Single-Image HDR Reconstruction by Multi-Exposure Generation –Supplementary Material–

Phuoc-Hieu Le[1] Quynh Le*[1,2] Rang Nguyen[1] Binh-Son Hua[1]

[1]VinAI Research [2]University of California San Diego

Abstract

In this supplemental document, we show additional comparisons on HDR merging methods (Sec. 1), on multiple-exposure images predicted by the network (Sec. 2) and then provide details about the proposed network, including network architecture (Sec. 3.1) and the inference process (Sec. 3.2).

1. On HDR Merging Methods

We compare different HDR merging methods including the commercial software Photomatix [4], OpenCV's merging based on Debevec and Malik [2] and Robertson *et al*. [10]. The comparison results are shown in Tab. 1 with the evaluation on HDR images. Additional to results of Photomatix reported in the main paper, we provide evaluations on tone-mapped images when the HDR merging method is Debevec and Malik [2] (Tab. 2) and Robertson *et al*. [10] (Tab. 3). From these results it can be seen that our method is not sensitive to HDR merging methods as in most cases, our method outperform previous methods.

Note that in the HDR domain, HDR-VDP-2 is a more preferred metric, which shows that Photomatix has better performance in HDR reconstruction. This aligns with our investigation as well, as we found that output images from Photomatix has less visual artifacts than those from Debevec and Malik [2] and Robertson *et al*. [10].

2. Additional Results and Comparisons

Quantitative comparisons on multi-exposure images. We use EV_0 as input and compare the predicted bracketed images as shown in Tab. 4. As can be seen, our model's scores are better than other methods. This result can be interpreted as that the overall structure of images using our model is well reconstructed as well as perceptually similar to ground-truths.

Qualitative comparisons on multi-exposure images. A significant advantage of our method compared to the previous works is that our approach allows the synthesis of images at arbitrary exposure values. We demonstrate this capability in Fig. 3, where the exposure value can be 0.75, 1.5, and 2.25. In the included video, our network can predict a smooth change in the virtual exposures of a scene.

Qualitative comparisons on tone-mapped images. We give further qualitative evaluation of tone-mapped images on diverse scenes as shown in Fig. 5 (natural scenes), Fig. 6 (outdoor scenes), and Fig. 7 (indoor scenes). Our method can reconstruct the HDR that matches well to the reference in color and contrast as well as produce minimal artifacts compared to DrTMO [3], DRHDRI [6], and SingleHDR [7] method.

Application: virtual bracketed images. Our method can be used as an application for virtually changing the exposure for an input image. Figure 4 illustrates the generated bracketed images by ours along with the references. We can see that the predicted stack matches the ground truth very well in terms of color and contrast. Our model can also generate images with smooth changes in the exposure values that are not defined in the training dataset. Please refer to supplementary materials for a video demonstration.

3. Details of the Proposed Network

3.1. Network Architecture

As our goal is to generate different exposure images from the input image, this can be seen as an image-to-image translation task. We adopt the U-Net [11] like architecture with the encoder-decoder module, which has shown good performance in this task. When the data goes to the next level, the size of feature maps is reduced by half, vertically and horizontally, and conversely doubled. Then, the abstracted feature map is reassembled with the previous feature maps for creating the desired output through a structure that in-

*This work is done when Quynh Le was a resident of the AI Residency program at VinAI Research.

HDR Reconstruction Method	Photomatix [4]				Debevec and Malik [2]				Robertson *et al*. [10]			
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	HDR-VDP-2 (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	HDR-VDP-2 (↑)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	HDR-VDP-2 (↑)
Lee *et al*. [6]	19.56	0.7920	0.2096	53.86 ± 4.46	21.07	0.9017	0.1362	39.21 ± 2.94	**19.20**	**0.8363**	0.2012	39.03 ± 2.79
Endo *et al*. [3]	21.60	0.8493	0.1592	54.56 ± 4.29	24.27	0.9243	0.1404	39.32 ± 2.93	15.27	0.7342	0.2664	39.09 ± 2.84
Ours	**23.74**	**0.8916**	**0.1231**	**55.69 ± 5.01**	**25.67**	**0.9434**	**0.0802**	**39.42 ± 2.95**	16.34	0.8035	**0.1722**	**39.30 ± 2.90**

Table 1: Quantitative comparisons on HDR images using different HDR merging methods.

