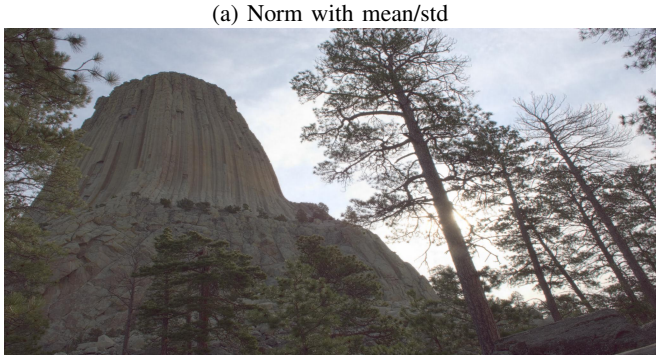
Supplementary Material: Deep Tone Mapping Operator for High Dynamic Range Images

I. Input HDR

In this paper, we feed the network with input HDR scaled with high precision [0,1] range. However, before adopting the simplistic scaling, we performed an experimental study using different normalization strategies to study the effect on tone mapped output, which includes a) mean-std normalization, where the input images are normalized by the input's mean and variance, b) min/max scaling, where the input is simply scaled in the range [0,1] while preserving the very high precision. We experimentally observed that generated images get very skewed results while forcing the input to have the same mean using (a). One example is shown in the Fig. 1 where we observe low illumination in dark regions and dark patchy patterns in the sky.

(a) Norm with mean/std



(b) Norm with [0,1] scaling

Fig. 1: Quantitative performance comparison of best performing DeepTMO with the target TMOs.

II. DeepTMO (Single-Scale) Architecture

In this section we specify the detailed architectural details of basic single-scale generator and discriminator.

A. Generator Architecture

$G^{(Front)}$ has first a convolution layer consisting of 64 filters kernels (or output channels) each of size 7×7 applied with a stride of (1,1) and padding (0,0). Next, there are four convolution layers with 128, 256, 512 and 1024 filter kernels respectively each with a size 3×3 and stride (2,2) and padding (1,1). Each of these four layers are followed by the batch norm with batch size = 1 (also called instance normalization [1]) and Relu [2]. Following this, we have $G^{(Res)}$ which is a set of 9 residual blocks, each of which contains two 3×3 convolutional layers, both with 1024 filter kernels. Next, for $G^{(Back)}$ there are four de-convolutional or transposed convolution layers with 512,256,128,64 filter kernels, each having a filter size of 3×3 and fraction strides of $\frac{1}{2}$. Both these layers have instance normalization and Relu added after the convolution. Finally, there is another convolution layer of size 7×7 and stride 1 followed by a tanh activation function at the end.

B. Discriminator Architecture

Discriminator architecture consists of 4 convolution layers of sizes 4×4 and stride (2,2). From first to the last, the number of filter kernels is 64, 128, 256 and 512 respectively. Each of the convolutional layer is appended with an instance normalization (except the first layer) and then leaky ReLU [3] activation function (with slope 0.2). Finally, a convolutional layer is applied at the end to yield a 1 dimensional output which is followed by a sigmoid function.

III. DeepTMO-R With/Without Skip Connections

We additionally explored the cGAN based network from [4] for our tone-mapping task and name it as DeepTMO-R. The generator architecture of DeepTMO-R is shown in 2. We use the same discriminator as of DeepTMO single-scale. DeepTMO-R design is altered by adding skip-connections between each layer i and layer $n-i$, n being the total number of encoder-decoder layers and called as DeepTMO-S, which as a result concatenates all the channels at layer i with layer $n-i$. Various past HDR reconstruction methods, have used skip connections [5] for generating HDR scenes from single exposure [6] or multi-exposure [7] LDR images. The basic idea had been that since, both LDR and HDR scenes are different renderings of the same underlying structure, at a particular scale, their structures are also more or less aligned. Hence, it is possible to effectively transmit low-level details from input to output scenes,circumventing the bottleneck of the encoder-decoder architecture.

