

Unpaired Learning for High Dynamic Range Image Tone Mapping

Yael Vinker Inbar Huberman-Spiegelglas Raanan Fattal

{yael.vinker, inbar.huberman1, raanan.fattal}@mail.huji.ac.il

School of Computer Science and Engineering
 The Hebrew University of Jerusalem, Israel

Abstract

High dynamic range (HDR) photography is becoming increasingly popular and available by DSLR and mobile-phone cameras. While deep neural networks (DNN) have greatly impacted other domains of image manipulation, their use for HDR tone-mapping is limited due to the lack of a definite notion of ground-truth solution, which is needed for producing training data.

In this paper we describe a new tone-mapping approach guided by the distinct goal of producing low dynamic range (LDR) renditions that best reproduce the visual characteristics of native LDR images. This goal enables the use of an unpaired adversarial training based on unrelated sets of HDR and LDR images, both of which are widely available and easy to acquire.

In order to achieve an effective training under this minimal requirements, we introduce the following new steps and components: (i) a range-normalizing pre-process which estimates and applies a different level of curve-based compression, (ii) a loss that preserves the input content while allowing the network to achieve its goal, and (iii) the use of a more concise discriminator network, designed to promote the reproduction of low-level attributes native LDR possess.

Evaluation of the resulting network demonstrates its ability to produce photo-realistic artifact-free tone-mapped images, and state-of-the-art performance on different image fidelity indices and visual distances.

1. Introduction

High dynamic range (HDR) photography gained a considerable popularity in the last decades among both professional and non-professional photographers. HDR capabilities are available in many DSLR cameras as well as in mainstream mobile smartphones, capable of providing 10- and 12-bits of color depth. Printing or displaying these images on conventional low dynamic range (LDR) devices require a tone-mapping step for reducing their dynamic range. The latter revealed itself as a non-trivial task and drew a consid-



Figure 1. Tone-mapped HDR images produced by our method. The four exposures on the left of each image portray the very high dynamic range in the original scenes.

erable research effort.

Unlike many image processing and restoration tasks, there is no ground-truth “solution” for tone-mapping HDR images, and the different methods developed over the years aim for different goals. Early approaches use global tone-reproduction curves (TRC) that make a better use of the output dynamic range than linear bracketing [5, 24, 47]. These curves avoid over- and under-exposed pixels, but their contrastive nature leads to a severe reduction in local contrasts.

Consequently, more modern approaches focus their goal on preserving, or enhancing, the local contrasts using detail separation and enhancement techniques [6, 8, 9, 33, 35, 39, 48]. These methods produce highly-detailed images, but applying high levels of compression remains a challenge in terms of avoiding edge-related artifacts or achieving balanced levels of contrast and an overall photo-realistic appearance.

Deep neural networks (DNNs) have greatly impacted various image processing tasks, such as super resolution [20], and deblurring [43], by using large training sets containing ground-truth examples. In the absence of ground-truth data, the prevalent approach in HDR tone-mapping is using existing tone-mapping operators (TMOs) to produce a panel of image labels, and narrowing it down using an image quality index to obtain the final training examples [34, 38].

The trained network is expected to reproduce the best *available* result on each image, but not surpass the performance of its underlining TMO algorithms. Moreover, the quality indices used reward for fulfilling a small number of regularities and hence they bias the training towards overfitting these attributes. Some approaches incorporate manual supervision, but they still rely on the indices in their final selection [55], or make a subjective decision by picking a single annotator [30].

In this paper we describe a new DNN-based TMO which is trained to produce images that bear the visual characteristics found in native LDR images. By formulating this distinct goal as an adversarial training, we replace the need for obtaining paired training examples with the plenitude of available high-quality LDR images.

Unsupervised adversarial training is a rather delicate process, prone to various instabilities. In order to achieve an effective and successful training, our method uses the following several new steps and components: (i) Before feeding the input HDR image to the network, we map its luminance through a range compression curve that reduces its variance and fixes its range. In order to train and apply our method on images from arbitrary sources, we use an adaptive compression level which we estimate on the basis of each input image.

Moreover, (ii) we incorporate a structure-preservation loss that penalizes for changes beyond local adjustments in brightness and contrast. This term ensures the input image content is preserved and no mode-collapsing occurs.