Tone-mapping Method	Reinhard *et al*. [9]			Photomatix [4]		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Lee *et al*. [6]	27.97	0.9584	0.1118	23.85	0.9282	0.1362
Endo *et al*. [3]	29.99	0.9618	0.1257	25.98	0.9451	0.1482
Ours	**32.08**	**0.9819**	**0.0508**	**28.01**	**0.9663**	**0.0723**

Table 2: Quantitative comparisons on tone-mapped images with existing methods using Debevec and Malik [2] merging algorithm.

Tone-mapping Method	Reinhard *et al*. [9]			Photomatix [4]		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
Lee *et al*. [6]	**25.29**	0.9271	0.1628	21.27	0.8758	0.1986
Endo *et al*. [3]	22.49	0.9271	0.1984	**23.49**	0.8967	0.2350
Ours	23.89	**0.9522**	**0.1004**	20.34	**0.9015**	**0.1350**

Table 3: Quantitative comparisons on tone-mapped images with existing methods using Robertson *et al*. [10] merging algorithm.

creases the width and height of the feature map. In this structure, we add skip-connections between encoder and decoder layers so that the characteristics of low-level features are reflected in the output. The down-sampling block consists of a convolutional layer followed by one batch normalization layer [5] and ReLU activation function. The up-sampling block contains a sub-pixel convolution layer [1] and one convolutional layer instead of a deconvolution layer or resized convolution [8] as Aitken *et al*. [1] showed that the sub-pixel convolution has more modeling power with the same computation as resize convolution [8]. The convolutional layer's output then gets passed onto the batch normalization layer and one Leaky ReLU activation function. The same architecture is used for all three sub-networks except for the output convolutional layer of these.

In our implementation, our choice of U-Net for each sub-network consists of 7 levels. Each level has two convolution layers with a kernel size of 3×3, a stride of 1, and padding of 1. The input is first extracted into 16 and 32 features in HDR Encoding Net and Up/Down-Exposure Net, respectively. The number of features then doubled at each level until reaching 256 features for HDR Encoding Net and 512 for Up/Down-Exposure Net. In the decoder, the extracted features get channel-wise concatenation from the previous level encoder at each level's start. The last convolution layer in each sub-network applies 1×1 kernel to combine feature maps. We define each sub-network to produce a 3-channel output.

HDR Encoding Net. In the *HDR Encoding Net*, the last layer contains one convolutional layer followed by $\tanh$ activation function and normalization. Mathematically, given the output features from the convolutional layer as F, our network $\mathcal{N}_1$ output is:

$$\hat{X} = \frac{1}{3}\left(\tanh(F) + I + 1\right) \tag{1}$$

with I is the input image for this network, and $\hat{X}$ is the sensor exposure representation of its image. The $\tanh$ activation function's output gets added with the input image before feeding into the following network. As $\tanh$'s output value is within the range of $[-1, 1]$ thus can be seen as a global adjustment on the input image I. With this, the network can better find a representation that is suitable to generate different exposure images. Then, as the physical property of sensor irradiance can only have positive values, we normalize the output to scale negative values after the previous adjustment. We experimented and decide to use the $\tanh$ activation function instead of ReLU as it showed stable training and faster convergence under the use of $\tanh$.

Up/Down-Exposure Net. As for Up and Down-Exposure Net, the two sub-networks have to output the longer and shorter exposures, respectively. For this reason, the last layer

