2019 IEEE Winter Conference on Applications of Computer Vision

Multi-scale Dense Networks for Deep High Dynamic Range Imaging

Qingsen Yan
Northwestern Polytechnical University
The University of Adelaide
yqs@mail.nwpu.edu.cn

Dong Gong
Northwestern Polytechnical University
The University of Adelaide
edgong01@gmail.com

Pingping Zhang
Dalian University of Technology
The University of Adelaide
jssxzhpp@mail.dlut.edu.cn

Qinfeng Shi
The University of Adelaide
javen.shi@adelaide.edu.au

Jinqiu Sun *
Northwestern Polytechnical University
sunjinqiu@nwpu.edu.cn

Ian Reid
The University of Adelaide
ian.reid@adelaide.edu.au

Yanning Zhang
Northwestern Polytechnical University
ynzhang@nwpu.edu.cn

Abstract

Generating a high dynamic range (HDR) image from a set of sequential exposures is a challenging task for dynamic scenes. The most common approaches are aligning the input images to a reference image before merging them into an HDR image, but artifacts often appear in cases of large scene motion. The state-of-the-art method using deep learning can solve this problem effectively. In this paper, we propose a novel deep convolutional neural network to generate HDR, which attempts to produce more vivid images. The key idea of our method is using the coarse-to-fine scheme to gradually reconstruct the HDR image with the multi-scale architecture and residual network. By learning the relative changes of inputs and ground truth, our method can produce not only artificial free image but also restore missing information. Furthermore, we compare to existing methods for HDR reconstruction, and show high-quality results from a set of low dynamic range (LDR) images. We evaluate the results in qualitative and quantitative experiments, our method consistently produces excellent results than existing state-of-the-art approaches in challenging scenes.

1. Introduction

High dynamic range (HDR) imaging is a fundamental and essential problem in computer vision and image processing, which is effective to capture and display real-world lighting. Due to the limitation of the imaging devices, standard digital cameras and most monitors have limited dynamic range, leading to under-exposed or over-exposed regions in an image. The under-exposed regions are too dark to show the details of the scenes, and the over-exposed regions appear saturated in the image. To address this issue, specialized camera hardware has been proposed to capture HDR images or videos directly [25, 13]. However, high-quality HDR cameras are not affordable for the general public. A recent, more easy-to-use approach is to generate an HDR image from low dynamic range (LDR) images [4, 27, 33, 32]. Since each exposure can be designed to capture a specific dynamic range, HDR imaging can recover a latent image from a stack of differently exposed images. For image display, tone mapping is used to show on the display device. This method produces appealing images and makes all existing details visible in a static scene. If the stack is captured in a dynamic scene with moving objects, HDR imaging is a more challenging task and generates results with ghosting artifacts. This is a serious limitation since real-world scenes often have moving objects.

Actually, removing ghost artifacts from HDR imaging has been the subject of extensive research. Aligning the input images to a reference image before merging the aligned images is the common way to generate an HDR image. The most successful approaches to align the images are based on optical flow [21, 3]. However, the HDR result is determined by the performance of optical flow, thus ghosting artifacts still exist in the standard HDR imaging results since images are not aligned to the reference very well. With the development of deep learning, many networks [17, 5, 2, 6] have been proposed to produce an HDR image with differ-

*Corresponding author

978-1-7281-1975-5/19/$31.00 ©2019 IEEE
DOI 10.1109/WACV.2019.00012

Figure 1. We propose a deep network to produce HDR images which contain more details from three input LDR images of a complicated scene. Images with different exposures are shown on the left, and our result after tonemapping is shown in the middle after Matlab function *tonemap*. The right columns show comparisons with other methods. Our proposed method using the coarse-to-fine scheme produces appealing results, which are better than other approaches both visually and numerically.

ent strategies. For example, Kalantari *et al.* [17] propose to merge aligned inputs and exclude artifacts using deep neural network. The result of directly predicted the HDR image has higher performance but has artifacts sometimes. U-Net [28] is one of the most famous networks for low-level computer vision. We try to use this network to capture more informative features with skip connection. Unfortunately, the upsampling layer will lead to inevitable color distortion or blurring artifacts, as shown in Figure 1.

In this paper, we propose a novel multi-scale dense network to generate HDR, which attempts to produce vivid artificial-free image. The multi-scale scheme is an effective and practical approach to generate image details. It has been very successful in both traditional optimization-based methods and recent neural-network based approaches [30, 37, 24] for image reconstruction. In the well-established multi-scale HDR imaging method, Mertens [23] adopts multi-scale methods for image fusion, and has higher performance and appealing results. We thus embrace the multi-scale processing scheme which avoids the use of upsampling layers to improve HDR image quality. On the coarse scale, the network predicts the global information (such as color, context) of HDR image from LDR images. A medium branch learns how to generate the middle level details of output with considering neighborhoods pixels. The final scale is used to keep details of LDR images and predicts the high-frequency information that is not captured by the raw image. Compared with the progressive network, our proposed network is designed to capture multi-scale information independently. To exploit coarse and middle level information while preserving fine level information at the same time, the outputs of the three scale are concatenated as input to refine network. In general, we use three sub-networks on different scales to capture the high-level information, middle level features and local details. The results show that the proposed network can generate results with better quality than existing state-of-the-art HDR reconstruction methods (Figure 1). In addition, to reused the features, the Dense Convolutional Network (DenseNet) [15] is also integrated into the U-Net architecture in each scale.