Finally, (iii) we describe the use of an ensemble of relatively shallow discriminator networks in order to better match the low-level attributes of native LDR images, and suppress the edge-related artifacts that plague existing TMOs. We demonstrate our method’s ability to efficiently produce naturally-looking and artifact-free LDR renditions of highly challenging HDR scenes. A quantitative evaluation over benchmark HDR images reveals its superior performance over established image quality metrics and visual distances.

2. Related Work

We briefly review here the different tone-mapping approaches developed over the years, with an emphasis on recent DNN-based methods.

Tone-Mapping Algorithms. Tone mapping algorithms are typically divided into global and local operators. Global operators map pixels based on their tone and apply the same curve across the image. Notable examples use linear scaling [52], non-linear tone-curve [47] and logarithmic tone-curve [21, 5]. The class of S-shaped curves is associated with responses in the human visual system (HVS) [31] and is used in [36, 41]. Computational efficiency and lack of over- and under-exposed pixels are the key advantages

of global operators. However, their non-local reasoning leads to a sub-optimal local treatment characterized by a loss of contrast. Local operators address this shortcoming by adapting their operation based on the local content of the image. Some works [23, 35, 39] derive their operation from the local adaptation mechanism of the HVS. However, the use of spatially-invariant filtering to analyze the image leads to unwanted halo effects around edges. To minimize these effects, different types of edge-aware filters were suggested, such as the anisotropic diffusion [48], robust averaging [4], bilateral filter [6], weighted least-squares [8], local-Laplacian pyramid [33], and a multiscale decomposition [12]. Another recent work by Liang *et al.* [25] uses a l_0 and l_1 sparsity to decompose image details. An additional approach tackles the problem by operating at the fine scale of image gradients [9, 44], and over a low-passed measure of contrast [28]. This research effort led to a significant improvement in the ability to tone-map HDR images while preserving their local contrasts and minimizing edge-related artifacts. Nevertheless, as noted above, high compression levels may still result in artifacts or compromised realism due to contrast imbalances.

DNN-based Tone-Mapping. A number of approaches were suggested for utilizing DNNs for HDR tone-mapping, despite the lack of ground-truth training data. Hou *et al.* [17] train a network to reproduce the input HDR image from its log-transformed luminance map, using a VGG-perceptual loss [20]. This method operates under the assumption that the compressive effect of the logarithm will persist under this norm and given the limited network capacity used. Their network must be trained specifically for each input image. Gharbi *et al.* [11] use a network architecture inspired by the bilateral grid to perform various image enhancement operations given input and output example pairs. The emphasis of this approach is on reproducing the enhancements in real-time.

Several recent methods [3, 30, 32, 34, 38, 55] train their tone-mapping network using conditional-GAN [18, 50], to obtain a more abstract metric of matching distributions. This conditional framework, however, requires pairs of input HDR and a corresponding tone-mapped LDR. These methods differ in the way they obtain their paired data and the additional similarity losses they use. Patel *et al.* [34], Rana *et al.* [38] and Cao *et al.* [3] construct their training pairs by applying a list of tone-mapping algorithms over each example HDR image and choosing the result with the highest tone mapped quality index (TMQI) [53] as its ground-truth label. Panetta *et al.* [32] train over a supervised low-light images dataset. These methods aid the convergence of their adversarial training using an additional reconstruction loss, Rana *et al.* and Cao *et al.* use the VGG-perception loss, Patel *et al.* use an L1 loss, and Panetta *et al.* use a combination of this loss with a gradient profile loss

