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

Qingsen Yan[1,2†], Dong Gong[2†], Qinfeng Shi[2],
Anton van den Hengel[2], Chunhua Shen[2], Ian Reid[2], Yanning Zhang[1*]
[1]School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2]The University of Adelaide, Australia
https://donggong1.github.io/ahdr

Abstract

Ghosting artifacts caused by moving objects or misalignments is a key challenge in high dynamic range (HDR) imaging for dynamic scenes. Previous methods first register the input low dynamic range (LDR) images using optical flow before merging them, which are error-prone and cause ghosts in results. A very recent work tries to bypass optical flows via a deep network with skip-connections, however, which still suffers from ghosting artifacts for severe movement. To avoid the ghosting from the source, we propose a novel attention-guided end-to-end deep neural network (AHDRNet) to produce high-quality ghost-free HDR images. Unlike previous methods directly stacking the LDR images or features for merging, we use attention modules to guide the merging according to the reference image. The attention modules automatically suppress undesired components caused by misalignments and saturation and enhance desirable fine details in the non-reference images. In addition to the attention model, we use dilated residual dense block (DRDB) to make full use of the hierarchical features and increase the receptive field for hallucinating the missing details. The proposed AHDRNet is a non-flow-based method, which can also avoid the artifacts generated by optical-flow estimation error. Experiments on different datasets show that the proposed AHDRNet can achieve state-of-the-art quantitative and qualitative results.

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.

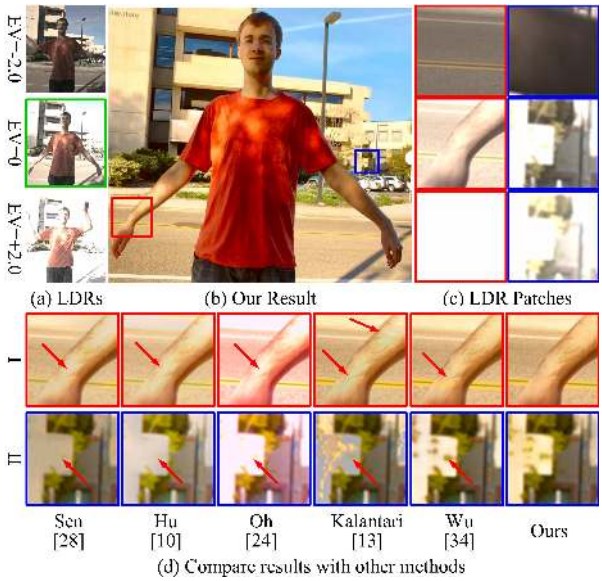


Figure 1. LDR images with different exposures are shown in (a), and our result after tonemapping is shown in (b). The areas of the images that exhibit both large-scale movement and saturation are displayed in (c). The proposed AHDRNet generates an HDR image with less ghosting artifacts and more details in saturated regions (See zoomed-in patches in (d)).

Some specialized hardware devices [24, 35] have been proposed to produce HDR images directly, but they are usually too expensive to be widely adopted. As a result, computational HDR imaging methods have drawn more attention. The most common strategy is to take a series of LDR images at different exposures and then merge them into an HDR image [2, 21, 23, 29, 39]. In multiple exposure meth-

*† The first two authors contributed equally to this work. This work was partially supported by NSFC (61871328), ARC (DP140102270, DP160100703). Q. Yan and Y. Zhang were partially supported by National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology. Q. Yan was supported by a scholarship from CSC.

ods, one of the LDR images is usually considered as the reference image (shown with the green border in Figure 1 (a)). Although these methods often generate high-quality HDR results when the scene and camera are completely static, they will suffer from significantly ghosting and blurring artifacts when there is motion between the input images.

Global image misalignments can be compensated for using homographies [33, 34, 37]. However, the ghosting artifacts caused by moving objects and the missing details due to saturation are complex to overcome. To tackle the ghosting issue, some methods first carry out a more detailed alignment of the LDR images before merging [9, 14, 31]. A variety of alignment procedures have been applied (*e.g*. optical flow [15, 16, 46]), but they still suffer from the artifacts due to the estimation error. To avoid this alignment error, some methods [25, 28] proposed to reject the misaligned moving components as outliers directly. However, pixel-accurate identification of moving objects is difficult to achieve robustly, particularly when relying on simple pixel level characteristics (*e.g*. pixel color [28]).

Inspired by the successes of the deep neural networks (DNNs) in many image restoration tasks [18, 7, 40, 8], some deep learning-based approaches [15, 37, 38] have been proposed recently to improve the HDR image composition process. In [15], a DNN is proposed to merge the LDR images after an optical flow based alignment process. However, the DNN cannot handle the distortions caused by the inevitable optical flow estimation error (See Kalantari *et al.*'s method in Figure 1 (d)). In [37], the HDR imaging task is treated as an image translation problem. Although the model can produce satisfactory results in some examples, it still suffers from ghosting artifacts when there are large-scale movements between the images. The DNN-based methods can hallucinate some details in regions with saturation, but the existing methods cannot handle large areas of saturation, particularly when there is also occlusion.

