HDR Image Reconstruction Based on Deep Learning

Di Qiu
Peking University
China
qiudi@stu.pku.edu.cn

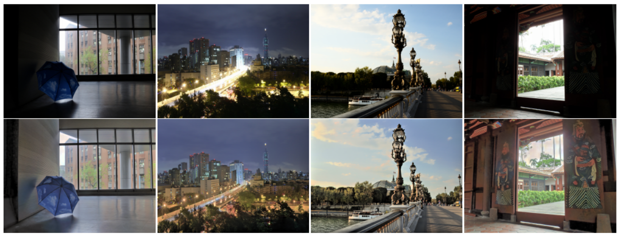


Figure 1. Some examples of HDR images reconstruction.The upper row is input LDR images,and the lower row is the reconstructed HDR images.

Abstract

This paper introduces five papers with regard to HDR images reconstruction using deep learning methods.Three of them are muti-image reconstruction and two of them are single-image reconstruction.I analyze these methods and gives a original table to summarize their characteristics and differences.Besides,I notice that HDR images reconstruction still faces many challenges,and I propose some aspects that can be researched in future work.

1. Introduction

The dynamic range of natural luminance values varies over several orders of magnitude. However, most digital photography sensors can only measure a limited fraction of this range. The resulting low dynamic range (LDR) images thus often have over or underexposed regions and don't reflect the human ability to see details in both bright and dark areas of a scene. High dynamic range (HDR) imaging has been developed to compensate for these limitations, and ideally aims to generate a single image that represents a broad range of illuminations,which are now routinely used in many applications including photo realistic image synthesis and a range of post-processing operations; for an overview see [2].

The majority of the work in the field of HDR imaging has focused on displaying HDR content and on the creation of such content from a series of LDR images, see for instance [8]. And recent years, several methods [24,26,31,33] that focus on recovering an HDR image from only a single LDR input,which is called Single-image HDR reconstruction,have been also developed.

In this paper,I present three methods [20, 39, 40] to merge several LDR images captured at different exposures based on deep learning and two methods [9, 28] for single-image HDR reconstruction.The first approach proposed by KALANTARI [20] uses a convolutional neural network to generate the HDR image from a set of images aligned with optical flow.And the second approach proposed by Wu *et al*. [39] formulate HDR imaging as an image translation problem without optical flows.The third one is achieved by a novel attention-guided end-to-end deep neural network (AHDRNet),which is proposed by Yan *et al*. [40].The fourth and the fifth is about single-image HDR reconstruction,which the former [9] is by deep convolutional neural networks (CNNs) and the latter [28] is by learning to reverse the camera pipeline.These approaches solve problems to a great extent,such as alleviating ghosting artifacts,recovering missing information in under-exposed regions and hallucinating the contents in saturated regions,which contribute a lot to the reconstruction of HDR images.

2. Related Work

High dynamic range imaging has been the subject of extensive research over the past decades. One class of techniques captures HDR images in a single shot by modifying the camera hardware. For example, a few methods use a beam-splitter to split the light to multiple sensors [37]. Several approaches propose to reconstruct HDR images from coded per pixel exposure [15] or modulus images [43]. These methods produce high-quality results on dynamic scenes since they capture the entire image in a single shot. Unfortunately, they require cameras with a specific optical system or sensor, which are typically custom made and expensive and, thus, not available to the general public.

Another category of approaches reconstructs HDR images from LDR images. Since bracketed exposure images can be easily captured with standard digital cameras, these methods are popular and used in widely available devices such as smartphone cameras. They categorize these approaches into two general classes and discuss them next.

Multi-image HDR reconstruction. The most common technique for creating HDR images is to fuse a stack of bracketed exposure LDR images [8]. To handle dynamic scenes, image alignment and post-processing are required to minimize artifacts [23,29,36]. Conventionally, HDR reconstruction has been performed by non-learning-based brightness enhancement through filtering or light-source detection.Bogoni [7] estimated motion vectors using optical flow and used parameters to warp pixels in the exposures. Kang *et al*. [21] transformed intensities of LDR images to the luminance domain using exposure time information and computed the optical flow to find corresponding pixels among the LDR images. Sen *et al*. [34] proposed a patch-based energy minimization approach that integrates alignment and HDR reconstruction in a joint optimization. Hu *et al*. [17] optimized image alignment based on brightness and gradient consistencies on the transformed domain. Hafner *et al.* [13] proposed an energy-minimization approach which simultaneously calculates HDR irradiance and displacement fields. However, non-learning-based approaches cannot estimate physically accurate amounts of light due to the lack of knowledge about real HDR images; thus, the quality of the estimated HDR images is limited. Recent methods apply CNNs to fuse multiple flow-aligned LDR images [19] or unaligned LDR images [39]. These methods have the advantage that they can exploit information extracted from training data to identify and compensate for image regions that do not meet the assumptions underlying the HDR process.