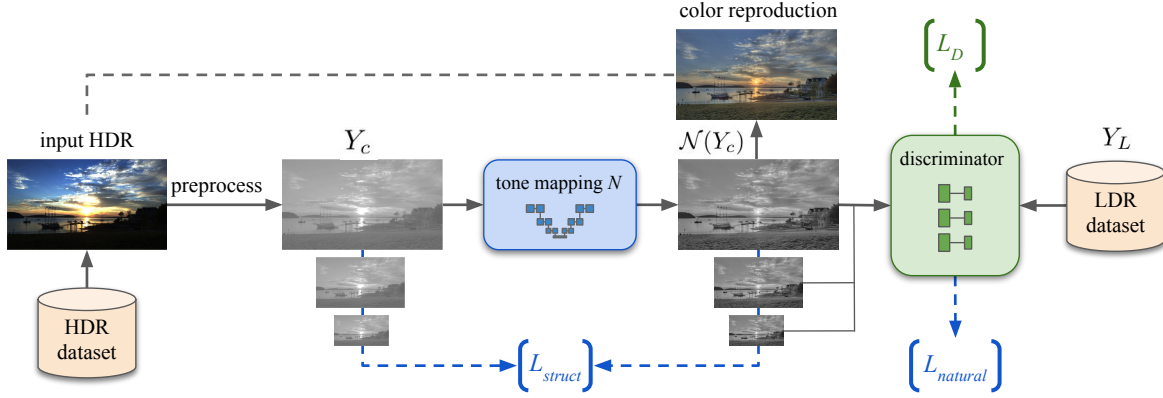


Figure 2. Tone-Mapping Pipeline. Given an input HDR image, we estimate the level of compression needed for mapping its luminance Y into a fixed-range map, Y_c . The latter is fed to the tone mapping network \mathcal{N} as well as used for defining the structure-preservation loss. The network is also trained to minimize an adversarial loss with respect to native LDR images. The final output image is produced from the resulting luma, $\mathcal{N}(Y_c)$, and the chromaticity of the input image. We mark in blue and green the components that participate in the training of the network \mathcal{N} and \mathcal{D} respectively.

for better considering edge information. Zhang *et al.* [55] incorporate a manual supervision by providing three photographers with a tone-mapping toolbox which they can use to obtain the most compelling result for each training image. The image with the highest TMQI score is then used as the training label. Rico *et al.* [30] use the MIT-Adobe 5K dataset [2] which was also produced by multiple (five) photographers that were allowed to apply different retouching operations to each training image. Rico *et al.* ended up using the tone-mapped results of only one of the experts as the training label, and they also employ a VGG-based reconstruction loss.

By contrast to these methods, we alleviate the need for paired training examples by training our network to reproduce the visual attributes of native LDR images.

Let us finally mention the line of work on low-light image enhancement methods. While these methods are not designed to cope with HDR images, they are also expected to manipulate (increase) the brightness of images. As these manipulations apply considerably weaker changes to the input image than HDR TMOs, global curves-based approaches are sufficient. Indeed, Guo and Li *et al.* [14] predict an image-specific curve mapping using deep learning. Wang *et al.* [49] use retouched images to train a network to deal with underexposed images, and Jiang *et al.* [19] train an unsupervised GAN for this purpose. Unlike our method, this work does not take any particular measure to allow it to apply significant levels of dynamic range compression.

3. Method

We begin with a brief overview of our new tone-mapping pipeline, which is summarized in Figure 2, and proceed with a detailed description of each of its steps.

Similarly to other TMOs [9, 41, 48], we perform the

tone-mapping over the luminance channel Y of the input HDR in YUV color-space.

We rely on the power of adversarial training to match the visual appearance of native LDR images. This type of training is however a delicate process, prone to various instabilities. To ensure stable training, at the first step of our pipeline, we bring the input luminance Y into a fixed range by mapping it through a TRC that applies the proper amount of compression. This amount is estimated from the discrepancy between Y 's histogram and a canonical LDR luminance distribution.

The resulting image is then fed into the tone mapping network, \mathcal{N} , which is trained to reproduce the visual attributes of native LDRs and remove the biases created by the TRC applied. This is achieved by an adversarial training with respect to a dataset of native LDR images. We augment this training process with a loss that preserves the structural integrity of the input image by restricting the action of \mathcal{N} to local changes in brightness and contrast. This loss also ensures no mode-collapsing occurs.

Finally, the output color image is recovered by the procedure used in [41, 48, 9].

Adaptive Curve-Based Compression. The luminance range of different HDR images can differ by orders of magnitude, which brings the need for applying different levels of compression. This variability in the input appears to challenge \mathcal{N} and undermine the convergence of its training. Consequently, as a pre-process step, we map the input luminances Y to a fixed range using a global curve mapping. The mapping ranges from a linear transformation to a severe log-like curve, depending on the level of compression required by the input Y . This family of TRCs is given by

$$Y_c(\mathbf{x}) = \log \left(\lambda \frac{Y(\mathbf{x})}{\max(Y)} + \varepsilon \right) / \log(\lambda + \varepsilon), \quad (1)$$



Figure 3. Operation of the Tone-Mapping Network. The figure shows an input HDR, a color rendition of Y_c before being inputted to the network, and the action of \mathcal{N} over Y_c . The latter reproduces a significant amount of contrasts which are missing in Y_c . The figure also compares the results obtained by two different discriminator networks. The deep DCGAN architecture leads to spurious brightened and darkened regions in the image, as indicated by the red frame. The use of an ensemble of shallower discriminator networks achieves a more consistent tone-mapping.

where $\max(Y)$ refers to the maximal value across the image pixels \mathbf{x} , the parameter λ sets the applied level of compression, and $\varepsilon = 0.05$. The resulting luminance values are restricted to $[0, 1]$ regardless of λ .

