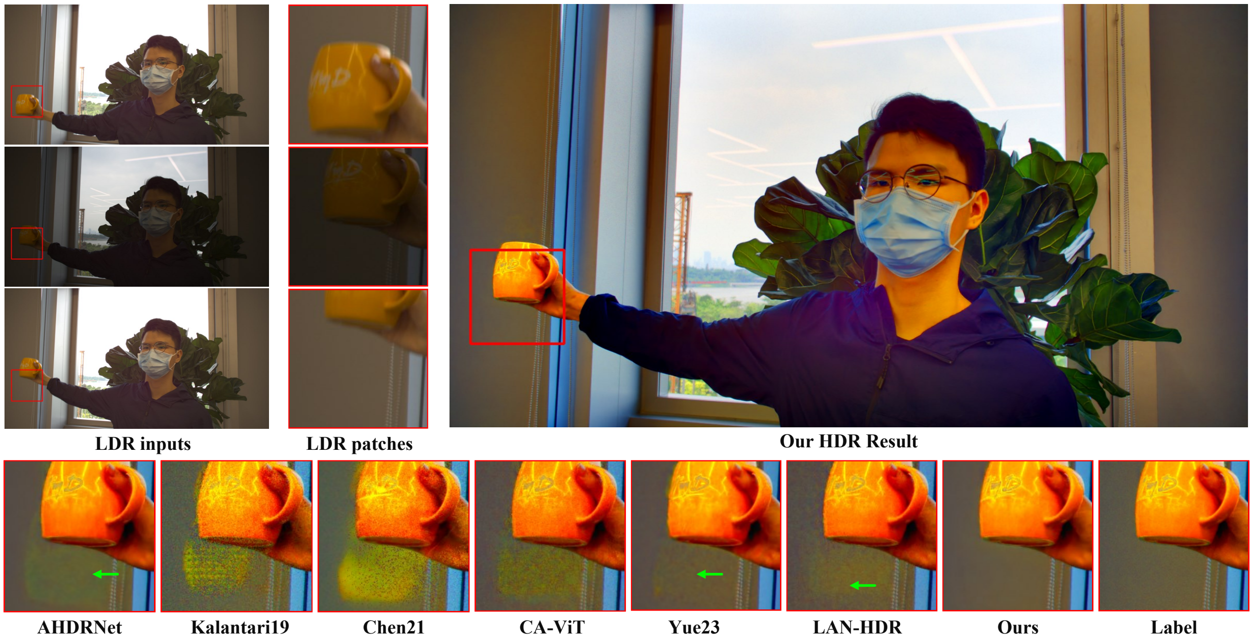

Figure S6. Visual comparison on the *dynamic set* of the Chen21 dataset [1]. Please zoom in for more details.

Figure S7 shows the visual results on the *static set* of the Chen21 dataset [1]. Obviously, when the reference frame is low-exposure, our method is able to restore more and better details than other methods and generates less noisy artifacts for the under-exposed regions.

