Towards Real-World HDRTV Reconstruction: A Data Synthesis-based Approach Supplementary Material

Zhen Cheng[1⋆], Tao Wang[2⋆], Yong Li[2], Fenglong Song[2✉], Chang Chen[2], and Zhiwei Xiong[1✉]

[1] University of Science and Technology of China
mywander@mail.ustc.edu.cn,zwxiong@ustc.edu.cn
[2] Huawei Noah's Ark Lab
{wangtao10,liyong156,songfenglong,chenchang25}@huawei.com

This supplementary document is organized as follows:

Sec. 1 provides more results and implementation details for the TMOs used in our paper.

Sec. 2 provides more details about the datasets we collect and capture for our experiments.

Sec. 3 provides more experimental details about the user study.

Sec. 4 provides some visual results for SDRTVs synthesized from unlabeled real-world HDRTVs.

Sec. 5 provides some visual results for the ablation study.

Sec. 6 provides an intuitive understanding of our HTMP loss function.

Sec. 7 provides additional implementation details of both our method and the baseline HDRTV reconstruction networks.

1 More results and details for used TMOs

We collect 31 traditional TMOs and one deep learning based TMO, *i.e.*, UTMNet [39] and use them in the main paper for both statistical analysis in Sec. 3 and experimental comparisons in Sec. 5. Specifically, we use traditional ones implemented in two open-source libraries FFmpeg [3] and HDR Toolbox [4] while the TMO Liang [28] is implemented by the official code [5]. For UTMNet, we use its official implementation and pre-trained model [6].

Among these methods, Youtube [1] and Davinci [2] are implemented using FFmpeg with the default 3DLUTs used for their softwares, which can map the HDRTVs to SDRTVs straightforwardly. For the other TMOs originally used for linear radiance maps, we need to linearize the input HDRTVs first. In specific, for the TMOs which process color information independent of the luminance

⋆ Equal contribution. This work was done when Zhen Cheng was an intern in Huawei Noah's Ark Lab.

[3] https://www.ffmpeg.org/

[4] https://github.com/banterle/HDR_Toolbox

[5] https://github.com/zhetongliang/L1-L0-Tone-mapping

[6] https://github.com/yael-vinker/unpaired_hdr_tmo

channel using either color channel rescaling strategy [3,9,11,12,22,25,36,38,27,39] or self-designed chromatic processing strategies [26,31,32,28], we only exploit the PQ EOTF [19] for linearization. For the methods which process all the channels together, we apply a gamut transformation as described in BT.2407 [20] for linearization.

We list these TMOs in Table 1 together with evaluation metrics, *i.e.*, PSNR, SSIM [41], CIEDE-2000 [37] and TMQI [43] averaged on our RealHDRTV dataset. We also list the results of our method as a reference.

	PSNR↑	SSIM↑	CIEDE↓	TMQI↑		PSNR↑	SSIM↑	CIEDE↓	TMQI↑
Clip*	13.82	0.719	18.68	0.7477	KimKautz [22]	17.42	0.753	13.58	0.8189
Linear* [42]	16.46	0.758	15.15	0.7353	Krawczyk [25]	13.61	0.708	19.78	0.7776
Gamma*	20.90	0.809	8.46	0.7638	Kuang [26]	20.92	0.717	9.35	0.7804
Hable* [15]	23.27	0.840	6.38	0.7822	Logarithmic	16.86	0.725	14.52	0.7801
Mobius*	15.62	0.748	15.84	0.7702	Mertens [31]	17.69	0.696	12.81	0.7806
Reinhard* [34]	17.02	0.766	13.73	0.7759	Normalize	20.38	0.757	9.94	0.7968
Davinci* [2]	17.01	0.747	14.09	0.7903	Raman [32]	20.97	0.627	9.52	0.7759
Youtube* [1]	22.99	0.824	6.83	0.7940	ReinhardDevlin [33]	11.51	0.625	26.98	0.7512
Ashikhmin [3]	13.42	0.681	21.48	0.7992	Reinhard [34]	19.94	0.776	10.65	0.8194
BestExposure	19.78	0.775	10.32	0.8110	Schlick [36]	20.13	0.782	10.65	0.8136
BruceExpoBlend [5]	14.58	0.688	16.40	0.7881	Tumblin [38]	17.08	0.748	12.48	0.8341
Chiu [9]	13.49	0.686	19.42	0.7437	WardGlobal [42]	12.30	0.578	22.79	0.6780
Drago [11]	18.27	0.753	13.37	0.8156	WardHistAdj [27]	19.59	0.739	10.52	0.8361
Durand [12]	9.99	0.618	30.38	0.7181	Liang [28]	16.21	0.676	14.81	**0.8807**
Exponential	9.81	0.595	32.46	0.7320	UTMNet [39]	15.77	0.681	16.14	0.8747
Gamma	11.31	0.557	26.17	0.7222	Ours	**24.54**	**0.844**	**5.80**	0.7988

