Deep High Dynamic Range Imaging of Dynamic Scenes

NIMA KHADEMI KALANTARI, University of California, San Diego
RAVI RAMAMOORTHI, University of California, San Diego

Fig. 1. We propose a learning-based approach to produce a high-quality HDR image (shown in middle) given three differently exposed LDR images of a dynamic scene (shown on the left). We first use the optical flow method of Liu [2009] to align the images with low and high exposures to the one with medium exposure, which we call the reference image (shown with blue border). Note that, we use reference to refer to the LDR image with the medium exposure, which is different from the ground truth HDR image. Our learning system generates an HDR image, which is aligned to the reference image, but contains information from the other two images. For example, the details on the table are saturated in the reference image, but are visible in the image with the shorter exposure. The method of Kang et al. [2003] is able to recover the saturated regions, but contains some minor artifacts. However, the patch-based method of Sen et al. [2012] is not able to properly reproduce the details in this region because of extreme motion. Moreover, Kang et al.'s method introduces alignment artifacts which appear as tearing in the bottom inset. The method of Sen et al. produces a reasonable result in this region, but their result is noisy since they heavily rely on the reference image. Our method produces a high-quality result, better than other approaches both visually and numerically. See Sec. 4 for details about the process of obtaining the input LDR and ground truth HDR images. The full images as well as comparison against a few other approaches are shown in the supplementary materials. The differences in the results presented throughout the paper are best seen by zooming into the electronic version.

Producing a high dynamic range (HDR) image from a set of images with different exposures is a challenging process for dynamic scenes. A category of existing techniques first register the input images to a reference image and then merge the aligned images into an HDR image. However, the artifacts of the registration usually appear as ghosting and tearing in the final HDR images. In this paper, we propose a learning-based approach to address this problem for dynamic scenes. We use a convolutional neural network (CNN) as our learning model and present and compare three different system architectures to model the HDR merge process. Furthermore, we create a large dataset of input LDR images and their corresponding ground truth HDR images to train our system. We demonstrate the performance of our system by producing high-quality HDR images from a set of three LDR images. Experimental results show that our method consistently produces better results than several state-of-the-art approaches on challenging scenes.

CCS Concepts: • **Computing methodologies** → **Computational photography**;

Additional Key Words and Phrases: high dynamic range imaging, convolutional neural network

ACM Reference format:
Nima Khademi Kalantari and Ravi Ramamoorthi. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *ACM Trans. Graph.* 36, 4, Article 144 (July 2017), 12 pages.
DOI: http://dx.doi.org/10.1145/3072959.3073609

1 INTRODUCTION

Standard digital cameras typically take images with under/over-exposed regions because of their sensors' limited dynamic range. The most common way to capture high dynamic range (HDR) images using these cameras is to take a series of low dynamic range (LDR) images at different exposures and then merge them into an HDR image [Debevec and Malik 1997]. This method produces spectacular images for tripod mounted cameras and static scenes, but generates results with ghosting artifacts when the scene is dynamic or the camera is hand-held.

Generally, this problem can be broken down into two stages: 1) aligning the input LDR images and 2) merging the aligned images into an HDR image. The problem of image alignment has been extensively studied and many powerful optical flow algorithms have been developed. These methods [Liu 2009; Chen et al. 2013] are typically able to reasonably align images with complex non-rigid motion, but produce artifacts in the regions with no correspondences (see Fig. 2). These artifacts usually appear in the HDR results, which are obtained by merging the aligned images during the second stage.

Our main observation is that the artifacts of the alignment can be significantly reduced during merging. However, this is a complex

process since it requires detecting the regions with artifacts and excluding them from the final results. Therefore, we propose to learn this complex process from a set of training data. Specifically, given a sequence of LDR images with low, medium, and high exposures, we first align the low and high exposure images to the medium exposure one (reference) using optical flow. We then use the three aligned LDR images as the input to a convolutional neural network to generate an HDR image that approximates the ground truth HDR image. Note that, the reference refers to the LDR image with medium exposure and is different from the ground truth HDR image. As seen in Fig. 1, the input LDR images can be of dynamic scenes with a considerable motion between them. To explore this idea, we present and compare three different system architectures and compute the required gradients for end-to-end training of each architecture.

One challenge is that we need a large number of scenes to properly train a deep network, but such a dataset is not available. We address this issue by proposing an approach to create a set of LDR images with motion and their corresponding ground truth image (Sec. 4). Specifically, we generate the ground truth HDR image using a set of three bracketed exposure images captured from a static scene. We then capture another set of three bracketed exposure images of the same scene with motion. Finally, we replace the medium exposure from the dynamic set with the corresponding image from the static set (see Fig. 7). We create a dataset of 74 training scenes with this approach and substantially extend it with data augmentation.

Experimental results demonstrate that our method is robust and handles challenging cases better than state-of-the-art HDR reconstruction approaches (see Fig. 1). In summary, our work makes the following contributions:

- We propose the first machine learning approach for reconstructing an HDR image from a set of bracketed exposure LDR images of a dynamic scene (Sec. 3).
- We fully explore the idea by presenting three different system architectures and comparing them extensively (Sec. 3.2).
- We introduce the first dataset suitable for learning HDR reconstruction, which can facilitate future learning research in this domain (Sec. 4). In addition, our dataset can potentially be used to compare different HDR reconstruction approaches. Note that, existing datasets, such as the one introduced by Karaduzovic et al. [2016], contain limited scenes and are not suitable for training a deep CNN.

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 [Tocci et al. 2011; McGuire et al. 2007]. Several approaches propose to reconstruct HDR images from coded per-pixel exposure [Heide et al. 2014; Hajisharif et al. 2015; Serrano et al. 2016] or modulus images [Zhao et al. 2015]. 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 a stack of bracketed exposure 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. We categorize these approaches into three general classes and discuss them next.

2.1 Rejecting Pixels with Motion

These approaches start by registering all the input images globally. The static pixels will have the same color across the stack and can be merged into HDR as usual. If a pixel is moving, these methods detect it and reject it. Different approaches have different ways of detecting the motion.

Khan et al. [2006] compute the probability that a given pixel is part of the background and assign weights accordingly. Jacobs et al. [2008] detects moving pixels by computing local entropy of different images in the stack. Pece and Kautz [2010] compute median threshold bitmaps for each image to generate a motion map. Zhang and Cham [2012] propose to detect movement by analyzing the image gradient. Several approaches predict the pixel colors of an image in another exposure and compare them to the original pixel colors to detect motion [Grosch 2006; Gallo et al. 2009; Raman and Chaudhuri 2011]. Heo et al. [2010] assign a weight to each pixel by computing a Gaussian-weighted distance to a reference pixel color.

Granados et al. [2013] detects the consistent subset of pixels across the image stack and then solves a labeling problem to produce a visually pleasing HDR result. Detecting the inconsistent pixels with a bidirectional approach has been investigated by Zheng et al. [2013] and Li et al. [2014]. Rank minimization has also been used [Lee et al. 2014; Oh et al. 2015] to reject outliers and reconstruct the final HDR image. However, these methods are not able to handle moving HDR content as they simply reject their corresponding pixels.

2.2 Alignment Before Merging

These approaches first align the input images and then merge them into an HDR image. Several methods have been proposed to perform rigid alignment using translation [Ward 2003] or homography [Tomaszewska and Mantiuk 2007]. However, they are unable to handle moving HDR content.

Bogoni [2000] estimates local motion using optical flow to align the input images. Kang et al. [2003] use a variant of the optical flow method by Lucas and Kanade [1981] to estimate the flow and propose a specialized HDR merging process to reject the artifacts of the registration. Jinno and Okuda [2008] pose the problem as a Markov random field to estimate a displacement field. Zimmer et al. [2011] find optical flow by minimizing an energy function consisting of gradient and smoothness terms. Hu et al. [2012] align the images by finding dense correspondences using HaCohen et al.'s method [2011]. Gallo et al. [2015] propose a fast motion estimation approach for images with small motion. These approaches use simple merging methods to combine the aligned LDR images, and thus, are not able to avoid alignment artifacts in challenging cases.

2.3 Joint Alignment and Reconstruction

The approaches in this category perform the alignment and HDR reconstruction in a unified optimization system. Sen et al. [2012]

propose a patch-based optimization system to fill in the missing under/over-exposed information in the reference image from the other images in the stack. Hu et al. [2013] propose a similar patch-based system, but include camera calibration as part of the optimization. Although these two methods are perhaps the state of the art in HDR reconstruction, patch-based synthesis produces unsatisfactory results in challenging cases where the reference has large over-exposed regions or is significantly under-exposed (Figs. 1, 13, 14, 16 and Table 1).