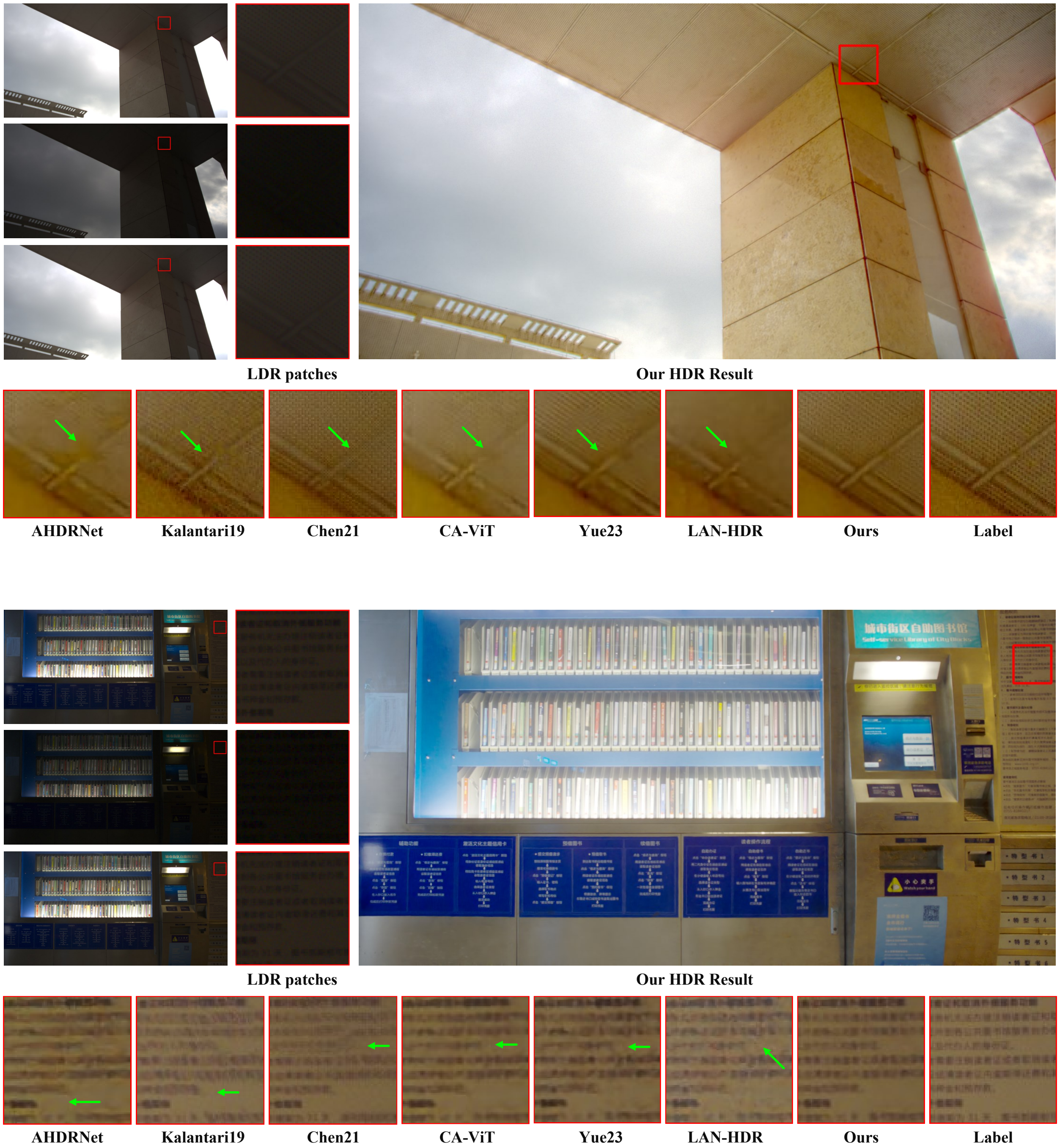

Figure S7. Visual comparison on the *static set* of the Chen21 dataset [1]. Please zoom in for more details.

3.3. More Results on Unlabeled Sequences of the Chen21 dataset

The *unlabeled sequences* of the Chen21 dataset [1] are captured in practical scenarios, which contain uncontrolled dynamic scenes and provide more diverse motion patterns for qualitative evaluation. We provide the visual results on the *unlabeled sequences* in Figure S8, Figure S9, and Figure S10.