The main contributions can be summarized as:

- We adopt three independent networks to capture the multi-scale information from a sequence of bracketed exposure LDR images for reconstructing HDR image.
- In order to use the features effectively and decrease the numbers of parameters, we incorporate dense connections into U-Net to connect each layer to every other layer in a feed-forward fashion.
- We take advantages of the multi-scale HDR image to gradually reconstruct the final of HDR image.
- We propose a multi-scale loss function that not only constraints the estimated HDR images are similar to ground truth in different scales, but also guarantees the quality of the final result.

2. Related Work

A number of methods to HDR imaging have been presented in the literature. Those methods can coarsely be divided into two classes from different applications (static scenes and dynamic scenes). The static scenes commonly are captured by estimating the camera response function and recovering the latent HDR image [4, 11]. However, the static scene approaches introduce other limitations such as blurring or ghosting artifacts in dynamic scenes.

Existing methods to reduce the ghosting artifacts for HDR imaging can be categorized into two general classes. The first class is **alignment before merging**, which aligns the different images before merging them into an HDR image. The optical flow (OF) is the most widely used approach in aligning image fileds. For example, to enhance the dynamic range, Bogoni [1] estimated local unconstrained motion vectors using optical flow, and warped other exposed images into alignment. Kang *et al.* [18] improved optical

flow based methods to compensate scene and camera movement, and proposed a specific HDR merging process. Zimmer *et al.* [38] computed the optical flow by minimizing an energy function approach which takes into account the varying exposure conditions. However, the overexposure or underexposure regions have less information and may lead to error displacement fields.

Meanwhile, **patch matching** based methods were employed in HDR imaging to handle dense correspondence [29, 14]. For instance, Sen *et al.* [29] proposed a patch-based energy-minimization system to integrate LDR images and reconstruction HDR images. The algorithm preserves the high-quality information in the reference image, and also fills in the missing under-exposed or over-exposed information from other input LDR images. Hu *et al.* [14] added camera calibration as a part of the optimization. The resulting method obtains a latent image which is similar to the reference image but exposed like source image according to intensity mapping function. Although patch matching based methods can generate excellent results of HDR images, the computation complexity is very huge, reducing the model efficiency.

The second class is **moving objects** based methods, which rejects moving objects by using lower weights to erase the influence of this regions. The key difference between these methods is how to detect the ghost regions. Grosch [10] detected ghost regions using the divergence of predicted pixel colors and the original pixel colors to detect motion. Jacobs *et al.* [16] employed local entropy of different input images to detect ghost regions. Assuming the linearity of the image intensity with exposure times, Gallo *et al.* [7] used deviation of two exposure pixel values of adjacent images to measure ghost effects. Heo *et al.* [34] computed a Gaussian-weighted distance between the color of the reference image and other input images, then refined the image based on a global energy minimization system. Zhang and Cham [36] detected motion by analyzing the gradient maps between exposures image, due to the magnitude and orientation are different in motions object and saturation regions. Oh *et al.* [26] presented a rank minimization algorithm which simultaneously aligns LDR images and detects outliers for the robust HDR generation. Lee *et al.* [20] adopted rank minimization in HDR de-ghosting. Yan *et al.* [31] proposed a ghost-free HDR image synthesis algorithm which utilizes a sparse representation framework. However, the performances of those de-ghosting algorithms mainly depend on the accuracy of the ghost detection.

Recently, **deep learning** based methods have been proposed for low-level computer vision tasks [9, 37, 17]. There are also many works to generate HDR images. Eilertsen *et al.* [5] reconstructed a high-quality HDR image from single exposed LDR image with an auto-encoder network. Endo [6] synthesized LDR images from one LDR image with a deep neural network, then reconstructed a HDR image by merging the synthesized LDR images. Cai *et al.* [2] proposed a convolutional neural network (CNN) to train a SICE enhancer for a single image. Most of the works are devoted to generate HDR images from a single LDR image, due to they don't consider the dynamic scenes. However, generating HDR images from a sequence of LDR images is more vulnerable in the dynamic scenes. Kalantari *et al.* [17] utilized the optical flow to align LDR images to the reference image, then trained a deep CNN to generate HDR images. Compared with previous works, the results have sharper details. The main reason is that they used a weighted average of the aligned HDR images. Different from Kalantari's method, our work focuses on collecting information from multi-scale LDR images, and generates artificial free and colorful HDR images.

