End-to-End Differentiable Learning to HDR Image Synthesis for Multi-exposure Images

Jung Hee Kim[*1], **Siyeong Lee**[*2], **Suk-Ju Kang**[1]
[1] Department of Electronic Engineering Sogang University, Seoul, Korea
[2] NAVER LABS, Bundang, Korea
kjhe129@sogang.ac.kr, siyeong.lee@naverlabs.com, sjkang@sogang.ac.kr

Abstract

Recently, high dynamic range (HDR) image reconstruction based on the multiple exposure stack from a given single exposure utilizes a deep learning framework to generate high-quality HDR images. These conventional networks focus on the exposure transfer task to reconstruct the multi-exposure stack. Therefore, 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. In other words, our differentiable HDR synthesis layer helps the deep neural network to train to create multi-exposure stacks while reflecting the precise correlations between multi-exposure images in the HDRI process. In addition, our network uses the image decomposition and the recursive process to facilitate the exposure transfer task and to adaptively respond to recursion frequency. The experimental results show that the proposed network outperforms the state-of-the-art quantitative and qualitative results in terms of both the exposure transfer tasks and the whole HDRI process.

Introduction

Recently, various applications use high dynamic range imaging (HDRI) technique because it provides better aesthetic appreciation than ordinary imaging techniques with a limited dynamic range (Sen and Aguerrebere 2016). Moreover, HDRI aims to restore under-exposed and over-exposed regions, so that the reconstructed high dynamic range (HDR) images convey much information such as image details, irrespective of the illuminance change. Especially, recent vision systems have used HDRI to improve their performance in terms of robustness and consistency (e.g., passing through the tunnel). In this context, many approaches such as fusing the multi-exposure stack (Debevec and Malik 2008), implementing the event cameras (Wang et al. 2019) have been introduced to generate high-quality images with gamings (Khaldieh et al. 2018) and sports (Weber 2015).

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 (Eilertsen et al. 2017; Marnerides et al. 2018; Liu et al. 2020) and multi-exposure stack-based synthesis methods (Endo, Kanamori, and Mitani 2017; Lee, An, and Kang 2018a,b). 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 (Endo, Kanamori, and Mitani 2017). There have been many attempts to solve the data quantity problem by crawling image pairs from the internet (Kim, Oh, and Kim 2019) or generating synthetic image pairs (Liu et al. 2020). 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 (Lee, An, and Kang 2018a,b). However, they suffer from severe local inversion artifacts due to the limitations of networks being trained only with the ground truth multi-exposure stack's supervision. Therefore, the conventional approaches had difficulties training the network in an end-to-end manner to reflect the whole HDRI process.

We propose the differentiable HDR synthesis process, which enables the end-to-end training procedure and alleviates the generation of the local inversion artifacts. We also incorporate the image decomposition approach to disentangle an exposure transfer task and the recurrent network to gradually increase or decrease the exposure level to reconstruct a multi-exposure stack from a single exposure image. In summary, our contributions are three-fold as follows:

- We propose a novel framework with a differentiable HDRI synthesis method. To overcome the conventional limitations of multi-exposure stack-based HDR synthesis, we applied the differentiable CRF function, which converts discrete pixel intensity values into luminance values in the standard HDRI. By back-propagating the gradient of the loss between the network's outputs and ground truth HDR images explicitly, the networks can escape the local optimum which only focuses on the exposure transfer task, so that generates high-quality HDR images without

[*]Equal contribution

the local inversion artifacts.

- 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 individually.
- We propose a recurrent approach in the multi-exposure stack generation to efficiently utilize the recursive process. Our network learns to generate sequential images with multiple exposures in the recurrent structure as the recursive process requires to maintain gradients until the entire multi-exposure stack is generated.

Related Works

Radiometric Calibration

Recovering the scene luminance with given LDR images and reconstructing HDR images requires estimating the intensity-to-luminance mapping function of the individual camera. The estimating process of the mapping function is called the radiometric calibration. The commonly used radiometric calibration estimates mapping function, which is the camera response function (CRF), from a given multi-exposure stack and corresponding exposure values. Based on the assumption about the shape of the CRF, most approaches can be categorized into two classes: parametric and non-parametric methods.