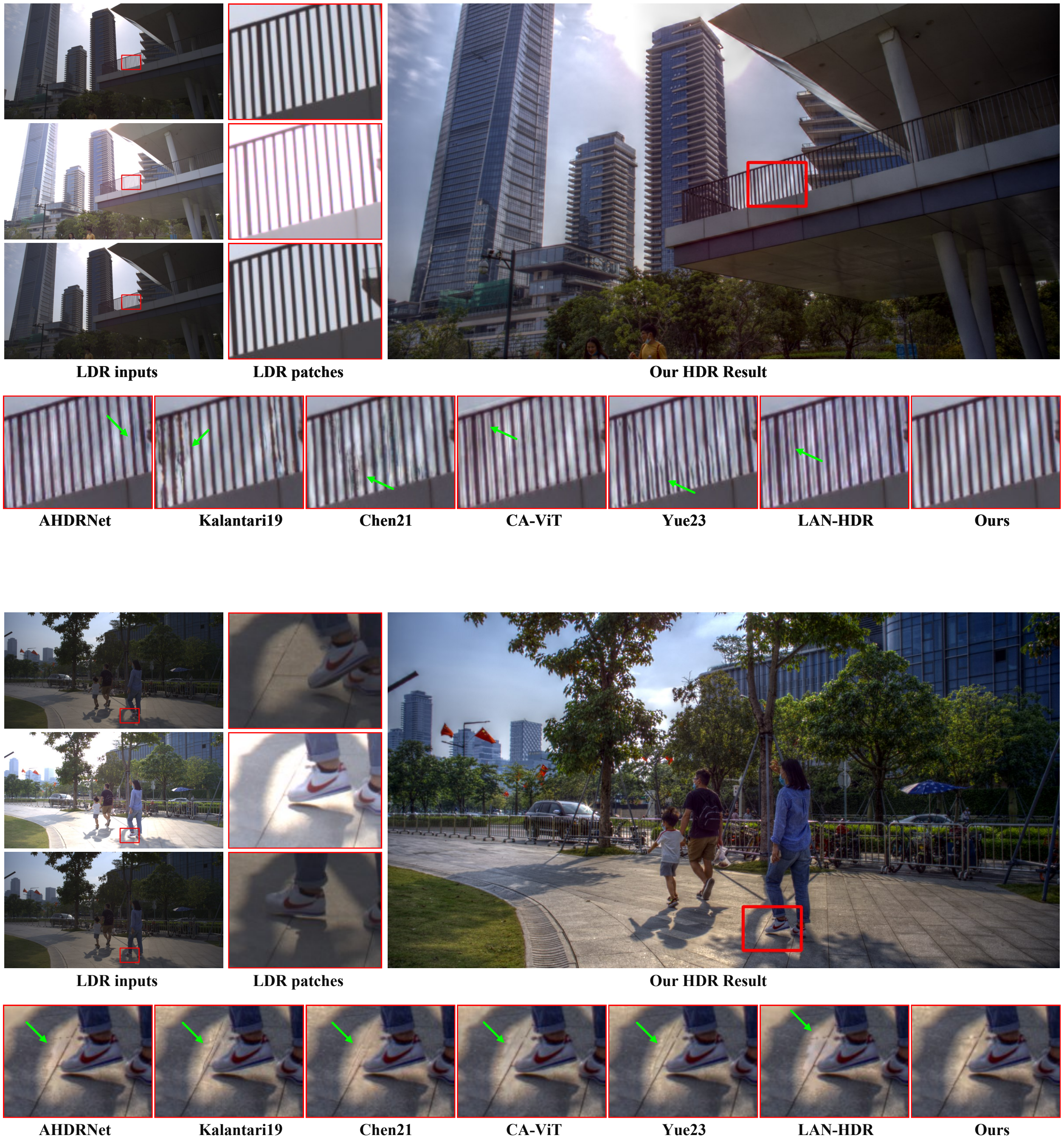

Figure S8. Visual comparison on *unlabeled sequences* of the Chen21 dataset [1], where the reference frames are high-exposure. Please zoom in for more details.