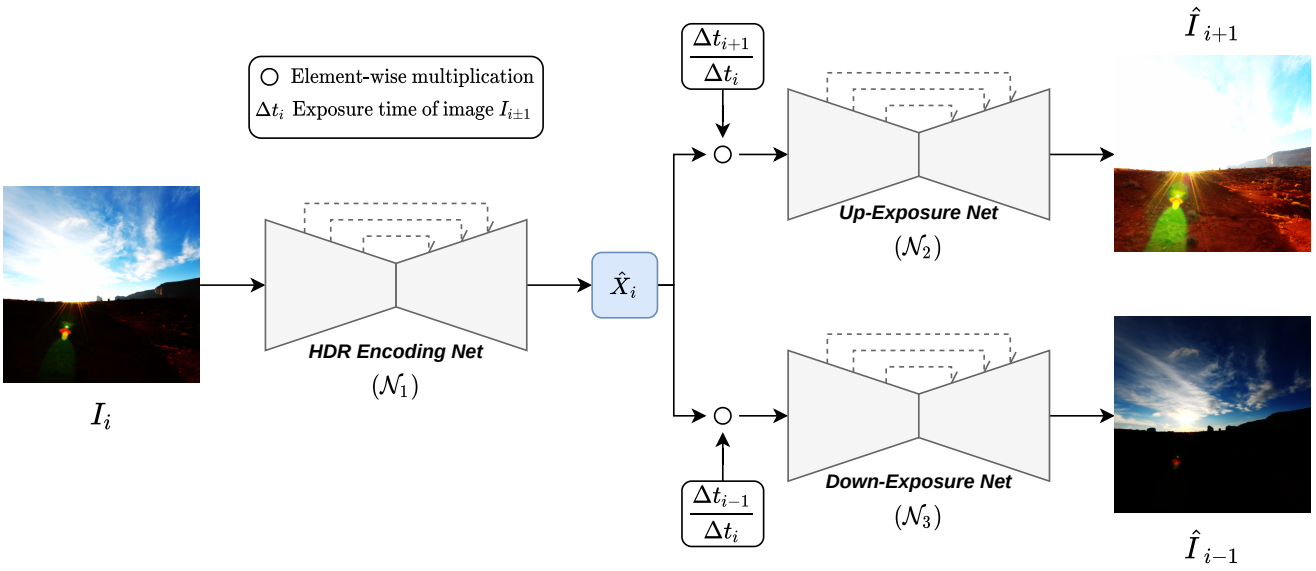


Figure 1: At inference, we take a single image as input and predict multiple exposures by passing the image to Up-Exposure and Down-Exposure Net and varying the exposure time Δt. The HDR image can be reconstructed from multiple exposures and tone-mapped using traditional methods.

is the combination between a convolutional layer followed by the *normalized* $\tanh$ activation function that is defined as:

$$\tanh_{\text{norm}}(x) = \frac{1}{2}\left(\tanh\left(x\right) + 1\right) \tag{2}$$

The activation output of the function will lie within $[0, 1]$, which can be seen as the normalized image. We opt for this function as $\tanh$ has stronger gradients than the usual sigmoid $\sigma(x)$ which helps speed up the learning process. We then can use other reconstruction loss functions to optimize the network.

Masked Regions. As mentioned before, our model takes input with masked over- and under-exposed regions. The input images first convert from RGB into YUV color space. Then based on the luma component Y of the image, we identify which pixel is over- or under-exposed. Particularly, we denote $\Lambda\left(.\right)$ as the indicator function for the well-exposed regions of image:

$$I_i^{'} = I_i \odot \Lambda\left(I_i\right),\ i = 1, 2 \tag{3}$$

The function outputs a soft mask within the range $[0, 1]$ that helps to define how well-exposed each pixel is. The value of 1 indicates that the pixel is well-exposed. Conversely, the 0 value is assigned to the pixels that are under-exposed or saturated. Mathematically, $\Lambda\left(I\right)$ is formulated as:

$$\begin{aligned}
\Lambda_1\left(I\right) &= 1 - \frac{\max\left(0, (1-\gamma) - I\right)}{1-\gamma} \\
\Lambda_2\left(I\right) &= 1 - \frac{\max\left(0, I-\gamma\right)}{1-\gamma} \\
\Lambda\left(I\right) &= \max\left(\Lambda_1\left(I\right), \Lambda_2\left(I\right)\right)
\end{aligned} \tag{4}$$

where γ is the threshold to determine whether a pixel is over/under-exposed or not. In our implementation, we choose $\gamma = 0.05$. An example of the masks is shown in Fig. 2.

3.2. Inference process

At inference, we take the input image and pass it to the network to generate multiple exposures. Our network allows us to vary the exposure time to obtain an image bracket with exposures from EV_{-2} to EV_{+2}, respectively. We then apply traditional HDR reconstruction to merge the predicted exposures into the final HDR image. Figure 1 demonstrates our inference process.