We propose an attention-guided deep neural network (AHDRNet) for HDR imaging (See Figure 2). The neural network learns the relationships between input LDR images and HDR output. Previous methods [15, 37] take stacked LDR images, or LDR image feature maps, as the input to the merging process, which mixes the misaligned image components at an early stage of the network, making it difficult to obtain ghost-free HDR results. Considering that ghosting is primarily an artifact of object motion and misalignments [15], we propose the learnable attention modules to guide the merging process. The attention modules generate soft attention maps to evaluate the importance of different image regions for obtaining the required HDR image. They are expected to highlight the features complementary to the reference image and exclude regions with motion and severe saturation. The LDR image features with attention guidance are then fed to the merging network to generate the HDR image. We construct the merging network using dilated residual dense blocks (DRDBs), which are achieved by employing the dilated convolution layers in the residual dense block (RDB) proposed in [43]. The RDBs help to make full use of information from different convolutional layers, thus preserving more details from the input LDR images. The dilated convolutions enlarge the receptive fields, helping to recover the details contaminated by saturation and moving objects. The main contributions of the paper can be summarized as:

- We propose a new attention-guided network for ghost-free HDR imaging. It has all of the benefits of a neural network model, and overcomes one of the primary problems in HDR imaging is that it is robust to large misalignments of image pixels and saturation.
- We propose a network based on dilated residual dense blocks to merge the attention guided feature maps from LDR images. The dilated residual dense blocks can simultaneously preserve the image details and enlarge the receptive fields, allowing the network to hallucinate the contents in saturated regions and produce HDR images with rich details.
- Extensive experiments on different datasets validate the superiority of the proposed AHDRNet. We also conduct ablation studies to quantify the roles of different components in our model.

2. Related Work

The primary relevant works are as follows.

Methods relying on pixel rejection These approaches label each pixel as belonging to a static region or a moving object based on the assumption that the images are globally registered. Grosch [9] defined an error map that uses the color difference of inputs to get the ghost-free HDR image. Jacobs *et al.* [14] detected ghost regions based on a weighted variance measure. Pece and Kautz [27] computed the median threshold bitmap for input images to detect motion regions. Heo *et al.* [11] roughly detected motion regions by joint probability densities and these regions are refined using energy minimization based on graph-cuts methods. Zhang and Cham [42] proposed quality measures based on image gradients to generate a weighting map over the inputs. Rank minimization [19, 25] has also been used to detect motion regions and reconstruct HDR images. Even it is achieved to the required pixel accuracy, rejecting pixels reduces the information available to reconstruct the HDR image, which often leads to missing details (See Oh's method [25] in Figure 1).

Methods relying on registration These approaches reconstruct each HDR region by searching for the best matching region in LDR images. This is achieved using pixel (optical flow methods) or patch (patch-based methods) based

dense correspondences. Bogoni [1] estimated motion vectors using optical flow and used parameters to warp pixels in the exposures. Kang *et al.* [16] 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.* [30] proposed a patch-based energy minimization approach that integrates alignment and HDR reconstruction in a joint optimization. Hu *et al.* [12] optimized image alignment based on brightness and gradient consistencies on the transformed domain. Hafner *et al.* [10] proposed an energy-minimization approach which simultaneously calculates HDR irradiance and displacement fields. This approach improves robustness, but fails for large motions, doesn't learn by examples, and makes no attempt to compensate for saturation.

Deep learning based methods Many deep learning approaches [3, 15, 37] have been developed. Eilertsen *et al.* [3] proposed a deep autoencoder network to predict HDR values from one image. Endo [4] synthesized multiple LDR images from one LDR image with the deep-learning-based approach, then reconstructed an HDR image by merging them. Kalantari *et al.* [15] used optical flow to align the input images to the reference image, then employed a convolutional neural network to obtain the HDR image. Wu *et al.* [37] proposed a network that can learn to translate multiple LDR images into a ghost-free HDR image. 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. Each method addresses an important issue, but none has the flexibility and robustness that the proposed attention-based approach enables (See Figure 1).

Attention mechanisms in deep learning methods Attention has shown to be a pivotal development in deep learning and has been used in many computer vision applications. Lu *et al.* [20] proposed a novel adaptive attention model with a visual sentinel for image captioning. Fan *et al.* [5] stacked latent attention for multiple multimodal reasoning tasks. Zhao *et al.* [44] proposed a diversified visual attention network to address the problem of fine-grained object classification. Each has achieved the hitherto impossible performance and robustness by allowing models to focus on only the relevant information.

3. Attention-guided Network for HDR Imaging

Given a series of LDR images of a dynamic scene $(I_1, I_2, ..., I_k)$ with different exposures, the target of HDR imaging is to recover an HDR image H aligned to a prescribed reference image I_r (selected from the input LDR images). All of the images I_i and H are RGB images with three channels. Following the settings in [15, 37], we use three LDR images (I_1, I_2, I_3) (sorted by their exposure lengths), *i.e.* $k = 3$, and let the middle exposure image I_2 be the reference image.