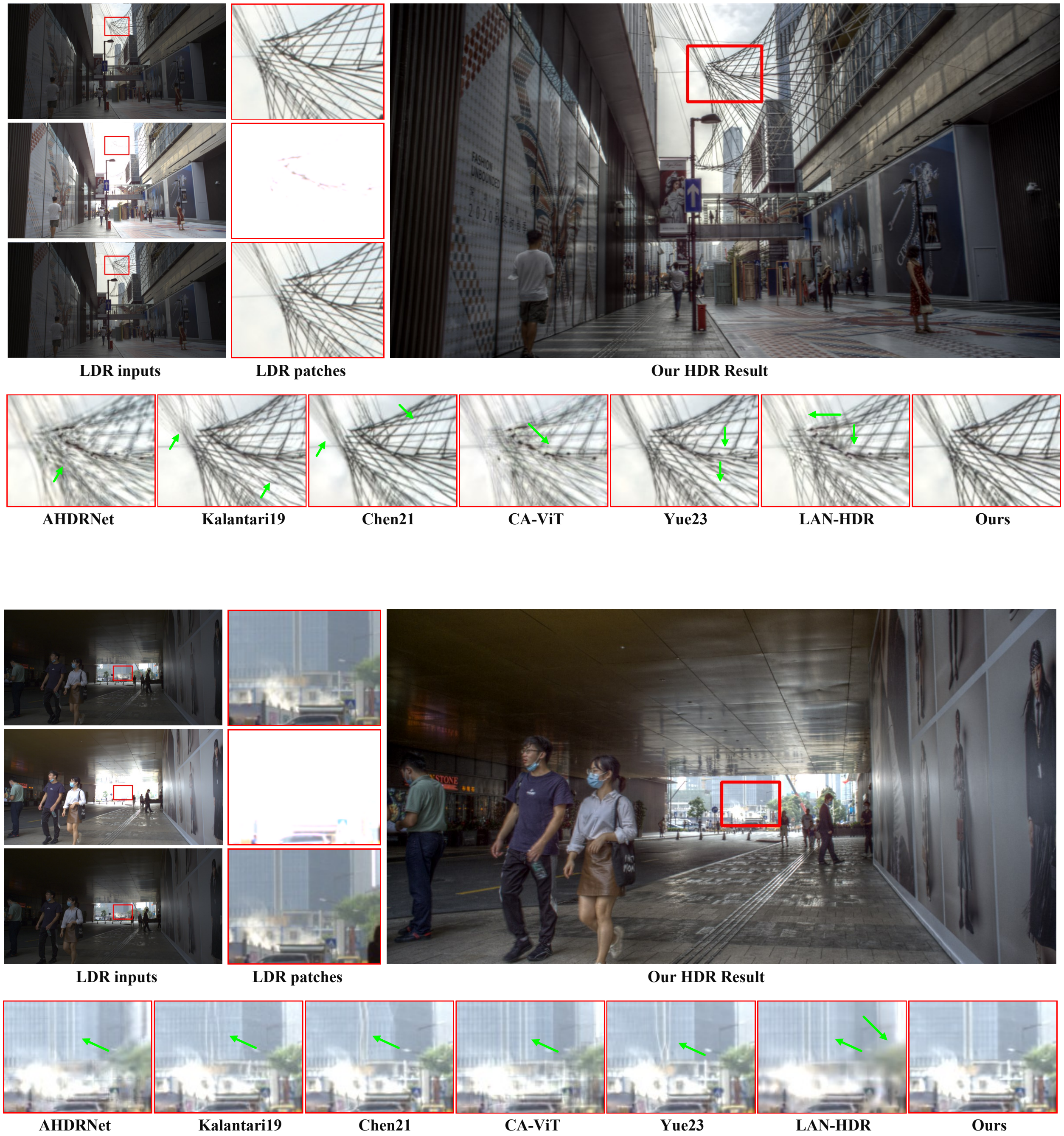

Figure S9. Visual comparison on *unlabeled sequences* of the Chen21 dataset [1], where the reference frames are high-exposure. Please zoom in for more details.