The parametric methods assume the CRF to have a specific and analytic functional form, such as a gamma function(Mann and Picard 1994), or a polynomial function (Mitsunaga and Nayar 1999). Furthermore, Grossberg and Nayar (Grossberg and Nayar 2003) modeled CRF using the principal component analysis (PCA) to collect vectors from a large number of real CRFs. Besides, PCA based modeling methods were incorporated into recent deep learning methods (Li and Peers 2017; Liu et al. 2020). However, parametric approaches suffer from making explicit assumptions on the analytical form of the CRFs, which is not adequate for monotonic modern camera configurations (Chen, McCloskey, and Yu 2019).

Non-parametric methods focus on estimating the CRF in a discrete function with the lookup table structure. Debevec and Malik (Debevec and Malik 2008) proposed a least-square formulation with the smoothness constraints to recover CRF in discrete function form. Lee *et al.*(Lee et al. 2012) utilized the observation that images in the multi-exposure stack are linearly dependent on reconstructing HDR images. Badki *et al.*(Badki, Kalantari, and Sen 2015) proposed a radiometric calibration method to compensate significant motions in images using a random sample consensus (RANSAC)-based method. Furthermore, a recent deep learning-based approach (Endo, Kanamori, and Mitani 2017; Lee, An, and Kang 2018b) applied the most commonly used non-parametric method: Debevec and Malik's approach recovering the CRF. However, since the non-parametric radiometric calibration recovers a CRF as a discrete function and non-differentiable form, the whole end-to-end network implementation considering the multi-exposure stack has been limited.

Deep Learning-based HDR Reconstruction

Direct HDR reconstruction The recent development of deep neural networks has imposed on learning the direct mapping function between a single LDR image and a target HDR image. Direct methods generate the HDR image without fusing the image stack of different exposures, thereby removing the ghosting artifacts because a spatially aligned multi-exposure image stack is not required. Eilerstsen *et al.*(Eilertsen et al. 2017) focused on restoring saturated regions of the under-exposed LDR image to recover the luminance map, which is combined with the input LDR image to reconstruct the HDR image. Marnerides *et al.*(Marnerides et al. 2018) proposed a CNN model that trains to infer a direct mapping function between LDR and HDR images. Khan *et al.*(Khan, Khanna, and Raman 2019) implemented feedback structure in reconstructing the HDR image by iteratively refining the HDR reconstruction result. To overcome the dataset quantity challenge, Kim *et al.*(Kim, Oh, and Kim 2019) and Liu *et al.*(Liu et al. 2020) utilized the dynamic range constrained dataset, which consists of images crawled and extracted from the Internet, and the virtual dataset, respectively. However, since the datasets have diverse dynamic ranges, the normalization process and standardization process for the images become difficult. Due to the undetermined dynamic range of images, the models might be trained in the wrong direction, on account of the gap between virtually generated images and real images.

Multi-exposure stack HDR synthesis Multi-exposure stack HDR synthesis methods incorporated the deep neural network to generate a multi-exposure image stack. The dataset quantity problem of the direct methods can be compensated with a multi-exposure stack method as an arbitrary number of images with different exposures can be used as the training set. The ambiguity of the LDR-to-HDR mapping relation is avoided by focusing on the intermediate task of generating a multi-exposure stack. Endo *et al.*(Endo, Kanamori, and Mitani 2017) and Lee *et al.*(Lee, An, and Kang 2018a,b) focused on reconstructing the multi-exposure stack from a single LDR image to synthesize a target HDR image. However, these approaches caused the generation of severe local inversion artifacts on reconstructed HDR image as the methods were not trained in the end-to-end structure, where the pixel-wise relations were not imposed.