A. Experimentation

We performed the training and testing for both, DeepTMO-R and DeepTMO-S in a similar fashion to that of DeepTMO and also on the high-resolution images of size $1024x2048$.

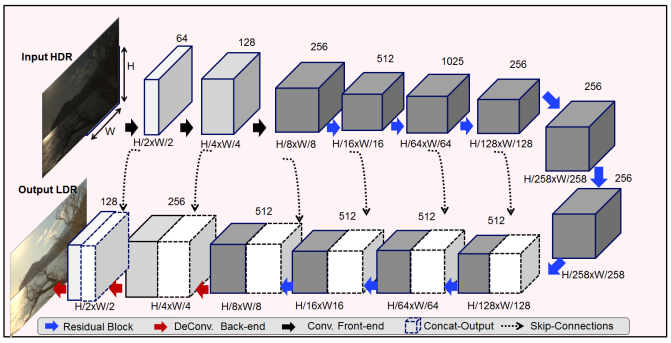
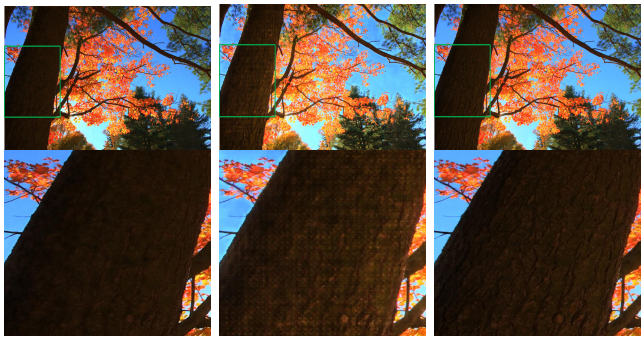


Fig. 2: DeepTMO-R and DeepTMO-S generator architecture.



DeepTMO-R DeepTMO-S DeepTMO

Fig. 3: While the DeepTMO-R simply results in blurred outputs in the bark of tree, the DeepTMO-S tries to refine them but is faced by *checkerboard* artifacts [8], [9]. The DeepTMO provides best results amongst the three methods while preserving the fine details, contrast and sharpness in the image.

In Fig. 3, we show an simple case of a daytime natural scene where the three architecture provide results with some prominent visible effects. From the cropped insets, we see that DeepTMO-R results in blurry effect on the textured bark of the tree, similar to previous example. DeepTMO-S on the other hand, doesn't produce any blurriness, but instead, we notice pronounced repetitive checkerboard artifacts. Such artifacts have been recently discussed in deep-learning based image rendering problems [8], [9] and are mainly caused due to no direct relationship among intermediate feature maps generated in de-convolutional layers. Nevertheless, it is still an open problem. DeepTMO, on the other hand, gives us sharper and checkerboard free images while preserving the fine-details too.

IV. DeepTMO (Multi-Scale) Architecture

A. Multi-Scale Generator Architecture

$G_1^{(F)}$ here consists of 5 convolution layers, with the number of output channels from first till last being 64,128,256,512,1024 respectively. $G_1^{(R)}$ has 9 residual blocks each having 1024 as the number of output channel. For $G_1^{(B)}$, we have 5 transposed convolution layers with output channels 512,256,128,64 and 1. All the component layers have similar nomenclature as used in the component layers of G, including the first layer of $G_1^{(F)}$ and the last layer in $G_1^{(B)}$.

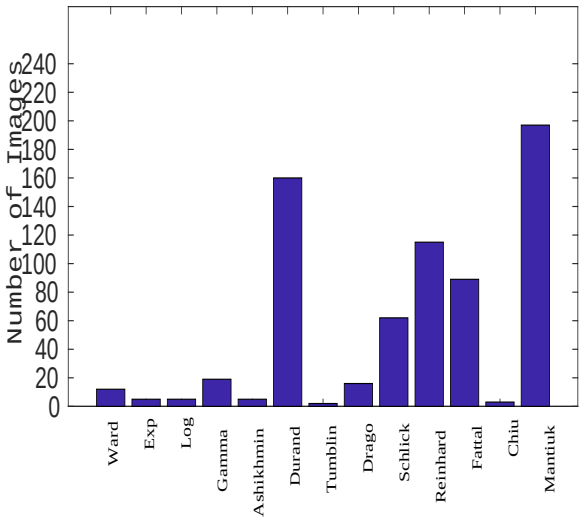


Fig. 4: Distribution of best tone mapped output on training dataset of 700 Images.