Before feeding the LDR images to the network, we first map the input LDR images $\{I_i\}$ to the HDR domain relying on gamma correction [15, 37] to generate a corresponding set of $\{H_i\}$:

$$H_i = I_i^{\gamma}/t_i, \ \forall i = 1, 2, 3, \tag{1}$$

where $\gamma > 1$ denotes the gamma correction parameter and t_i denotes the exposure time of the image I_i. We set $\gamma = 2.2$ in this work. As suggested in [15], we 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. Intuitively, the LDR images L_i help to identify the noisy and saturated regions, while the H_i facilitate the detection of the alignments[15]. Given (X_1, X_2, X_3) as input, the proposed AHDRNet obtains the HDR image by

$$H = f(X_1, X_2, X_3; \theta), \tag{2}$$

where $f(\cdot)$ denotes the proposed HDR network, and θ is the network parameters. The attention mechanism works as part of the end-to-end AHDRNet network $f(\cdot)$. Note that the input images of the proposed model can be the original images without any alignment preprocessing.

3.1. Overview of the AHDRNet Architecture

Unlike the previous methods [15, 37] 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. As shown in Figure 2, the AHDRNet consists of two major subnetworks, *i.e.* the *attention network* (for feature extraction) and the *merging network* (for HDR image estimation).

The *attention network* first separately extracts features from each LDR image relying on the corresponding convolutional encoders. Then, we apply specific attention maps on the *non-reference images* to identify the beneficial features. The attention maps are obtained via the attention modules according to the feature maps from the reference image and each non-reference image. Considering that the target of the model is to generate the HDR image with the scene consistent to the reference image, the motivation of applying attention on the non-reference images is to identify the misaligned components before merging the features for alleviating the ghosting artifacts.

The *merging network* takes the features extracted with the attention guidance as input and estimates the HDR image relying on a series of dilated residual dense blocks (DRDBs) and the global residual learning (GRL) strategy. The DRDBs and GRL help to utilize the image features effectively and obtain the HDR image with plausible details.

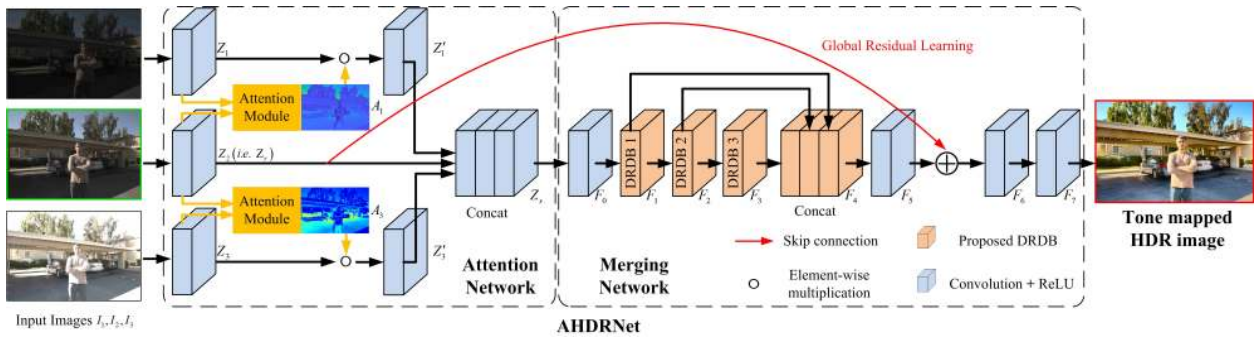


Figure 2. The architecture of the proposed AHDRNet. The network consists of an *attention network* for feature extraction and a *merging network* for predicting the HDR image. The attention module is used to exclude the harmful components caused by misalignment and saturation or highlight the useful details. 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. The final HDR result is obtained by tonemapping. All the feature maps have 64 channels, and the kernel size is 3. The visualized map is a presentation of the averaged attention feature A_i.

The merging network fuses the features from the LDR images and hallucinates the details in the regions contaminated by the saturation and misaligned moving objects.

3.2. Attention Network for Feature Extraction

Given three 6-channel input images $X_i, i = 1, 2, 3$ corresponding to the three LDR images, the attention network first uses a shared encoding layer to extract feature maps $Z_i, i = 1, 2, 3$ with 64 channels from three inputs. For clarity, we define notations X_r and Z_r to indicate X_2 and Z_2 corresponding to the reference LDR image in some special context. As shown in Figure 2, to obtain the attention maps for the non-reference images, we 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:

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

A_i has the same size as Z_i. The values in A_i are in the range $[0, 1]$. Details of the *attention modules* are provided below. The predicted attention maps are used to attend the features of the non-reference images via:

$$Z_i' = A_i \circ Z_i,\ i = 1, 3, \tag{4}$$

where $\circ$ denotes the point-wise multiplication and Z_i' denotes the feature maps with attention guidance.