End-to-End Differentiable Learning to HDR Image Synthesis

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

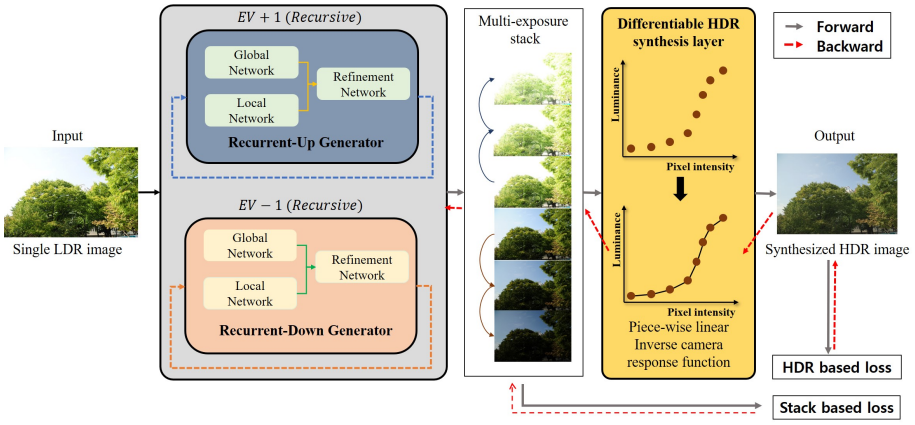

Figure 1: The overall structure of the proposed framework. Our model consists of recurrent-up and recurrent-down networks with the differentiable HDR synthesis layer. Given an input LDR image, the multi-exposure image stack is generated with recursions. Then, the generated stack is synthesized to reconstruct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

network in the end-to-end structure. We also describe our recurrent network that restores details in saturated regions of the multi-exposure stack by incorporating the image decomposition approach.

Differentiable HDR Synthesis Layer

Debevec and Malik (Debevec and Malik 2008) proposed the HDRI pipeline that estimates the CRF using the non-parametric radiometric calibration, which is commonly used. Given LDR images with different exposures, estimating the CRF or inverse CRF is modeled as the least-square problem as follows:

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

where O denotes an objective function, g denotes an inverse CRF, and Z_{ij} as a pixel intensity value of i-th pixel with j-th exposure value. Z_{min} and Z_{max} indicate minimum and maximum intensity values of given LDR images. N and P are the number of images and exposure values of the stack, and i, and j are their corresponding indices, respectively. E_i denotes the luminance value of i-th pixel and EV_j denotes the j-th exposure value. The exposure value can substitute exposure time with a fixed aperture and ISO value. The second term of the objective function regularizes the CRF to be smoothened with the hyperparameter λ. By minimizing the objective function, we can obtain the discrete CRF of g, which maps 8-bit pixel intensity values to 32-bit luminance values. With the recovered inverse CRF g, the pixel intensity value can be remapped to the luminance value as follows:

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

The scene luminance is remapped with 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 $g = [p_0, p_1, \cdots, p_N]$ with N 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 (Grossberg and Nayar 2003), we linearize the function with the prior assumptions of the CRF having the characteristic of monotonically increasing with the shape of the non-linear curve. We reformulate the function with a piece-wise linear form to back-propagate the difference between the function value and the one before, as shown in Eq. (3). The simple linearization method enables the propagation of gradients to each pixel of the multi-exposure stack with the chain rule (Goodfellow, Bengio, and Courville 2016). The gradients from the loss of luminance values flow to pixel intensity values of each image, which imposes 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.

Furthermore, we implemented the polynomial curve fitting approach proposed by Mitsunaga and Nayar(Mitsunaga and Nayar 1999) in the differentiable HDR synthesis layer. Polynomial curves can be differentiated; however, as modern cameras' CRF have monotonic and resembling shape (Chen, McCloskey, and Yu 2019), the higher-order models are not necessary. Therefore, we focused on the piece-wise linearization approach with further experiments. We verified our differentiable HDR layer outputs to reproduce the identical results with the MATLAB HDR Toolbox (Banterle et al.

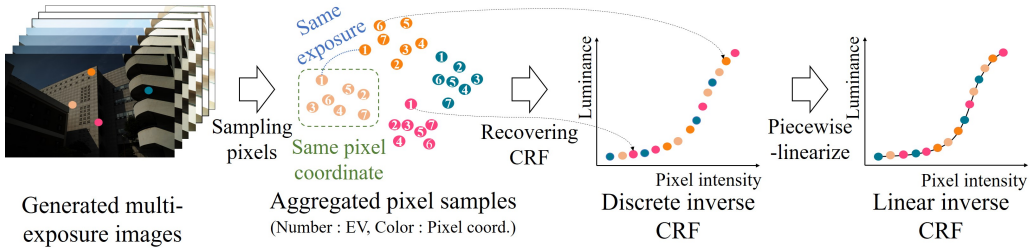

Figure 2: Conceptual diagram of the proposed piece-wise linearization for the CRF. We sample pixels from the multi-exposure stack to aggregate pixels of the same coordinate with different exposure values. We then estimate the inverse CRF with Eq. (1) and convert the function into a differentiable linear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