References

[1] Andrew Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

[2] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings*

		Ours	Endo [3]	Lee [6]
$EV_{-2.0}$	PSNR	**27.25**	24.40	19.40
	SSIM	**0.9350**	0.9022	0.7848
	LPIPS	**0.1032**	0.1966	0.2214
$EV_{-1.5}$	PSNR	**28.13**	25.01	—
	SSIM	**0.9511**	0.9205	—
	LPIPS	**0.0898**	0.1877	—
$EV_{-1.0}$	PSNR	**30.12**	26.65	23.38
	SSIM	**0.9667**	0.9386	0.9077
	LPIPS	**0.0732**	0.1764	0.1323
$EV_{-0.5}$	PSNR	**33.92**	30.23	—
	SSIM	**0.9823**	0.9569	—
	LPIPS	**0.0509**	0.1602	—
$EV_{+0.5}$	PSNR	**33.50**	32.65	—
	SSIM	0.9629	**0.9645**	—
	LPIPS	**0.0379**	0.1286	—
$EV_{+1.0}$	PSNR	**31.15**	30.61	25.28
	SSIM	**0.9613**	0.9552	0.9290
	LPIPS	**0.0553**	0.1335	0.1207
$EV_{+1.5}$	PSNR	**29.74**	29.19	—
	SSIM	**0.9541**	0.9442	—
	LPIPS	**0.0734**	0.1405	—
$EV_{+2.0}$	PSNR	**29.04**	28.12	21.84
	SSIM	**0.9468**	0.9324	0.8840
	LPIPS	**0.0899**	0.1469	0.1940

Table 4: Quantitative results on inferred bracketed images using EV_0 as the input. The — symbol indicates *Not Available* due to the proposed model by Lee *et al*. [6] can only produce images with EV differ by one value.

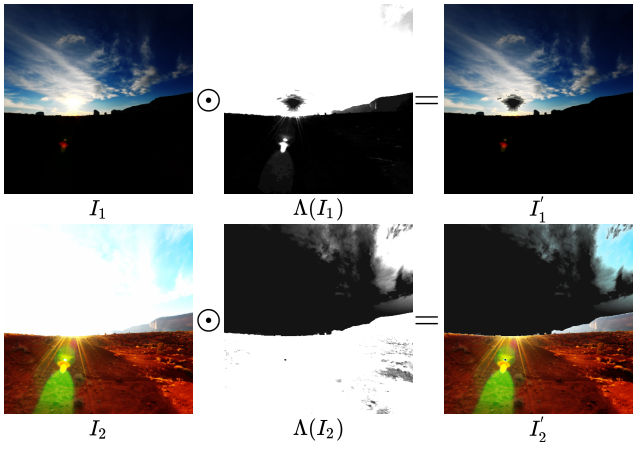
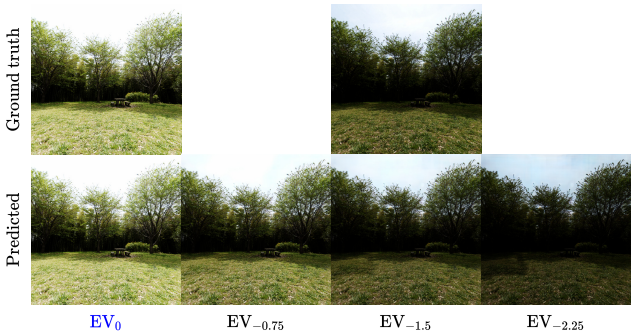
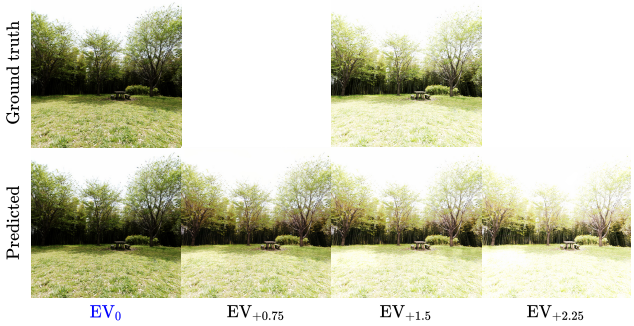


Figure 2: An example of the masks that are applied to the input images before passing through our model.



(a) Down-exposed inferred images.



(b) Up-exposed inferred images.

Figure 3: Our method can generate virtual exposures at arbitrary exposure value not in the training set, e.g., $EV_{\pm 0.75}$ and $EV_{\pm 2.25}$. Blue text indicates the input image (EV_0).

of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '97, page 369–378, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

[3] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6), nov 2017.

[4] HDRsoft. Photomatix.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 448–456. JMLR.org, 2015.