Instead of stacking the original feature maps Z_i's for HDR merging, we stack the reference feature map Z_r (*i.e.* Z_2) and the features of the non-reference images Z_1' and Z_3' for merging. The attention network thus obtains a stack of features with the guidance of the reference as Z_s

$$Z_s = \text{Concat}(Z_1', Z_2, Z_3'), \tag{5}$$

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

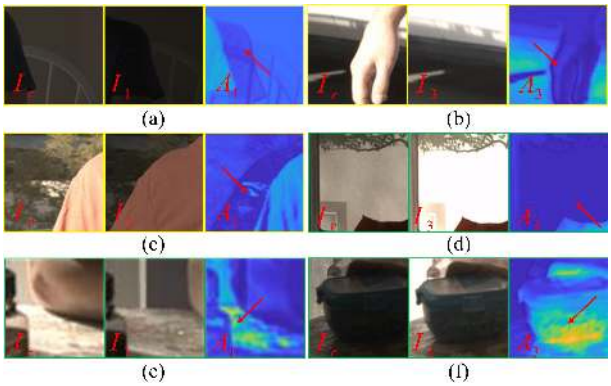


Figure 3. Example image patches and the corresponding attention maps. In (a)-(f), from left to right: the reference image, one non-reference image, and the attention map applied on the non-reference image. (a), (b) and (c) display attention maps for the significantly misaligned regions. (d), (e) and (f) show the attention maps can highlight useful regions.

Since the HDR imaging process centers on the reference image, the attention maps are predicted and applied according to the reference. As shown in Figure 3, the attention maps can suppress the misaligned (See (a) (b) and (c)) and saturated regions (See (d)) in the non-reference images, which avoids the harmful features getting into the merging process and thus alleviates the ghosts from the source. When some regions in the reference are saturated (See (e)) or noisy (See (f)), the attention maps can also highlight useful features in the non-reference images. More studies in Section 4.2.1 further prove the effectiveness of the proposed attention mechanism in HDR imaging.

Attention module The attention modules $a_i(\cdot), i = 1, 3$ in Eq. (3) are two small CNNs. The structure of the attention modules is shown in Figure 4. The attention module first concatenates the input feature maps Z_i and Z_r and obtains the attention map after two convolution (Conv) layers. Each Conv layer applies 64 3×3 layers. The two Conv layers are

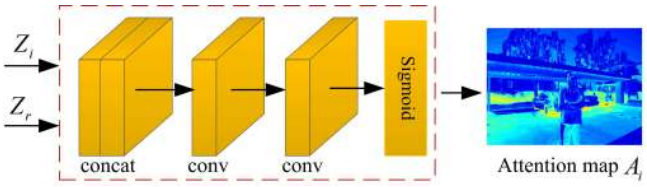


Figure 4. The attention unit first concatenates the two inputs and then obtains attention maps via two Conv layers, which restricts the output in [0,1] using a sigmoid activation.

followed by a ReLU activation and a sigmoid activation, respectively. As a result, the attention module can obtain the 64-channel attention map A_i with values in the range $[0, 1]$.

3.3. Merging Network for HDR Image Estimation

The merging network takes the stacked feature map Z_s and the reference image feature map Z_r as input. In the design of the merging network, we take account of the characteristics of the HDR imaging problem and use the basic structure of the residual dense network in [43] as the reference. As shown in Figure 2, the network consists of several convolution layers, dilated residual dense blocks and several skip connections. The generated feature maps at different layers are noted as $F_j, j = 0, 1, ..., 7$.

Given the stacked feature Z_s, the merging network first obtains a 64-channel feature map after a Conv layer, and then feed it into three DRDBs, which results in three corresponding feature maps F_1, F_2 and F_3. Instead of using the RDB proposed in [43], we proposed to use the RDBs with dilated convolution (DRDB) for HDR imaging. The details of DRDB can be found in the following. By applying 3×3 Conv on the concatenated feature map F_4, we generate the merged and transferred feature map F_5.

Global residual learning with the reference features Before reconstructing the HDR image from F_5, inspired by the super-resolution methods [18, 43], we apply a global residual learning strategy to obtain feature maps by

$$F_6 = F_5 + Z_r, \tag{6}$$

where Z_r is the shallow feature map of the reference image. The merging network thus tends to learn the residual features. In the proposed AHDRNet, we have the shallow feature map Z_r containing the pure information from the reference image. We thus apply the global residual learning with the reference feature maps. We consider that the feature map F_6 contains enough information to reconstruct the HDR image. Empirical studies in Section 4.2.1 show the effectiveness of the global residual learning strategy.

After two convolution layers (followed by activations), we estimate the HDR image $\widehat{H}$ in the HDR domain. The final HDR image is displayed via the tonemapping operation (See Section 3.4).

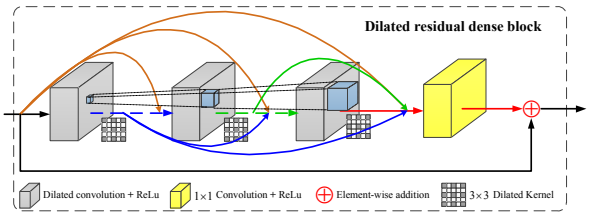


Figure 5. Illustration of dilated residual dense block structure with three convolution layers. We adopt a residual dense block [43] as its backbone and each convolution layer can be substituted by dilated convolution. By using dilated residual dense blocks, the receptive field at each block is expanded.

Dilated residual dense block Since the reconstruction of some local areas of the HDR images cannot get enough information from the LDR images due to the occlusion of moving objects and saturation, the merging network requires larger receptive field for hallucinating details. We thus apply the 2-dilated convolutions [41] in the residual dense block (RDB) [43]. As shown in Figure 5, the proposed dilated residual dense block (DRDB) consists of a series of Conv layers followed by ReLU activations and dense concatenation based skip-connections. Each Conv layer takes the concatenation of all the feature maps from previous layers as input. In contrast to the dense block proposed in [13], the RDB and DRDB apply a local residual skip-connection between the input and output of a block. More details of the RDB can be found in [43]. In our implementation, we use 6 Conv layers in each DRDB. The empirical ablations studies in Section 4.2.1 show the effectiveness of the DRDBs.