3 ALGORITHM

Given a set of three LDR images of a dynamic scene (Z_1, Z_2, Z_3), our 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, $\mathcal{I} = \{I_1, I_2, I_3\}$, 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. These non-rigid alignment approaches are able to reasonably register the LDR images with complex non-rigid motion, but often produce artifacts around the motion boundaries and on the occluded regions (Fig. 2). 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.

Our main observation is that the alignment artifacts from the first stage can be significantly reduced through the HDR merge in the second stage. This is in fact a challenging process and there has been significant research on this topic, even for the case when the images are perfectly aligned. Therefore, we propose to model this process with a learning system.[1] Inspired by the recent success of deep learning in a variety of applications such as colorization [Cheng et al. 2015; Iizuka et al. 2016] and view synthesis [Flynn et al. 2016; Kalantari et al. 2016], we propose to model the process with a convolutional neural network (CNN).

3.1 Overview

In this section, we provide an overview of our approach (shown in Fig. 3) by explaining different stages of our system.

Preprocessing the Input LDR Images. If the LDR images are not in the RAW format, we first linearize them using the camera response function (CRF), which can be obtained from the input stack of images using advanced calibration approaches [Grossberg and Nayar 2003; Badki et al. 2015]. We then apply gamma correction ($\gamma = 2.2$) on these linearized images to produce the input images to our system, Z_1, Z_2, Z_3. The gamma correction basically maps the images into a domain that is closer to what we perceive with our eyes [Sen et al. 2012]. Note that, this process replaces the original CRF with the gamma curve which is used to map images from LDR to the HDR domain and vice versa.

[1] We also experimented with learning the alignment process, but the system had similar performance as the optical flow method, since most artifacts could be reduced through the merging step.

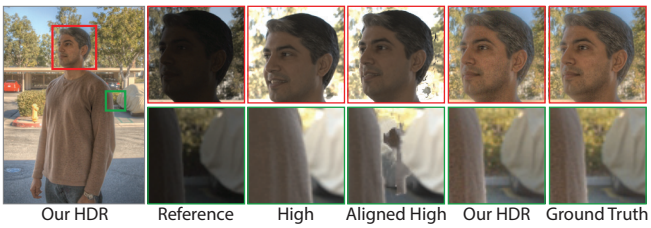

Fig. 2. We use the optical flow method of Liu [2009] to align the images with high and low exposures (only high is shown here) to the reference image. As shown in the top inset, optical flow methods are able to reasonably align the images where there are correspondences. However, in the regions with no correspondence (the bottom row), they produce artifacts. Our learning-based system is able to produce a high-quality HDR image by detecting these regions and excluding them from the final results.

Alignment. Next, we produce aligned images by registering the images with low (Z_1) and high (Z_3) exposures to the reference image, Z_2. For simplicity, we explain the process of registering Z_3 to Z_2, but Z_1 can be aligned to Z_2 in a similar manner. Since optical flow methods require brightness constancy to perform well, we first raise the exposure of the darker image to the brighter one. In this case, we raise the exposure of Z_2 to match that of Z_3 to obtain the exposure corrected image. Formally, this is obtained as $Z_{2,3} = \text{clip}(Z_2 \Delta_{2,3}^{1/\gamma})$, where the clipping function ensures the output is always in the range $[0, 1]$. Moreover, $\Delta_{2,3}$ is the exposure ratio of these two images, $\Delta_{2,3} = t_3/t_2$, where t_2 and t_3 are the exposure times of the reference and high exposure images.

We then compute the flow between Z_3 and $Z_{2,3}$ using the optical flow algorithm by Liu [2009]. Finally, we use bicubic interpolation to warp the high exposure image Z_3 using the calculated flow. This process produces a set of aligned images $\mathcal{I} = \{I_1, I_2, I_3\}$ which are then used as the input to our learning-based HDR merge component to produce the final HDR image, H. An example of aligned images can be seen in Fig. 9.

HDR Merge. The main challenge of this component is to detect the alignment artifacts and avoid their contribution to the final HDR image. In our system, we use machine learning to model this complex task. Therefore, we need to address two main issues: the choice of 1) model, and 2) loss function, which we discuss next.

1) Model: We use convolutional neural networks (CNNs) as our learning model and present and compare three different system architectures to model the HDR merge process. We discuss them in detail in Sec. 3.2.

2) Loss Function: Since HDR images are usually displayed after tonemapping, we propose to compute our loss function between the tonemapped estimated and ground truth HDR images. Although powerful tonemapping approaches have been proposed, these methods are typically complex and not differentiable. Therefore, they are not suitable to be used in our system. Gamma encoding, defined as $H^{1/\gamma}$ with $\gamma > 1$, is perhaps the simplest way of tonemapping in image processing. However, since it is not differentiable around zero, we are not able to use it in our system.

Therefore, we propose to use μ-law, a commonly-used range compressor in audio processing, which is differentiable (see Eq. 5) and suitable for our learning system. This function is defined as:

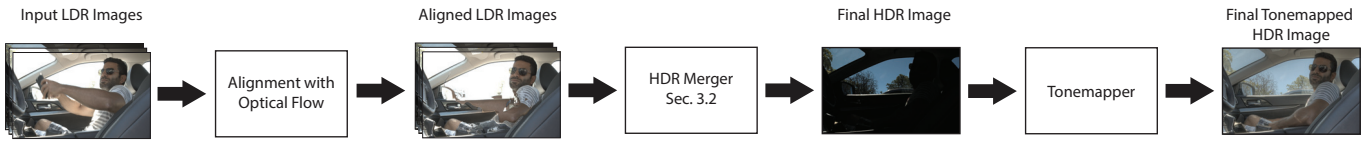

Fig. 3. In our approach, we first align the input LDR images using the optical flow method of Liu [2009] to the reference image (medium exposure). We then use the aligned LDR images as the input to our learning-based HDR merge system to produce a high-quality HDR image which is then tonemapped to produce the final image.

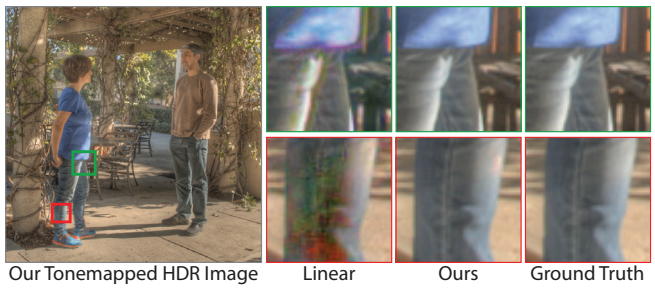

Fig. 4. We compare the result of training our system using the loss function in Eq. 2 in the linear and tonemapped (indicated as "Ours") domains. Tonemapping boosts the pixel values in the dark regions, and thus, optimization in the tonemapped domain gives more emphasis to these darker pixels in comparison with the optimization in the linear domain. Therefore, optimizing in the linear domain often produces results with discoloration, noise, and other artifacts in the dark regions, as shown in the insets.

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

where μ is a parameter which defines the amount of compression, H is the HDR image in the linear domain, and T is the tonemapped image. In our implementation, H is always in the range $[0, 1]$ and we set μ to 5000. In our approach, we train the learning system by minimizing the ℓ^2 distance of the tonemapped estimated and ground truth HDR images defined as:

$$E = \sum_{k=1}^{3} \left(\hat{T}_k - T_k\right)^2, \tag{2}$$

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

Note that we could have chosen to instead train our system by computing the error in Eq. 2 directly on the estimated ($\hat{H}$) and ground truth (H) HDR images in the linear domain. Although this system produces HDR images with small error in the linear HDR domain, the estimated images typically demonstrate discoloration, noise, and other artifacts after tonemapping, as shown in Fig. 4.

3.2 Learning-Based HDR Merge

The goal of the HDR merge process is to take the aligned LDR images, I_1, I_2, I_3, as input and produce a high-quality HDR image, H. Intuitively, this process requires estimating the quality of the input aligned HDR images and combining them based on their quality. For example, an image should not contribute to the final HDR result in the regions with alignment artifacts, noise, or saturation.

Generally, we need the aligned images in both the LDR and HDR domains to measure their quality. The images in the LDR domain are required to detect the noisy or saturated regions. For example, a simple rule would be to consider all the pixels that are smaller than 0.1 and larger than 0.9, noisy and saturated, respectively. Moreover, the images in the HDR domain could be helpful for detecting misalignments by, for example, measuring the amount of deviation from the reference image.