Table 1. Evaluation metrics on both fidelity and color difference between the SDRTVs synthesized by different methods and the ground truth ones on our RealHDRTV dataset. The superscript $*$ denotes that the TMO is implemented with FFmpeg while the others except Liang [28] and UTMNet [39] are implemented with HDR Toolbox. Gray background indicates the results shown in the main paper.

2 Details about our datasets

For the training of our SDRTV data synthesis network, we collect a dataset $\mathcal{H}$ containing 3679 HDRTVs (BT.2020 with PQ OETF [19]) and a dataset $\mathcal{S}$ containing 3603 SDRTVs (BT.709 with gamma OETF [18]) from public datasets [23]. Specifically, we collect 27 4K HDR10 videos and 20 4K SDR videos and use FFmpeg to extract the frames. Note that we only consider the most common format with PQ OETF, for other formats such as HLG [21] and Dolby Vision [10], people are encouraged to collect their own HDRTV dataset to train our network and generate the SDRTV-HDRTV pairs specified to their desired formats.

For the paired evaluation, we captured SDRTV-HDRTV pairs with a smartphone camera. It has two modes "SDR" and "HDR10" for 8K video acquisition. The resultant videos are with the same formats with our collected training data. While building the scenes, we involve dolls, fruits and plants with various and vivid color (as shown in Fig. 1) to maximize the advantages of the wide color gamut with HDRTVs. Moreover, we also controll the lighting conditions and directions by several light sources to create high-light and low-light regions (as

shown in Fig. 1(c) and (d)) to enlarge the differences between SDRTVs and HDRTVs.

To avoid possible misalignment, we only capture indoor scenes such as dolls in a table of a meeting room as shown in Fig. 1(a), (c) and (d) or controlled static scenes such as an empty hall as shown in Fig. 1(b). We also use a professional steady tripod (SIRUI R-2204[7]) to avoid the possible vibration of the environment. After that, we get 93 pairs of 8K SDR and 8K HDR10 videos and extract their frames using FFmpeg. However, they still suffer from obvious misalignment with more than 10 pixels, we then cut out the regions with obvious motions and light condition changes and use a global 2D translation to align the cropped image pairs follow the procedures of [6]. Specifically, we compute and match SIFT key points [30] and estimate a homography using RANSAC [13]. The estimated homography is used to shift the SDRTVs through interpolation for alignment. Afterwards, we crop each image pair into 4K images and remove the pairs which are still with obvious misalignment and finally get 97 4K image pairs with misalignment no more than 1 pixel for evaluation.

In addition, as we noted in the main paper, although we set careful acquisition environment and applied post-processing to avoid misalignment, the sub-pixel shifts between the SDRTVs and their HDRTV versions still exist, thus the full-reference metrics such as PSNR and SSIM with our RealHDRTV dataset may be not as high as those with perfectly-aligned datasets.