3.4. Training Loss

As described in Section 3.3, the proposed AHDRNet predicts the HDR image $\widehat{H}$ in the HDR domain. Since the HDR images are usually displayed after tonemapping, training the network on the tonemapped images is more effective than training directly in the HDR domain [15]. Given an HDR image H in HDR domain, we compress the range of the image using μ-law:

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)}, \tag{7}$$

where μ is a parameter defining the amount of compression and $\mathcal{T}(H)$ denotes the tonemapped image. In this work, we always keep H in the range $[0, 1]$ and set $\mu = 5000$. The tonemapper in Eq. 7 is differentiable, which is very suitable for training the network.

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

$$\mathcal{L} = \|\mathcal{T}(\widehat{H}) - \mathcal{T}(H)\|_1. \tag{8}$$

We also tested the ℓ_2 loss used in previous work [15, 37] and noticed that ℓ_1 loss is more powerful for preserving details (See Section 4.2.2), which is consistent with the observation in [45].

3.5. Implementation Details

In our implementation, we apply 64 3×3 features in the Conv layers, which are followed by ReLU activations, if not specified otherwise. We set the stride size for all Conv layers as 1 and keep the feature map size using zero padding. We define the output layer to produce 3-channel images. The growth rate of all DRDBs is 32. The last Conv layer in each DRDB applies 1×1 convolution to compress the feature maps.

For training, we use Adam optimizer [17] and set the batch size and learning rate as 8 and 1×10^{-5}, receptively. Given training images, we randomly crop the 256×256 patches for training. All weights of the network are initialized using Xavier method [6]. We implement our model using PyTorch [26], which takes takes 0.32s to process a 1500×1000 image with an NVIDIA GeForce 1080 Ti GPU.

4. Experiments

4.1. Experimental Settings

Training data We train the AHDRNet on the HDR dataset [15] which includes 74 samples for training and 15 samples for testing. For each sample, three different LDR images are captured with exposure biases of $\{-2, 0, +2\}$ or $\{-3, 0, +3\}$. Transformations on the cropped patches are applied as data augmentation to alleviate overfitting.

Testing data We test the proposed AHDRNet on the Kalantari's dataset [15] and the datasets without ground truth, such as Sen's [30] and Tursun's [36] datasets.

Evaluation Metrics We conduct evaluations with four metrics as the following. We compute the PSNR values for images after tonemapping using μ-law (PSNR-μ), Matlab function *tonemap* (PSNR-M), and linear (PSNR-L) domains. We also conducted a quantitative evaluation by computing the HDR-VDP-2 [22].

4.2. Ablation Studies

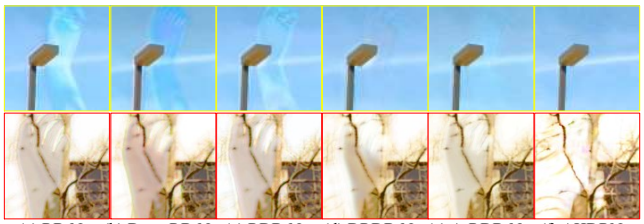
4.2.1 Study on the Model Architecture

We investigate the architecture of AHDRNet and validate the importance of different individual components in the whole AHDRNet. We achieve this ablation study by comparing the proposed AHDRNet and the following variants of AHDRNet:

- **AHDRNet**. The full model of the AHDRNet.
- **DRDB-Net** (*i.e.* AHDRNet w/o attention). We remove the attention module in this variant, in which the feature maps Z_i's are directly stacked and fed to the merging network.
- **A-RDB-Net** (*i.e.* AHDRNet w/o dilation). We do not use dilated convolution in this variant of AHDRNet.
- **RDB-Net** (*i.e.* AHDRNet w/o attention and dilation). This variant of AHDRNet does not contain the attention operation and dilated convolution layers.
- **RB-Net** (*i.e.* AHDRNet w/o attention, dilation and densely connection). This baseline is a merging network based on the residual block (RB). We replace the DRDBs as the same number of RBs.
- **Deep-RB-Net**. More RBs are used to approach the model compressibility of the RDB-Net.

Table 1. Quantitative comparisons of different models. All scores are the average across all testing images.

	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
RB-Net	39.8648	28.3548	38.0044	60.1905
Deep-RB-Net	41.1788	29.5414	38.9679	60.2724
RDB-Net	41.2058	29.4335	38.9747	60.5107
DRDB-Net	42.7345	31.4169	39.7800	60.8740
A-RDB-Net	43.0536	32.2025	40.5105	61.6362
w/o GRL	42.5313	32.9552	40.7558	62.2827
AHDRNet	**43.6172**	**33.0429**	**41.0309**	**62.3044**



(a) RB-Net (b) Deep-RB-Net (c) RDB-Net (d) DRDB-Net (e) A-RDB-Net (f) AHDRNet

Figure 6. Visual results of AHDRNet and its baseline variants.