Logarithm curves are often used for tone-reproduction given their similarity to the compressive response to light that takes place in the human visual system, a.k.a. Weber-Fechner law [5]. We note however that unlike existing TRC-based methods we use an adaptive compression level, where different values of λ are used on different images. In Section 3.1 below we describe the way we determine this parameter based on the histogram of Y .

Natural Appearance. Global TRCs, as the ones applied in Eq. 1, are notorious for their inability to produce sufficient levels of local contrasts. To avoid this shortcoming as well as the question of what amount of contrasts should appear in the tone-mapped rendition, we adopt the goal of producing tone-mapped images that best resemble LDR images capturing native LDR scenes.

We leverage the remarkable ability of adversarial training to reproduce a distribution given example images. More specifically, we train a discriminator network \mathcal{D} to distinguish between $\mathcal{N}(Y_c)$ and a set of high-fidelity native LDR images. By training \mathcal{N} to fool \mathcal{D} , its output acquires the desired appearance. The significant effect of the resulting \mathcal{N} is demonstrated in Figure 3.

This unpaired training scheme alleviates the need to collect or produce example tone-mapped images. As noted earlier, unlike [34, 38] who rely on existing tone mapping algorithms, our training allows us to surpass their performance. Being unsupervised, our training spares the manual effort assumed in [30, 55], as well as scores better on objective image fidelity measures.

The architecture of the discriminator network plays a central role in the success of this approach. The multi-layered DCGAN architecture prototype [37] has the ability to correlate between features at multiple scales and discriminate based on high-level semantic content. Consequently, this architecture is often used for training adversarial gen-

erative networks. We, however, do not train \mathcal{N} to conceive new images from scratch, but specialize on removing biases in Y_c with respect to regular LDR images.

Since these differences relate to local contrasts and edge-related effects, we set the focus of \mathcal{D} to this level of image modelling by limiting its depth to two convolutional layers. In order to identify discrepancies at multiple spatial scales, we follow the approach in [50], and train an ensemble of discriminator networks, each one applied at a different image resolution, using the following loss

$$L_{\mathcal{D}} = \sum_{k \in \{0,1,2\}} \left(\mathbb{E}_{Y_L \sim LDR} [\mathcal{D}_k(\downarrow^k Y_L) - 1]^2 + \mathbb{E}_{Y \sim HDR} \left[\mathcal{D}_k \left(\downarrow^k \mathcal{N}(Y_c) \right) \right]^2 \right), \quad (2)$$

where Y_L denotes the luminance channel of the LDR training images, and Y_c is the compressed luminance computed from Y using Eq. 1. The \downarrow^k denotes a bicubic $\times 2^k$ image downscaling operator, and \mathcal{D}_k are the discriminator networks applied at the corresponding image scales. These networks have an identical architecture that we describe in Section 3.2 along with the architecture of \mathcal{N} . Finally, we use \mathcal{D}_k to improve the ability of \mathcal{N} to match the natural appearance of LDR images by training it to minimize

$$L_{natural} = \sum_{k \in \{0,1,2\}} \mathbb{E}_{Y \sim HDR} \left[\mathcal{D}_k \left(\downarrow^k \mathcal{N}(Y_c) \right) - 1 \right]^2. \quad (3)$$

Note that both $L_{\mathcal{D}}$ and $L_{natural}$ correspond to a least-squares GAN training losses [29]. Figure 2 summarizes the computational graph related to these adversarial losses.

Finally, Figure 3 shows the stability related artifacts produced by the use of a potent deep DCGAN-styled discriminator compared to the more consistent results obtained by our ensemble of shallower networks.

