# Differentiable HDR Image Synthesis using Multi-exposure Images

**Jung Hee Kim** [*1], **Siyeong Lee** [*2], **Suk-Ju Kang** [1]
kjhe129@sogang.ac.kr, siyeong.lee@naverlabs.com, sjkang@sogang.ac.kr
[*] Equal contribution, [1] Sogang University, [2] NAVER LABS

## Abstract

Recently, high dynamic range (HDR) image reconstruction based on the multiple exposure stack from a given single exposure utilizes a deep learning framework. These conventional networks focus on the exposure transfer task to reconstruct the multi-exposure stack. However, they often fail to fuse the multi-exposure stack into a perceptually pleasant HDR image as the inversion artifacts occur. We tackle the problem in stack reconstruction-based methods by proposing a novel framework with a fully differentiable high dynamic range imaging (HDRI) process. By explicitly using the loss, which compares the network's output with the ground truth HDR image, our framework enables a neural network that generates the multiple exposure stack for HDRI to train stably. Our differentiable HDR synthesis layer helps the deep neural network to train to reconstruct multi-exposure stacks while reflecting the precise correlations between multi-exposure images in the HDRI process. In addition, our network facilitates the characteristic of the exposure transfer task to adaptively respond to recursion frequency. The experimental results show that the proposed network outperforms the state-of-the-art results.

## 1 Introduction

Deep neural networks, especially convolutional neural networks (CNNs), have shown their significant role in reconstructing the HDR image. Two primary approaches exist in reconstructing the HDR image: direct reconstruction methods [7, 16, 14] and multi-exposure stack-based synthesis methods [8, 12, 13]. Direct reconstruction aims to recover a HDR image (32bits/pixel) from a given single low dynamic range (LDR) image (8bits/pixel). In this case, a large number of LDR-HDR image pair data is required to train a deep neural network [8, 11, 14]. On the other hand, HDR synthesis with the multi-exposure stack focuses on transferring exposures to generate the multi-exposure stack accurately. These approaches alleviate the dataset quantity problem as they require much fewer scenes with multi-exposure stack [12, 13]. However, they suffer from severe local inversion artifacts due to the limitations of networks being trained only with the supervision of the ground truth multi-exposure stack. Therefore, the conventional multi-exposure stack-based approaches had difficulties training the network in an end-to-end manner to reflect the whole HDRI process.

We propose a novel differentiable HDR image synthesis process, which enables the end-to-end training procedure and alleviates the generation of the local inversion artifacts. Specifically, we propose a novel framework with a differentiable HDRI synthesis method. To overcome the conventional limitations of multi-exposure stack-based HDR synthesis, we modify the discrete camera response function (CRF), which converts pixel intensity values into luminance values in the standard HDRI, to be differentiable with the linear approximation technique. Moreover, we incorporate the image decomposition method for reconstructing the HDR image to focus on preserving the image details in exposure transfer tasks. We disentangle exposure transfer tasks with the two-pathway approach, which adjusts the global tone and reconstructs the local structure of the image separately. In addition we propose a recurrent

Workshop on Differentiable Vision, Graphics, and Physics in Machine Learning at NeurIPS 2020.

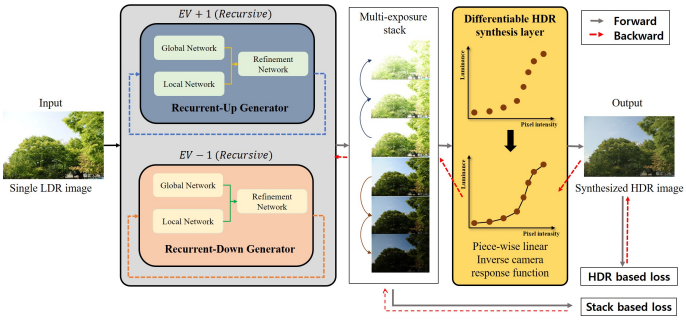


Figure 1: The overall structure of the proposed framework. Our model consists of recurrent-up and recurrent-down networks with the differentiable HDR synthesis layer. Given an input LDR image, the multi-exposure image stack is generated with recursions. Then, the generated stack is synthesized to reconstruct the HDR image with the estimated camera response function using Eq. (1).

approach in the multi-exposure stack generation to facilitate the recursive process by conveying the hidden state vectors.

## 2 Differentiable HDR Image Synthesis using Multi-exposure Images

This section describes our differentiable HDR synthesis framework that trains both the exposure transfer process for multi-exposure stack generation and the HDR image synthesis in end-to-end structure, as shown in Fig. 1. We first generate the multi-exposure stack with the recursive process using the recurrent-up and recurrent-down networks. We then synthesize the stack with the differentiable HDR synthesis layer to reconstruct the HDR image and train the network in the end-to-end structure.