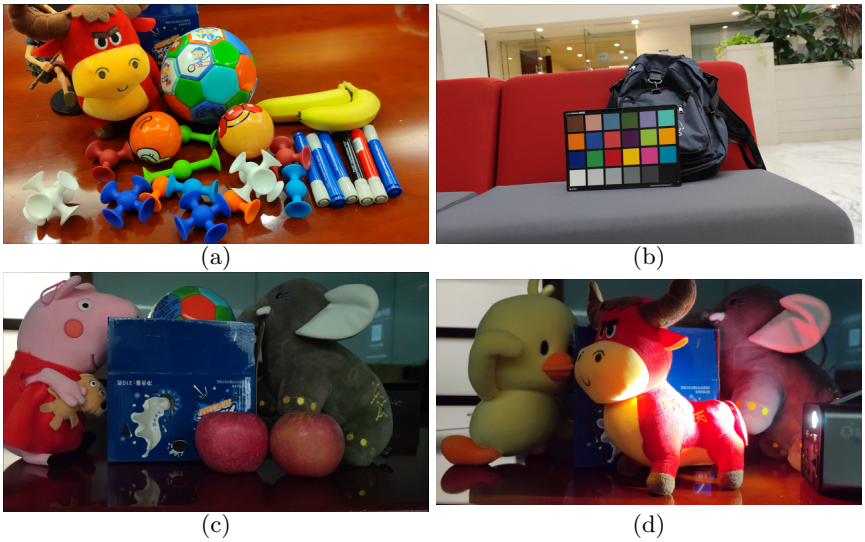

Fig. 1. SDRTV examples of our RealHDRTV dataset.

[7] https://www.siruiusa.com/

3 Details about the user study

For the user study, we invite 11 professional photographers for preference testing. Specifically, for each input SDRTV, we feed it into 3 HDRTVNet-AGCM [7] networks trained with 3 paired SDRTV-HDRTV datasets synthesized by Youtube [1], Hable [15] and Ours and get 3 versions of output HDRTVs. Every two of them are merged to a video with the HDR10 format (BT.2020 with PQ OETF [19]) with an all-white scanning line as shown in Fig. 2 and we can slide the progress bar of the video to change the position of the scanning line for a better one-to-one comparison. Note that the video frame shown in Fig. 2 is rendered to be properly displayed with ordinary SDR-TVs and printed papers and the rendering tool we used here is the madVR[8] equipped to the video player MPC-HC[9].

Fig. 2. A visual example for the videos used for pair-wise comparisons during the user study.

During the testing, we display the videos on an HDR-TV (EIZO ColorEdege CG319X[10] with a peak brightness of 1000 *nits* and the display mode set to BT.2020 color gamut and PQ OETF) in a darkroom and instruct the participants to take the following factors into considerations: (1) whether there are obvious artifacts and unnatural color, (2) whether the overall chromaticity and contrast are natural, (3) whether the details in extreme-light regions are recovered properly. Each participant is given enough time to move the scanning line for comparison and decide which one is better for him/her. Note that for each

[8] http://madvr.com/
[9] https://mpc-hc.org/
[10] https://www.eizo.com/products/coloredge/cg319x/

image, the orders of HDRTV versions are randomized so that the participants can not make intentional preference. After the testing, we collect their preference chart and summarize them into a preference matrix following the same procedure as in [8]. As shown in the main paper, our results achieves much obvious preference over most participants compared with every baseline, which indicates the superiority of our method at modeling realistic degradations.

4 Visual results for SDRTVs synthesized from unlabeled real-world HDRTVs

In this section we provide visual results for SDRTVs synthesized from unlabeled real-world HDRTVs in Fig. 3. The input HDRTVs shown here are from the HDRTV1K dataset [7]. We can see that while the baseline TMOs suffer from information over-preservation (Fig. 3(a) and Fig. 3(b)) and obvious artifacts (Fig. 3(c) and Fig. 3(d)), our method can drop out the extreme-light details (*e.g.*, the clouds in the red rectangle of Fig. 3(a) and the lines of the building in the red rectangle of Fig. 3(b)) and avoid the artifacts such as wrong structures (*e.g.*, the words in the red rectangle of Fig. 3(c)) and color banding (*e.g.*, the sky in the red rectangle of Fig. 3(d)), resulting more realistic and natural SDRTV results.