Structure Preservation. Mapping the luminance dynamic range into a narrower range inevitably changes the brightness of objects. This process can take place without changing the shape of the objects, in terms of the location

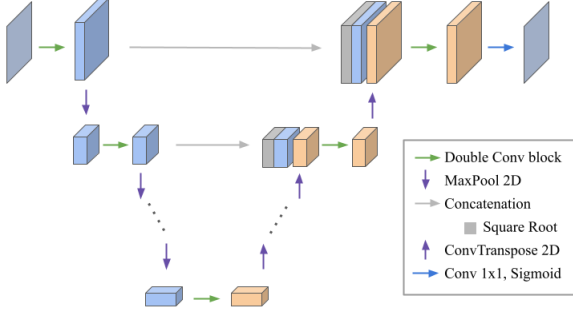


Figure 4. U-Net architecture that we use in our tone-mapping network \mathcal{N} . The skip connections at each layer pass the activations as well as their square-root as indicated by the gray boxes.

and orientation of their edges. However, the adversarial training in Eq. 3 does not require $\mathcal{N}(Y_c)$ to bear any similarity to the input Y , or Y_c . To avoid this pitfall as well as any mode-collapsing during training, we promote the co-occurrence of brightness-normalized variations in the input Y_c and the output $\mathcal{N}(Y_c)$ images. We derive this measure based on the Pearson correlation inside small image patches of two images I and J , by

$$\rho(I, J) = \frac{1}{n_p} \sum_{p_I, p_J} \frac{\text{cov}(p_I, p_J)}{\sigma(p_I)\sigma(p_J)}, \quad (4)$$

where p_I and p_J are all the 5-by-5 pixel patches in I and J respectively. The covariance, cov , and standard deviation, σ , are also computed inside these patches. This measure is normalized by the total number of patches used, n_p .

We use this measure to preserve the content and structural integrity of the input at multiple spatial scales by minimizing

$$L_{struct} = \sum_{k \in \{0,1,2\}} \rho(\downarrow^k Y_c, \downarrow^k \mathcal{N}(Y_c)), \quad (5)$$

which we use in combination with the adversarial loss in Eq. 3 to train \mathcal{N} .

We finally note that the structural similarity index (SSIM) [51] also analyzes the images in the basis of its patches. However, it is not invariant to changes in the patch brightness and contrast. Hence, this index is irrelevant for our purpose as $\mathcal{N}(Y_c)$ is expected to undergo major changes in both brightness and contrast.

Color Reproduction. At inference time we recover an output color image using the formula used in [9, 41, 48]. Specifically, each RGB color channel C_{in} of the input HDR image is mapped independently to produce the corresponding output channel by, $C_{out} = (C_{in}/Y)^s \mathcal{N}(Y_c)$. We use the default color saturation parameter $s = 0.5$.

3.1. Compression Level Estimation

The TRC we use in Eq. 1 allows us to apply different levels of compression, depending on the amount needed for

each image. We explain here the way we determine this level.

When collecting training images from the wild, or applying our TMO over images from arbitrary sources, the meta-data related to the scale of their luminance values may not be available. Hence, determining λ based on the maximal Y value can be unreliable. The option of using the dynamic range, computed by the ratio between the largest and smallest input luminances, may be unstable as the noise level regulating this estimate is unavailable in such circumstances.

To avoid these indeterminacies and train, and apply, our method over arbitrary images, we derive the following estimate for λ . Specifically, since λ controls the TRC in Eq. 1, we search for the value that brings Y_c closest to a canonical LDR image behaviour. As the luminance distribution of HDR images is considerably richer than the one of LDR images, we use these distributions as the descriptors for measuring this proximity.

Formally, we search for the λ that minimizes the following cross-entropy

$$\min_{\lambda} - \sum_l H_l(Y_c) \log(H_l(LDR)), \quad (6)$$

where the histogram of Y_c , denoted by $H(Y_c)$, is a function of λ , and $H(LDR)$ denotes the histogram of native LDR images. The latter was computed by averaging the histogram of 900 high-quality images from the DIV2k dataset [1]. We used 20 bins indexed by l . Finally, we solve the optimization in Eq. 6 using a stochastic search [45].

Finally, in Section 4.3 we report an ablation study demonstrating the contribution of this estimate compared to using the raw luminance values in the images we collected for training and testing our method.

3.2. Implementation Details

Tone-mapping Network Architecture. The network \mathcal{N} consists of a U-Net architecture [40] with four levels, each containing two sets of 3-by-3 convolution, ReLU, and max-pooling. The number of filters is doubled at every level, starting with 32 filters. We use $\times 2$ pooling factor and a convolution-transpose for unpooling. At the output layer, the ReLU operator is replaced by a sigmoid. This architecture is depicted in Figure 4.