$G_2^{(F)}$ has two consecutive convolution layers, with output channels 32 and 64 respectively. The last feature map of $G_1^{(B)}$ (with 64 output channels) is then element-wise summed with the output feature map of $G_2^{(F)}$, to provide corresponding input to the residual block $G_2^{(R)}$. $G_2^{(R)}$ consists of 3 residual blocks each with 64 output channels. Following this we have two deconvolution layer in $G_2^{(B)}$ with 32 and 3 output channels respectively. Again the structure of all the component-wise layers is similar to G.

V. Training Dataset

We provide the training distribution of target tone-mapped images on training dataset in Fig. 4 considering 13 TMOs. Note that while training we have used several data augmentation techniques. These target tone-mapped scenes have been selected using the default parametric settings.

VI. Dataset Source

Dataset is collected from the following sources: [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]

References

[1] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[2] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[3] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models."

[4] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[5] 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*. Springer, 2015, pp. 234–241.

[6] G. Eilertsen, R. Wanat, R. K. Mantiuk, and J. Unger, "Evaluation of Tone Mapping Operators for HDR-Video," *Computer Graphics Forum*, 2013.

[7] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017)*, vol. 36, no. 6, nov 2017.

[8] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[9] H. Gao, H. Yuan, Z. Wang, and S. Ji, "Pixel deconvolutional networks," *arXiv preprint arXiv:1705.06820*, 2017.

[10] F. Xiao, J. M. DiCarlo, P. B. Catrysse, and B. A. Wandell, "High dynamic range imaging of natural scenes," in *In Tenth Color Imaging Conference: Color Science, Systems, and Applications*, 2002.

[11] W. J. Adams, J. H. Elder, E. W. Graf, J. Leyland, A. J. Lugtigheid, and A. Muryy, "The southampton-york natural scenes (syns) dataset: Statistics of surface attitude," 2016.

[12] Pfstools. (2007) Pfstools image database. [Online]. Available: http://pfstools.sourceforge.net/hdr/gallery.html

[13] M. Database. (2004) Mpi hdr image database. [Online]. Available: http://resources.mpi-inf.mpg.de/hdr/gallery.html

[14] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997.

[15] ETHyma. (2015) Ethyma database for high dynamic range images. [Online]. Available: http://ivc.univ-nantes.fr/en/databases/ETHyma/

[16] G. Krawczyk. (2006) Mpi hdr video database. [Online]. Available: http://resources.mpi-inf.mpg.de/hdr/video/

[17] A. A. Rad, L. Meylan, P. Vandewalle, and S. Süsstrunk, "Multidimensional image enhancement from a set of unregistered and differently exposed images," in *Computational Imaging V, San Jose, CA, USA, January 29-31, 2007*, 2007. [Online]. Available: http://lcavwww.epfl.ch/alumni/meylan/

[18] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays," 2014. [Online]. Available: http://spiedigitallibrary.org

[19] M. Azimi, A. Banitalebi-Dehkordi, Y. Dong, M. T. Pourazad, and P. Nasiopoulos, "Evaluating the performance of existing full-reference quality metrics on high dynamic range (hdr) video content," in *International Conference on Multimedia Signal Processing (ICMSP)*, 2014.

[20] A. Rana, G. Valenzise, and F. Dufaux, "Evaluation of feature detection in HDR based imaging under changes in illumination conditions," in *IEEE International Symposium on Multimedia, ISM 2015, Miami,USA, December, 2015*, 2015, pp. 289–294.