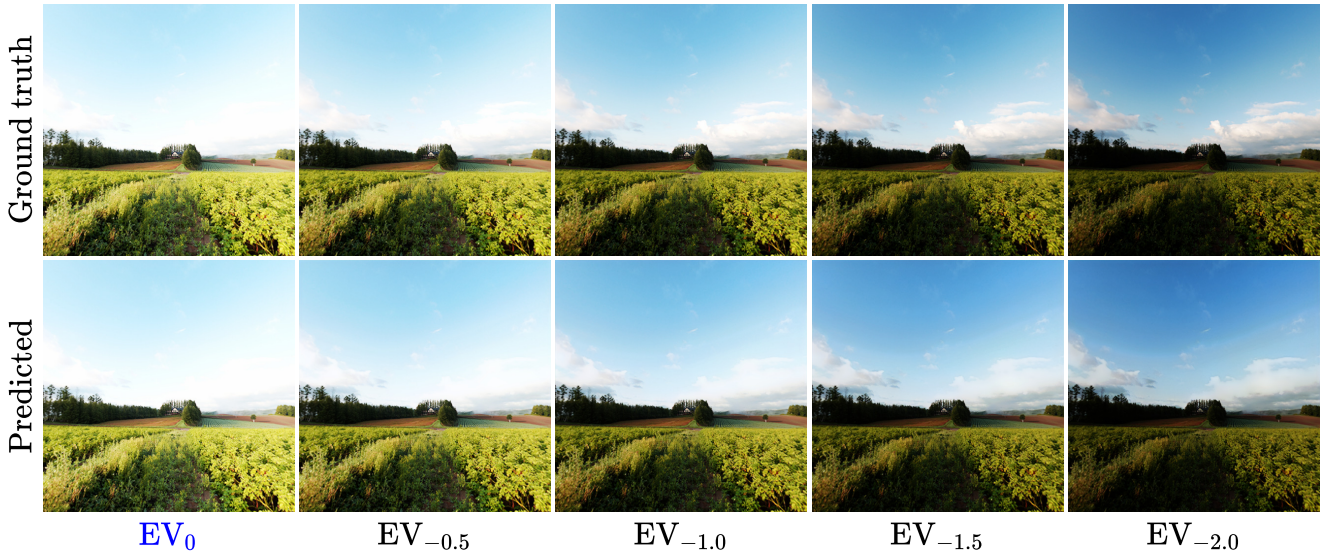
[6] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018.

[7] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1651–1660, 2020.

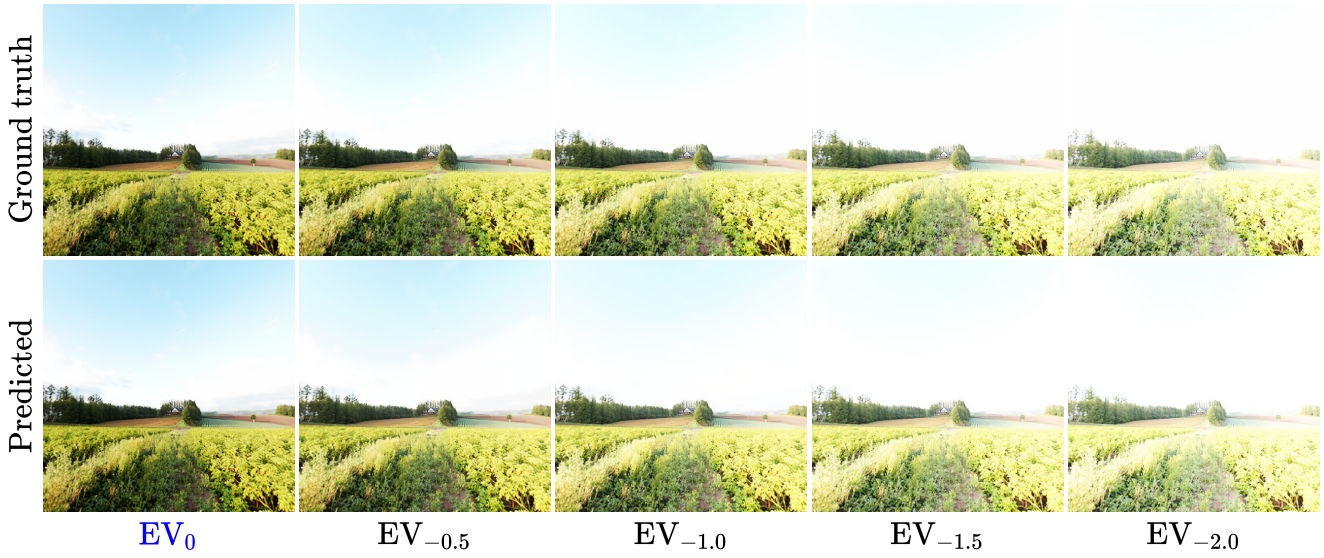
[8] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[9] E. Reinhard and K. Devlin. Dynamic range reduction inspired by photoreceptor physiology. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):13–24, 2005.