**Differentiable HDR synthesis layer** Debevec and Malik [5] proposed the HDRI pipeline that estimates the CRF using the non-parametric radiometric calibration, which is commonly used. From a given multi-exposure stack, the CRF or inverse CRF estimation and the luminance value can be obtained as follows:

$$O = \sum_{i}^{N}\sum_{j}^{P}[g(Z_{ij}) - \ln E_i + EV_j]^2 + \lambda \sum_{z=Z_{min}+1}^{Z_{max}-1} g''(z)^2, \tag{1}$$

$$\ln E_i = g(Z_{ij}) - EV_j, \tag{2}$$

where $O$ denotes an objective function, $g$ denotes an inverse CRF, and $Z_{ij}$, $E_i$, $EV_j$ denote the pixel intensity value of $i$-th pixel with $j$-th exposure value, the luminance value of $i$-th pixel, and the $j$-th exposure value, respectively. $Z_{min}$ and $Z_{max}$ indicate minimum and maximum intensity values of given LDR images. $N$ and $P$ are the numbers of images and exposure values of the stack. The second term of the objective function regularizes the CRF to be smoothened with the hyperparameter $\lambda$. By minimizing the objective function, we can obtain the discrete inverse CRF of $g$. With the recovered inverse CRF $g$, the pixel intensity value can be remapped to the luminance value as Eq. (2). However, as inverse CRF has the form of the non-differentiable function, we transform the inverse CRF with a linear approximation technique.

Let an inverse CRF be $g = [p_0, p_1, \cdots, p_K]$ with $K$ denoting the maximum intensity value of multi-exposure images. We define the derivative of the linearized function $\hat{g}$ as follows:

$$\frac{\partial \hat{g}}{\partial Z_{ij}} = \begin{cases} g(0), & \text{if } Z_{ij} = 0 \\ g(Z_{ij}) - g(Z_{ij} - 1), & \text{otherwise.} \end{cases} \tag{3}$$

Fig. 2 illustrates our approach to piecewise-linearize the inverse CRF. With the sampled pixels using the Grossberg and Nayar's method [9], we piece-wise linearize the function, as shown in Eq. (3) regarding the prior assumptions of the inverse CRF [5]. The gradients from the loss of luminance values impose constraints on the generated multi-exposure stack to have correlated values with Eq. (3). Hence, our novel framework enables the networks to accomplish both the multi-exposure stack generation task and the HDR synthesis task, with the optimal objective of reconstructing high-quality HDR images. See Appendix A for the explicit results that demonstrates the consistency and the robustness of our method. In addition, see Appendix D for the implementation of the differentiable HDR synthesis layer.

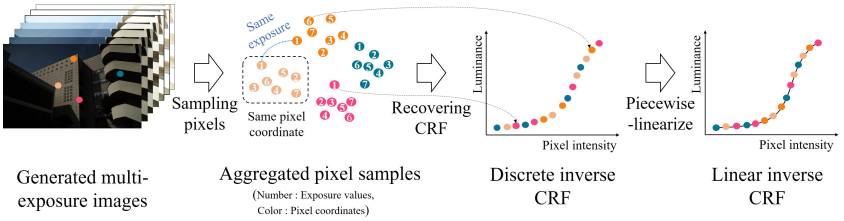


Figure 2: Conceptual diagram of the proposed piece-wise linearization for the CRF. We sample pixels from the multi-exposure stack to aggregate pixels of the same coordinate with different exposure values. We then estimate the inverse CRF with Eq. (1) and convert the function into a differentiable linear form.

**Recursive multi-exposure stack generation** We incorporate the recursive generation of the multi-exposure image stack with the prior knowledge of the exposure manifold space [13]. Our model utilizes recurrent-up and recurrent-down networks, which contains three sub-networks of U-Net structures [21] : the global, local, and refinement networks. The global and local networks are constructed with 5-level and 4-level structures, respectively, with 2 convolutional layers for each level. In addition, we implement the Swish activation [19] on each convolution layer and the convolutional gated recurrent unit (Conv-GRU) [22] in the bottleneck of the global and local networks. We impose the global and local networks to focus on adaptively responding to the number of recursions, and the refinement network to focus on integrating the global and local components, which are histograms and gradient-based edge structures of a target LDR image, respectively. See Appendix B for The detailed structure and the training process.