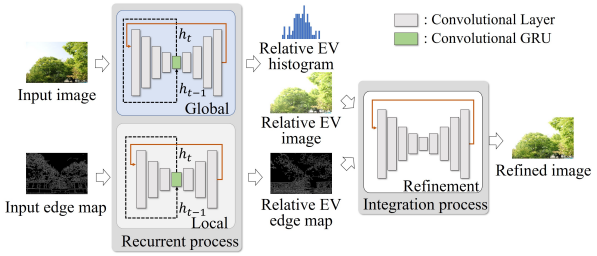

Figure 3: Sub-networks architecture. The global network focuses on minimizing the difference of histograms between the generated and target EV image, and the local network focuses on generating gradient-based edge structures. We facilitate the hidden state h_t of t-the recursion to feed into the bottleneck layers of the global and local networks for the recurrent process. We then concatenate the input image, relative EV image, and edge map to feed into the refinement network to focus on the integration process.

2017) and to generate a pixel-wise gradient.

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 (Lee, An, and Kang 2018b). We propose the recurrent-up and recurrent-down networks to be distinct from conventional methods (Lee, An, and Kang 2018a,b). Since the process is defined as a recursive process, we implement the convolutional gated recurrent unit (Conv-GRU) (Siam et al. 2017) to construct the recurrent network. In addition, as multi-exposure images have different over-exposed and under-exposed regions regarding their exposure values, we decompose the exposure transfer task into two path-ways. From a given single image, our model learns the global tone and local details with the global network and local network, respectively. With decomposed images, the refinement network integrates global and local components to generate fine-tuned images.

Fig. 3 shows the structures of sub-networks in our model. Our recurrent-up and recurrent-down networks contain three sub-networks of U-Net structures (Ronneberger, Fischer, and Brox 2015) to transfer exposures to the images with the relative up and down EVs: 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. We implemented the Swish activation (Ramachandran, Zoph, and Le 2017) on each convolution layer to alleviate a gradient vanishing problem in recurrent models. Note that the refinement network shares the same structure with the global network except for the Conv-GRUs on bottleneck layers. 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 global tones and gradient-based edge structures of a target LDR image, respectively. The image decomposition approach resolves the complexity of a single network (Lee, An, and Kang 2018b), as the adaptive response and integration are learned separately for the network to explicitly learn each corresponding task.

The recurrent-up (or recurrent-down) network exploits the same weights for transferring exposures, even with the recurrent state that differs from the exposure value of an input. However, both the recurrent-up and recurrent-down networks should adaptively produce the over-exposed and under-exposed images corresponding to the exposure value of an input. Therefore, we use the conditional instance normalization to standardize feature maps of different exposure values. The normalization transforms a feature map, X, of which the shape is $C \times H \times W$, into a normalized map Y by using two learnable parameters of γ_e and β_e with the target exposure value of e, which are in $\mathbb{R}^C$. The normalized map is formulated as $Y = \frac{\gamma_e(X-\mu)+\beta_e}{\sigma}$, where μ and σ are the mean and the standard deviation of X taken across spatial axes, respectively. In other words, our networks select the scale and shift factors according to the exposure value of an input LDR image. By using the conditional instance normalization, we can assist the network to focus on detecting subtle differences between the estimated and target images. Thus, we implemented a conditional instance normalization layer on the decoding layers of each level.

Training

Our model is designed to facilitate the recurrent structure, which shifts the exposure level of the image gradually,

as presented in Fig. 1. Specifically, the recurrent-up and recurrent-down networks are trained separately with a given single LDR image to generate the multi-exposure stack recursively. For the sub-networks, the global and local networks are trained in advance for 10k iterations, then we jointly trained the entire network, including the refinement network. The loss functions are defined independently with each sub-network. 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 (Canny 1986) of $\sigma = 2$. The refinement network is trained with L_1 loss (L_1), the contextual bilateral loss (L_{CoBi}) (Zhang et al. 2019), and the HDR loss (L_{HDR}). For the HDR loss, we used a tone-mapped HDR loss with μ-law to stabilize the training process (Yan et al. 2019). Note that L_{CoBi} 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 (Simonyan and Zisserman 2014) with the bilateral filtering. Overall loss functions are formulated as follows:

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

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

$$\begin{aligned} L_{refine} &= \lambda_4 L_1 + \lambda_5 L_{HDR} + \lambda_6 L_{CoBi} \\ &= \frac{\lambda_4}{N \cdot E} \sum_{e}^{E} \sum_{i}^{N} |\hat{I}_i^e - I_i^e| + \frac{\lambda_5}{N} \sum_{i}^{N} |log \frac{1 + \mu \hat{H}_i}{1 + \mu H_i}| \\ &+ \frac{\lambda_6}{M} \sum_{j}^{M} \min_k (\mathbb{D}_{p_j, q_k} + w_s \mathbb{D}'_{p_j, q_k}), \end{aligned} \tag{6}$$

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 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, and μ is the compression parameter of the HDR image, where we set the value with 5000. $\mathbb{D}_{p,q}$ 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.

Experimental Results

Datasets We trained our model on the VDS dataset (Lee, An, and Kang 2018a), where the training set has 48 multi-exposure stacks, and the testing set has 48 stacks. In addition, we evaluated our model on the stacks of the HDR-Eye dataset (Lee, An, and Kang 2018a; Liu et al. 2020; Nemoto et al. 2015), which is widely used for the performance evaluation. To perform evaluations on more real image dataset, we conducted experiments with the RAISE dataset (Dang-Nguyen et al. 2015). Input images were upscaled or downscaled into 256 $\times$ 256 pixel resolutions by the Lanczos interpolation method (Ken 1990), and all 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 (Yong et al. 2020) with the learning rate of $1e^{-4}$. The momentum parameters of β_1 and β_2 were set to 0.5 and 0.999, respectively. We trained our model with a batch size of 1. 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 (Liu et al. 2020; Mantiuk et al. 2011; Marnerides et al. 2018). The experiments were conducted under the same process provided with the state-of-the-art method (Marnerides et al. 2018; Liu et al. 2020). We scaled the target and generated HDR image to match the 0.1 and 99.9 percentiles before measuring the HDR-VDP-2 score. We have set the hyperparameters of HDR-VDP-2 score as the color encoding of RGB-BT.709 and 30 pixels per one visual degree. We also assessed the quality of estimated multi-exposure stacks with peak signal-to-noise ratios (PSNR), structure similarity (SSIM), and multi-scale SSIM (MS-SSIM). We then used the Reinhard tone mapping operator (Reinhard et al. 2002) for the visualization.

Comparison With The State-of-the-art Methods

The comparison evaluations were performed with 6 recent deep learning-based methods, both direct methods(HDRCNN (Eilertsen et al. 2017), ExpandNet (Marnerides et al. 2018), FHDR (Khan, Khanna, and Raman 2019), SingleHDR (Liu et al. 2020)) and multi-exposure stack-based methods (DrTMO (Endo, Kanamori, and Mitani 2017), Deep recursive HDRI (Lee, An, and Kang 2018b)) as benchmarks. The interchangeability of training datasets between 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, DrTMO, FHDR (2-iteration), and SingleHDR.

HDR quality assessment When measuring the quality of generated HDR images, we applied the method of Debevec and Malik (Debevec and Malik 2008) with the stack-based methods. The size of training datasets across different methods was imbalanced, as shown in Table 1. Compared to other models, our method is trained with much fewer scenes