Attention module. The attention module is a very effective mechanism for HDR image de-ghosting task. As shown in Figure 6, compared with RDB-Net, A-RDB-Net can alleviate the ghosting artifacts due to the attention module. A similar result can be observed with DRDB-Net and AHDR-Net. Although DRDB-Net can remove ghosting artifacts, it tends to generate artifacts in saturated regions (the bottom patch of Figure 6). The proposed AHDRNet with attention module can eliminate ghosting artifacts while retaining the background information (See Figure 3). As shown by the quantitative results in Table 1, AHDRNet and A-RDB-Net acquire a better improvement than the DRDB-Net and RDB-Net.

Dilated residual dense blocks. Compared with DRDB-Net, the RB-Net results have visible ghosts (See Figure 6 (a) and (d)). Even the results of Deep-RB-Net that has more RBs cannot remove ghosting artifacts. Hence, increasing the depth of the network is not a practical approach to en-

hance HDR image quality. On the other hand, the DRDB-Net with the same network depth can capture more contents and alleviate ghosts. The performance of DRDB-Net in Table 1 is better than RB-Net and RDB-Net.

Dilated convolution. To demonstrate the capability of dilated convolution, we compare the RDB-Net and DRDB-Net. As displayed in Figure 6 (c) and (d), the results of DRDB-Net alleviate ghosting artifacts compared with RDB-Net. The results show that a larger receptive field is helpful to suppress the ghosting artifacts and hallucinate the missing details. Furthermore, the proposed AHDRNet can completely remove ghosts. The quantitative comparisons in Table 1 show that the models with dilated convolution can obtain high values on PSNR metrics.

Global residual learning. We also study the performance of global residual learning (GRL) strategy. Quantitative comparisons of the results are shown in Table 1. Since GRL helps to transfer information from front layers, the model with GRL can bring better performance.

4.2.2 Study on Training Loss Function

In this experiment, we compare the performances of our method with different loss functions. Quantitative comparisons of the results are shown in Table 2, which implies that the ℓ_1 loss is more powerful for preserving details as discussed in [45]. We thus train our model using ℓ_1 loss.

Table 2. Quantitative comparisons of different loss functions.

	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
ℓ_2 loss	43.0630	31.7921	40.6798	62.0169
ℓ_1 loss	**43.6172**	**33.0429**	**41.0309**	**62.3044**

4.3. Comparison with the State-of-the-art Methods

We evaluate the proposed method and compare with previous state-of-the-art methods on a variety of datasets. Specifically, we compare the proposed method with two patch-based methods [30, 12], the method based on motion detection [25], the flow-based approach with DNN merger [15] and the DNN method without optic flow [37]. In addition, we compare with single frame HDR imaging methods [4, 3]. For all methods, we employed the codes provided by the authors. The same training dataset and setting are used for deep learning methods. Furthermore, we also apply the proposed AHDRNet with the input images aligned via estimated optical flow [32] (referred to as Ours+OF).

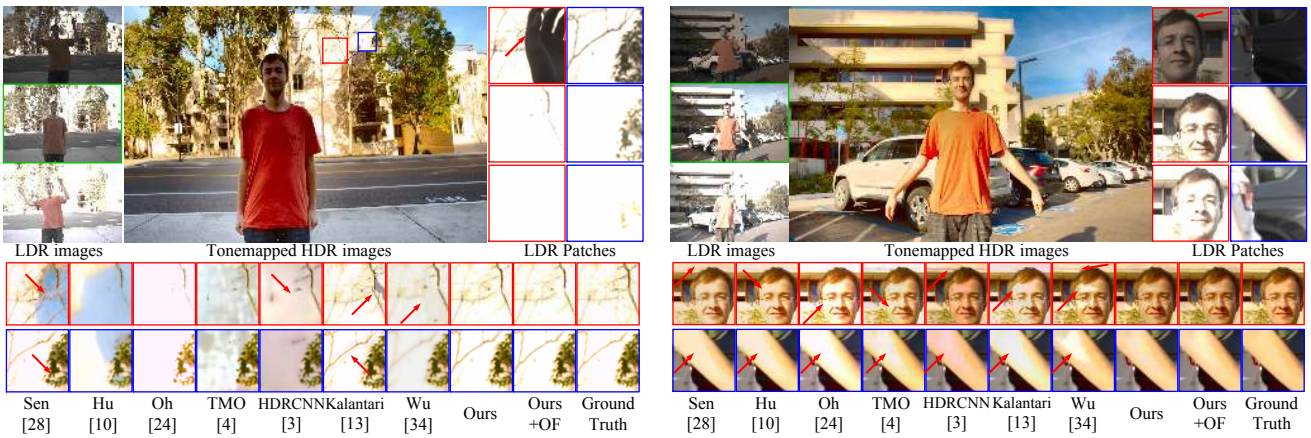
4.3.1 Evaluation on Kalantari *et al.*'s [15] Dataset