**Training** The recurrent-up and recurrent-down networks are trained separately with a given single LDR image to generate the multi-exposure stack recursively. Specifically, the global network is trained with the pixel-wise $L_1$ loss ($L_1$) and histogram loss ($L_{hist}$) to constraint the network to generate the image with a similar global tone to the target image. The local network is trained with pixel-wise $L_1$ loss ($L_{edge}$) on edge maps computed with Canny edge detector [2] of $\sigma = 2$. The refinement network is trained with $L_1$ loss ($L_1$), the contextual bilateral loss ($L_{CoBi}$) [27], and the HDR loss ($L_{HDR}$). We used a tone-mapped HDR loss with $\mu$-law to stabilize the training process [25]. Note that $L_{CoBi}$ [27] alleviates the ghosting artifacts due to the misaligned images by minimizing the distances between the matching features extracted from the 3-rd and 4-th layer of the pre-trained VGG-19 network [23] with the bilateral filtering. Overall loss functions are formulated as follows:

$$L_{global} = \lambda_1 L_1 + \lambda_2 L_{hist} = \frac{\lambda_1}{N \cdot E} \sum_{e}^{E} \sum_{i}^{N} |\hat{I}_i^e - I_i^e| + \frac{\lambda_2}{L \cdot E} \sum_{e}^{E} \sum_{l}^{L} |cnt_l(\hat{I}^e) - cnt_l(I^e)| \quad (4)$$

$$L_{local} = \lambda_3 L_{edge} = \frac{\lambda_3}{N \cdot E} \sum_{e}^{E} \sum_{i}^{N} |\hat{E}_i^e - edge(I_i^e)| \quad (5)$$

$$L_{refine} = \lambda_4 L_1 + \lambda_5 L_{HDR} + \lambda_6 L_{CoBi} \quad (6)$$

$$= \frac{\lambda_4}{N \cdot E} \sum_{e}^{E} \sum_{i}^{N} |\hat{I}_i^e - I_i^e| + \frac{\lambda_5}{N} \sum_{i}^{N} |log \frac{1 + \mu \hat{H}_i}{1 + \mu H_i}| + \frac{\lambda_6}{M} \sum_{j}^{M} \min_k (\mathbb{D}_{p_j, q_k} + w_s \mathbb{D}'_{p_j, q_k}) \quad (7)$$

where $N$, $E$, $L$, and $M$ denote the number of pixels, exposure values, intensity levels, and features respectively, and for all the equations, $\hat{\cdot}$ represents the prediction of the network. $I_i^e$ denotes the $i$-th pixel value in image $I$ of the exposure value $e$, and $cnt_l(\cdot)$ indicates the number of pixels which has a rounded down intensity $l$ in the input image $I$. $edge(\cdot)$ extracts gradient-based edge maps from the image $I$, and $E_i$ denotes the $i$-th pixel value in predicted edge map. $H_i$ is a pixel luminance in the HDR image, which is derived from the Eq. (2), and $\mu$ is the compression parameter of the HDR image, where we set the value with 5000. $\mathbb{D}_{\cdot,\cdot}$ indicates the sum of cosine distances between all the matched features of $p$ and $q$, and $\mathbb{D}'_{p,q}$ indicates spatial coordinate distance. Note that $j$ and $k$ indicate indices of the matched feature of $p$ and $q$ respectively. We set the hyperparameters $\lambda_1 = \lambda_3 = \lambda_4 = \lambda_5 = 1$ and $\lambda_2 = \lambda_6 = 0.1$ in our experiments to stably train the networks. See Appendix D for the detailed implementation of the histogram loss.

Table 1: Quantitative comparison of proposed and conventional HDR reconstruction methods. We measured the HDR-VDP-2 score [15] for synthesized HDR images.

| Method | Training dataset quantity | VDS | HDR-Eye | RAISE |
|---|---|---|---|---|
| | | $m \pm \sigma$ | $m \pm \sigma$ | $m \pm \sigma$ |
| **Proposed** | **48 scenes** | **58.807±5.413** | **55.914±1.917** | **59.493±3.420** |
| HDRCNN | 3,700 scenes | 53.031±4.957 | 50.804±5.790 | 57.154±3.642 |
| DrTMO | 1,043 scenes | 55.227±4.662 | 51.800±5.933 | 57.645±4.028 |
| Deep recursive HDRI | **48 scenes** | 56.347±3.492 | 52.832±2.944 | 57.570±3.697 |
| ExpandNet | 1,013 scenes | 44.720±9.432 | 50.428±4.493 | 54.717± 1.998 |
| SingleHDR | 10,289 scenes | 55.237±4.487 | 54.509±3.714 | 59.304± 3.541 |

## 3 Experimental Results

**Datasets** We trained our model on the VDS dataset [12], where the training set has 48 multi-exposure stacks, and the testing set has 48 stacks. In addition, we evaluated our model on the HDR-Eye dataset [18], and the RAISE dataset [4]. Input images were upscaled or downscaled into $256 \times 256$ pixel resolutions by the Lanczos interpolation method [10], and all the LDR images were in the sRGB color space.