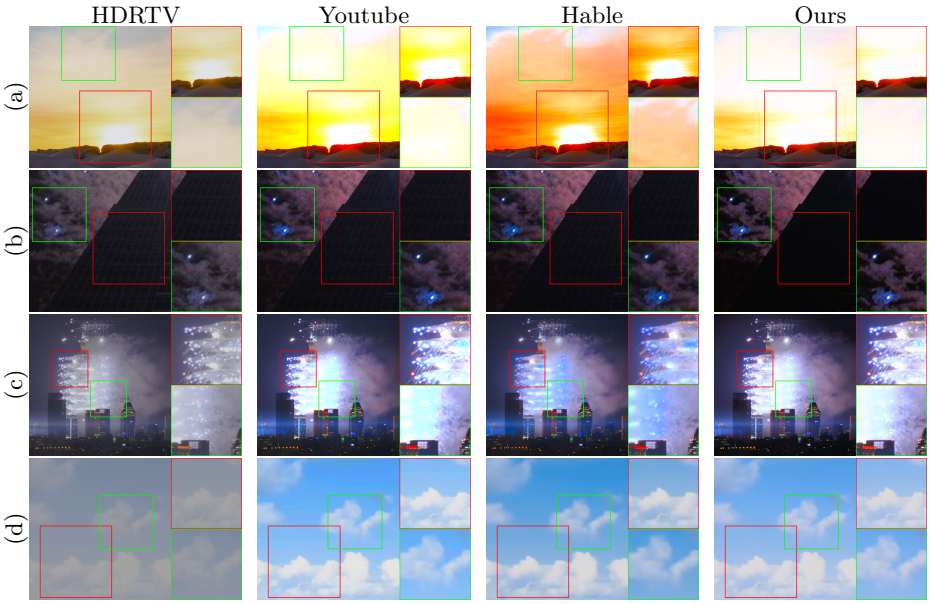

Fig. 3. Visual comparisons on the SDRTVs synthesized by Youtube [1], Hable [15] and Ours as well as the input HDRTVs. The input HDRTVs are from the unlabeled real-world HDRTV dataset HDRTV1K [7]. Zoom in the figure for a better visual experience.

5 Visual results for the ablation study

In this subsection, we provide some visual examples for the ablation study shown with numerical results in the main paper. The test HDRTVs here are from the HDRTV1K [7] dataset. Note that although there are no ground truth SDRTVs here, we can still recognize the differences between the results and judge whether it is realistic according to the aspects we discussed in Sec.4.2 of the main paper.

5.1 Network design

Condition. We provide a visual example for the ablation on the condition network N_c in Fig. 4. We can see that, without the condition network or with the input HDRTV as the condition, there are obvious highlight discontinuities and wrong structures. But we can see that our result is much more continuous without any unexpected structures. Such results show that the conditioned tone mapping results together with the condition network can guide the network to learn better degradation modeling.

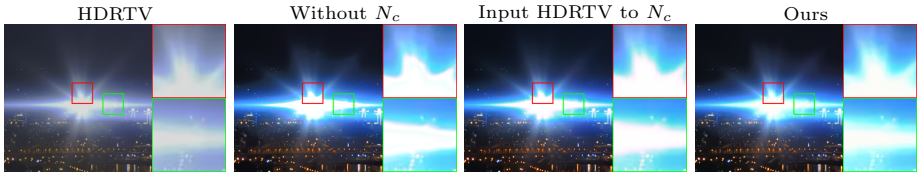

Fig. 4. Visual comparison for ablation on the condition network. Zoom in the figure for a better visual experience.

Streams. We provide an example for the ablation on the network streams, *i.e.*, the global stream N_g and the local stream N_l, in Fig. 5. We can see that while the global stream N_g is unable to drop out the details at the high-light region marked in a red rectangle, the local stream N_l can make up this problem. Also, while the local stream is unable to transform the color of the sunset clouds marked in a green rectangle, the global stream can do it. Thus, merging these two streams together forms a better modeling of real-world SDRTV data synthesis.

5.2 Loss function

$\mathcal{L}_{htmp}$ and $\mathcal{L}_{adv}$. We provide an example for the ablation on $\mathcal{L}_{htmp}$ and $\mathcal{L}_{adv}$ in Fig. 6. With only $\mathcal{L}_{adv}$, the output SDRTV only changes its style and does not have any aspects of realistic SDRTVs. But with only $\mathcal{L}_{htmp}$, the network produces natural cloud in the sky and drops out information in the low-light region. Moreover, with the help of $\mathcal{L}_{adv}$, these details are emphasized to be more realistic. Such visual results show a clear advantage combining these two loss functions together to generate more realistic SDRTVs.

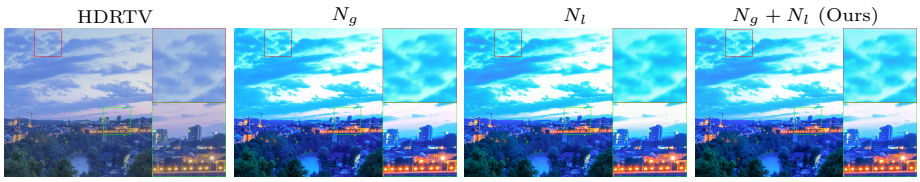

Fig. 5. Visual comparison for ablation on the network streams. Zoom in the figure for a better visual experience.