Single-image HDR reconstruction. Single-image HDR reconstruction does not suffer from ghosting artifacts but is significantly more challenging than the multi-exposure counterpart. Early approaches estimate the density of light sources to expand the dynamic range [1, 3–6] or apply the cross-bilateral filter to enhance the input LDR images [18, 25]. With the advances of deep CNNs [14, 35], several methods have been developed to learn a direct LDR-to-HDR mapping [30, 41, 42]. Given a single input LDR image, Endo use an auto-encoder [16] to generate a set of LDR images with different exposures. These images are then combined to reconstruct the final HDR image. Lee *et al*. [26] chain a set of CNNs to sequentially generate the bracketed LDR images. Later, they propose *et al.* [26]to handle this application through a recursive conditional generative adversarial network (GAN) *et al.* [12] combined with a pixel-wise l_1 loss.In contrast to these approaches, a few methods [9, 30, 41] directly reconstruct the HDR image without generating bracketed images. [9]use a network with U-Net architecture to predict the values of the saturated areas, whereas linear non-saturated areas are obtained from the input. [30]present a novel dedicated architecture for end-to-end image expansion. [41] reconstruct HDR image for image correction application. They train a network for HDR reconstruction to recover the missing details from the input LDR image, and then a second network transfers these details back to the LDR domain.

3. Deep High Dynamic Range Imaging of Dynamic Scenes

KALANTARI *et al.* [20]first used the optical flow method of Liu [27] to align the images with low and high exposures to the one with medium exposure, which is called the reference image,and they proposed a convolutional neural network which can avoid artifacts to a great extent to generate the HDR image from a set of aligned images.The contributions of the work are as follows:

- They propose the first machine learning approach for reconstructing an HDR image from a set of bracketed exposure
- They fully explore the idea by presenting three different system architectures and comparing them extensively. LDR images of a dynamic scene.
- They introduce the first dataset suitable for learning HDR reconstruction, which can facilitate future learning research in this domain. In addition, their dataset can potentially be used to compare different HDR reconstruction approaches.And we can see later in the survey that the dataset is indeed widely used in research.

3.1. Algorithm

Given a set of three LDR images of a dynamic scene (Z_1, Z_2, Z_3), the goal is to generate a ghost-free HDR image, H, which is aligned to the medium exposure image Z_2 (reference). This process can be broken down into two stages of 1) alignment and 2) HDR merge. During alignment, the LDR images with low and high exposures, defined with Z_1 and Z_3, respectively, are registered to the reference image, denoted as Z_2. This process produces a set of aligned images, I = $\{I_1, I_2, I_3\}$, where $I_2 = Z_2$. These aligned images are then combined in the HDR merge stage to produce an HDR image, H.

Extensive research on the problem of image alignment (stage 1) has resulted in powerful techniques over the past decades,but often produce artifacts around the motion boundaries and on the occluded regions.Since the aligned images are used during the HDR merge (stage 2) to produce the final HDR image, these artifacts could potentially appear in the final result.

The authors observed that the alignment artifacts from the first stage can be significantly reduced through the HDR merge in the second stage and inspired by the recent success of deep learning in a variety of applications,they proposed to model this process with a convolutional neural network (CNN).

Preprocessing the Input LDR Images. They first linearize LDR images using the camera response function (CRF) if they are not in the RAW format,then apply gamma correction (γ = 2.2) on them to produce the input images to their system, Z_1, Z_2, Z_3.

Alignment. They first raise the exposure of the darker image to the brighter one since optical flow methods require brightness constancy to perform well.Then They compute the flow between Z_3 and Z_2whose brightness has been adjusted using the optical flow algorithm by Liu [27].Finally, they use bicubic interpolation to warp the high exposure image Z_3 using the calculated flow.This process produces a set of aligned images I = $\{I_1, I_2, I_3\}$.

HDR Merge. The main challenge of this component is to detect the alignment artifacts and avoid their contribution to the final HDR image.In their system, they use machine learning to model this complex task. Therefore, there are two main issues to be addressed: the choice of 1) model, and 2) loss function. Let's see loss function first because there are some mathematical symbols introduced and will appear in the introduction of models.

*1)Loss Function:*Since HDR images are usually displayed after tonemapping, they choose to compute the loss function between the tonemapped estimated and ground truth HDR images.And they propose to use μ-law, a commonly-used range compressor in audio processing, which is differentiable and suitable for their learning system.This function is defined as:

$$T = \frac{log(1+\mu H)}{log(1+\mu)} \tag{1}$$

where μ is a parameter which defines the amount of compression and is set to 5000 in their implementation, H is the HDR image in the linear domain, and T is the tonemapped image. They train the learning system by minimizing the l^2 distance of the tonemapped estimated and ground truth HDR images defined as

$$E = \sum_{K=1}^{3} (\hat{T}_k - T_k)^2 \tag{2}$$

where $\hat{T}_k$ and T_k are the estimated and ground truth tonemapped HDR images and the summation is over color channels.

2)Model:They propose three different architectures:direct, weight estimator (WE) architecture,weight and image estimator (WIE) architecture.