3. The Proposed Method

The overall framework of the proposed network is shown in Figure 2. The network is designed to introduce more information via a multi-scale dense connection architecture. As introduced in Section 1, we reconstruct an HDR image from a sequence of LDR inputs. More specifically, we use three LDR images of a dynamic scene (L_1, L_2, L_3), and the middle exposure image (L_2) is set as the reference. Our goal is to generate a ghost-free HDR image H, which is aligned with the reference image L_2. The information of the HDR image either merges the contents of LDR images or generated from the network. In this paper, we not only focus on how to remove ghost artifacts but also generate more vivid colorful HDR images.

The proposed method is divided into the data preprocessing and merge phases. Following [17], in the data preprocessing, the aligned LDR images I_1, I_2, I_3 are first created based on optical flow [21], then HDR images H_1, H_2, H_3 are generated by gamma correction.

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

where, t_i denotes the exposure time of i-th image I_i, γ is set as 2.2 in our experiments. Then we concatenate LDR and HDR images $X = \{I_1, I_2, I_3, H_1, H_2, H_3\}$ as inputs to the proposed network. LDR images can be used to detect the misalignment and saturation region, while HDR images are facilitated to generate the final HDR images H.

In the field of image reconstruction, the multi-scale scheme is a very successful strategy in both traditional optimization-based methods and recent neural-network based approaches [30, 37, 24, 35]. Therefore, we believe this strategy is also helpful to generate HDR images. Based on this motivation, we use multi-scale strategy to reconstruct HDR images and then refine them.

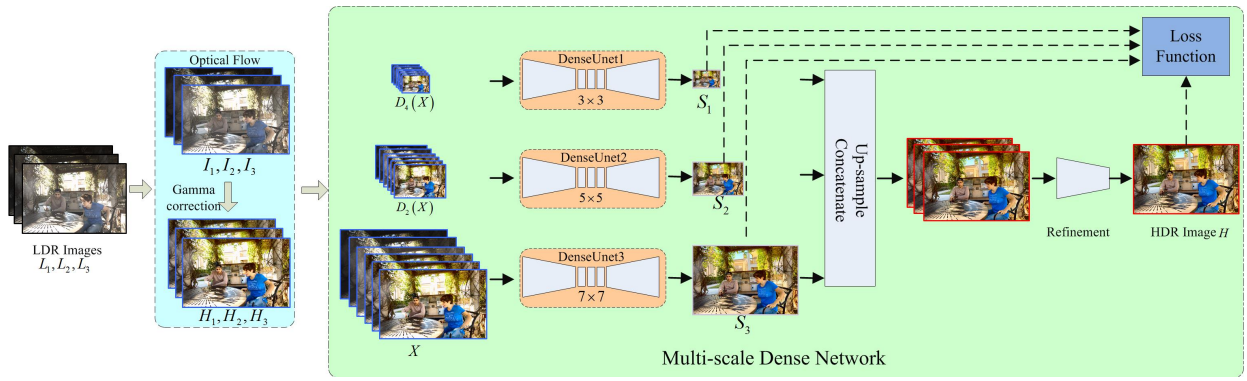


Figure 2. Our proposed framework. Firstly, the LDR images are aligned based on optical flow. Then, gamma correction is performed to change the pixel range. These images are feed into the multi-scale dense networks to generate the multi-scale HDR images. Afterwards, we utilize a shallow network to refine the multi-scale predictions. The final result is tonemapped before it can be displayed. S_i is the result of HDR images in the i-th scale, H denotes the result of final HDR image.

To enrich the image information, our method takes a sequence of LDR and HDR images as inputs at different scales, and adopts three subnetworks to obtain corresponding HDR images (global, middle and local). To uniform the resolution of the multi-scale prediction, we upsample the results from subnetworks by bilinear interpolation. The final HDR images are generated using a shallow refinement network. The refinement network takes the concatenation of the upsampled HDR images as inputs. Therefore, the HDR merge process can be expressed as:

$$H = g(D_4(X), D_2(X), X), \tag{2}$$

where g defines the mapping from the inputs to HDR image. $D(X)_4$ and $D(X)_2$ denote the downsampled inputs of X with scale 4 and 2.