[10] Mark A Robertson, Sean Borman, and Robert L Stevenson. Dynamic range improvement through multiple exposures. In



(a) Down-exposure inferred images.



(b) Up-exposure inferred images.

Figure 4: Predicted multi-exposure images. The predicted stack matches the ground truth very well in terms of color and contrast. Input is EV_0.

Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348), volume 3, pages 159–163. IEEE, 1999.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

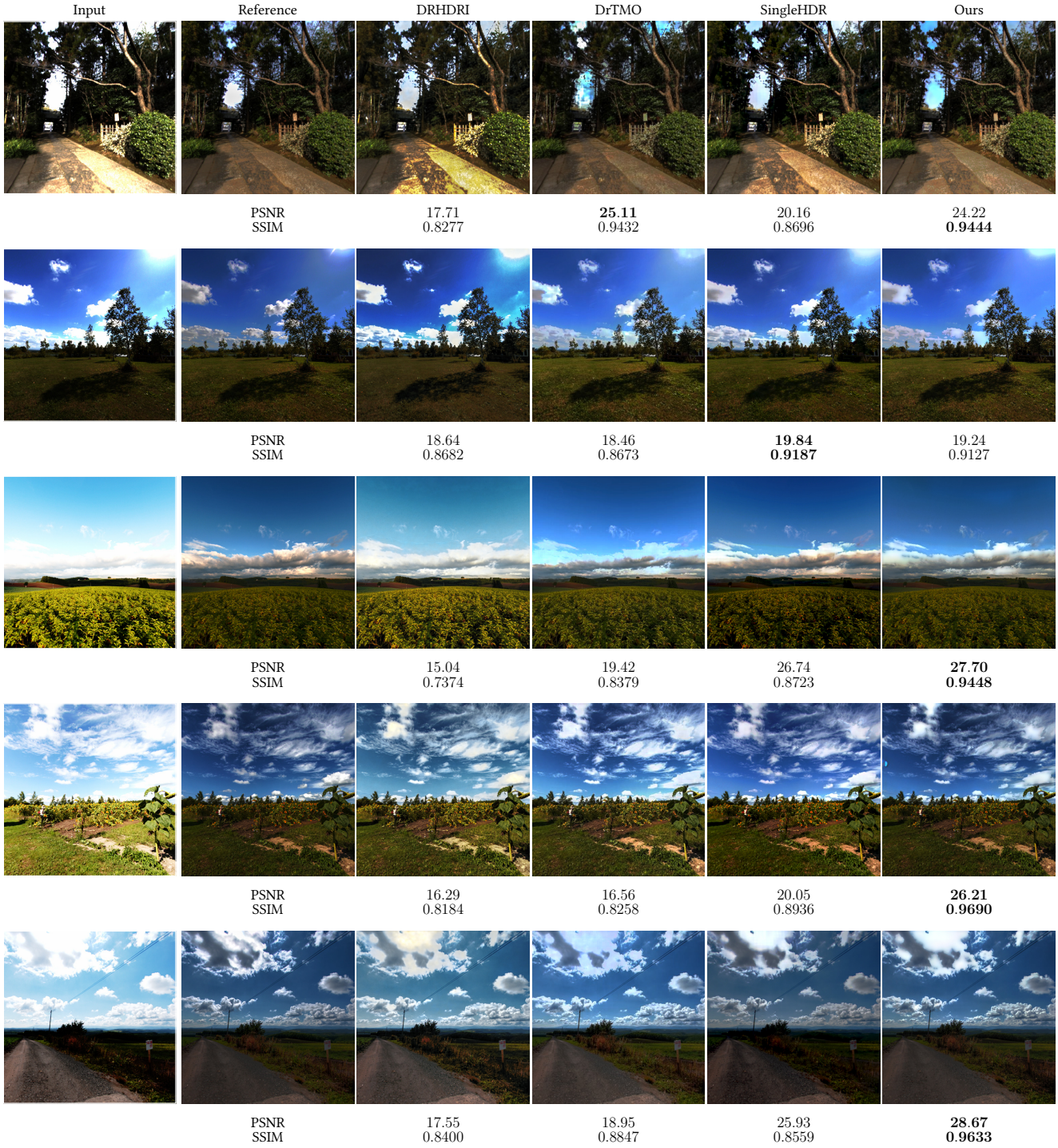


Figure 5: Tone-mapped HDR images comparison between ours, DrTMO [3], Deep Recursive HDRI [6], and SingleHDR [7] on nature scenes.