Direct. In this architecture,they model the entire HDR merge process using a CNN.The CNN takes a stack of aligned images in the LDR and HDR domains as input, $\{I, H\}$ and outputs the final HDR image, H.The goal of training is to find the optimal network weights, w, by minimizing the error between the estimated and ground truth tonemapped HDR images,defined in Eq.(2).To compute the derivative, of the error with respect to the network weights,they use the chain rule to break down this derivative into three terms as:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial w} \tag{3}$$

The second term is the derivative of the tonemapping function, defined in Eq.(1),and can be computed as:

$$\frac{\partial \hat{T}}{\partial \hat{H}} = \frac{\mu}{log(1+\mu)} \frac{1}{1+\mu \hat{H}} \tag{4}$$

Weight Estimator (WE). The existing techniques typically compute a weighted average of the aligned HDR images to produce the final HDR result:

$$\hat{H}(p) = \frac{\sum_{j=1}^{3} \alpha_j(p) H_j(p)}{\sum_{j=1}^{3} \alpha_j(p)}, \quad \text{where} \quad H_j(p) = \frac{I_j^Y}{t_j}. \tag{5}$$

Here, the weight $\alpha_j(p)$ basically defines the quality of the j^{th} aligned image at pixel p and needs to be estimated from the input data.

They propose to learn the weight estimation process using a CNN. In this case, the CNN takes the aligned LDR and HDR images as input, $\{I, H\}$, and outputs the blending weights, α.

The this derivative is broken down into four terms as:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial \alpha} \frac{\partial \boldsymbol{\alpha}}{\partial w} \tag{6}$$

Since the estimated HDR image in this case is obtained using Eq.(5), we can compute the third term as

$$\frac{\partial \hat{H}}{\partial \alpha_i} = \frac{H_i(p) - \hat{H}(p)}{\sum_{j=1}^{3} \alpha_j(p)} \tag{7}$$

This architecture is more constrained than the direct architecture and easier to train. Therefore, it produces high-quality results with significantly fewer residual artifacts.

Weight and Image Estimator (WIE). In this architecture they relax the restriction of the previous architecture by allowing the network to output refined aligned images in addition to the blending weights.They use Eq.(5) to compute the final HDR image using the refined images, $\tilde{I}_i$, and the estimated blending weights, α_i.The derivative can be calculated as follows:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial \{\boldsymbol{\alpha}, \tilde{I}\}} \frac{\partial \{\boldsymbol{\alpha}, \tilde{I}\}}{\partial w} \tag{8}$$

They propose to perform the training in two stages so that the convergence can be faster.In the first stage, they force the network to output the original aligned images as the refined ones, by minimizing the ℓ^2 error of the output of the network and the original aligned images.In the second stage, they perform a direct end-to-end training and further optimize the network by synthesizing refined aligned images.

Network Architecture. In their system, the networks have a decreasing filter size starting from 7 in the first layer to 1 in the last layer. All the layers with the exception of the last layer are followed by a rectified linear unit (ReLU). For the last layer, they use sigmoid activation function so the output of the network is always between 0 and 1. And they use a fully convolutional network, so our system can handle images of any size.

Discuss. The direct architecture is the simplest among the three, but in rare cases leaves small residual alignment artifacts in the results. The WE architecture is the most constrained one and is able to better suppress the artifacts in these rare cases. Finally, similar to the direct architecture, the WIE architecture is able to synthesize content that is not available in the aligned LDR images. However, the direct and WIE architectures slightly overblur images in dark regions to suppress the noise.Therefore, they believe the WE is the most stable architecture and produces results with the best visual quality.

3.2. Dataset

Training deep networks usually requires a large number of training examples. In this case, each training example should consist of a set of LDR images of a dynamic scene and their corresponding ground truth HDR image. Unfortunately, most existing HDR datasets either lack ground truth images *et al*. [38], are captured from static scenes *et al*. [10], or have a small number of scenes with only rigid motion [22].

To overcome this problem, they created their own training dataset of 74 different scenes and substantially extend it through data augmentation,which also brings benefits to the follow-up studies.

Capturing Process. The goal is to produce a set of LDR images with motion and their corresponding ground truth HDR image. First,they capture a static set with different exposures and use a simple triangle weighting scheme, similar to the method of Debevec and Malik [8], to merge them into a ground truth HDR image.Next, they capture a dynamic set to use as our input by asking the subject to move.After discarding scenes containing unacceptable motions,they got 74 training scenes.

Data Augmentation. They used color channel swapping and geometric transformation (rotating 90 degrees and flipping) with 6 and 8 different combinations, respectively,finally increasing the number of scenes to 740.

Patch Generation. They break down the training images into overlapping patches of size 40 × 40 with a stride of 20 and select the proper training patches,which results in around 1,000,000 selected patches.

3.3. Limitations

The main limitation of this approach is that the network takes a specific number of images as the input because the model is trained by a set of three input images.Another limitation is that in some cases, because of the camera motion, the low and high exposure images do not have information at the boundaries of the image,and the approach will result in being noisy or saturated.

4. Deep High Dynamic Range Imaging with Large Foreground Motions

Different from KALANTARI [20]'s using optical flow to align images, this paper [39] proposes the first non-flow-based deep framework for HDR imaging of dynamic scenes with large-scale foreground motions,and resolve the issues that in some cases color artifacts and geometry distortions appears using KALANTARI's approach due to the unreliability of the optical flow. And it can hallucinate plausible details in largely saturated regions with large foreground motions,and recovers highlight regions greatly.Also,it can be easily extended with more inputs, and with different reference images, not limited to the medium exposure LDR,which is a limitation of [20].

4.1. Approach

They formulate the problem of HDR imaging as an image translation problem and focus on handling large foreground motions.

Network Architecture. The framework is essentially a symmetric encoder-decoder architecture, with two variants, *Unet* whose more details can be seen in [32] and *ResNet* which is similar to Image Transformation Networks proposed in [19] and the middle layers is replaced with residual blocks.