The combination of linear layers (convolution) and ReLU operators span a space of piecewise affine transformations. In order to express a smoother luminance mapping, we concatenate the activations passing through the skip connections with their square-root, as indicated in Figure 4. This step doubles the activations dimension, but the convolution that follows reduces this dimension. As shown in the ablation study in Section 4.3, these compressive pointwise transformations improve the photo-realism of our tone-mapped images.

Discriminator Architecture. We use three separate discriminators which do not share their weights. The networks have an identical architecture consisting of four layers. The first two are 4-by-4 convolution layers, with 16 and 32 filters respectively, and use LeakyReLU activations. Both levels have $\times 2$ strided pooling. The third layer uses an unstrided 1-by-1 convolution, followed by an output fully-connected layer with a sigmoid activation.

Datasets. We train the network on HDR images taken from the HDR+ dataset [15], which contains indoor and outdoor scenes at different lighting conditions. We use 1000 images from this dataset for training and another 1000 images for testing. Each image is cropped and rescaled to provide two 256-by-256 images. The native LDR images, used for the adversarial loss in Eq. 2, were taken from the DIV2k dataset [1], which contains 1000 high-quality images. We randomly split this dataset into two halves for training and testing. Here again, we augment each set by cropping every image into two, and rescaling them to 256-by-256 pixels.

Training Details. We implemented our model using the Pytorch framework and trained it on a single GeForce GTX1080Ti GPU using an Adam optimizer with a learning rate of 10^{-4} for \mathcal{N} and $\times 1.5$ this rate for \mathcal{D} . We pre-trained \mathcal{D} to discriminate between Y_c and LDR images for 50 epochs, and then trained both networks over 300 epochs with a learning-rate decay factor of half every 50 epochs. The pre-training stage is meant to stabilize the GAN training. The tone-mapping network \mathcal{N} consists of about 4.5M parameters, and the discriminators a total of 31k parameters. Training these networks took approximately 3 and a half hours. **Running Times.** It takes 0.5 seconds for our trained tone-mapping operator to process a 1333-by-2000 pixels image using a GPU.

4. Results

We report the quantitative and qualitative evaluation of our tone-mapping network next to the state-of-the-art algorithms of Farbman *et al.* [8], Ferradans *et al.* [10], Gu *et al.* [12], Mai *et al.* [27], Shan *et al.* [42], Shibata *et al.* [44], Paris *et al.* [33], Liang *et al.* [25], Khan *et al.* [22] and Zhang *et al.* [56]. The default parameters published by the authors were used for testing these methods.

We also compare to DNN-based methods of Zhang *et al.* [55], Rana *et al.* [38], Panetta *et al.* [32] and Cao *et al.* [3], by running our method on the same sets of test images these methods used in their evaluation.

4.1. Quantitative Evaluation

In this comparison we used the tone-mapped image quality index (TMQI) [53] and the blind TMQI (BTMQI) [13], which are commonly used for assessing TMOs. These indices are evaluated over 105 benchmark HDR images from the HDR Photographic Survey dataset [7]. In cases where

TMO	TMQI	BTMQI
Ferradans <i>et al.</i> [10]	0.836	4.563
Mai <i>et al.</i> [27]	0.856	3.958
Shibata <i>et al.</i> [44]	0.87	3.578
Gu <i>et al.</i> [12]	0.871	3.878
Shan <i>et al.</i> [42]	0.874	3.625
Zhang <i>et al.</i> [56]*	0.88	3.76
Rana <i>et al.</i> [38]*	0.88	–
Farbman <i>et al.</i> [8]	0.886	3.602
Liang <i>et al.</i> [25]	0.887	3.691
Khan <i>et al.</i> [22]*	0.889	–
Ma <i>et al.</i> [26]	0.895	3.868
Cao <i>et al.</i> [3]*	0.9	–
Paris <i>et al.</i> [33]	0.906	2.988
Ours	0.919	2.89
Zhang <i>et al.</i> [55]**	0.874	3.519
Ours*	0.902	3.24
Panetta <i>et al.</i> [32]***	0.873	3.44
Ours*	0.883	3.041

Table 1. Mean TMQI scores (higher is better) and BTMQI (lower is better), both computed over the same 105 images from the HDR Survey dataset [7]. In (*) these scores were taken directly from the respective paper. The scores in (**) were computed over 304 images from the HDRI Haven dataset [54], which were used to evaluate [55], and in (***) over the 456 images which were used for the evaluation in [32].

the competitors’ original code was not available, we used their test images to evaluate our method using these scores.