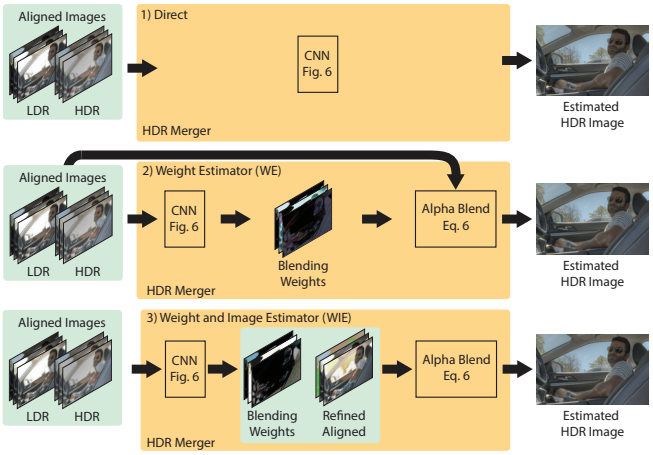

Fig. 5. Each row demonstrates a different architecture for learning the HDR merge process. The top row shows the architecture where we model the entire process using a CNN. We constrain the problem for the other two architectures (middle and bottom rows) by using the knowledge from existing techniques. See the text in Sec. 3.2 for more details.

Therefore, the HDR merge process can be formally written as:

$$H = g(\mathcal{I}, \mathcal{H}), \tag{3}$$

where g is a function which defines the relationship of the HDR image, H, to the inputs. Here, $\mathcal{H}$ is the set of aligned images in the HDR domain, H_1, H_2, H_3. Note that these are obtained from the aligned LDR images, I_i, as: $H_i = I_i^{\gamma}/t_i$, where t_i is the exposure time of the i^{th} image.[2] As discussed earlier, the HDR merge process, which is defined with the function g, is complex. Therefore, we propose to model it with a learning system and present and compare three different architectures for this purpose (see Fig. 5).

We start by discussing the first and simplest architecture (direct), where the entire process is modeled with a single CNN. We then use knowledge from the existing HDR merge techniques to constrain the problem in the weight estimator (WE) architecture by using the network to only estimate a set of blending weights. Finally, in the weight and image estimator (WIE) architecture, we relax some of the constraints of the WE architecture by using the network to output a set of refined aligned LDR images in addition to the blending weights. Overall, the three architectures produce high-quality results, but have small differences which we discuss later.

[2]During the preprocessing step, a gamma curve is used to map the images from linear HDR domain to the LDR domain, and thus, we raise the LDR images to the power of gamma to take them to the HDR domain.

1) Direct. In this architecture, we model the entire HDR merge process using a CNN, as shown in Fig. 5 (top). In this case, the CNN directly parametrizes the function g in terms of its weights. The CNN takes a stack of aligned images in the LDR and HDR domains as input, $\{\mathcal{I}, \mathcal{H}\}$ and outputs the final HDR image, H.

The estimated HDR image is then tonemapped using Eq. 1 to produce the final tonemapped HDR image (see Fig. 3). 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. In order to use gradient descent based techniques to train the system, we need to compute the derivative of the error with respect to the network weights. To do so, we use the chain rule to break down this derivative into three terms as:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial w}. \tag{4}$$

The first term is the derivative of the error function in Eq. 2 with respect to the estimated tonemapped image. Since our error is quadratic, this derivative can be easily computed. The second term is the derivative of the tonemapping function, defined in Eq. 1, with respect to its input. Since we use μ-law function as our tonemapping function, this derivative can be computed as:

$$\frac{\partial \hat{T}}{\partial \hat{H}} = \frac{\mu}{\log(1+\mu)} \frac{1}{1+\mu\hat{H}}. \tag{5}$$

Finally, the last term is the derivative of the network output with respect to its weights which can be calculated using backpropagation [Rumelhart et al. 1986].

Overall, the CNN in this simple architecture models the entire complex HDR merge process, and thus, training the network with a limited number of scenes is difficult. Although this architecture is able to produce high-quality results, in some cases it leaves residual alignment artifacts in the final HDR images, as will be shown later in Fig. 9 (top row). In the next architecture, we use some elements of the previous HDR merge approaches to constrain the problem.

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

$$\hat{H}(p) = \frac{\sum_{j=1}^{3} \alpha_j(p) H_j(p)}{\sum_{j=1}^{3} \alpha_j(p)}, \quad \text{where} \quad H_j(p) = \frac{I_j^{\gamma}}{t_j}. \tag{6}$$

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. Previous HDR merging approaches calculate these weights by, for example, the derivative of inverse CRF [Mann and Picard 1995], a triangle function [Debevec and Malik 1997], or modeling the camera noise [Granados et al. 2010]. Unfortunately, these methods assume that the images are perfectly aligned and do not work well on dynamic scenes. To handle the alignment artifacts, Kang et al. [2003] propose to use a Hermite cubic function to weight the other images based on their distance to the reference.

We propose to learn the weight estimation process using a CNN. In this case, the CNN takes the aligned LDR and HDR images as input, $\{\mathcal{I}, \mathcal{H}\}$, and outputs the blending weights, $\boldsymbol{\alpha}$. We then compute a weighted average of the aligned HDR images using these estimated weights (see Eq. 6) to produce the final HDR image.

To train the network in this architecture, we need to compute the derivative of the error with respect to the network's weights. We use the chain rule to break down this derivative into four terms as:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial \boldsymbol{\alpha}} \frac{\partial \boldsymbol{\alpha}}{\partial w}. \tag{7}$$

Note that, the last term is basically the derivative of the network's output with respect to its weights and can be calculated using backpropagation [Rumelhart et al. 1986]. Here, the only difference with respect to Eq. 4 is the third term. This term, $\partial\hat{H}/\partial\boldsymbol{\alpha}$, is the derivative of our estimated HDR image with respect to the blending weights, $\alpha_1, \alpha_2, \alpha_3$. Since the estimated HDR image in this case is obtained using Eq. 6, we can compute this derivative as:

$$\frac{\partial \hat{H}}{\partial \alpha_i} = \frac{H_i(p) - \hat{H}(p)}{\sum_{j=1}^{3} \alpha_j(p)}. \tag{8}$$

This architecture is more constrained than the direct architecture and easier to train. Therefore, it produces high-quality results with significantly fewer residual artifacts (see Fig. 9). Moreover, this architecture produces the final HDR results using only the original content of the aligned LDR images. Therefore, it should be used when staying faithful to the original content is important.

3) Weight and Image Estimator (WIE). In this architecture we relax the restriction of the previous architecture by allowing the network to output refined aligned images in addition to the blending weights. Here, the network takes the aligned LDR and HDR images as input and outputs the weights and the refined aligned images, $\{\boldsymbol{\alpha}, \tilde{\mathcal{I}}\}$. We use Eq. 6 to compute the final HDR image using the refined images, $\tilde{I}_i$, and the estimated blending weights, α_i.

Again we can compute the derivative of the error with respect to the network weights using the chain rule as:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial \hat{T}} \frac{\partial \hat{T}}{\partial \hat{H}} \frac{\partial \hat{H}}{\partial \{\boldsymbol{\alpha}, \tilde{\mathcal{I}}\}} \frac{\partial \{\boldsymbol{\alpha}, \tilde{\mathcal{I}}\}}{\partial w}. \tag{9}$$

The only difference with respect to Eq. 7 lies in the third term, $\partial\hat{H}/\partial\{\boldsymbol{\alpha}, \tilde{\mathcal{I}}\}$, as the network in this case outputs refined aligned images in addition to the blending weights.

The derivative of the estimated HDR image with respect to the estimated blending weights, $\partial\hat{H}/\partial\boldsymbol{\alpha}$, can be estimated using Eq. 8. To compute $\partial\hat{H}/\partial\tilde{\mathcal{I}}$ we can use the chain rule to break it down into two terms as:

$$\frac{\partial \hat{H}}{\partial \tilde{I}_i} = \frac{\partial \hat{H}}{\partial \tilde{H}_i} \frac{\partial \tilde{H}_i}{\partial \tilde{I}_i}. \tag{10}$$

Here, the first term is the derivative of the estimated HDR image with respect to the aligned images in the HDR domain. The relationship between $\hat{H}$ and $\tilde{H}_i$ is given in Eq. 6, and thus, the derivative can be computed as:

$$\frac{\partial \hat{H}}{\partial \tilde{H}_i} = \frac{\alpha_i}{\sum_{j=1}^{3} \alpha_j}. \tag{11}$$

Finally, the second term in Eq. 10 is the derivative of the refined aligned images in the HDR domain with respect to their LDR version. Since the HDR and LDR images are related with a power function (see Eq. 6), this derivative can be computed with the power rule as:

$$\frac{\partial \tilde{H}_i}{\partial \tilde{I}_i} = \frac{\gamma}{t_i} \tilde{I}_i^{\gamma-1}. \tag{12}$$

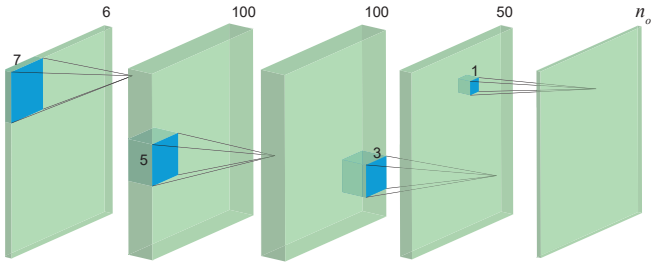

Fig. 6. We use a network with four fully convolutional layers and decreasing kernel sizes as our model. We use sigmoid as the activation function for the last layer and use rectified linear unit (ReLU) for the rest of the layers. We use the same network in our three different system architectures with the exception of the number of outputs which is different in each case.