The overall architecture can be conceptually divided into three components: encoder, merger and decoder. Instead of duplicating the whole network, which may defer the merging, they separate the first two layers as encoders for each exposure inputs. After extracting the features, the network learns to merge them, mostly in the middle layers, and to decode them into an HDR output, mostly in the last few layers.

The encoding layers are convolution layers with a stride of 2, while the decoding layers are deconvolution layers kernels with a stride of 1/2. The output of the last deconvolution layer is connected to a flat-convolution layer to produce the final HDR. All layers use 5 × 5 kernels, and are followed by batch normalization (except the first layer and the output layer) and leaky ReLU (encoding layers) or ReLU (decoding layers). The channel numbers are doubled each layer from 64 to 512 during encoding and halved from 512 to 64 during decoding.And they used the dataset provided by [20]for training and testing.

Processing Pipeline and Loss Function They denote the set of input LDRs by $\mathcal{I} = \{I_1, I_2, I_3\}$, sorted by their exposure biases. They first map them to $\mathcal{H} = \{H_1, H_2, H_3\}$ in the HDR domain, and use simple gamma encoding for this mapping:

$$H_i = \frac{I_i^{\gamma}}{t_i}, \gamma > 1 \tag{9}$$

where t_i is the exposure time of image I_i .The values of I_i, H_i and H are bounded between 0 and 1.

They then concatenate $\mathcal{I}$ and $\mathcal{H}$ channel-wise into a 6-channel input and feed it directly to the network as suggested in [20].The network f is thus defined as:

$$\hat{H} = f(\mathcal{I}, \mathcal{H}) \tag{10}$$

where $\hat{H}$ is the estimated HDR image, and is also bounded between 0 and 1.And the loss funnction is defined as: [20].The network f is thus defined as:

$$\mathcal{L}_{Unet} = \|\mathcal{T}(\hat{H}) - \mathcal{T}(H)\|_2 \tag{11}$$

where H is the ground truth HDR image and $\mathcal{T}(H)$ is μ-law,as Eq.(1).

4.2. Limitations

Since they are focused on handling large foreground motions, they align the backgrounds of the LDR inputs using homography transformation.Without background alignment, the network tends to produce blurry edges where background is largely misaligned.And homography is not always perfect,for example,if there is existence of parallax effects in saturated regions.What's more,recovering massive saturated regions with minimal number of input LDRs is still a challenge for the network.

5. Attention-guided Network for Ghost-free High Dynamic Range Imaging

In this paper [40], Yan propose an attention-guided deep neural network (AHDRNet) for HDR imaging.The neural network learns the relationships between input LDR images and HDR output.The attention modules generate soft attention maps to evaluate the importance of different image regions for obtaining the required HDR image.By doing this, they overcome one of the primary problems in HDR imaging is that it is robust to large misalignments of image pixels and saturation.

5.1. Attention-guided Network for HDR Imaging

Following the settings in [20, 39], they use three LDR images (I_1, I_2, I_3)(sorted by their exposure lengths),and use gamma correction to generate a corresponding set of $\{H_i\}$.As suggested in [20], they concatenate images I_i and H_i along the channel dimension to obtain the 6-channel tensors $X_i = [I_i, H_i]$, i = 1, 2, 3 as the input of the network.

5.2. AHDRNet Architecture

Unlike the previous methods [20, 39] that stack the input images X_i or the extracted feature maps in the early stage of the network for merging, the proposed AHDRNet obtains the attention maps by comparing the encoded image features and then merges features with the guidance of the attention maps.The AHDRNet consists of two major subnetworks, the attention network (for feature extraction) and the merging network (for HDR image estimation).

Attention network.. The attention module is used to exclude the harmful components caused by misalignment and saturation or highlight the useful details. The network first uses a shared encoding layer to extract feature maps Z_i , i = 1,2,3 with 64 channels from three inputs.They feed the features Z_i , i = 1,3 of the non-reference images to the convolutional attention module $a_i(\cdot)$, i = 1,3 along with the reference image feature map Z_r, and then obtain the attention maps A_i for the non-reference images:

$$A_i = a_i\left(Z_i, Z_r\right), i = 1, 3 \tag{12}$$

The predicted attention maps are used to attend the features of the non-reference images via:

$$Z_i' = A_i \circ Z_i, i = 1, 3 \tag{13}$$

where ∘ denotes the point-wise multiplication and Z_j' denotes the feature maps with attention guidance.

Then they stack the images:

$$Z_s = \mathrm{Concat}\left(Z_1', Z_2, Z_3'\right) \tag{14}$$

where Concat$(\cdot)$ denotes the concatenation operation. Z_s will be used as the input of the merging network.

The attention modules $a_i(\cdot)$, i = 1, 3 in Eq. (14) are two small CNNs and followed by a ReLU activation and a sigmoid activation, respectively.

Merging network. The network consists of several convolution layers, dilated residual dense blocks and several skip connections.By using dilated residual dense blocks, the receptive field at each block is expanded. And the network is showed in Fig2.