The Fréchet Inception distance (FID) [16] is a benchmark score for measuring the distance between distributions of visual descriptors extracted from the Inception network [46]. We use this distance to measure the resemblance of our tone-mapped images and native LDR images. This distance, however, requires a large number of images ($\geq 10,000$) which are currently not available in the domain of HDR imaging. Instead, we use a variant of this distance, which we call *pixFID*, and can be reliably evaluated using less images. In this test we collected 1000 image from the HDR+ dataset [15], and another 1000 test LDR images from the DIV2k dataset [1]. At the Appendix at the Supplemental Material we describe this distance and show its consistency with the FID.

Tone-Mapped Image Quality Indices. Our network is trained to reproduce the visual appearance of native LDR images while preserving the structural content of the input. Both the TMQI and BTMQI reward for these qualities and indeed, as Table 1 reports, our method achieves the best scores, with a non-negligible margin in the TMQI, next to a list of top-scoring tone-mapping algorithms and the recent DNN-based approaches of Zhang *et al.* [55], Rana *et al.* [38], Panetta *et al.* [32] and Cao *et al.* [3]. Note that the latter DNN-based methods are not the next top scor-

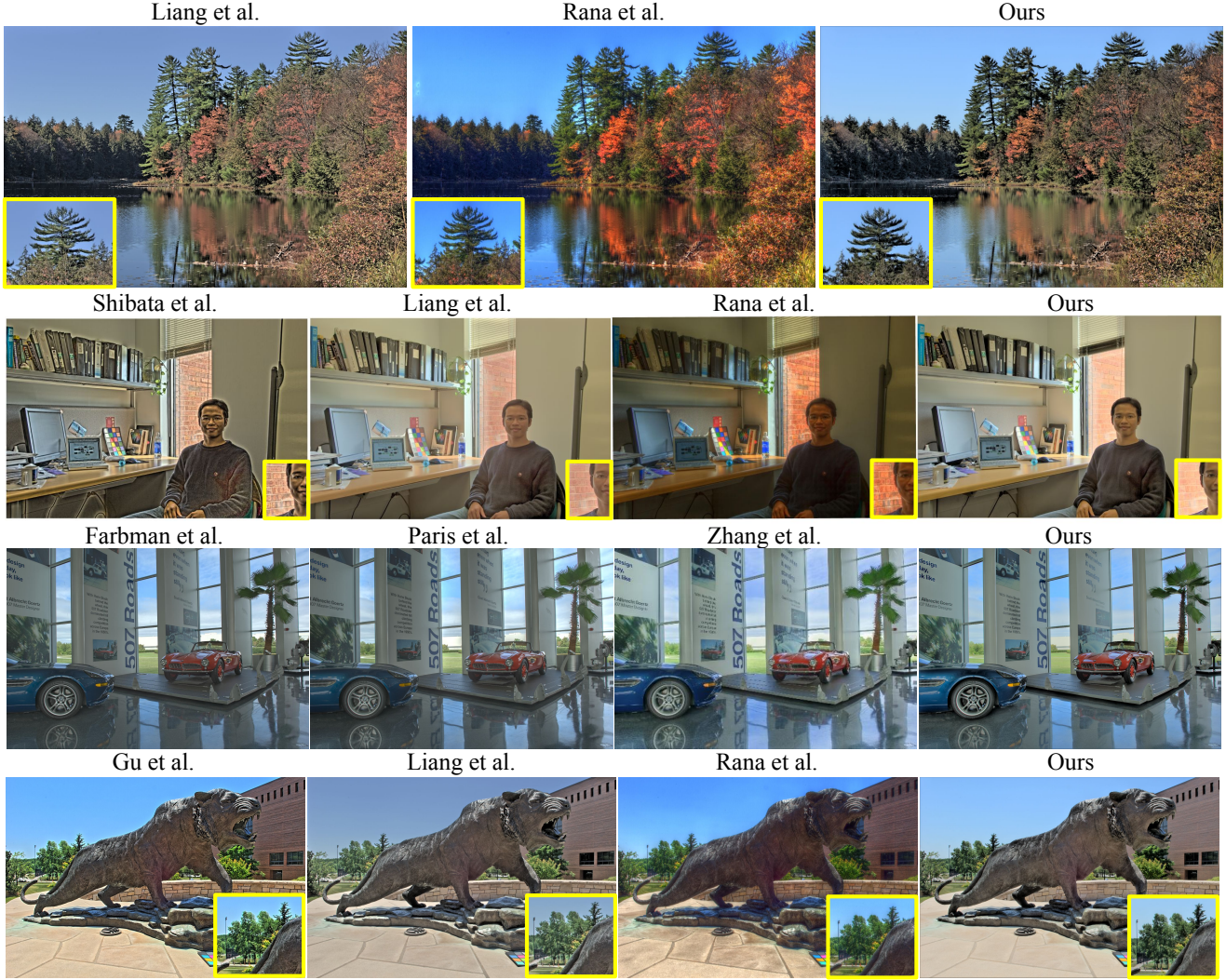


Figure 5. Tone-mapped images produced by different methods.

ing TMOs in the list. Hence, beyond pushing the state-of-the-art in HDR tone-mapping, our approach does so using DNNs which, as we note below, are available in a more efficient platform.

Fréchet Inception Distance. Table 2 reports the pixFID values obtained by the different methods where our network receives the lowest (best) value. This suggests our tone-mapped images bear the closest statistical resemblance to native LDR images. Note that the recent algorithm of Liang *et al.* is our closest competitor in this score. This score was not evaluated over the DNN-based methods due to the unavailability of their code.

4.2. Visual Evaluation

Figure 5 provides a side-by-side comparison of tone-mapped images produced by the methods achieving the top scores in the tests above.

The method of Shibata *et al.* appears to produce a proper compression level but also over-emphasized fine details and thin halos. The results of Liang *et al.*’s method show less edge-related artifacts and an adequate compression, but contain a fairly modest amount of local and global contrasts. The method of Rana *et al.* produces halo-free images which are rich in contrast, but also contain overly dark regions due to an insufficient amount of compression. Both the algorithms of Farbman *et al.