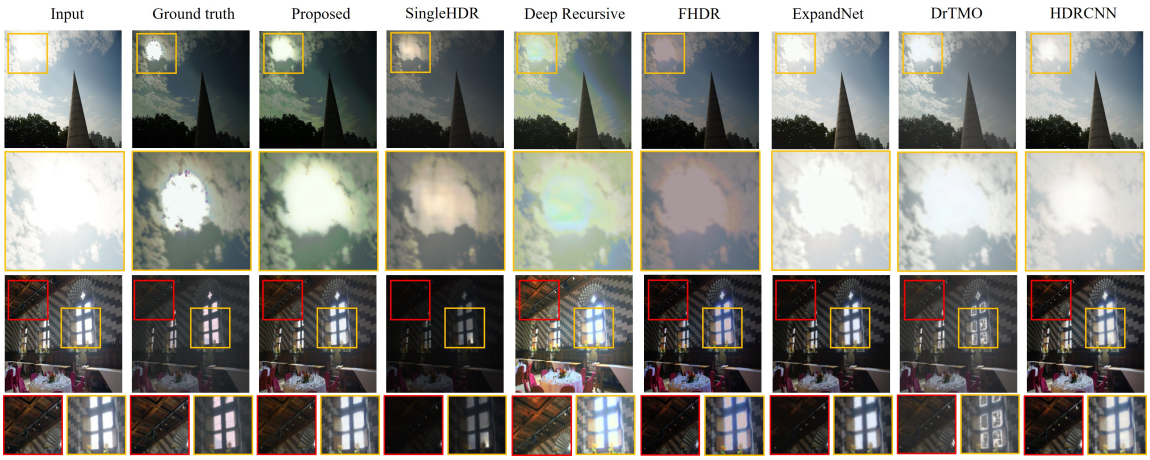

Figure 4: Comparison of tone-mapped HDR images from 6 different HDR reconstruction approaches on VDS, HDR-Eye, and RAISE datasets. The loss of image details in over-exposed and under-exposed regions occurs with the SingleHDR, FHDR, ExpandNet, and HDRCNN. The DrTMO and Deep recursive HDRI, which are stack-based methods, suffer from the local inversion artifacts. Nonetheless, our method reduces local inversion artifacts and preserves image details and contrasts in over-exposed regions.

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
FHDR	39,460 scenes	57.708±6.373	53.815±3.603	59.144±2.764
SingleHDR	10,289 scenes	55.237±4.487	54.509±3.714	59.304± 3.541

Table 1: Quantitative comparison of proposed and conventional HDR reconstruction methods. We measured the HDR-VDP-2 score (Mantiuk et al. 2011) for synthesized HDR images.

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 the data efficiency.

Multi-exposure stack reconstruction We verified the relations between the multi-exposure stack reconstruction and the HDR reconstruction. Specifically, we evaluated PSNR, SSIM, and MS-SSIM results of reconstructed stacks by our method and the previous stack-based method (Lee, An, and Kang 2018b). The previous approach (Lee, An, and Kang 2018b) focused on reconstructing the multi-exposure stack, and hence, reproducing stacks with high PSNRs, SSIMs, and MS-SSIMs. However, with the results of Fig. 5 and Table 2, our method reproduced similar PSNR, SSIM, and MS-SSIM with the previous method, but achieved much higher HDR-VDP-2 scores. The results indicate that focusing on the exposure transfer task might lead to suboptimal generation performances. Furthermore, our method does not include any adversarial loss; however, as the direct relation between pixel values was imposed during the training, we achieved the result of the highest quality, thereby providing higher HDR-VDP-2 scores.

Ablation Studies

We evaluated the effectiveness of the individual components in our model on the VDS dataset, as shown in Table 3. We added modules incrementally on the U-Net structure (Ronneberger, Fischer, and Brox 2015), which is a baseline of our model with 5-level and 2 convolutional layers for each level, and evaluated with the HDR-VDP-2 score. The overall results show that our method using all modules improved 9.305 and 4.483 with HDR-VDP-2 score and PSNR, respectively.

Recurrent network First, we added the recurrent module, the Conv-GRU (Siam et al. 2017), to be located in the bottleneck layer. We utilized the hidden state of each recurrent network to convey the important state variables, such as recursion numbers to the network. Table 3 shows that recurrent module could increase both the HDR reconstruction performance with the HDR-VDP-2 score and multi-exposure stack reconstruction with PSNR by 2.842 and 1.788, respectively.

	Method	PSNR (dB) $m \pm \sigma$	SSIM $m \pm \sigma$	MS-SSIM $m \pm \sigma$
Relative EV +1	**Proposed**	30.292±3.725	0.952±0.050	0.989±0.009
	Deep recursive HDRI (Lee, An, and Kang 2018b)	30.142±2.873	0.955±0.036	0.986±0.010
Relative EV -1	**Proposed**	30.403±3.601	0.940±0.038	0.985±0.011
	Deep recursive HDRI (Lee, An, and Kang 2018b)	30.483±3.836	0.936±0.044	0.982±0.014

Table 2: Quantitative comparison of stack reconstruction results. Relative EV+1 indicates the average value of three recursive recurrent-up results and Relative EV-1 indicates the average value of three recurrent-down results.