The direct end-to-end training of this network is challenging and usually the convergence is very slow. Therefore, we propose to perform the training in two stages. In the first stage, we force the network to output the original aligned images as the refined ones, i.e., $\tilde{I} = I$, by minimizing the ℓ^2 error of the output of the network and the original aligned images. This stage constrains the network to generate meaningful outputs and produce results with similar performance as the WE architecture.

In the second stage, we simply perform a direct end-to-end training and further optimize the network by synthesizing refined aligned images. Therefore, this architecture is able to produce results with the best numerical errors (see Table 1). However, as shown in Figs. 11 and 12, this additional flexibility in comparison to the WE architecture comes at the cost of producing slightly overblurred results in dark regions.

Network Architecture. As shown in Fig. 6, we propose to use a CNN with four convolutional layers similar to the architecture proposed by Kalantari et al. [2016]. We particularly selected this architecture, since they were able to successfully model the process of generating a novel view image from a set of aligned images, which is a similar but different problem. In our 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, we use sigmoid activation function so the output of the network is always between 0 and 1. We use a fully convolutional network, so our system can handle images of any size. Moreover, the final HDR image at each pixel can usually be obtained from pixel colors of the aligned images at the same pixel or a small region around it. Therefore, all our layers have stride of one, i.e., our network does not perform downsampling or upsampling.

We use the same network in the three system architectures, but with different number of output channels, n_o. Specifically, this number is equal to 3 corresponding to the color channels of the output HDR image in the direct architecture. In the WE architecture the network outputs the blending weights, $\alpha_1, \alpha_2, \alpha_3$, each with 3 channels, and thus, $n_o = 9$. Finally, for the network in the WIE architecture $n_o = 18$, since it outputs the refined aligned images, $\tilde{I}_1, \tilde{I}_2, \tilde{I}_3$, each with 3 color channels, in addition to the blending weights.

Discussion. In summary, the three architectures produce high-quality results, better than state-of-the-art approaches (Table 1),

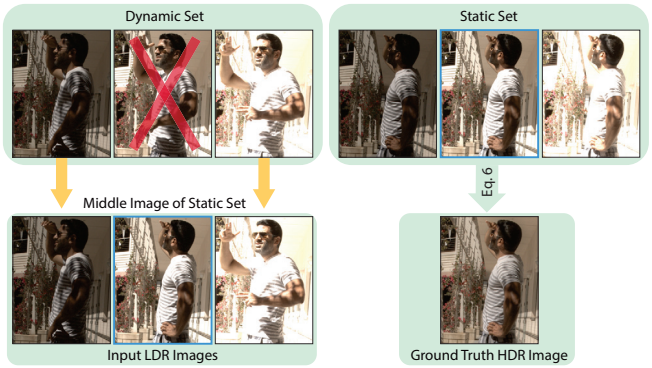

Fig. 7. We ask a subject to stay still and capture three bracketed exposure images on a tripod which are then combined to produce the ground truth image. We also ask the subject to move and capture another set of bracketed exposure images. We construct our input set by taking the low and high exposure images from this dynamic set and the middle exposure image from the static set.

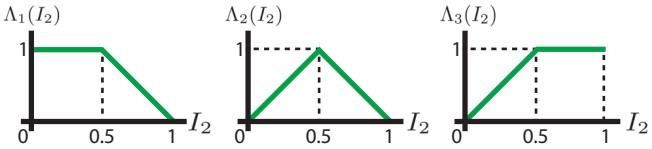

Fig. 8. The triangle functions that we use as the blending weights to generate our ground truth HDR images.

but have small differences. 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, as will be shown later in Figs. 11 and 12. Therefore, we believe the WE is the most stable architecture and produces results with the best visual quality.

4 DATASET

Training deep networks usually requires a large number of training examples. In our 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 [Tursun et al. 2015, 2016], are captured from static scenes [Funt and Shi 2010], or have a small number of scenes with only rigid motion [Karaduzovic-Hadziabdic et al. 2016]. We could potentially use the HDR video dataset of Froehlich et al. [2014] to produce our training sets. However, the number of distinct scenes in this dataset is limited, making it unsuitable for training deep networks.

To overcome this problem, we create our own training dataset of 74 different scenes and substantially extend it through data augmentation. Next, we discuss the capturing mechanism, data augmentation, and the process to generate our final training examples.

Capturing Process. The goal is to produce a set of LDR images with motion and their corresponding ground truth HDR image. For

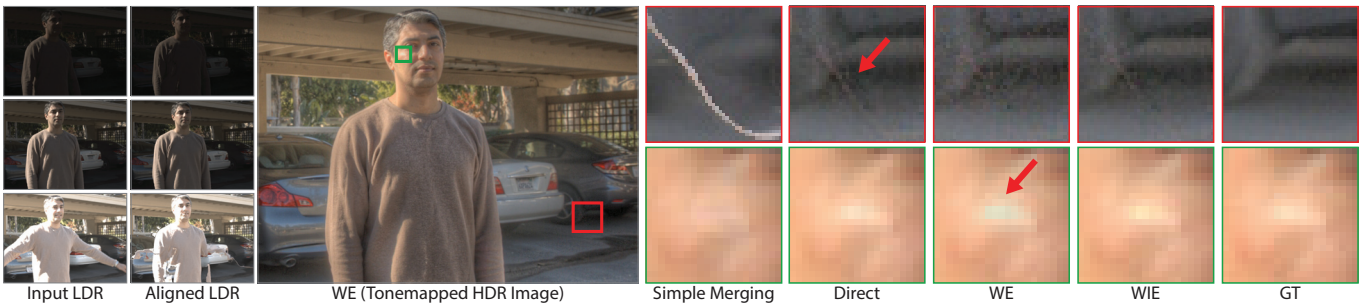

Fig. 9. We compare the result of our three architectures on the two insets indicated by the green and red boxes. We also show the result of simply merging the aligned LDR images (shown on the left) into an HDR result. The direct architecture sometimes leaves the residual alignment artifacts in the final results, while the other two architectures are more effective in suppressing these artifacts, as shown in the top inset. Moreover, the direct and WIE architectures are able to synthesize content, and thus, can reduce noise (top inset) and recover small highlights (bottom inset). In comparison, the WE architecture produces the final HDR results using the content of the aligned LDR images, and thus, is more constrained to the available content. Note that, we have adjusted the brightness and contrast of the top inset to make the differences visible.

this process, we consider mostly static scenes and use a human subject to simulate motion between the LDR images.

To generate the ground truth HDR image, we capture a static set by asking a subject to stay still and taking three images with different exposures on a tripod (see Fig. 7). Since there is no motion between these captured LDR images, we use a simple triangle weighting scheme, similar to the method of Debevec and Malik [1997], to merge them into a ground truth HDR image using Eq. 6. The weights in this case are defined as:

$$\alpha_1 = 1 - \Lambda_1(I_2), \quad \alpha_2 = \Lambda_2(I_2), \quad \alpha_3 = 1 - \Lambda_3(I_2), \tag{13}$$

where Λ_1, Λ_2, and Λ_3 are shown in Fig. 8. Although more sophisticated merging algorithms, such as Granados et al.'s approach [2010], can be used to produce the ground truth HDR image, we found that the simple triangle merge is sufficient for our purpose.

Next, we capture a dynamic set to use as our input by asking the subject to move and taking three bracketed exposure images either by holding the camera (to simulate camera motion) or on a tripod (see Fig. 7). Since in our system, the estimated HDR image is aligned to the reference image (middle exposure), we simply replace the middle image from the dynamic set with the one from the static set. Therefore, our final input set contains the low and high exposed images from the dynamic set as well as the middle exposed image from the static set.

We captured all the images in RAW format with a resolution of 5760×3840 and using a Canon EOS-5D Mark III camera. To reduce the possible misalignment in the static set, we downsampled all the images (including the dynamic set) to the resolution of 1500×1000. To ensure diversity of the training sets, we captured our bracketed exposure images separated by two or three stops.