3.1. Structure of the Dense U-Net for HDR imaging

The encoder-decoder structure is commonly used to map an image to a certain output image. U-Net [28] is an extension of the encoder-decoder structure, which introduces skip-connections to connect the encoder part and the decoder part. The skip-connections can benefit the gradient propagation and accelerate model convergence. The encoder part extracts high-level features, and the decoder part generates the corresponding predictions. This structure is a suitable solution to generate more information in HDR imaging, such as over-exposed or under-exposed regions. To reuse the multi-scale features, based on the U-Net structure, we introduce the Dense Convolutional Network (DenseNet), which connects each layer to other layers in a feed-forward fashion [15]. More specifically, the small scale is denoted by Dense U-Net1, the kernel size of which is 3×3. Thus the coarsest-scale network has a large enough receptive field to see the whole patch. The middle and high scale are define by Dense U-Net2 (kernel size is 5×5) and Dense U-Net3 (kernel size is 7×7), respectively. Based

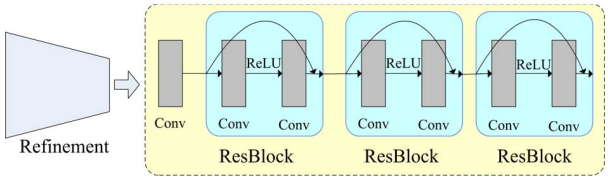


Figure 4. Details of the proposed refinement network. The refinement network includes three ResBlocks. Each of them contains two convolutional layers and one ReLU activation.

on this design, the multi-scale information can be obtained to reconstruct different scales HDR images. There are nine dense blocks and different transition blocks in each scale, as shown in Figure 3. All the building blocks consist of 16 channels, and the final output consists of 3 channels to fit the HDR image. The transition layer is added after each dense block to enhance the ability of representations. The network architecture of the three scales is the same. The only difference is the kernel size of convolution layer. Similar to U-Net, for dense block6, features with the same size are concatenated together as the input. Mathematically, each stream can be represented as:

$$In_6 = cat[DB3, DB4, DB5], \tag{3}$$

where cat denotes a concatenation operator, DB_i, $i = 1, ...9$ indicates the output features of the i-th dense block, In_6 denotes the input of dense block6. And the input of dense block8 and block9 can be denoted by

$$In_8 = cat[DB7, DB2], \tag{4}$$

$$In_9 = cat[DB8, DB1]. \tag{5}$$

3.2. Refinement Module

In order to concatenate the multi-scale predictions, we adopt the upsampling layers. However, the upsampling op-

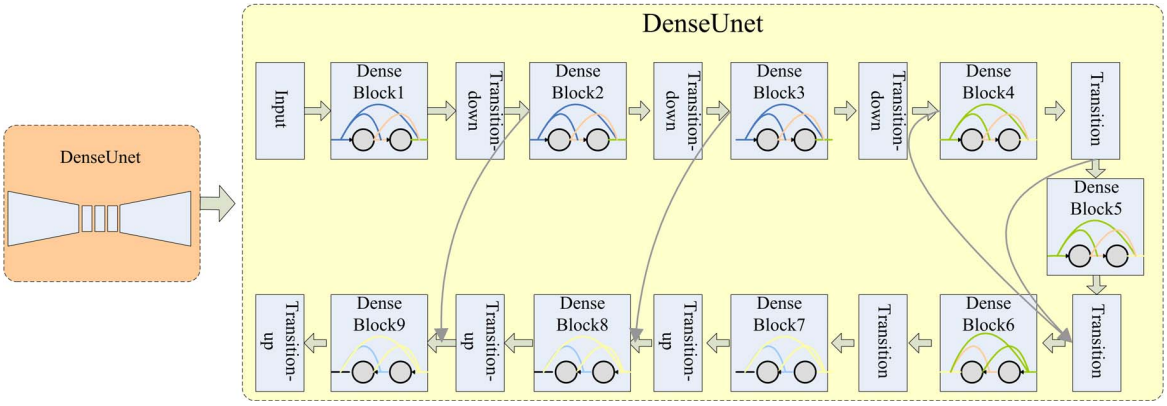


Figure 3. Details of the proposed Dense U-Net. Our network includes dense blocks, transition-down layers, transition layers and transition-up layers. The network architecture of three scales is the same, and the only difference is the kernel size of convolution layer.

erations inevitably bring in blurring artifacts of HDR images. To address this problem, we propose a shallow network to refine the multi-scale results. More specifically, we first upsample the result of lower scale S_1 and middle scale S_2 with factors 4 and 2 respectively (Figure 2). Then those images $(D_4(X), D_2(X), X)$ are concatenated together as the inputs of the refinement network. The refinement network contains one convolution layer followed by several ResBlocks. The first convolution layer is used to fuse the multi-scale images. The ResBlocks are used to refine the results and produce the final HDR image. We adopt the ResBlock in [12], which contains skip-connections from the early layers to the later layers. The short connections help to preserve the detailed high-frequency information in the early layers. As shown in Figure 4, our refinement network includes three ResBlocks, each of which contains two convolutional layers and one ReLU. All convolutional layers have the same number of kernels. The kernel is 3×3 with padding 1. This refinement process produces more details for HDR imaging.