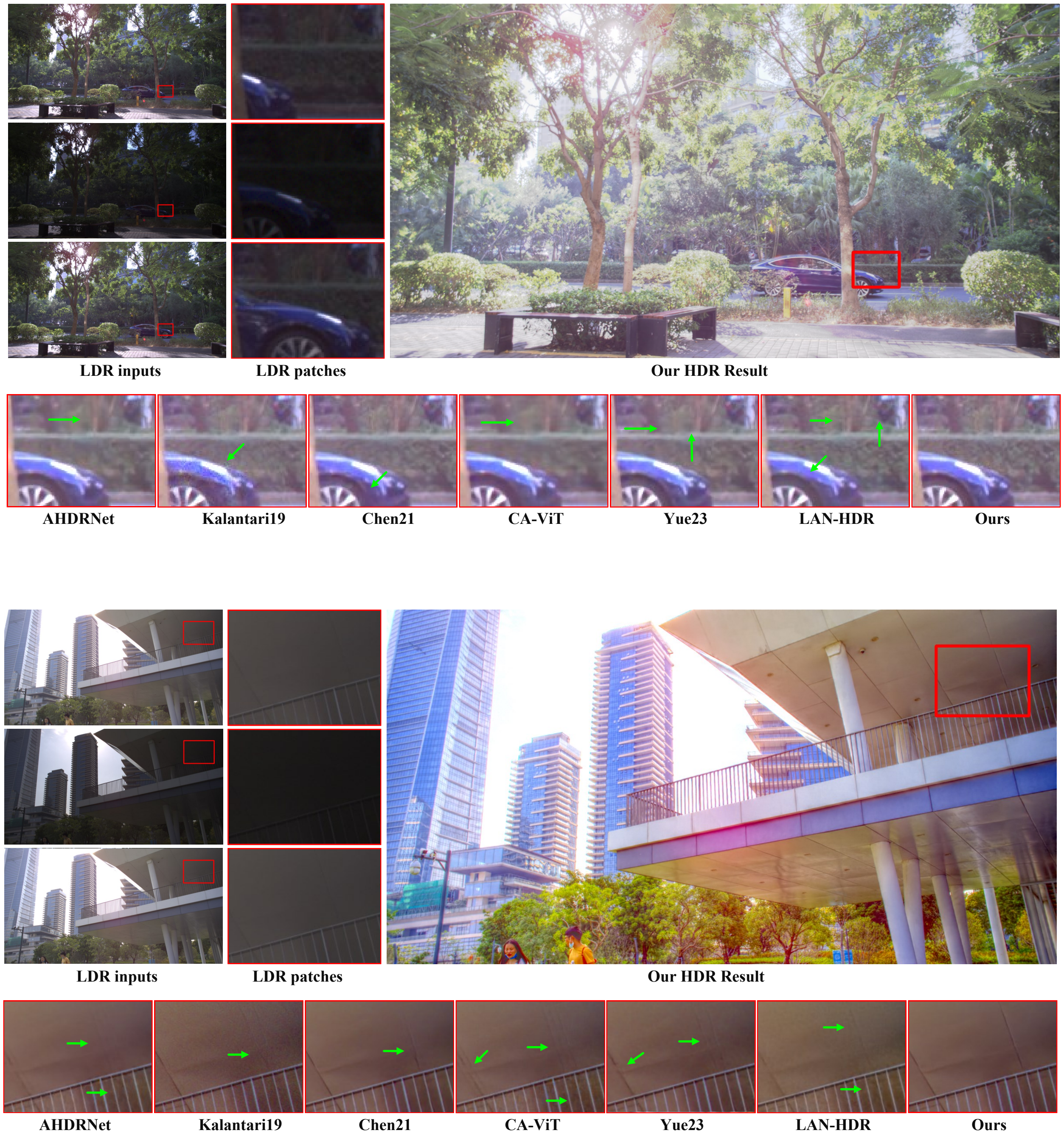

Figure S10. Visual comparison on *unlabeled sequences* of the Chen21 dataset [1], where the reference frames are low-exposure. Please zoom in for more details.