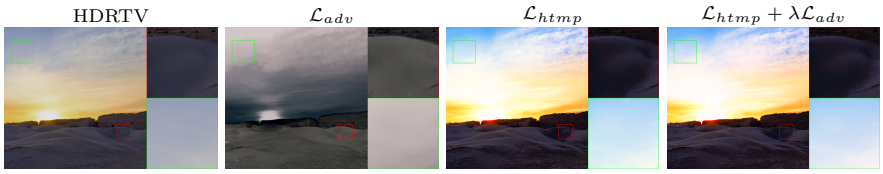

Fig. 6. Visual comparison for ablation on the loss function. Zoom in the figure for a better visual experience.

TMOs used for $\mathcal{L}_{htmp}$. In the main paper, we've conducted ablations on the TMOs used for the HTMP loss, here we show a visual example in Fig. 7. We can see that Linear leads to globally dark results while losing the low-light details as GT does. While μ-law leads to clear structures same as the input HDRTV, it makes the network output containing too much information at the low-light regions and under-saturated color due to its luminance stretching. Meanwhile, Youtube will leads the network to generated over-saturated color, which is complement with μ-law. With our HTMP loss integrating these prior tone mapping results via a region-aware weighting scheme, our network can drop out the information from the input HDRTVs selectively and get more accurate color. This indicates the effectiveness of our hybrid tone mapping prior loss again.

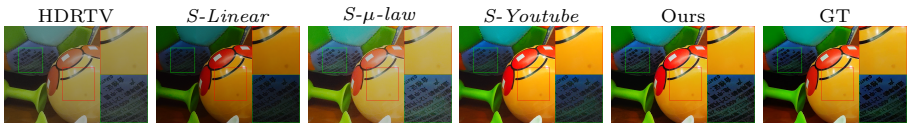

Fig. 7. Visual comparison for ablation on the TMOs used for $\mathcal{L}_{htmp}$. Zoom in the figure for a better visual experience.

6 Flowchart of our HTMP loss

In this section we provide a flowchart of our HTMP loss for a more intuitive understanding of the loss function in Fig. 8.

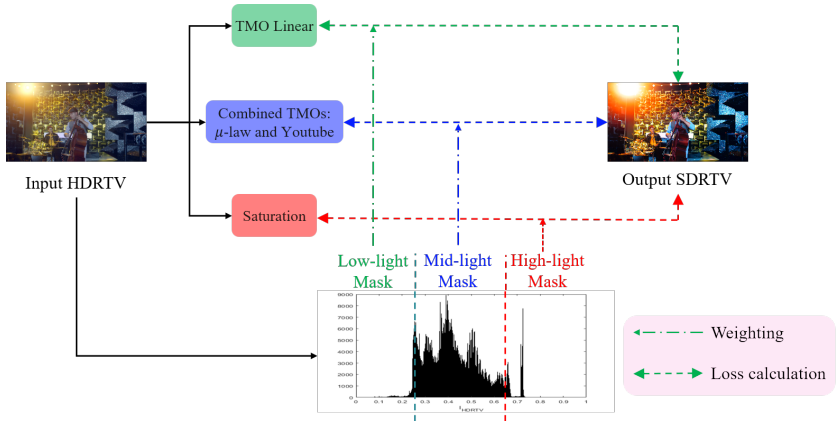

Fig. 8. The detailed flowchart of our hybrid tone mapping prior (HTMP) loss.

7 Implementation details

In this section we provide more implementation details for both our data synthesis framework and the HDRTV reconstruction networks used for comparison in Table 1 in the main paper.

7.1 Details of HA-conv

We use highlight-aware convolution blocks (HA-conv) [14] to build the local adjustment stream in our two-stream network, its structure is shown in Fig. 9. It uses an extreme-light mask generation branch to enable the extreme-light awareness of extracted features by element-wise product and exploits an Inception Block to get more non-local features. Such block has been validated effective in SVBRDF estimation from a single image [14] which needs to be aware of the high-light regions. In our work, we consider it as a better local transformation filter than simple convolution. On SDRTV data synthesis task (our conditioned two-stream network), we validate its effectiveness through ablation studies.