3.3. Training Loss

To display HDR images on the screen, Kalantari [17] proposed to use μ-law function, a commonly-used range compressor in audio processing, which is defined as:

$$T = \frac{log(1+\mu H)}{log(1+\mu)}, \tag{6}$$

where $\mu = 5000$ defines the amount of compression, H is the HDR image, and T is the tonemapped image. The L_1-norm loss is employed for each scale, between the estimated and ground truth HDR images (downsampled to the same size using bilinear interpolation) as

$$\mathcal{L} = \sum_{i=1}^{n} \|T(S_i) - \hat{T}_i\|_1 + \|T(H) - \hat{T}\|_1, \tag{7}$$

where $T(S_i)$ and $\hat{T}_i$ are the estimated and ground truth HDR images in the i-th scale, respectively. $T(H)$ and $\hat{T}$ denote the final output and ground truth tonemapped HDR images. The loss not only ensures the estimated HDR images are similar with ground truth in different scales, but also guarantee the quality of the final result. We have also tried L_2-norm loss, however, we find that the L_1-norm is better to generate sharp and clear results.

4. Experiments

Our experiments are conducted on a PC with Intel i7 CPU (12 GB of memory) and an NVIDIA GeForce GTX 1080Ti GPU. We implement our framework based on the PyTorch platform. To evaluate the effectiveness of our network, different baseline network structures are tested. For fairness, all experiments are conducted on the same dataset with the same training configuration.

4.1. Dataset

To train our model, we adopt the public dataset in [17]. The dataset consists of training samples from 74 different scenes and testing samples from 15 different scenes. For each scene, three different LDR images with motion were taken. To generate the ground truth HDR image, Kalantari *et al.* [17] capture a static set by asking a subject to stay still and taking three images with different exposures on a tripod. Then the captured images are used to produce ground truth HDR with reference to the middle exposure image. To reduce the possible misalignment in the static set, all the images are resized to the resolution of 1500×1000. Due to the limitation of the dataset (only 74 training samples), it is hard to train a deep learning-based model if we directly feed the full-size image to the network. To avoid overfitting, data augmentation is performed. More specifically, we randomly crop and flip the patches with 256×256 as training images.

Networks	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
One Scale	41.2959	30.6563	40.9213	60.1513
Two Scale	41.4630	31.1986	40.9269	60.1344
Proposed	**42.2263**	**31.5845**	**41.0170**	**60.2991**

Table 1. Quantitative comparisons of the results on the testing set. All scores are the average across 15 testing images.

4.2. Training Details

For model training, the Adam algorithm [19] is used with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and batch size=1. The learning rate is set to 10^{-4} and divided by 10 after 60,000 epoch. According to our experiments, 80,000 epochs are enough for convergence, which takes about two days. All weights are initialized by the Xavier method [8]. The parameters described above are fixed for all experiments. Our method takes roughly 0.7 seconds to product the final HDR image H of size 1000×1500. In this paper, all the HDR results are tonemapped by Matlab function *tonemap* with same parameters.

4.3. Evaluation Metrics

To evaluate the predicted results, four metrics are considered. 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], which can measure the visual difference based on human perception rather than the element-wise difference between two images. The parameters of the diagonal display size and viewing distance in HDR-VDP-2 algorithm are set as 40 inches and 1 meter.

4.4. Study on Multi-scale Strategies

To evaluate the effectiveness of the multi-scale strategies, we design several baseline models. The **One Scale** model uses the single scale Dense U-Net, and only takes a single-scale image as input at its original resolution. The **Two Scale** model uses two Dense U-Nets, and the inputs include the image of the original resolution and downsampled one with factor 2. The results of different models are shown in Table 1. As we can see the multi-scale strategy is very effective for HDR imaging. The results are consistently improved when we use two-scale network. Compared with the **One Scale** structure, the improvements of the **Two Scale** are rather minor. However, the proposed network with three scales produces better performance. The main reason is that multi-scale information has been effectively incorporated.

Loss	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
MSE loss	41.81	31.44	40.79	**60.59**
L_1 loss	**42.23**	**31.58**	**41.02**	60.30

Table 2. Quantitative comparisons of the results with different losses. All the scores are averaged over 15 testing scenes.

4.5. Study on Training Loss Function

To analyze the effects of different training losses, we evaluate the performance with different loss functions. We additionally train our model with Mean Squared Error (MSE) loss and L_1-norm loss, respectively. The MSE loss is defined by

$$\mathcal{L}_{MSE} = \sum_{i=1}^{n} \|T(S_i) - \hat{T}_i\|_F^2 + \|T(H) - \hat{T}\|_F^2. \quad (8)$$