We captured more than 100 scenes, while ensuring that each scene is generally static. However, we still had to discard a quarter of these scenes mostly because they contained unacceptable motions (e.g., leaves, human). These motions could potentially produce ghosting in the ground truth images and negatively affect the performance of the training. We note that slight motions are unavoidable, but they are rare and treated as outliers during training.

Data Augmentation. To avoid overfitting, we perform data augmentation to increase the size of our dataset. Specifically, we use color channel swapping and geometric transformation (rotating 90 degrees and flipping) with 6 and 8 different combinations, respectively. This process produces a total of 48 different combinations of data augmentation, from which we randomly choose 10 combinations to augment each training scene. Our data augmentation process increases the number of training scenes from 74 to 740.

Patch Generation. Finally, since training on full images is slow, we break down the training images into overlapping patches of size 40×40 with a stride of 20. This process produces a set of training patches consisting of the aligned patches in the LDR and HDR domains as well as their corresponding ground truth HDR patches. We then select the training patches where more than 50 percent of their reference patch is under/over-exposed, which results in around 1,000,000 selected patches. This selection is performed to put the main focus of the networks on the challenging regions.

5 RESULTS

We implemented our approach in MATLAB and used MatConvNet [Vedaldi and Lenc 2015] for efficient implementation of the convolutions in our CNNs. To train our network in all three architectures, we first initialized their weights using the Xavier approach [Glorot and Bengio 2010]. We then used ADAM solver to optimize the networks' weights with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a learning rate of 0.0001. We performed the training in all three architectures for 2,000,000 iterations on mini-batches of size 20, which took roughly two days on an Intel Core i7 with 64 GB of memory and a GeForce GTX 1080 GPU. Our method takes roughly 30 seconds to generate the final HDR image from three input LDR images of size 1000×1500. Specifically, it takes 28.5 seconds to align the images using the optical flow method of Liu [2009] and 1.5 seconds to evaluate the network and generate the final HDR result. The HDR results demonstrated here are all tonemapped with Photomatix [2017] to properly show the HDR details in each image.

Comparison of the Three Architectures. We begin by comparing our three system architectures (Sec. 3.2) in Fig. 9. We also show the result of simple triangle merging (Eqs. 6 and 13) to demonstrate the ability of our method to hide the alignment artifacts. As seen, all three architectures are able to suppress artifacts and produce

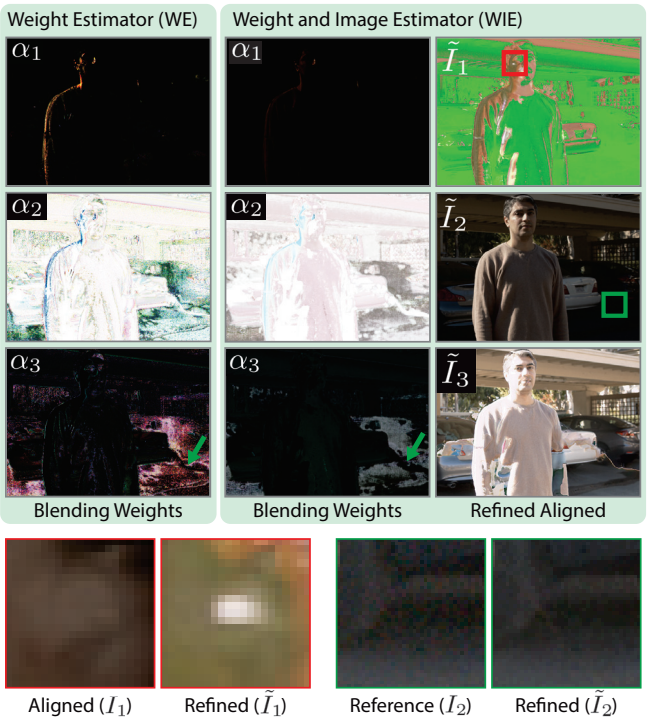

Fig. 10. We show the outputs of the network in WE and WIE architectures for the image in Fig. 9. As seen, the blending weights produced by the two architectures have similar patterns. The weight α_1 is responsible for drawing information from the low exposure image, and thus, has large values in the bright regions. In contrast, α_3 is large in the dark regions to utilize the information available in the high exposure image. Moreover, the two architectures assign small weights to the regions with artifacts (indicated by green arrows) to avoid introducing these artifacts to the final results. Finally, we show the refined aligned images for the WIE architecture on the right. Note that, since our training is end-to-end, the network sometimes produces invalid content in the regions that do not contribute to the final results, e.g., green areas in $\tilde{I}_1$. As shown in the red inset, our network in this architecture is able to hallucinate the highlight in the refined image, $\tilde{I}_1$, and consequently, reconstruct the highlight in the final HDR image (bottom row Fig. 9). Moreover, in the regions where the high exposure image contains alignment artifacts, our network synthesizes a refined image with slightly less noise than the reference image (green inset).

high-quality HDR results. However, they have small differences which comes from their design differences.

Overall, the direct architecture is the most simple and straightforward one among the three. However, since training the network in this architecture is difficult, it produces results with residual alignment artifacts in some cases (top inset in Fig. 9). In comparison, the other two architectures are more constrained, and thus, are able to better suppress the artifacts in these cases. Specifically, the weight estimator (WE) architecture is the most constrained one and produces the final HDR results using only the content of the original aligned LDR images. Therefore, if the fidelity to the content is of major concern, this architecture should be used. Finally, the weight and image estimator (WIE) is slightly less constrained and is able to synthesize content which is not available in the aligned images. Therefore, similar to the direct architecture, WIE is able to reduce noise and recover small highlights in some cases.

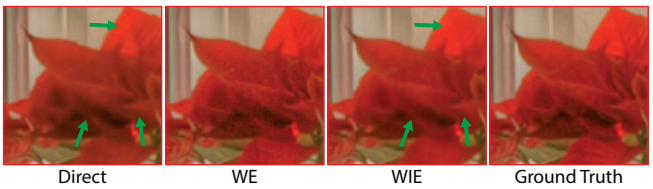

Fig. 11. We show the result of our three architectures on an inset taken from Fig. 1. The direct and WIE architectures overblur the fine details of the flower to remove the noise. The WE architecture keeps the details, but is slightly more noisy.

Fig. 12. The direct and WIE architectures reproduce highlights at the top inset, but slightly overblur the fine structures of the lady's hair. This can be seen better by toggling back and forth between the images in the supplementary materials.

In Fig. 10, we demonstrate the output of the networks in the WE and WIE architectures. As expected, in both networks the predicted blending weights, α_i, measure the quality of each aligned image. For example, the weight for the low exposure image (α_1) has large values in the highlights and bright regions, while the weight for the high exposure image (α_3) has large values in the dark regions. It is worth noting that our network in both cases avoids introducing artifacts to the final results by assigning small weights to the regions with artifacts, such as the ones shown with green arrows in the bottom row. Furthermore, as discussed, our network in the WIE architecture is able to hallucinate small highlights (red inset) and reduce the noise through reconstruction of the refined aligned images (green inset).

However, because of this additional flexibility, the WIE and direct architectures reduce noise through overblurring, as shown in Fig. 11. In contrast, the WE architecture is faithful to the content and produces results that are slightly better visually, but more noisy. Figure 12 shows another case, where the direct and WIE architectures are able to recover the highlights in the region where alignment fails, but overblur the fine details of the lady's hair. Overall, while all three architectures produce high-quality results, we believe the WE architecture produces results with slightly better visual quality.

Comparison on Test Scenes with Ground Truth. Next, we compare our three architectures against several state-of-the-art techniques. Specifically, we compare against the two patch-based methods of Hu et al. [2013] and Sen et al. [2012], the motion rejection method of Oh et al. [2015], and the flow-based approach of Kang et al. [2003].

We used authors' code for all the approaches, except for Kang et al.'s method that we implemented ourself since the source code is not available. Note that, we used the optical flow method of Liu [2009] (same as ours) to align the input LDR images in Kang et al.'s approach. Furthermore, the method of Oh et al. is a motion rejection approach which has a mechanism to align the images by estimating homography through an optimization process. However, we provide

	Kang (2003)	Sen (2012)	Hu (2013)	Oh (2015)	Ours Direct	Ours WE	Ours WIE
PSNR-T	39.10	40.75	35.49	32.19	42.92	42.74	43.26
HDR-VDP-2	64.46	63.43	60.86	61.31	67.45	66.63	67.50
PSNR-L	39.97	37.95	30.40	34.43	41.69	41.25	41.60

Table 1. Quantitative comparison of our three system architectures against several state-of-the-art methods. The PSNR-T and PSNR-L refer to the PSNR (dB) values calculated on the tonemapped (using Eq. 1) and linear images, respectively. All the values are averaged over 15 test scenes and larger values mean higher quality.