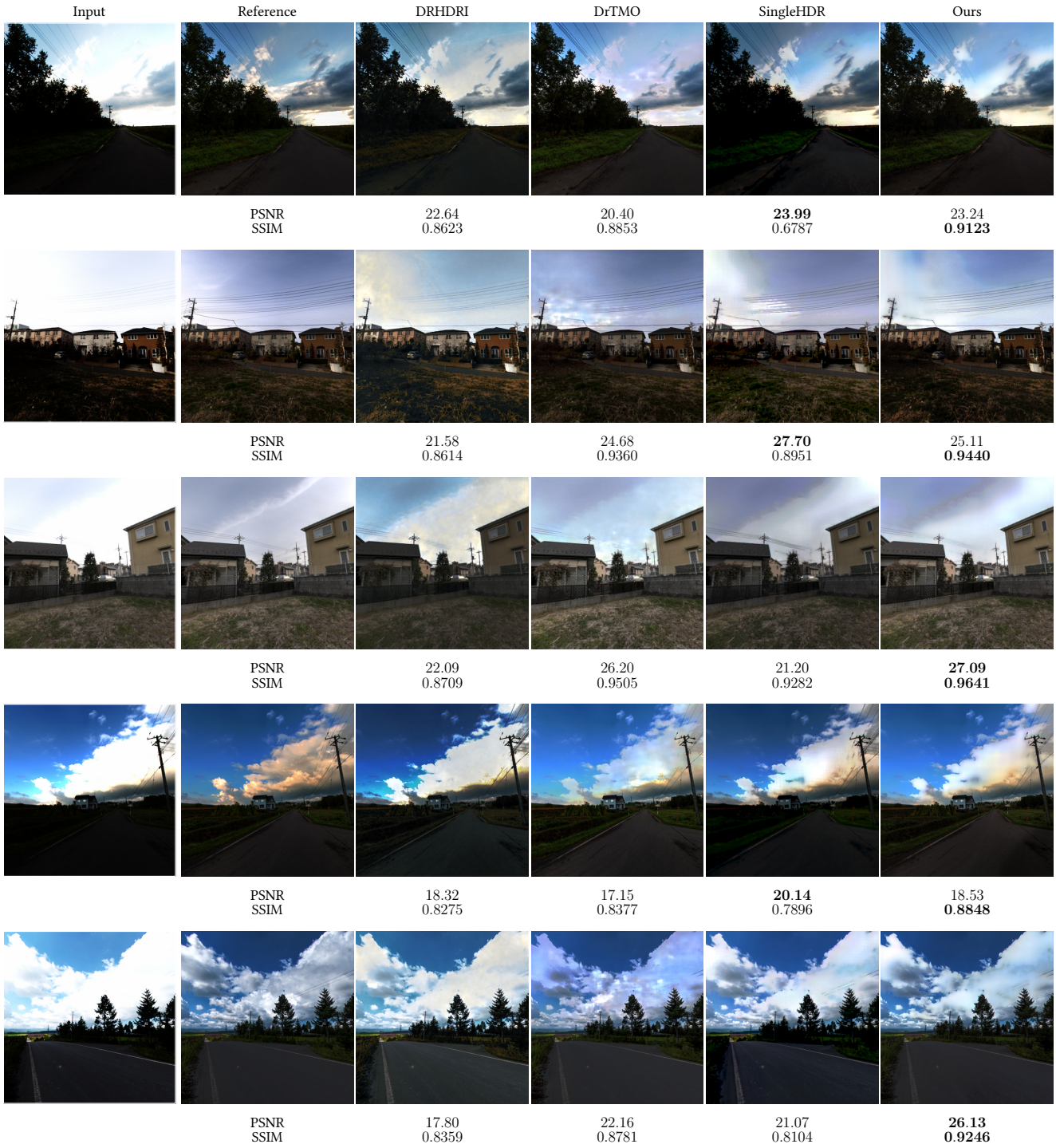


Figure 6: Tone-mapped HDR images comparison between ours, DrTMO [3], Deep Recursive HDRI [6], and SingleHDR [7] on outdoor scenes.