We compare our method with several state-of-the-art methods on the testing data of [15] (Figure 7 (a) and (b)), which contains some challenging samples with saturated background and foreground motions. The patch-based methods (Sen *et al.* [30] and Hu *et al.* [12]) cannot find corresponding patches and produce artifacts (See the result in Figure 7 (a)). The method of Oh *et al.* [25] cannot recover the details in the saturated areas. Since the single image methods TMO [4] and HDRCNN [3] only use the single reference image, they can avoid the ghosting artifacts, but are unable to reconstruct the sharp results and produces color distortion. The method of Kalantari *et al.* [15] products artifacts (See the red block in Figure 7 (a)), there have two main reasons: misalignment of optical flow and the limitation of their merging process. Wu *et al.*'s method [37] generates over smooth results, and cannot completely remove the ghosting artifacts (See the red block in Figure 7 (a) and (b)). Since our method uses the attention map (Figure 3) to select the useful regions and remove harmful components, it suppresses the ghosting artifacts and recovers the occluded or saturated details. (See the blue block in Figure 7 (b)). The proposed AHDRNet can produce high-quality results while taking the aligned images as inputs (See the results of Ours+OF) since the proposed attention module can also handle the artifacts caused by the error of alignment or optical flow estimation.

Table 3. Quantitative comparison of proposed network with several state-of-the-art methods. Red color indicates the best performance and blue color indicates the second best result.

	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
Sen [30]	40.9453	30.5507	38.3147	55.7240
Hu [12]	32.1872	25.5937	30.8395	55.2496
Oh [25]	27.351	22.6311	27.1119	46.8259
TMO [4]	8.2123	21.4368	8.6846	44.3944
HDRCNN [3]	14.0925	25.8217	13.1116	47.7399
Kalantari [15]	42.7423	32.0458	**41.2158**	60.5088
Wu [37]	41.6377	31.0998	40.9082	60.4955
Ours	**43.6172**	**33.0429**	41.0309	**62.3044**
Ours + OF	**43.9764**	**32.7785**	**42.2883**	**62.1296**

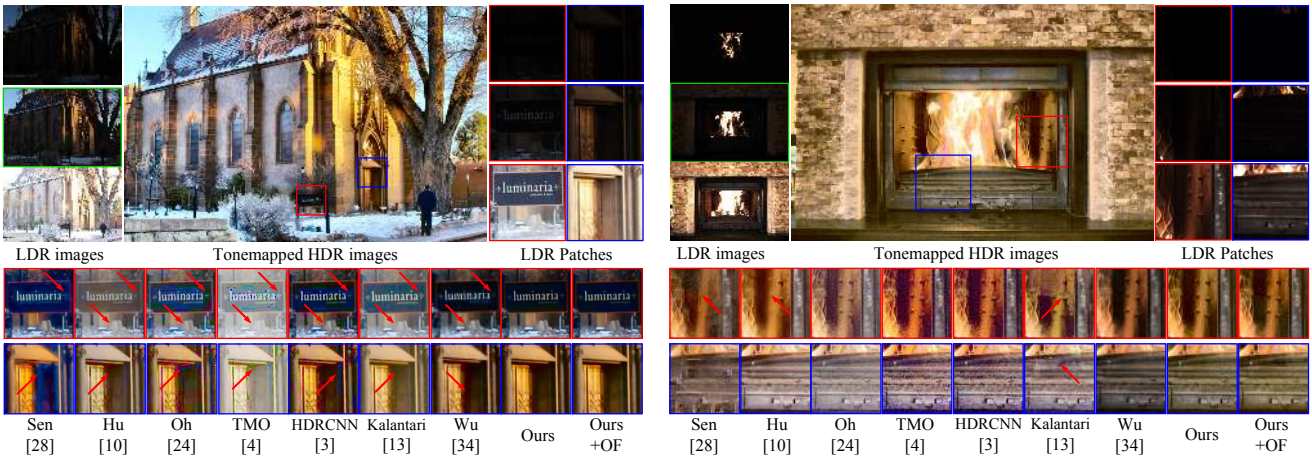
As the ground truth is available for this testing set, we can conduct the quantitative evaluations and comparisons. Results are shown in Table 3. All the values are averaged over 15 testing images. The proposed AHDRNet method produces better numerical performance than other methods. The result is best in terms of PSNR-μ and PSNR-M, showing the effectiveness of the our model. The proposed method (*i.e.* Ours+OF) can produce slightly better or competitive results with the optical flow based alignment as preprocessing. With same alignment process, our method (Ours+OF) produces better results than [15], which shows that our model can also help to handle the artifacts introduced by alignment error.



(a) Testing data (Building) in [15]

(b) Testing data (Parking) in [15]

Figure 7. Visual comparisons on the testing data from Kalantari *et al*. [15]. The top half part shows the input LDR images, LDR image patches, and the HDR image produced by the proposed method. We compare the zoomed-in local areas of the HDR images estimated by our methods and the compared methods. The propose network can produce a high-quality HDR image, especially saturated and object motions regions.



(a) Example from Sen *et al.*'s dataset [30]

(b) Example from Tursun *et al.*'s dataset [36]

Figure 8. Visual comparisons on the datasets without ground truth. The AHDRNet obtains results with sharper details and less artifacts.