3.4. More Visual Results on Our Real-HDRV

Figure S11 shows the visual results of different methods on our Real-HDRV. As seen, our method can recover the details for the under-exposed areas without introducing ghosting artifacts, while other competing methods typically generate severe ghosting artifacts. Similar improvement can also be observed for the over-exposed areas (see the 2^{nd} scene in Figure S11).

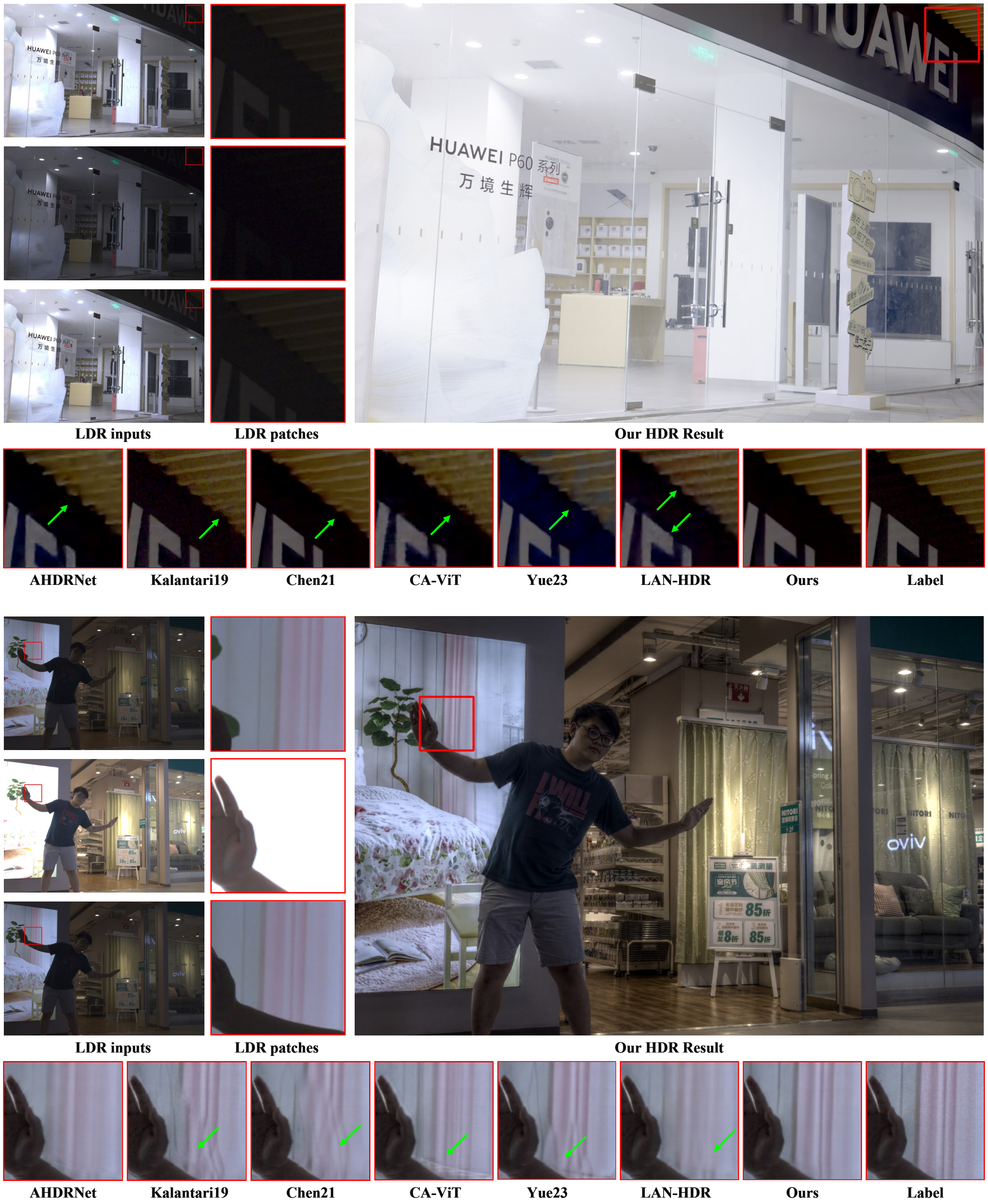

Figure S11. Visual comparison between our method and other methods on our Real-HDRV. Please zoom in for more details.