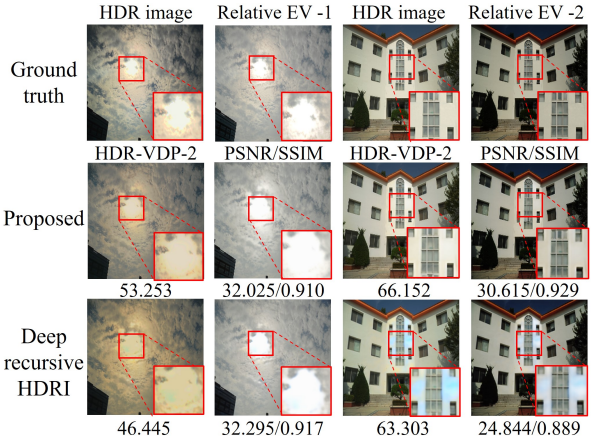

Figure 5: Case analysis of correlations between the multi-exposure stack reconstruction and the HDR reconstruction on the VDS dataset. The experiment was conducted with Lee *et al.*(Lee, An, and Kang 2018b) and our method. The result shows that two factors (stack reconstruction accuracy, HDR reconstruction accuracy) have a weak correlation (suboptimal, optimal).

Conditional instance normalization We demonstrated the effectiveness of the conditional instance normalization layer with a comparison experiment with the instance normalization layer (Ulyanov, Vedaldi, and Lempitsky 2016). We confirmed that the conditional instance normalization layer decreases the standard deviation of the reconstruction error.

Image decomposition We decomposed input images into global and local components. To verify the effectiveness of our structure, we compared the PSNR result of the decomposition network with that of the baseline network, as shown in Table 3. We trained both networks for the same iterations, and the quantitative result of PSNR shows that decomposition decreases the reconstruction error.

Differentiable HDR synthesis layer The proposed differentiable HDR synthesis layer could reconstruct the target HDR image without any learnable parameters in the layer. The mean of HDR-VDP-2 score was significantly increased by up to 3.265, and the standard deviation was decreased by up to 1.270. Hence, the differentiable HDR synthesis layer guided the network to generate the high-quality HDR image while stabilizing the training process.

Method	HDR-VDP-2	PSNR (dB)
Baseline	49.502±6.519	25.864±3.013
+ Recurrent network	52.344±6.852	27.652±3.189
+ Conditional instance normalization	53.020±5.110	27.996±2.779
+ Image decomposition	54.548±6.455	28.542±3.500
+ Differentiable HDR synthesis layer	57.813±5.185	29.592±3.596
+ Contextual bilateral loss	58.807±5.413	30.347±3.663

Table 3: Performance of various configurations on the VDS dataset (Lee, An, and Kang 2018a)

Contextual bilateral loss To enhance the perceptual quality of the generated multi-exposure stack, we added contextual bilateral loss (Zhang et al. 2019) to fine-tune our networks. This loss alleviated the limitations of using ghosting artifacts induced by applying L_1 loss on the misaligned image dataset. Table 3 shows that contextual bilateral loss fine-tunes the outputs of networks.

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 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.

Acknowledgements

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP2021-2018-0-01421) 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).

Ethics Statement

As we focus on theoretical grounds for restoring the HDR image, it has both positive and negative sides. On the negative side, our method is prone to be exploited like face generation and autonomous system attack (Müller 2020), because our method restores lost information based on the context of an input image. However, realistic images can be easily acquired from LDR images taken by a standard camera through our method. Depending on the applications using this method (e.g., autonomous driving, etc.), users should consider foreseeable potential risks, where the contents of the output images may be changed by adversarial attacks for the recurrent-up or recurrent-down networks.

References

Badki, A.; Kalantari, N. K.; and Sen, P. 2015. Robust radiometric calibration for dynamic scenes in the wild. In *2015 IEEE International Conference on Computational Photography (ICCP)*, 1–10. IEEE.

Banterle, F.; Artusi, A.; Debattista, K.; and Chalmers, A. 2017. *Advanced High Dynamic Range Imaging (2nd Edition)*. Natick, MA, USA: AK Peters (CRC Press). ISBN 9781498706940.

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

Chen, C.; McCloskey, S.; and Yu, J. 2019. Analyzing Modern Camera Response Functions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1961–1969. IEEE.

Dang-Nguyen, D.-T.; Pasquini, C.; Conotter, V.; and Boato, G. 2015. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, 219–224.