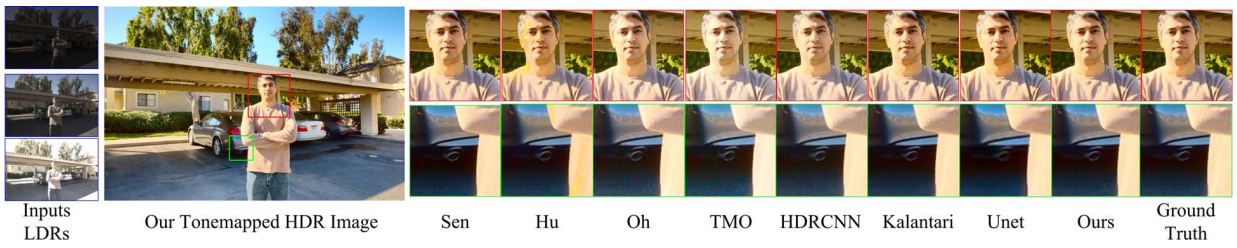
As we all know, the L_1 loss can generate more sharp images but easily affected by noises, while the MSE loss may capture the smooth result. The results of L_1 and MSE loss on the testing dataset are shown in Table 2. Although the MSE loss results in higher HDR-VDP-2 values, the L_1 loss achieves higher values in terms of the other three quality assessment metrics. Our goal is to show the HDR image on the screen. We focus on the metric PSNR-μ and PSNR-M. Hence, we choose L_1 loss as the training loss function.

4.6. Comparison with Other Methods

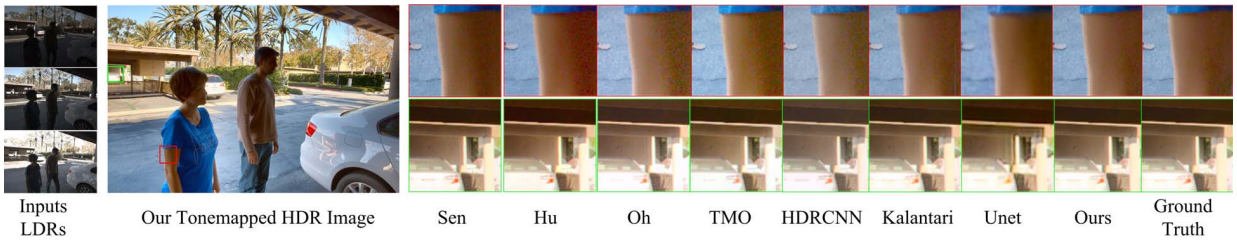
We compare our proposed model with several state-of-the-art methods. Specifically, we compare against two patch-based methods [29, 14], the method based on motion region detection [26], the flow-based approach with CNN merger [17]. For a fair comparison, we reproduce the code of [17] by PyTorch. In addition, we also compare with two single image HDR imaging methods [6, 5].

4.6.1 Qualitative Comparison

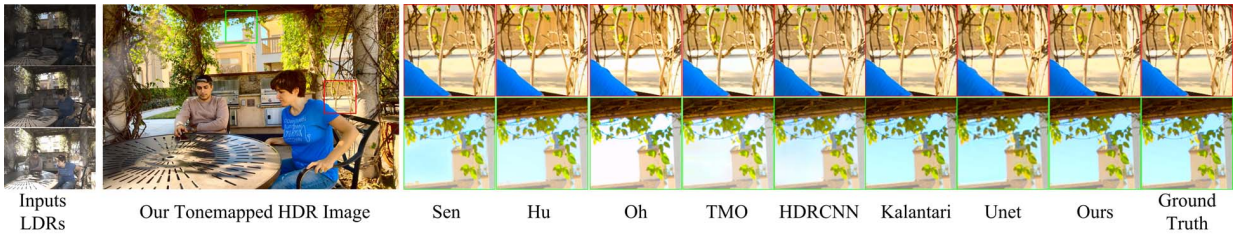
Figure 5 shows the visual comparison of our approach against other state-of-the-art methods on the testing set. The left column shows three LDR images, the middle images are our tonemapped HDR results, and the right nine columns show the detailed results of our approach and other methods. The numbers at the bottom present the PSNR of the tonemapped images using the tonemapping function in MATLAB. Compared with other methods, our method obtains more information in saturated regions. Figure 6 shows the corresponding HDR-VDP-2 visibility probability maps. The HDR images generated from single images have large differences compared to the ground truth. Meanwhile, they often have noises in predicted results. The main reason maybe that they heavily rely on the reference image. While our method can achieve much better results.



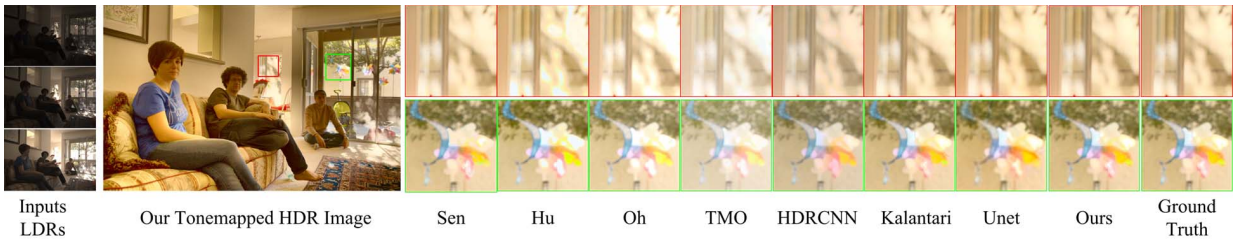
(a)Manstanding



(b) Garden



(c) Barbeque



(d) Talking

Figure 5. Comparison of our approach against several state-of-the-art methods on the testing set.