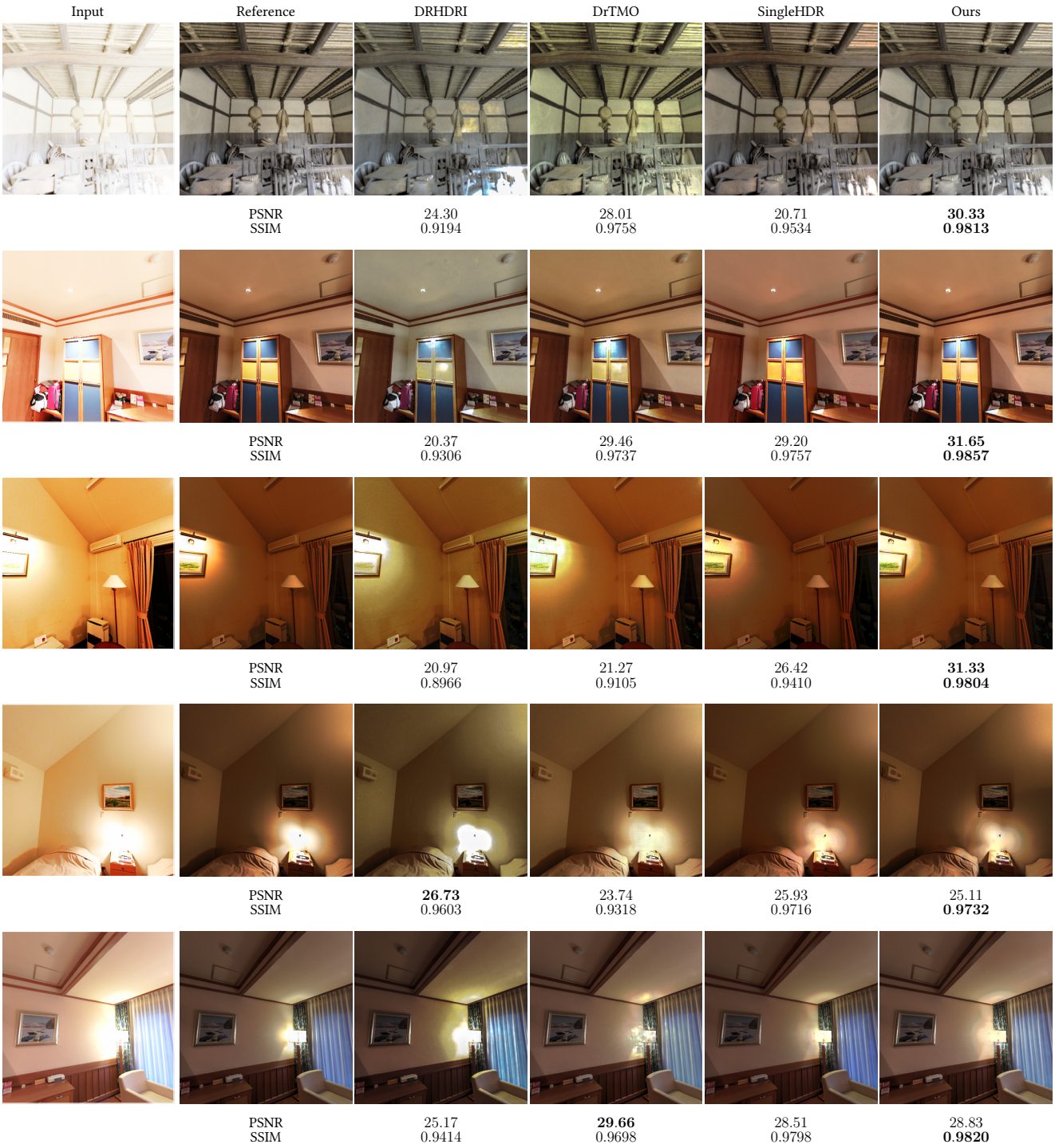


Figure 7: Tone-mapped HDR images comparison between ours, DrTMO [3], Deep Recursive HDRI [6], and SingleHDR [7] on indoor scenes.