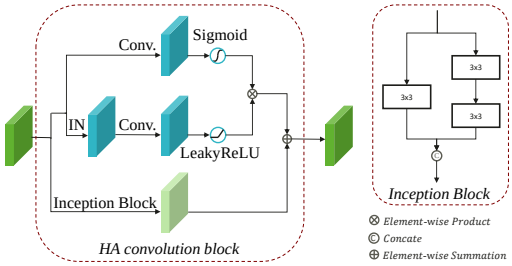

Fig. 9. The detailed structure of highlight-aware convolution block (HA-conv) proposed in [14] and used in the local adjustment stream in our two-stream network.

7.2 Implementation details of our data synthesis framework

To make the adversarial training more stable, we take a two-step training strategy to train our data synthesis network. At first, we use the HDRTV dataset $\mathcal{H}$ and our HTMP loss only to pre-train the network. For our HTMP loss, we follow the previous works [7,35,29] and set the truncation hyper-parameters a and b as 95 and 30, respectively. During the training, we set the patch size as 512×512 with random crop and set the initial learning rate as 2×10^{-4}. We decrease the learning rate with a factor of 0.1 every 80 epochs and end the training after 200 epochs. After the pre-training, we add the adversarial loss with the loss weight λ as 0.01 and finetune the generator network with a learning rate of 2×10^{-5} for another 200 epochs. Adam optimizer [24] and Kaiming initialization [17] are

adopted for both stages. We implement the network with the Pytorch framework. With two NVIDIA V100 GPUs, the pre-training and finetuning need 15 and 17 hours, respectively.

7.3 Implementation details of the baseline HDRTV reconstruction networks

As for JSI-Net [23], we use the official code [11] provided by the authors and only changes the training and testing data with tuning hyper-parameters for convergence. Considering the perception-distortion trade-off [4], in order to make only fidelity comparison, we do not show the results of JSI-GAN in our paper. However, as the significant performance gains in terms of both numerical and visual comparisons shown in the paper, our data synthesis framework is expected to also benefit the generalization ability of perception-oriented networks. As for CSR-Net [16] and HDRTVNet [7], we implement them using the official codes [12] [13] and tune the hyper-parameters to get the best performance on the validation set. As for SpatialA3DLUT [40], because the authors do not share their codes, we implement it by ourselves until we achieve the same performance shown in their paper. After that, we train SpatialA3DLUT for our HDRTV reconstruction task with different synthesized datasets.

[11] https://github.com/JihyongOh/JSI-GAN
[12] https://github.com/hejingwenhejingwen/CSRNet
[13] https://github.com/chxy95/HDRTVNet

References

1. https://www.youtube.com. 1, 2, 4, 6
2. https://www.blackmagicdesign.com/products/davinciresolve/ 1, 2
3. Ashikhmin, M.: A tone mapping algorithm for high contrast images. In: Proceedings of the 13th Eurographics workshop on Rendering. pp. 145–156 (2002) 2
4. Blau, Y., Michaeli, T.: The perception-distortion tradeoff. In: CVPR. pp. 6228–6237 (2018) 11
5. Bruce, N.D.: Expoblend: Information preserving exposure blending based on normalized log-domain entropy. Computers & Graphics **39**, 12–23 (2014) 2
6. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Camera lens super-resolution. In: CVPR. pp. 1652–1660 (2019) 3
7. Chen, X., Zhang, Z., Ren, J.S., Tian, L., Qiao, Y., Dong, C.: A new journey from sdrtv to hdrtv. In: ICCV. pp. 4500–4509 (2021) 4, 6, 7, 10, 11
8. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In: CVPR. pp. 6306–6314 (2018) 5
9. Chiu, K., Herf, M., Shirley, P., Swamy, S., Wang, C., Zimmerman, K., et al.: Spatially nonuniform scaling functions for high contrast images. In: Graphics Interface. pp. 245–245. Canadian Information Processing Society (1993) 2
10. Dolby: Dolby vision: White paper (2016), http://www.dolby.com/us/en/technologies/dolby-vision/dolby-vision-white-paper.pdf 2
11. Drago, F., Myszkowski, K., Annen, T., Chiba, N.: Adaptive logarithmic mapping for displaying high contrast scenes. In: Computer graphics forum. vol. 22, pp. 419–426. Wiley Online Library (2003) 2
12. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. In: Proceedings of the 29th annual conference on Computer graphics and interactive techniques. pp. 257–266 (2002) 2
13. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981) 3
14. Guo, J., Lai, S., Tao, C., Cai, Y., Wang, L., Guo, Y., Yan, L.Q.: Highlight-aware two-stream network for single-image svbrdf acquisition. ACM Transactions on Graphics **40**(4), 1–14 (2021) 10
15. Hable, J.: Uncharted 2: Hdr lighting. In: Game Developers Conference. p. 56 (2010) 2, 4, 6
16. He, J., Liu, Y., Qiao, Y., Dong, C.: Conditional sequential modulation for efficient global image retouching. In: ECCV. pp. 679–695. Springer (2020) 11
17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: ICCV. pp. 1026–1034 (2015) 10
18. ITU-R: Parameter values for the hdtv standards for production and international programme exchange. Recommendation ITU-R BT pp. 709–5 (2002) 2
19. ITU-R: Parameter values for ultra-high definition television systems for production and international programme exchange. Recommendation ITU-R BT pp. 2020–2 (2015) 2, 4
20. ITU-R: Colour gamut conversion from recommendation itu-r bt.2020 to recommendation itu-r bt.709. Recommendation ITU-R BT pp. 2407–0 (2017) 2
21. ITU-R: High dynamic range television for production and international programme exchange. Recommendation ITU-R BT pp. 2390–0 (2020) 2