To verify the generalisation, we also compare our approach against other methods on Sen's dataset [29]. As shown in Figure 7, our method produces appealing results, and outperforms others on this dataset. For example, our method produces clear results in zoomed-in regions, while other methods lose details of saturated regions (such as zoomed-in regions of Figure 7 (a)) or prone generate artifacts (such as zoomed-in regions of Figure 7 (b)(c)). Compared with Kalantari *et al.* [17], our result products sharper edges in under-exposed regions (the arm and hand in Figure 7 (b)). Due to the multi-scale information are introduced, the color of our results looks more realistic and the network can hallucinate missing details more easily compared to other methods.

4.6.2 Quantitative Comparison

Table 3 shows the quantitative comparison averaged over the 15 test scenes. As can be seen, the methods [6] and [5] have poor performance, because of the single image does not have the different exposure information. From the experiments, we also find that global color consistency is a critical factor for HDR imaging. For example, results of [26] suffer from substantial differences from the ground truth. The main reason is that the global intensity of result is more higher than ground truth. However, the global information is introduced in our network. The results of the proposed method have better performance than other methods on PSNR-μ, PSNR-M, PSNR-L, HDR-VDP-2. Besides, the tonmapping algorithm has a significant impact on the

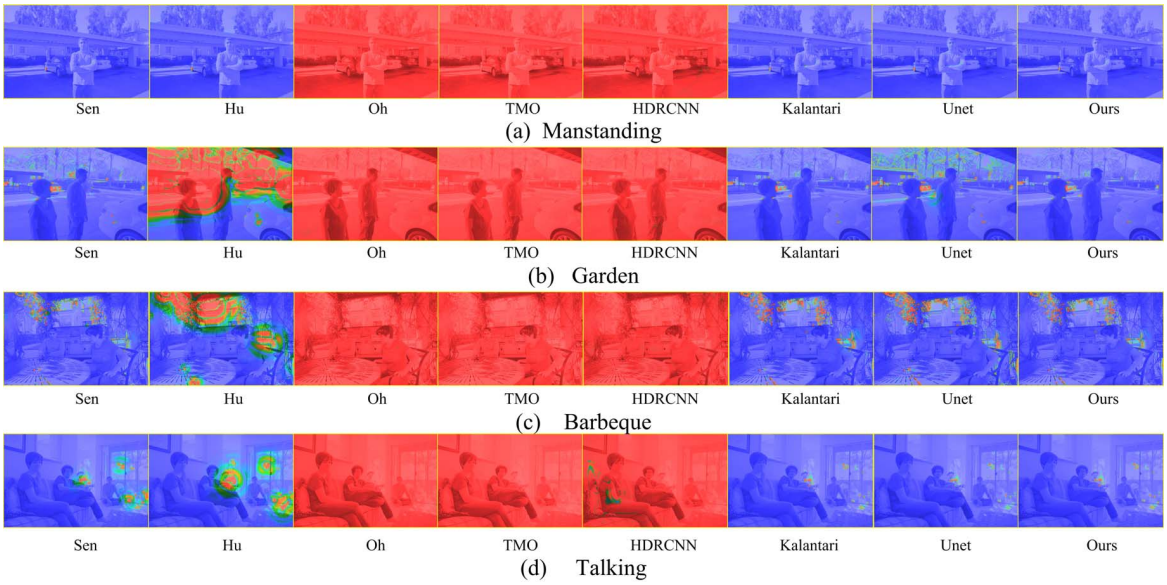


Figure 6. HDR-VDP-2 visibility probability maps. Red and blue denote major and minor differences, respectively.

Methods	PSNR-μ	PSNR-M	PSNR-L	HDR-VDP-2
Sen [29]	40.9453	30.5507	38.3147	55.7240
Hu [14]	32.1872	25.5937	30.8395	55.2496
Oh [26]	27.351	22.6311	27.1119	46.8259
TMO [6]	8.2123	21.4368	8.6846	44.3944
HDRCNN [5]	14.0925	25.8217	13.1116	47.7399
Kalantari [17]	42.0098	31.3868	40.7091	59.7234
U-Net	37.9675	27.6974	40.0354	60.0866
Ours	**42.2263**	**31.5845**	**41.0170**	**60.2991**

Table 3. Quantitative comparisons of the results on test set. All scores are the average across 15 testing images.

final result. However, under the same setting, our method achieves high performance in the terms of both PSNR-μ and PSNR-M.