4.3.2 Evaluation on the Datasets w/o Ground Truth

We compare the proposed AHDRNet with other methods on Sen's [30] and Tursun's [36] datasets which do not have ground truth. The results are shown in Figure 8 (a) and (b). The patch-based Sen *et al.*'s [30] and Hu *et al.*'s [12] methods produce artifacts in complex motion regions (zoomed-in patches in Figure 8 (b)), these methods cannot find corresponding patches in the non-reference images. As shown in Figure 8 (a) and (b), the single frame methods TMO [4] and HDRCNN [3] prone to generate serious noise and color distortion in the under-exposed regions. The method of Kalantari *et al.* [15] introduction artifacts (Figure 8 (b)) due to the alignment error. The results of Wu *et al.*'s method [37] miss details and have the obvious over smoothness in the results (Figure 8 (a) and (b)). In comparison, our proposed AHDRNet produces appealing results where the geometry distortion, color artifacts, and noise are significantly reduced compared with existing methods.

5. Conclusion

The multiple exposure methods for HDR imaging can achieve high-quality outputs that better correspond to the dynamic range of the human visual system but has been limited in its application due to ghosting and saturating artifacts. The attention-based neural network we proposed overcomes these limitations. Most notably, it can generate high-quality HDR images even in the presence of large image motion and saturation. It thus offers the prospect of more extensive applications of HDR imaging.

References

[1] L. Bogoni. Extending dynamic range of monochrome and color images through fusion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7–12, 2000.

[2] Debevec, E Paul, Malik, and Jitendra. Recovering high dynamic range radiance maps from photographs. In *Conference on Computer Graphics & Interactive Techniques*, pages 369–378, 1997.

[3] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafa K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics*, 36(6):178–193, 2017.

[4] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics*, 36(6):1–10, 2017.

[5] Haoqi Fan and Jiatong Zhou. Stacked latent attention for multimodal reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, 2018.

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, 9:249–256, 2010.

[7] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian Reid, Chunhua Shen, Anton Van Den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3806–3815, 2016.

[8] Dong Gong, Zhen Zhang, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, and Yanning Zhang. Learning an optimizer for image deconvolution. *arXiv preprint arXiv:1804.03368*, 2018.

[9] Thorsten Grosch. Fast and robust high dynamic range image generation with camera and object movement. In *IEEE International Conference of Vision, Modeling and Visualization*, 2006.

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

[11] YongSeok Heo, KyoungMu Lee, SangUk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *IEEE Asian Conference on Computer Vision (ACCV)*, pages 486–500, 2011.

[12] Jun Hu, O. Gallo, K. Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1163–1170, 2013.

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[14] Katrien Jacobs, Cline Loscos, and Greg Ward. Automatic high dynamic range image generation of dynamic environments. *Computer Graphics and Applications*, 28(2):84–93, 2008.

[15] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4):1–12, 2017.

[16] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High dynamic range video. *ACM Transactions on Graphics*, 22(3):319–325, 2003.

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[18] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1451–1460, 2017.

[19] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014.

[20] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Mann, Picard, S. Mann, and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of IS&T*, pages 442–448, 1995.

[22] 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. In *ACM Siggraph*, pages 1–14, 2011.

[23] Granados Miguel, Ajdin Boris, Wand Michael, Theobalt Christian, Seidel Hans-Peter, and P. A. Lensch Hendrik. Optimal hdr reconstruction with linear digital cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 215–222, 2010.

[24] S. K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 472–479, 2002.

[25] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2015.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[27] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *Visual Media Production*, pages 1–8, 2010.

[28] Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*, 27(12):1099–1114, 2011.

[29] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul E Debevec. *High dynamic range imaging, acquisition, display, and image-based lighting*. Princeton University Press, 2005.

[30] Pradeep Sen, Khademi Kalantari Nima, Yaesoubi Maziar, Darabi Soheil, Dan B Goldman, and Eli Shechtman. Ro-

bust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics*, 31(6):1–11, 2012.

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

[32] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

[33] Zygmunt L Szpak, Wojciech Chojnacki, Anders Eriksson, and Anton van den Hengel. Sampson distance based joint estimation of multiple homographies with uncalibrated cameras. *Computer Vision and Image Understanding*, 125:200–213, 2014.

[34] Zygmunt L Szpak, Wojciech Chojnacki, and Anton van den Hengel. Robust multiple homography estimation: An ill-solved problem. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2132–2141, 2015.

[35] Jack Tumblin, Amit Agrawal, and Ramesh Raskar. Why i want a gradient camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 103–110, 2005.

[36] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. An objective deghosting quality metric for hdr images. *Comput. Graph. Forum*, 35(2):139–152, 2016.

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

[38] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50, Jan 2019.

[39] Qingsen Yan, Jinqiu Sun, Haisen Li, Yu Zhu, and Yanning Zhang. High dynamic range imaging by sparse representation. *Neurocomputing*, 269(20):160–169, 2017.

[40] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *European Conference on Computer Vision (ECCV)*, pages 654–669, 2018.

[41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[42] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2012.

[43] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2472–2481, 2018.

[44] Bo Zhao, Xiao Wu, Jiashi Feng, Qiang Peng, and Shuicheng Yan. Diversified visual attention networks for fine-grained object classification. *IEEE Transactions on Multimedia*, 19(6):1245–1256, 2017.

[45] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

[46] Henning Zimmer, Andros Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, pages 405–414, 2011.