Fig. 13. Comparison of our approach against several state-of-the-art methods on one of the 15 test sets. See supplementary materials for the full images including the input LDR images.

our aligned images as the input to their method, which we found to significantly improve their results. To evaluate the results, we compute the PSNR values for images in the tonemapped (PSNR-T)[3] and linear (PSNR-L) domains. Note that, since we observe the HDR images after tonemapping, the PSNR values in the tonemapped domain better reflect the quality of the HDR images. However, we also show the PSNR values in the linear domain for completeness. Moreover, we measure the quality of the results using HDR-VDP-2 [Mantiuk et al. 2011], which is a visual metric specifically designed to evaluate the quality of HDR images.

Table 1 shows the result of this comparison averaged over 15 test scenes. Note that, none of the test scenes are included in the training sets and they are captured from different subjects. As can be seen, all our three architectures produce results with better numerical errors than the state-of-the-art techniques. Moreover, while all the architectures have similar numerical errors, the WE architecture is slightly worse. This is perhaps because this architecture is the most constrained, and thus, is not as flexible as the other architectures in minimizing the error. However, we believe the WE architecture is slightly more stable and produces results with higher visual quality, and thus, use it to produce the results in the rest of the paper.

[3]Note that, we use Eq. 1 as our tonemapping operator in this case, which is different from the operator used to show the final images. Since the operator in Eq. 1 does not clamp the images, the tonemapped images contain all the HDR information.

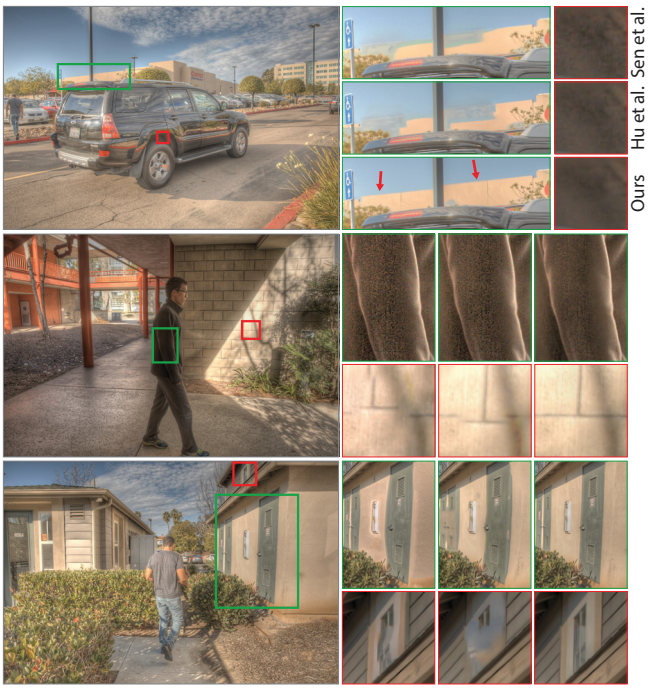

Fig. 14. Comparison of our approach against the patch-based methods of Sen et al. [2012] and Hu et al. [2013].

In Fig. 13, we compare our approach against other methods on one of these scenes, demonstrating three people in a dark room with bright windows. The first row of insets shows a region where the highlights need to be reconstructed from the low exposure image. The methods of Kang et al. and Oh et al. are able to recover the highlights despite having small artifacts as indicated by the arrows. The patch-based approaches of Sen et al. and Hu et al. are not able to find corresponding patches in the low exposure image, and thus, produce saturated highlights. Our approach is able to recover the highlights and produces an HDR image which is reasonably close to the ground truth. The second row demonstrates a region with significant motion, where the approaches by Kang et al. and Oh et al. are not able to avoid introducing the alignment artifacts in the final results. The methods of Sen et al. and Hu et al. are able to faithfully reconstruct the hands. However, they often heavily rely on the reference image, and thus, produce an overall noisy result. In contrast, our approach is able to avoid alignment artifacts, but draws information from the high exposure image and produces a relatively noise-free results.

Comparison on Natural Scenes. We compare our method against the patch-based approaches of Sen et al. and Hu et al. on three challenging test scenes in Fig. 14. Note that, we do not have ground truth images in these cases as we captured images of natural dynamic scenes. The top row shows a picture of an outdoor scene with a moving car. In this case, the patch-based approaches are not able to recover the top of the building, which is saturated in the reference image, because of the car's significant motion. Moreover, these two techniques produce noisy results in the dark regions because they heavily rely on the reference.

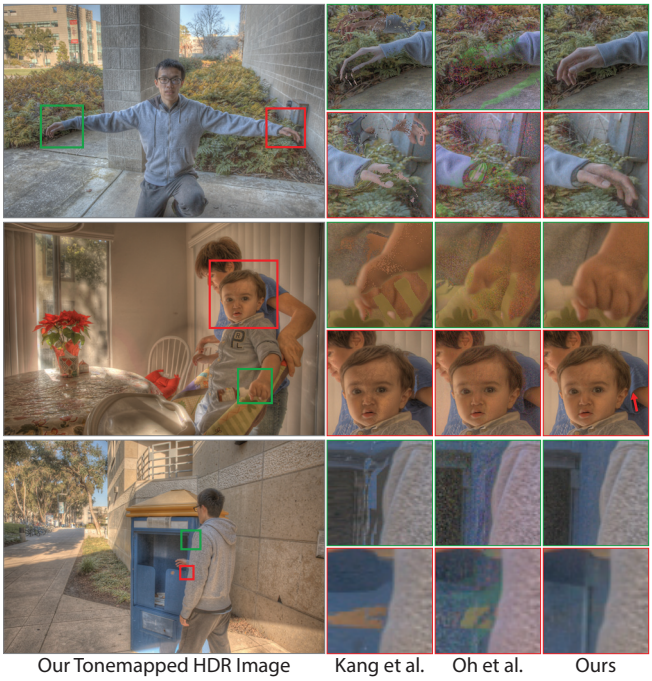

Fig. 15. Comparison of our approach against the approaches by Kang et al. [2003] and Oh et al. [2015].

The second row demonstrates a picture of a man walking in a dark hallway. The patch-based methods are not able to effectively suppress the noise in the top inset. Moreover, these approaches typically have problem with the structured regions, and thus, are not able to properly reconstruct the edges of the bricks in the bottom inset. Our method is able to reduce noise in the dark areas and properly reconstruct the saturated regions. Finally, the third row shows a picture of an outdoor scene on a bright day with a walking person. All the methods are able to plausibly reconstruct the moving person. However, this particular scene has large saturated regions in the reference image (see supplementary materials). Therefore, the patch-based approaches are not able to properly reconstruct the saturated regions due to insufficient constraints. On the other hand, our method produces a high-quality HDR image.

Figure 15 shows a comparison of our approach against the methods of Kang et al. and Oh et al. on three other test scenes. The top row shows an outdoor scene with a bright background where a man is sitting in a dark area. Here, the other approaches are not able to avoid alignment artifacts and generate results with duplicate (Kang et al.) or missing (Oh et al.) hands. However, our method is able to produce a noise-free high-quality HDR result. The second row shows a picture of a lady and a baby in a dark room with a bright window. The two other approaches are not able to properly reconstruct the baby's hand as alignment fails in this region because of the motion blur. Note that, only our approach is able to reconstruct the bright highlight on the lady's shirt and the baby's face without noise and other artifacts.

Finally, the third row demonstrates an outdoor scene with a large dynamic range and significant motion. Kang et al.'s method is not able to suppress the alignment artifacts around the motion boundaries. Similarly the method of Oh et al. introduces alignment artifacts

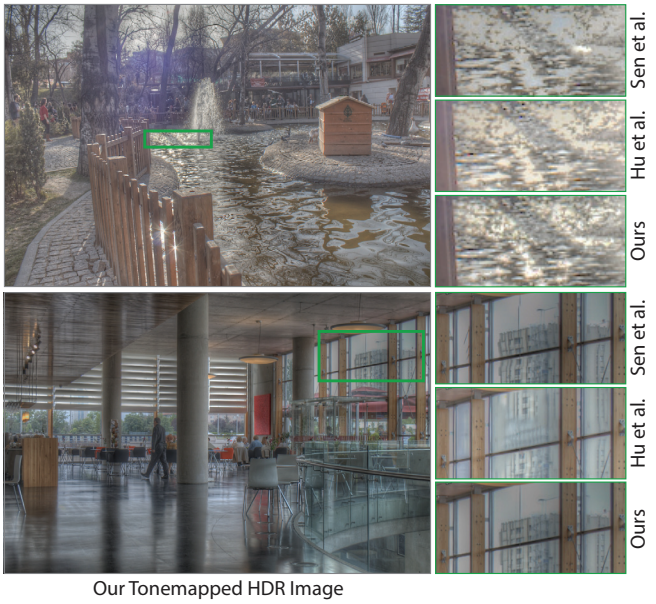

Fig. 16. Comparison against the patch-based methods of Sen et al. [2012] and Hu et al. [2013] on Tursun et al.'s scenes [2015; 2016].