5. Conclusions

In this paper, we present a multi-scale dense network to generate HDR images. Our method adopts a coarse-to-fine scheme, which contains three subnetworks to gradually reconstruct the HDR images. These subnetworks can generate multi-scale information which is very helpful to handle saturated and motion regions. Furthermore, we introduce dense connections into U-Net to reuse the features. By the refinement module with the multi-scale HDR images, our method can generate the artificial-free image with more details. Experimental results show that our method is excellent in the HDR imaging, both qualitatively and quantitatively.

Figure 7. Visual comparison of different approaches on Sen's dataset [29].

Acknowledgment

The work was partially supported by grants NSF of China (61231016, 61301193, 61303123, 61301192), Chang Jiang Scholars Program of China (100017GH030150, 15GH0301), Australian Research Council grants (CE140100016, FL130100102, DP160100703) and China Scholarship Council.

References

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

[2] J. Cai, S. Gu, and L. Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, PP(99):1–1, 2018.

[3] Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu. Large displacement optical flow from nearest neighbor fields. In *Computer Vision and Pattern Recognition*, pages 2443–2450, 2013.

[4] Debevec, E. Paul, Malik, and Jitendra. Recovering high dynamic range radiance maps from photographs. *Proc Siggraph*, 97:369–378, 1997.

[5] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger. Hdr image reconstruction from a single exposure using deep cnns. *Acm Transactions on Graphics*, 36(6), 2017.

[6] Y. Endo, Y. Kanamori, and J. Mitani. Deep reverse tone mapping. *Acm Transactions on Graphics*, 36(6):1–10, 2017.

[7] O. Gallo, N. Gelfand, W.-C. Chen, M. Tico, and K. Pulli. Artifact-free high dynamic range imaging. In *Proceedings of the IEEE International Conference on Computational Photography*, pages 1–7, 2009.

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

[9] D. Gong, J. Yang, L. Liu, Y. Zhang, I. Reid, C. Shen, A. V. D. Hengel, and Q. 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*, 2016.

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

[11] M. D. Grossberg and S. K. Nayar. Determining the camera response from images: What is knowable? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1455–1467, 2003.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pajak, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli. Flexisp: a flexible camera image processing framework. *Acm Transactions on Graphics*, 33(6):231, 2014.

[14] J. Hu, O. Gallo, K. Pulli, and X. Sun. Hdr deghosting: How to deal with saturation? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1170, 2013.

[15] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2261–2269, 2017.

[16] K. Jacobs, C. Loscos, and G. Ward. Automatic high dynamic range image generation of dynamic environments. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008.

[17] N. K. Kalantari and R. Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *Acm Transactions on Graphics*, 36(4):1–12, 2017.

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

[19] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

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

[21] C. Liu. *Beyond pixels: exploring new representations and applications for motion analysis*. Massachusetts Institute of Technology, 2009.

[22] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. 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] T. Mertens, J. Kautz, and F. V. Reeth. Exposure fusion. In *Conference on Computer Graphics and Applications*, pages 382–390, 2007.

[24] S. Nah, T. H. Kim, and K. M. Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–265, 2016.

[25] S. K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, pages 472–479 vol.1, 2002.

[26] T. H. Oh, J. Y. Lee, Y. W. Tai, and I. S. Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1219–1232, 2015.

[27] E. Reinhard, G. Ward, S. Pattanaik, and P. E. Debevec. *High dynamic range imaging : acquisition, display, and image-based lighting*. Princeton University Press,, 2005.

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

[29] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics*, 31(6):1–11, 2012.

[30] X. Tao, H. Gao, Y. Wang, X. Shen, J. Wang, and J. Jia. Scale-recurrent network for deep image deblurring. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[31] Q. Yan, J. Sun, H. Li, Y. Zhu, and Y. Zhang. High dynamic range imaging by sparse representation. *Neurocomputing*, 269:160–169, 2017.

[32] Q. Yan, Y. Zhu, and Y. Zhang. Robust artifact-free high dynamic range imaging of dynamic scenes. *Multimedia Tools and Applications*, doi: https://doi.org/10.1007/s11042-018-6625-x, 2018.

[33] Q. Yan, Y. Zhu, Y. Zhou, J. Sun, L. Zhang, and Y. Zhang. Enhancing image visuality by multi-

exposure fusion. *Pattern Recognition Letters*, doi: https://doi.org/10.1016/j.patrec.2018.10.008, 2018.

[34] H. YongSeok, L. KyoungMu, L. SangUk, M. Youngsu, and C. Joonhyuk. Ghost-free high dynamic range imaging. In *Proceedings of the IEEE International Conference on Asian Conference on Computer Vision*, pages 486–500, 2011.

[35] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. *In IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[36] W. Zhang and W.-K. Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2012.

[37] P. Zhou, B. Ni, C. Geng, J. Hu, and Y. Xu. Scale-transferrable object detection. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[38] H. Zimmer, A. Bruhn, and J. Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, pages 405–414, 2011.