Debevec, P. E.; and Malik, J. 2008. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, 1–10.

Eilertsen, G.; Kronander, J.; Denes, G.; Mantiuk, R. K.; and Unger, J. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (TOG)* 36(6): 1–15.

Endo, Y.; Kanamori, Y.; and Mitani, J. 2017. Deep reverse tone mapping. *ACM Trans. Graph.* 36(6): 177–1.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

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

Ken, T. 1990. Filters for Common Resampling Tasks, Graphics Gems I.

Khaldieh, A.; Ploumis, S.; Pourazad, M. T.; Nasiopoulos, P.; and Leung, V. 2018. Tone mapping for video gaming applications. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, 1–2. IEEE.

Khan, Z.; Khanna, M.; and Raman, S. 2019. FHDR: HDR Image Reconstruction from a Single LDR Image using Feedback Network. *arXiv preprint arXiv:1912.11463* .

Kim, S. Y.; Oh, J.; and Kim, M. 2019. 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*, 3116–3125.

Lee, J.-Y.; Matsushita, Y.; Shi, B.; Kweon, I. S.; and Ikeuchi, K. 2012. Radiometric calibration by rank minimization. *IEEE transactions on pattern analysis and machine intelligence* 35(1): 144–156.

Lee, S.; An, G. H.; and Kang, S.-J. 2018a. Deep chain HDRI: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access* 6: 49913–49924.

Lee, S.; An, G. H.; and Kang, S.-J. 2018b. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 596–611.

Li, H.; and Peers, P. 2017. CRF-net: Single image radiometric calibration using CNNs. In *Proceedings of the 14th European Conference on Visual Media Production (CVMP 2017)*, 1–9.

Liu, Y.-L.; Lai, W.-S.; Chen, Y.-S.; Kao, Y.-L.; Yang, M.-H.; Chuang, Y.-Y.; and Huang, J.-B. 2020. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. *arXiv preprint arXiv:2004.01179* .

Mann, S.; and Picard, R. 1994. Beingundigital'with digital cameras. *MIT Media Lab Perceptual* 1: 2.

Mantiuk, R.; Kim, K. J.; Rempel, A. G.; and Heidrich, W. 2011. 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.

Marnerides, D.; Bashford-Rogers, T.; Hatchett, J.; and Debattista, K. 2018. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, 37–49. Wiley Online Library.

Mitsunaga, T.; and Nayar, S. K. 1999. Radiometric self calibration. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, 374–380. IEEE.

Müller, V. C. 2020. Ethics of artificial intelligence and robotics .

Nemoto, H.; Korshunov, P.; Hanhart, P.; and Ebrahimi, T. 2015. Visual attention in LDR and HDR images. In *9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, CONF.

Ramachandran, P.; Zoph, B.; and Le, Q. V. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941* .

Reinhard, E.; Stark, M.; Shirley, P.; and Ferwerda, J. 2002. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 267–276.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Sen, P.; and Aguerrebere, C. 2016. Practical high dynamic range imaging of everyday scenes: Photographing the world as we see it with our own eyes. *IEEE Signal Processing Magazine* 33(5): 36–44.

Siam, M.; Valipour, S.; Jagersand, M.; and Ray, N. 2017. Convolutional gated recurrent networks for video segmentation. In *2017 IEEE International Conference on Image Processing (ICIP)*, 3090–3094. IEEE.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022* .

Wang, L.; Ho, Y.-S.; Yoon, K.-J.; et al. 2019. Event-based high dynamic range image and very high frame rate video generation using conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10081–10090.

Weber, K. 2015. 4K, HDR and further image enhancements for live image acquisition. In *SMPTE 2015 Annual Technical Conference and Exhibition*, 1–21. SMPTE.

Yan, Q.; Gong, D.; Shi, Q.; Hengel, A. v. d.; Shen, C.; Reid, I.; and Zhang, Y. 2019. Attention-Guided Network for Ghost-Free High Dynamic Range Imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1751–1760.

Yong, H.; Huang, J.; Hua, X.; and Zhang, L. 2020. Gradient Centralization: A New Optimization Technique for Deep Neural Networks. *arXiv preprint arXiv:2004.01461* .

Zhang, X.; Chen, Q.; Ng, R.; and Koltun, V. 2019. Zoom to learn, learn to zoom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3762–3770.