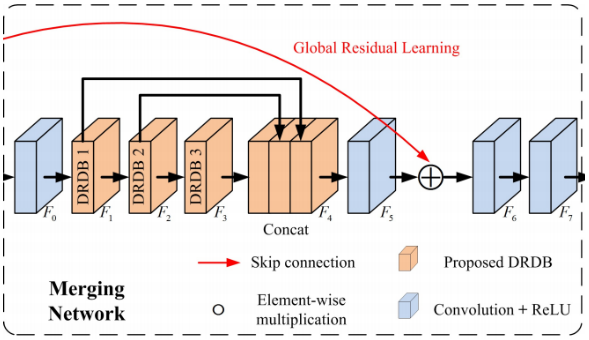


Figure 2. The merging network is constructed based on a series of dilated residual dense blocks (DRDBs). The global residual skip connection is used to boost the training.

5.3. Loss Function

In this method, they train the network by minimizing ℓ_1- norm based distance between the tonemapped estimated and the ground truth HDR images. The loss function is defined as:

$$\mathcal{L} = \|\mathcal{T}(\widehat{H}) - \mathcal{T}(H)\|_1 \tag{15}$$

where $\mathcal{T}(H)$ is μ-law,seen in Eq.(1)

6. HDR image reconstruction from a single exposure using deep CNNs

The Section 6 and Section 7 will focus on two approaches for single-image HDR reconstruction.In this paper,EILERTSEN *et al.* [9]propose a novel method for reconstructing HDR images from LDR input images, by estimating missing information in bright image parts, such as highlights, lost due to saturation of the camera sensor.It can reconstruct a high quality HDR image from an arbitrary single exposed LDR image, provided that saturated areas are reasonably small.And they propose a hybrid dynamic range autoencoder that is tailored to operate on LDR input data and output HDR images.

6.1. HDR Reconstruction Model

Problem formulation and constraints The final HDR reconstructed pixel $\hat{H}_i$,c with spatial index i and color channel c is computed using a pixel-wise blending with the blend value α_i,

$$\hat{H}_{i,c} = (1 - \alpha_i) f^{-1} (D_{i,c}) + \alpha_i \exp (\hat{y}_{i,c}) \tag{16}$$

where $D_{i,c}$ is the input LDR image pixel and $\hat{y}_{i,c}$ is the CNN output (in the log domain).The inverse camera curve f^{-1} is used to transform the input to the linear domain.

$$\alpha_i = \frac{\max (0, \max_c (D_{i,c}) - \tau)}{1 - \tau} \tag{17}$$

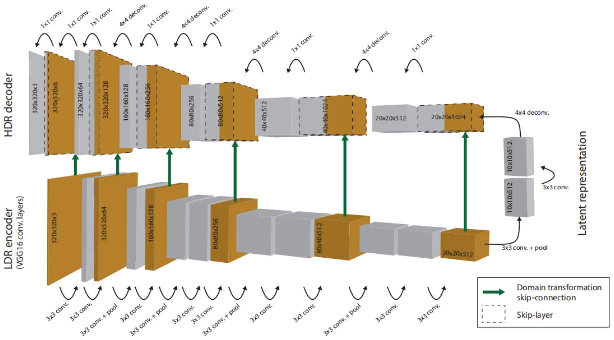


Figure 3. Fully convolutional deep hybrid dynamic range autoencoder network, used for HDR reconstruction. The encoder converts an LDR input to a latent feature representation, and the decoder reconstructs this into an HDR image in the log domain. The skip-connections include a domain transformation from LDR display values to logarithmic HDR, and the fusion of the skip-layers is initialized to perform an addition. The network is pre-trained on a subset of the Places database, and deconvolutions are initialized to perform bilinear upsampling. While the specified spatial resolutions are given for a 320 × 320 pixels input image, which is used in the training, the network is not restricted to a fixed image size.

Hybrid dynamic range autoencoder Autoencoder architectures transform the input to a low-dimensional latent representation, and a decoder is trained to reconstruct the full-dimensional data [16].The complete autoencoder design is depicted in Figure 3.More specifically, the complete LDR to HDR skip connection is defined as:

$$\tilde{\boldsymbol{h}}_i^D = \sigma \left(W \left[\log \left(f^{-1} \left(\boldsymbol{h}_i^E \right) + \epsilon \right) \right] + \boldsymbol{b} \right) \tag{18}$$

Adding the skip-connections enables a more optimal use of existing details.

HDR loss function In this system the HDR decoder is designed to operate in the log domain. Thus, the loss is formulated directly on logarithmic HDR values, given the predicted log HDR image $\hat{y}$ and the linear ground truth H,

$$\mathcal{L}(\hat{y}, H) = \frac{1}{3N} \sum_{i,c} \left| \alpha_i \left(\hat{y}_{i,c} - \log \left(H_{i,c} + \epsilon \right) \right) \right|^2 , \tag{19}$$

where N is the number of pixels.

As the visual system may indirectly perform such separation when inferring reflectance or discounting illumination [11].They therefor propose another, more flexible loss function that treats illuminance and reflectance separately.

$$\begin{aligned} \log \left(I_i^{\hat{y}} \right) &= \left(G_\sigma * L^{\hat{y}} \right)_i \\ \log \left(R_{i,c}^{\hat{y}} \right) &= \hat{y}_{i,c} - \log \left(I_i^{\hat{y}} \right) \end{aligned} \tag{20}$$

and the parameters are:

$$L_i^{\hat{y}} = \log\left(\sum_c w_c \exp\left(\hat{y}_{i,c}\right)\right), w = \{0.213, 0.715, 0.072\}$$
$$G_\sigma = \text{ Gaussian } \quad \text{low} \; - \; \text{pass} \quad \text{ filter }, \sigma = 2 \tag{21}$$

The illumination component I describes the global variations, and is responsible for the high dynamic range. The reflectance R stores information about details and colors.