4. Wider Applications of Our Proposed Dataset

Since our dataset provides data in RAW format and contains the per-frame HDR labels, it can be easily adjusted to make training data for different HDR tasks for future research (*e.g*., single-image HDR reconstruction [14], HDR deghosting [9], multi-exposure image fusion [8], etc.). The flexibility of our dataset enables it to alleviate the lack of datasets in HDR imaging. In this work, we extend our dataset to HDR deghosting. Specifically, following the strategy in [4], we organize the dynamic LDR images in the order of under-, medium- and over-exposed (LDR images with exposure values of {-2, 0, +2} or {-3, 0, +3}), and generate the HDR label for medium-exposed LDR image. Finally, we obtained 450 training pairs and 50 testing pairs.

Table S2. Comparison between different datasets. OD, ON, ID and IN denote outdoor daytime, outdoor nighttime, indoor daytime and indoor nighttime, respectively.

Dataset	Numbers (Training / Testing)	Scenes
Kalantari17 [4]	74 / 15	ID, OD
Tel23 [9]	108 / 36	ID, IN, OD
Ours	**450 / 50**	ID, IN, OD, ON

We further compare our dataset with the commonly adopted HDR deghosting dataset (Kalantari17 [4]) and the latest HDR deghosting dataset (Tel23 [9]). The comparison between different datasets is shown in Table S2. To further demonstrate the superiority of our dataset, we train representative HDR deghosting models (*i.e*., DeepHDR [10], AHDRNet [11], and ADNet [6]) on our dataset and Kalantari17 dataset [4] , and evaluate the performance of trained models on the Tel23 dataset [9]. The quantitative result is shown in Table S3. As seen, the models trained on our dataset acquire the higher scores in almost all the evaluation metrics, demonstrating the superiority of our dataset. These significant improvements stem from the large-scale data, diverse scenes, and diverse motion patterns in our dataset.

Table S3. Quantitative comparison for training on Kalantari17 dataset [4] or our proposed dataset, while evaluating on the Tel23 dataset [9]. ‘$\divideontimes$’ means the models are trained on Kalantari17 dataset [4]. ‘$\dagger$’ means the models are trained on our Real-HDRV. The better results are highlighted in bold.

Method	PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM	HDR-VDP-2
DeepHDR$^{\divideontimes}$ [10]	34.54	0.9594	29.13	0.9579	57.19
DeepHDR† [10]	**35.22**	**0.9648**	**29.83**	**0.9660**	**57.72**
AHDRNet$^{\divideontimes}$ [11]	34.06	0.9606	28.84	0.9620	56.74
AHDRNet† [11]	**35.48**	**0.9642**	**30.34**	**0.9695**	**58.52**
ADNet$^{\divideontimes}$ [6]	34.52	**0.9650**	29.33	0.9666	56.24
ADNet† [6]	**35.33**	0.9643	**30.18**	**0.9682**	**58.53**

References

[1] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[2] Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[3] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. *Computer Graphics Forum*, 2019.

[4] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 2017.

[5] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Transactions on Graphics*, 2013.

[6] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021.

[7] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[8] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Proceedings of 15th Pacific Conference on Computer Graphics and Applications (PG)*, 2007.

[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 *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[10] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[13] Huanjing Yue, Yubo Peng, Biting Yu, Xuanwu Yin, Zhenyu Zhou, and Jingyu Yang. Hdr video reconstruction with a large dynamic dataset in raw and srgb domains, 2023.

[14] Yunhao Zou, Chenggang Yan, and Ying Fu. Rawhdr: High dynamic range image reconstruction from a single raw image. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.