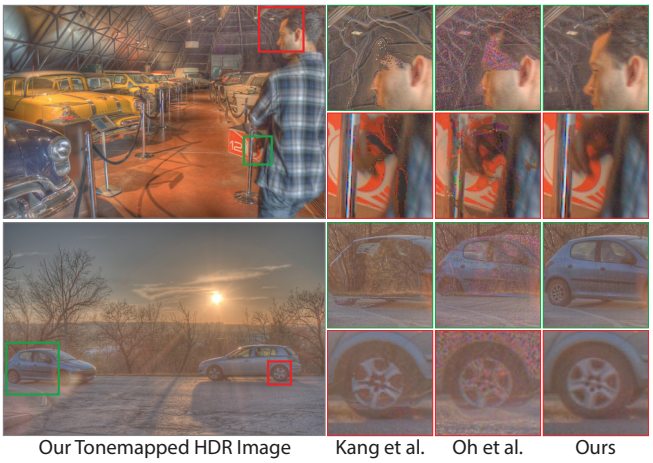

Fig. 17. Comparison against the approaches by Kang et al. [2003] and Oh et al. [2015] on Tursun et al.'s scenes [2015; 2016].

to the final results and is noisy. However, our method properly reconstructs the areas around the motion boundaries and produces a high-quality HDR result.

We also compare our approach against other methods on several scenes from Tursun et al. [2015; 2016]. These scenes have 9 images with one stop separation from which we select three images with two or three stop separations. Note that these scenes are captured using different cameras than the one we used to capture our training scenes. Figure 16 shows comparison of our approach against the methods of Sen et al. [2012] and Hu et al. [2013] on the Fountain (top) and Cafe (bottom) scenes. Since the motion of water in the Fountain scene is complex, the patch-based approaches are not able to find correspondences in these regions. Therefore, these methods are not able to recover the highlights on the water. The Cafe scene contains bright windows on the right, which are completely saturated in the reference image. Although other methods recover the

building, they produce results with crooked (Sen et al.) or replicated and blurred (Hu et al.) structures, which are common problems of patch-based synthesis.

In Fig. 17, we show comparison of our method against the approaches by Kang et al. [2003] and Oh et al. [2015] on the Museum1 (top) and Cars (bottom) scenes. Since the person walking in front of the camera in the Museum1 scene has motion blur, the warped images contain severe alignment artifacts due to inaccuracies in optical flow. As a result, the other approaches produce images with missing head and hand, while our method generates a high-quality HDR image. Similarly, optical flow is not able to align the fast moving cars in the Cars scene, and thus, other methods produce results with artifacts on the two cars.

6 DISCUSSION, LIMITATIONS, AND FUTURE WORK

As discussed in Sec. 2, there are approaches that capture the multiple exposures in a single shot by, for example, varying the per-pixel exposure [Heide et al. 2014; Hajisharif et al. 2015; Serrano et al. 2016]. Although these approaches inherently handle dynamic scenes, they require special types of sensors, which are not readily accessible. In contrast, bracketed exposure images, as used in our method, can be easily captured using standard digital cameras. This is perhaps why stack-based HDR imaging approaches are popular and implemented in commercial devices such as smartphone cameras.

It is worth noting that dynamic range can also be increased by combining multiple images captured with the same exposure time [Zhang et al. 2010; Hasinoff et al. 2016]. Comparing to bracketed exposure methods, the images in these techniques have similar content, and thus, alignment is generally simpler. However, these methods typically demonstrate increasing dynamic range by only two or three stops. The main reason is that, to increase the dynamic range by a large factor, these methods require capturing and processing an impractically large number of input images. For example, increasing the dynamic range by four or six stops, as we show in this paper, requires capturing 16 or 64 images, respectively. Therefore, bracketed exposure approaches, like ours, are more suitable for capturing scenes with large dynamic range.

The main limitation of our approach is that our network takes a specific number of images as the input. We demonstrated that our system is able to produce high-quality results with a set of three input images. Although we observed that three images are sufficient to capture the dynamic range of most scenes, it would be interesting to retrain our network for cases with more than three inputs (e.g., 5 or 7) and evaluate its performance. Moreover, investigating flexible network architectures to make the system independent of the number of inputs would be an interesting future research topic.

In this paper, we trained our networks on scenes with two and three stop separations. It is worth noting that our system is able to produce high-quality results on scenes with separations that it has not been trained on, e.g., the scene in Fig. 15 (middle) is captured at -2.66, 0, and +3.33 stops. However, to produce high-quality results on scenes with significantly different separations than two or three stops, our system needs to be retrained.

Another limitation of our method 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. In these situations, we simply use the content of the reference image to reconstruct the HDR image. Therefore, the final HDR image could appear noisy or saturated if the reference image is under/over-exposed in these regions. While all the other flow-based techniques have this limitation, the patch-based methods are usually able to perform hole-filling and synthesize the content of these regions. However, this is not a major limitation as the same patch-based hole-filling could be performed in a postprocess after reconstructing the HDR image with our system.

Since our goal is to handle alignment artifacts, we train our networks on images with significant motion. In this case, our system learns to properly merge the images in the aligned regions, while avoiding the artifacts in the regions with misalignments. As a result, we produce results that are slightly noisier than the images obtained by noise optimal merging approaches [Hasinoff et al. 2010; Granados et al. 2010] in the aligned regions. Considering the ability of our approach in avoiding significant alignment artifacts, we believe this is an acceptable sacrifice.

As discussed, while our WIE architecture produces the best results numerically (PSNR-T in Table 1), it sometimes overblurs the noisy regions producing results that are not visually pleasing, as shown in Fig. 11. In the future, it would be interesting to see if training the network in a perceptual way by, for example, using generative adversarial networks [Goodfellow et al. 2014], could improve the visual quality of the results.

Finally, in this paper we used optical flow to align the input images. However, the flow estimation could potentially be learned using an additional CNN and trained end-to-end to minimize the error between the estimated and ground truth HDR images. We performed a simple experiment to learn the final flow by providing a network with a homography field, optical flow, and a flow obtained by matching patches. However, the final HDR images generated with this system were generally similar to the ones generated by our system. This experiment suggests that the optical flow is perhaps the best among the three inputs to the network and the artifacts of the alignment can be easily avoided by the merge network. However, in this simple experiment, the network was basically selecting the best flow among the three input flows. In the future, it would be interesting to investigate the possibility of training a network, perhaps similar to the one proposed by Dosovitskiy et al. [2015] or Ilg et al. [2016], to estimate the flow from the input images.

7 CONCLUSION

We have presented the first learning-based technique to produce an HDR image using a set of LDR images captured from a dynamic scene. We use a convolutional neural network to generate the HDR image from a set of images aligned with optical flow. To properly train the network, we proposed a strategy to produce a set of input LDR images and their corresponding ground truth image. We present three architectures for our learning-based techniques and find through extensive comparison that using the knowledge from existing techniques in our learning system leads to improvement. Specifically, we found that using the network to estimate blending weights for combining aligned LDR images is slightly better than modeling the entire process with a network. This finding implies

that learning approaches could use elements of existing techniques to potentially solve complex problems more efficiently.

ACKNOWLEDGMENTS

This work was supported in part by ONR grant N000141512013, NSF grant 1617234, and the UC San Diego Center for Visual Computing.

REFERENCES

A. Badki, N. Khademi Kalantari, and P. Sen. 2015. Robust Radiometric Calibration for Dynamic Scenes in the Wild. In *IEEE ICCP*. 1–10.

Luca Bogoni. 2000. Extending Dynamic Range of Monochrome and Color Images through Fusion. In *IEEE ICPR*. 3007–3016.

Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu. 2013. Large Displacement Optical Flow from Nearest Neighbor Fields. In *IEEE CVPR*. 2443–2450.

Z. Cheng, Q. Yang, and B. Sheng. 2015. Deep Colorization. In *IEEE ICCV*. 415–423.

Paul E. Debevec and Jitendra Malik. 1997. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH*. 369–378.

A. Dosovitskiy, P. Fischery, E. Ilg, P. HÃ∂usser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*. 2758–2766.

John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. 2016. DeepStereo: Learning to Predict New Views from the WorldâĂŹs Imagery. In *IEEE CVPR*. 5515–5524.

Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. 2014. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. *SPIE* 9023 (2014), 90230X–90230X–10.

Brian Funt and Lilong Shi. 2010. The Rehabilitation of MaxRGB. *Color and Imaging Conference* 2010, 1 (2010), 256–259.

O. Gallo, N. Gelfand, W. Chen, M. Tico, and K. Pulli. 2009. Artifact-free High Dynamic Range Imaging. In *IEEE ICCP*. 1–7.

O. Gallo, A. Troccoli, J. Hu, K. Pulli, and J. Kautz. 2015. Locally non-rigid registration for mobile HDR photography. In *IEEE CVPRW*. 48–55.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, Vol. 9. 249–256.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative Adversarial Networks. *ArXiv e-prints* (June 2014). arXiv:stat.ML/1406.2661

M. Granados, B. Ajdin, M. Wand, C. Theobalt, H. P. Seidel, and H. P. A. Lensch. 2010. Optimal HDR reconstruction with linear digital cameras. In *IEEE CVPR*. 215–222.

Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. 2013. Automatic Noise Modeling for Ghost-free HDR Reconstruction. *ACM TOG* 32, 6, Article 201 (2013), 10 pages.

T. Grosch. 2006. Fast and Robust High Dynamic Range Image Generation with Camera and Object Movement. In *Vision, Modeling and Visualization*. 277–284.

Michael D. Grossberg and Shree K. Nayar. 2003. Determining the Camera Response from Images: What Is Knowable? *IEEE PAMI* 25, 11 (Nov. 2003), 1455–1467.

Yoav HaCohen, Eli Shechtman, Dan B. Goldman, and Dani Lischinski. 2011. Non-rigid dense correspondence with applications for image enhancement. *ACM TOG* 30, 4, Article 70 (2011), 10 pages.

Saghi Hajisharif, Joel Kronander, and Jonas Unger. 2015. Adaptive dualISO HDR reconstruction. *EURASIP Journal on Image and Video Processing* 2015, 1 (2015), 41.

S. W. Hasinoff, F. Durand, and W. T. Freeman. 2010. Noise-optimal capture for high dynamic range photography. In *CVPR*. 553–560.

Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst Photography for High Dynamic Range and Low-light Imaging on Mobile Cameras. *ACM TOG* 35, 6, Article 192 (Nov. 2016), 12 pages.

Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pająk, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli. 2014. FlexISP: A Flexible Camera Image Processing Framework. *ACM TOG* 33, 6, Article 231 (Nov. 2014), 13 pages.

Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. 2010. Ghost-free high dynamic range imaging. In *ACCV*, Vol. 4. 486–500.

Jun Hu, Orazio Gallo, and Kari Pulli. 2012. *Exposure Stacks of Live Scenes with Hand-Held Cameras.* Springer Berlin Heidelberg, Berlin, Heidelberg, 499–512.

J. Hu, O. Gallo, K. Pulli, and X. Sun. 2013. HDR Deghosting: How to Deal with Saturation?. In *IEEE CVPR*. 1163–1170.

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. Let There Be Color!: Joint End-to-end Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM TOG* 35, 4, Article 110 (July 2016), 11 pages.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. 2016. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. *CoRR* abs/1612.01925 (2016).

K. Jacobs, C. Loscos, and G. Ward. 2008. Automatic High-Dynamic Range Image Generation for Dynamic Scenes. *IEEE Computer Graphics and Applications* 28, 2 (Mar.-Apr. 2008), 84–93.

T. Jinno and M. Okuda. 2008. Motion blur free HDR image acquisition using multiple exposures. In *IEEE ICIP*. 1304–1307.

Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. 2016. Learning-based View Synthesis for Light Field Cameras. *ACM TOG* 35, 6, Article 193 (Nov. 2016), 10 pages.

Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. 2003. High dynamic range video. *ACM TOG* 22, 3 (2003), 319–325.

Kanita Karaduzovic-Hadziabdic, Jasminka Hasic Telalovic, and Rafal Mantiuk. 2016. Subjective and Objective Evaluation of Multi-exposure High Dynamic Range Image Deghosting Methods. In *EG*, T. Bashford-Rogers and L. P. Santos (Eds.). The Eurographics Association. https://doi.org/10.2312/egsh.20161007

E.A. Khan, A.O. Akyüz, and E. Reinhard. 2006. Ghost Removal in High Dynamic Range Images. In *IEEE ICIP*. 2005–2008.

C. Lee, Y. Li, and V. Monga. 2014. Ghost-Free High Dynamic Range Imaging via Rank Minimization. *IEEE SPL* 21, 9 (Sept 2014), 1045–1049.

Z. Li, J. Zheng, Z. Zhu, and S. Wu. 2014. Selectively Detail-Enhanced Fusion of Differently Exposed Images With Moving Objects. *IEEE Transactions on Image Processing* 23, 10 (Oct 2014), 4372–4382.

C. Liu. 2009. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis.* Doctoral thesis. Massachusetts Institute of Technology.

Bruce D. Lucas and Takeo Kanade. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision. In *IJCAI*. 674–679.

Steve Mann and Rosalind W. Picard. 1995. On Being 'undigital' With Digital Cameras: Extending Dynamic Range By Combining Differently Exposed Pictures. In *Proc. of Society for Imaging Science and Technology*. 442–448.

Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A Calibrated Visual Metric for Visibility and Quality Predictions in All Luminance Conditions. *ACM TOG* 30, 4, Article 40 (July 2011), 14 pages.

M. McGuire, W. Matusik, H. Pfister, B. Chen, J.F. Hughes, and S.K. Nayar. 2007. Optical Splitting Trees for High-Precision Monocular Imaging. *IEEE Computer Graphics and Applications* 27, 2 (march-april 2007), 32–42.

T. H. Oh, J. Y. Lee, Y. W. Tai, and I. S. Kweon. 2015. Robust High Dynamic Range Imaging by Rank Minimization. *IEEE PAMI* 37, 6 (2015), 1219–1232.

F. Pece and J. Kautz. 2010. Bitmap Movement Detection: HDR for Dynamic Scenes. In *CVMP*. 1–8.

Photomatix. 2017. Commercially-available HDR processing software. (2017). http://www.hdrsoft.com/.

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

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning Representations by Back-propagating Errors. *Nature* 323 (1986), 533–536.

Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. 2012. Robust patch-based HDR reconstruction of dynamic scenes. *ACM TOG* 31, 6, Article 203 (Nov. 2012), 11 pages.

Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. 2016. Convolutional Sparse Coding for High Dynamic Range Imaging. *CGF* 35, 2 (2016), 153–163.

Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. 2011. A versatile HDR video production system. *ACM TOG* 30, 4, Article 41 (July 2011), 10 pages.

A. Tomaszewska and R. Mantiuk. 2007. Image Registration for Multi-exposure High Dynamic Range Image Acquisition. In *WSCG*.

Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. 2016. An Objective Deghosting Quality Metric for HDR Images. *CGF* 35, 2 (2016), 139–152.

Okan Tarhan Tursun, Ahmet OÄ§uz AkyÃijz, Aykut Erdem, and Erkut Erdem. 2015. The State of the Art in HDR Deghosting: A Survey and Evaluation. *CGF* 34, 2 (2015).

Andrea Vedaldi and Karel Lenc. 2015. MatConvNet: Convolutional neural networks for Matlab. In *ACMMM*. 689–692.

Greg Ward. 2003. Fast, Robust Image Registration for Compositing High Dynamic Range Photographs from Hand-Held Exposures. *Journal of Graphics, GPU, and Game Tools* 8, 2 (2003), 17–30.

L. Zhang, A. Deshpande, and X. Chen. 2010. Denoising vs. deblurring: HDR imaging techniques using moving cameras. In *CVPR*. 522–529.

Wei Zhang and Wai-Kuen Cham. 2012. Gradient-Directed Multiexposure Composition. *IEEE TIP* 21, 4 (April 2012), 2318–2323.

H. Zhao, B. Shi, C. Fernandez-Cull, S. K. Yeung, and R. Raskar. 2015. Unbounded High Dynamic Range Photography Using a Modulo Camera. In *ICCP 2015*. 1–10.

J. Zheng, Z. Li, Z. Zhu, S. Wu, and S. Rahardja. 2013. Hybrid Patching for a Sequence of Differently Exposed Images With Moving Objects. *IEEE TIP* 22, 12 (Dec 2013), 5190–5201.

Henning Zimmer, AndrÃĺs Bruhn, and Joachim Weickert. 2011. Freehand HDR Imaging of Moving Scenes with Simultaneous Resolution Enhancement. *CGF* 30, 2 (April 2011), 405–414.