The resulting loss function using I and R is defined as:

$$\begin{aligned}\mathcal{L}_{IR}(\hat{\boldsymbol{y}}, H) = \tfrac{\lambda}{N} \textstyle\sum_i \left|\alpha_i \left(\log\left(I_i^{\hat{y}}\right) - \log\left(I_i^{\boldsymbol{y}}\right)\right)\right|^2 \\ + \tfrac{1-\lambda}{3N} \textstyle\sum_{i,c} \left|\alpha_i \left(\log\left(R_{i,c}^{\hat{y}}\right) - \log\left(R_{i,c}^{y}\right)\right)\right|^2\end{aligned} \tag{22}$$

6.2. Limitations

There is a content-dependent limit on how much missing information the network can handle which is generally hard to quantify.For example,structures and details of images with a large region with saturation in all color channels cannot be inferred.Besides,a situation where besides a similar loss of spatial structures, extreme intensities are underestimated.

There is also a limitation on how much compression artifacts that can be present in the input image. If there are blocking artifacts around highlights, these will impair the reconstruction performance to some extent.

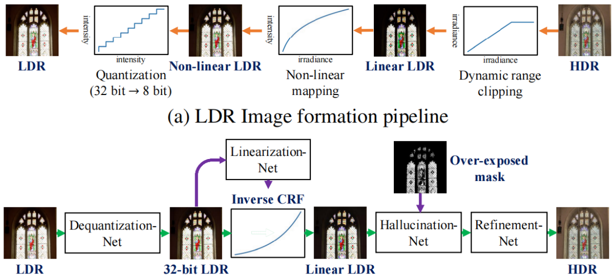
7. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline

In contrast to existing learning-based methods, the core idea of this paper [28] is to incorporate the domain knowledge of the LDR image formation pipeline into the model. They model the HDR-to-LDR image formation pipeline as the (1) dynamic range clipping, (2) non-linear mapping from a camera response function,(3) quantization, and then propose to learn three specialized CNNs to reverse these steps.By explicitly modeling the inverse functions of the LDR image formation pipeline, they significantly reduce the diffi-culty of training one single network for reconstructing HDR images and get fantastic results.

7.1. Learning to Reverse the Camera Pipeline

As shown in Fig 4,the process of converting one HDR image to one LDR image can be modeled by the following major steps:(1) Dynamic range clipping. (2) Non-linear mapping.(3) Quantization.

To learn the inverse mapping , they propose to decompose the HDR reconstruction task into three sub-tasks: dequantization, linearization, and hallucination, which model the inverse functions of the quantization, non-linear mapping, and dynamic range clipping, respectively.See in Fig 4.



(a) LDR Image formation pipeline

(b) Proposed method

Figure 4. The LDR Image formation pipeline and overview of single-image HDR reconstruction methods.

Dequantization The Dequantization-Net adopts a 6-level U-Net architecture. Each level consists of two convolutional layers followed by a leaky ReLU ($\alpha = 0.1$) layer.And they use the Tanh layer to normalize the output of the last layer to [-1.0, 1.0].Finally, they add the output of the Dequantization-Net to the input LDR image to generate the dequantized LDR image $\hat{I}_{\text{deq}}$.And the loss function is defined as:$\mathcal{L}_{\text{deq}} = \left\|\hat{I}_{\text{deq}} - I_n\right\|_2^2$.

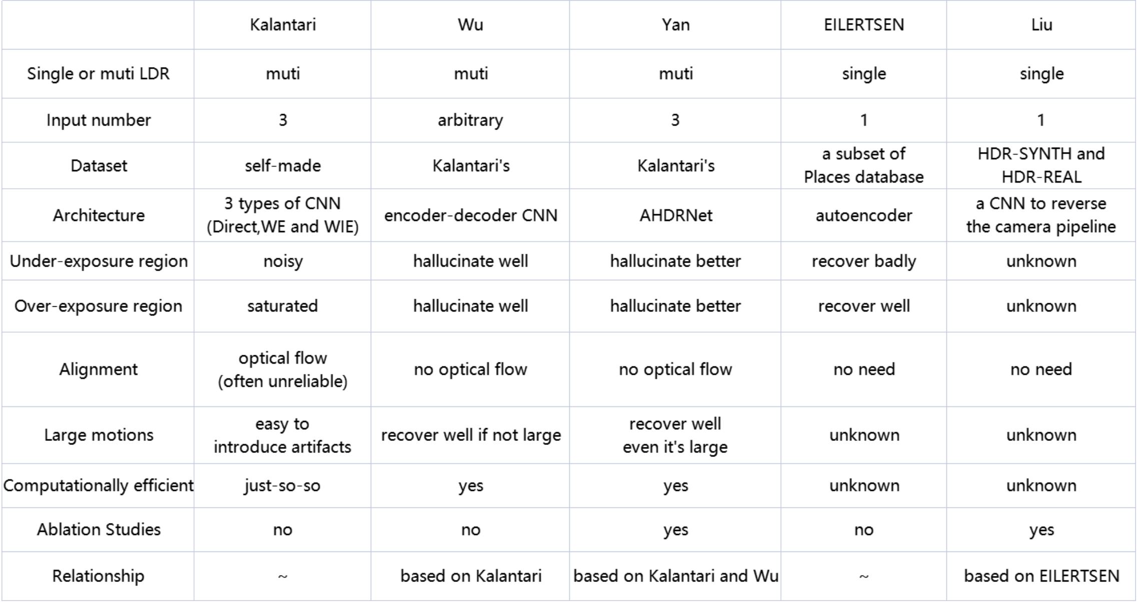
Linearization The goal of linearization is to estimate a CRF and convert a non-linear LDR image to a linear irradiance.To predict the inverse CRF, they train a Linearization-Net to estimate the weights from the input non-linear LDR image.They use the ResNet-18 [14] as the backbone of our Linearization-Net. To extract a global feature, they add a global average pooling layer after the last convolutional.They then use two fully-connected layers to generate K PCA weights and reconstruct an inverse CRF.