22. Kim, M.H., Kautz, J., et al.: Consistent tone reproduction. In: Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging. pp. 152–159. ACTA Press Anaheim (2008) 2
23. Kim, S.Y., Oh, J., Kim, M.: Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In: AAAI. vol. 34, pp. 11287–11295 (2020) 2, 11
24. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
25. Krawczyk, G., Myszkowski, K., Seidel, H.P.: Lightness perception in tone reproduction for high dynamic range images. In: Computer Graphics Forum. vol. 24, pp. 635–646. Citeseer (2005) 2
26. Kuang, J., Johnson, G.M., Fairchild, M.D.: icam06: A refined image appearance model for hdr image rendering. Journal of Visual Communication and Image Representation **18**(5), 406–414 (2007) 2
27. Larson, G.W., Rushmeier, H., Piatko, C.: A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Transactions on Visualization and Computer Graphics **3**(4), 291–306 (1997) 2
28. Liang, Z., Xu, J., Zhang, D., Cao, Z., Zhang, L.: A hybrid l1-l0 layer decomposition model for tone mapping. In: CVPR. pp. 4758–4766 (2018). https://doi.org/10.1109/CVPR.2018.00500 1, 2
29. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: CVPR. pp. 1651–1660 (2020) 10
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004) 3
31. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: 15th Pacific Conference on Computer Graphics and Applications. pp. 382–390. IEEE (2007) 2
32. Raman, S., Chaudhuri, S.: Bilateral filter based compositing for variable exposure photography. In: Eurographics. pp. 1–4 (2009) 2
33. Reinhard, E., Devlin, K.: Dynamic range reduction inspired by photoreceptor physiology. IEEE Transactions on Visualization and Computer Graphics **11**(1), 13–24 (2005) 2
34. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. pp. 267–276 (2002) 2
35. Santos, M.S., Ren, T.I., Kalantari, N.K.: Single image hdr reconstruction using a cnn with masked features and perceptual loss. ACM Transactions on Graphics **39**(4), 80–1 (2020) 10
36. Schlick, C.: Quantization techniques for visualization of high dynamic range pictures. In: Photorealistic rendering techniques, pp. 7–20. Springer (1995) 2
37. Sharma, G., Wu, W., Dalal, E.N.: The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research & Application **30**(1), 21–30 (2005) 2
38. Tumblin, J., Hodgins, J.K., Guenter, B.K.: Two methods for display of high contrast images. ACM Transactions on Graphics **18**(1), 56–94 (1999) 2
39. Vinker, Y., Huberman-Spiegelglas, I., Fattal, R.: Unpaired learning for high dynamic range image tone mapping. In: ICCV. pp. 14657–14666 (2021) 1, 2
40. Wang, T., Li, Y., Peng, J., Ma, Y., Wang, X., Song, F., Yan, Y.: Real-time image enhancer via learnable spatial-aware 3d lookup tables. In: ICCV. pp. 2471–2480 (2021) 11

41. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) 2
42. Ward, G.: A contrast-based scalefactor for luminance display. Graphics Gems **4**, 415–21 (1994) 2
43. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Transactions on Image Processing **22**(2), 657–667 (2012) 2