**Implementation** For training the recurrent-up and recurrent-down networks, we chose the gradient centralized Adam optimizer [26] with the learning rate of $1e^{-4}$. The momentum parameters of $\beta_1$ and $\beta_2$ were set to 0.5 and 0.999, respectively. Our model was trained on two GTX Titan X GPUs for four days to reach 80k iterations.

**Evaluation metrics** We evaluated the quality of HDR image reconstruction with the HDR-VDP-2 score [14–16]. The experiments were conducted under the same process provided with the state-of-the-art method [16, 14]. We scaled the target and generated HDR images to match the 0.1 and 99.9 percentiles before measuring the HDR-VDP-2 score.

**Comparison with the state-of-the-art methods** The comparison evaluations were performed with 6 recent deep learning-based methods, both direct methods (HDRCNN [7], ExpandNet [16], SingleHDR [14]) and multi-exposure stack-based methods (DrTMO [8], Deep recursive HDRI [13]) as benchmarks. The interchangeability of training datasets for direct methods and multi-exposure stack-based methods is limited as the direct methods need a large amount of LDR-HDR image pair datasets, and the multi-exposure stack-based method requires an adequate amount of images of different exposures. Therefore, we used pre-trained models for ExpandNet, HDRCNN, SingleHDR, and DrTMO.

The size of training datasets across different methods was imbalanced, as shown in Table 1. Compared to other models, our method was trained with much fewer scenes and outperformed both the direct and multi-exposure stack-based methods with favorable HDR-VDP-2 scores on three datasets. The result indicates that our method has a strong advantage in data efficiency. Moreover, we evaluated multi-exposure stack reconstruction results in the Appendix C to present the strength of our model toward exposure transfer tasks.

## 4 Conclusion

This paper presented a novel framework that generates both the multi-exposure stack and the HDR image. We proposed a differentiable HDR synthesis layer with a deep learning framework that converts the HDR synthesis process to be differentiable with the linear approximation technique. Hence, our approach enabled an entire network to be trained to reconstruct HDR images with direct supervision. Moreover, we used recurrent, and decomposition approaches for the multi-exposure stack generation with the purpose to disentangle the exposure transfer task. The results show that our framework achieved the state-of-the-art results for both direct and stack-based methods by removing the severe local inversion artifacts and restoring the details regardless of image conditions. For the future work, as we yielded impressive results regarding the relatively low PSNR, we will further analyze the relationship between the multi-exposure stack generation and the HDR image synthesis to optimize multiple tasks to be mutually complementary.

## 5 Acknowledgements

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP2020-2018-0-01421) supervised by the IITP(Institute for Information Communications Technology Planning Evaluation), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. 2020M3H4A1A02084899) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1D1A1B07048421)

## References

[1] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced High Dynamic Range Imaging (2nd Edition)*. AK Peters (CRC Press), Natick, MA, USA, July 2017. ISBN 9781498706940.

[2] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

[3] Can Chen, Scott McCloskey, and Jingyi Yu. Analyzing modern camera response functions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1961–1969. IEEE, 2019.

[4] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pages 219–224, 2015.

[5] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008.

[6] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. *arXiv preprint arXiv:1606.00704*, 2016.

[7] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):1–15, 2017.

[8] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6): 177–1, 2017.

[9] Michael D Grossberg and Shree K Nayar. Determining the camera response from images: What is knowable? *IEEE Transactions on pattern analysis and machine intelligence*, 25(11):1455–1467, 2003.

[10] Turkowski Ken. Filters for common resampling tasks, graphics gems i, 1990.

[11] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3116–3125, 2019.

[12] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018.

[13] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018.

[14] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. *arXiv preprint arXiv:2004.01179*, 2020.

[15] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011.

[16] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

[17] Tomoo Mitsunaga and Shree K Nayar. Radiometric self calibration. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 374–380. IEEE, 1999.

[18] Hiromi Nemoto, Pavel Korshunov, Philippe Hanhart, and Touradj Ebrahimi. Visual attention in ldr and hdr images. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, number CONF, 2015.

[19] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[20] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 267–276, 2002.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[22] Mennatullah Siam, Sepehr Valipour, Martin Jagersand, and Nilanjan Ray. Convolutional gated recurrent networks for video segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3090–3094. IEEE, 2017.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[25] 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 Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019.

[26] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient centralization: A new optimization technique for deep neural networks. *arXiv preprint arXiv:2004.01461*, 2020.

[27] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019.