To satisfy the constraint that a CRF/inverse CRF should be monotonically increasing, they adjust the estimated inverse CRF by enforcing all the first-order derivatives to be non-negative. Specifi- cally, they calculate the first-order derivatives by $g_1' = 0$ and $g_d' = g_d - g_{d-1}$ for $d \in [2, \cdots, 1024]$ and find the smallest negative derivative $g_m' = \min\left(\min_d\left(g_d'\right), 0\right)$.They then shift the derivatives by $\tilde{g}_d' = g_d' - g_m'$.The inverse CRF $\tilde{\mathbf{g}} = [\tilde{g}_1, \cdots, \tilde{g}_{1024}]$is then reconstructed by integration and normalization:

$$\tilde{g}_d = \frac{1}{\sum_{i=1}^{1024} \tilde{g}_i'} \sum_{i=1}^{d} \tilde{g}_i'. \tag{23}$$

The linear LDR image reconstruction loss is defined as $\mathcal{L}_{\text{lin}} = \left\|\hat{I}_{\text{lin}} - I_c\right\|_2^2$,and the inverse CRF reconstruction loss is defined as $\mathcal{L}_{\text{crf}} = \|\tilde{\mathbf{g}} - \mathbf{g}\|_2^2$ .And they train the Linearization-Net by optimizing $\mathcal{L}_{\text{lin}} + \lambda_{\text{crf}}\mathcal{L}_{\text{crf}}$.

Hallucination They adopt an encoder-decoder architecture with skip connections as our Hallucination-Net. The reconstructed HDR image is modeled by $\hat{H} = \hat{I}_{\text{lin}} + \alpha \cdot \mathcal{C}^{-1}\left(\hat{I}_{\text{lin}}\right)$, where $\hat{I}_{\text{lin}}$ is the image generated from the



	Kalantari	Wu	Yan	EILERTSEN	Liu
Single or muti LDR	muti	muti	muti	single	single
Input number	3	arbitrary	3	1	1
Dataset	self-made	Kalantari's	Kalantari's	a subset of Places database	HDR-SYNTH and HDR-REAL
Architecture	3 types of CNN (Direct,WE and WIE)	encoder-decoder CNN	AHDRNet	autoencoder	a CNN to reverse the camera pipeline
Under-exposure region	noisy	hallucinate well	hallucinate better	recover badly	unknown
Over-exposure region	saturated	hallucinate well	hallucinate better	recover well	unknown
Alignment	optical flow (often unreliable)	no optical flow	no optical flow	no need	no need
Large motions	easy to introduce artifacts	recover well if not large	recover well even it's large	unknown	unknown
Computationally efficient	just-so-so	yes	yes	unknown	unknown
Ablation Studies	no	no	yes	no	yes
Relationship	~	based on Kalantari	based on Kalantari and Wu	~	based on EILERTSEN

Figure 5. The original summary table of the five papers.

Linearization-Net and $\alpha = \max\left(0, \hat{I}_{\text{lin}} - \gamma\right)/(1 - \gamma)$ is the over-exposed mask with γ = 0.95.Since the missing values in the over-exposed regions should always be greater than the existing pixel values, they constrain the Hallucination-Net to predict positive residuals by adding a ReLU layer at the end of the network.And the loss function is $\mathcal{L}_{\text{hal}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}} + \lambda_{\text{tv}}\mathcal{L}_{\text{tv}}$,where $\mathcal{L}_{\text{hal}} = \|\log(\hat{H}) - \log(H)\|_2^2$,$\mathcal{L}_{\text{p}}$ is the perceptual loss [19] and $\mathcal{L}_{tv}$ is the total variation(TV) loss.

Joint training First,they train the three models respectively.After they converge,they jointly fine-tune the entire pipeline to reduce error accumulation by minimizing the combination of loss functions $\mathcal{L}_{\text{total}}$:

$$\lambda_{\text{deq}}\,\mathcal{L}_{\text{deq}} + \lambda_{\text{lin}}\,\mathcal{L}_{\text{lin}} + \lambda_{\text{crf}}\,\mathcal{L}_{\text{crf}} + \lambda_{\text{hal}}\,\mathcal{L}_{\text{hal}} + \lambda_{\text{p}}\mathcal{L}_{\text{p}} + \lambda_{\text{tv}}\,\mathcal{L}_{\text{tv}} \tag{24}$$

w=$\{1, 10, 1, 1, 0.001, 0.1\}$

Refinement Refinement-Net adopts the same U-Net architecture as the Dequantization-Net, which learns to refine the output of the Hallucination-Net by a residual learning and is proved effective.

7.2. Ablation studies

To demonstrate the effectiveness of explicitly reversing the camera pipeline, they train our entire model (including all sub-networks) from scratch without any intermediate supervisions and find the performance of such a model drops significantly.which shows that the stage-wise training is effective, and the performance improvement does not come from the increase of network capacity.

8. Future work

Existing methods have solve many problems and makes the results better and better. While the advantages of these methods are clear,they are yet to be perfect solution.I also observe some challenges in the HDR images construction.

8.1. Dataset

Both training and evaluation of HDR imaging algorithms require high quality annotated datasets. But creating a high quality HDR dataset with such features still poses several challenges.So I hold the view that introduce a robust and proper dataset for training deep network for HDR images reconstruction is significant.

8.2. Large motions recovered by single image

Notice that,Wu [39] reconstruct HDR images with large foreground motions brilliantly.However,the single-image reconstruction methods aren't able to make it well.So I consider this will be a interesting work for us to do.

9. Conclusion

In this paper,I present five methods to reconstruct HDR images by deep learning.They solve many problems such as large motions recovery,hallucinate well in the over-exposure region and introduce dataset that is proper for deep learning.Also,there are still many problems waiting for us to resolve.I am looking forward to more robust methods and prospect of more extensive applications of HDR imaging.

References

[1] Ahmet Oğuz Akyüz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bülthoff. Do hdr displays support ldr content? a psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 26(3):38–es, 2007. 2

[2] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017. 1

[3] Francesco Banterle, Kurt Debattista, Alessandro Artusi, Sumanta Pattanaik, Karol Myszkowski, Patrick Ledda, and Alan Chalmers. High dynamic range imaging and low dynamic range expansion for generating hdr content. In *Computer graphics forum*, volume 28, pages 2343–2367. Wiley Online Library, 2009. 2

[4] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006. 2

[5] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics*, pages 33–41, 2008. 2

[6] Francesco Banterle, Patrick Ledda, Kurt Debattista, Alan Chalmers, and Marina Bloj. A framework for inverse tone mapping. *The Visual Computer*, 23(7):467–478, 2007. 2

[7] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 7–12. IEEE, 2000. 2

[8] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 1, 2, 4

[9] 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. 1, 2, 6

[10] Brian Funt and Lilong Shi. The rehabilitation of maxrgb. In *Color and imaging conference*, volume 2010, pages 256–259. Society for Imaging Science and Technology, 2010. 4

[11] Alan Gilchrist and Alan Jacobsen. Perception of lightness and illumination in a world of one reflectance. *Perception*, 13(1):5–19, 1984. 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[13] David Hafner, Oliver Demetz, and Joachim Weickert. Simultaneous hdr and optic flow computation. In *2014 22nd International Conference on Pattern Recognition*, pages 2065–2070. IEEE, 2014. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 7

[15] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014. 1

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2, 6

[17] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1163–1170, 2013. 2

[18] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5):507–517, 2014. 2

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2, 4, 8

[20] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 1, 2, 4, 5

[21] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 2

[22] Kanita Karaduzovic-Hadziabdic, Jasminka Hasic Telalovic, and Rafal Mantiuk. Subjective and objective evaluation of multi-exposure high dynamic range image deghosting methods, 2016. 4

[23] Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *2006 International Conference on Image Processing*, pages 2005–2008. IEEE, 2006. 2

[24] Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. Fhdr: Hdr image reconstruction from a single ldr image using feedback network. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5. IEEE, 2019. 1

[25] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE, 2014. 2

[26] 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. 1, 2

[27] Ce Liu et al. *Beyond pixels: exploring new representations and applications for motion analysis*. PhD thesis, Massachusetts Institute of Technology, 2009. 2, 3

[28] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 1, 7

[29] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image*

Processing XXXIII, volume 7798, page 779812. International Society for Optics and Photonics, 2010. 2

[30] 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. 2

[31] Kenta Moriwaki, Ryota Yoshihashi, Rei Kawakami, Shaodi You, and Takeshi Naemura. Hybrid loss for learning single-image-based hdr reconstruction. *arXiv preprint arXiv:1812.07134*, 2018. 1

[32] 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. 4

[33] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *arXiv preprint arXiv:2005.07335*, 2020. 1

[34] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012. 2

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[36] Abhilash Srikantha and Désiré Sidibé. Ghost detection and removal for high dynamic range images: Recent advances. *Signal Processing: Image Communication*, 27(6):650–662, 2012. 2

[37] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *ACM Transactions on Graphics (TOG)*, 30(4):1–10, 2011. 1

[38] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for hdr images. In *Computer Graphics Forum*, volume 35, pages 139–152. Wiley Online Library, 2016. 4

[39] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 1, 2, 4, 5, 8

[40] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 1, 5

[41] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Image correction via deep reciprocating hdr transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2018. 2

[42] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4519–4528, 2017. 2

[43] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015. 1