* and Paris *et al.* produce artifacts-free images, but seem to apply an insufficient level of compression that leads to weaker contrasts. Zhang *et al.*’s results are better compressed and contain improved contrasts. Finally, the method of Gu *et al.* achieves a fair level of compression but suffers from over-emphasized fine details and minor halo effects. The tone-mapped images produced by our method do not seem to suffer from visual artifacts, and have an adequate amount of compression with no over-

TMO	pixFID
Shan <i>et al.</i> [42]	1.22
Shibata <i>et al.</i> [44]	1.20
Ferradans <i>et al.</i> [10]	1.165
Mai <i>et al.</i> [27]	1.141
Paris <i>et al.</i> [33]	1.107
Gu <i>et al.</i> [12]	1.105
Farbman <i>et al.</i> [8]	1.102
Liang <i>et al.</i> [25]	1.04
Ours	1.008

Table 2. Pixel Fréchet Inception Distance (lower is better) computed against native LDR images from the DIV2k dataset.

and under-exposed regions. The local and global contrasts appear fairly balanced, and consistent with ones found in contrast-rich native LDR images. Additional side-by-side comparison images can be found in the Supplemental Material of this paper.

User Study. Finally, in order to further assess the visual quality of our results, we conducted a user-study that follows the protocol used in [25]. Specifically, it consists of eight participants who were asked to rate between 1 to 8 images produced by the methods in [25, 33, 38, 12, 8] as well as ours. The images were displayed next to each other (at a random order). The participants were naive as to the purpose of the test and were asked to score the images based on how realistic and unprocessed they appear.

Our method obtained the highest mean score 6.43 with std 1.6, while the next top scoring method is [25] with mean score of 5.78 with standard-deviation of 2.04, and then [33] (4.39, 1.99), [8] (4.70, 1.98), [12] (2.43, 1.95) and [38] (2.14, 1.28). A plot summarizing these results can be found at the Supplemental Material.

4.3. Ablation Study

In order to demonstrate the contribution of different components described in Section 3, we evaluated the TMQI and FID scores obtained by switching them off. For reference, we remind that our method achieves pixFID of 1.008 and TMQI of 0.919.

Adaptive Compression Level. TMOs that use a TRC determine the amount of compression in one of two ways: (i) operate on the raw luminance values of the input image, i.e., omitting the division by $\max(Y)$ in Eq. 1, and allow their range to affect the compression [9, 6, 8], or (ii) as in [21] the input luminance is normalized to a fixed range, i.e., using a fixed amount of compression $\lambda = 5000$ in Eq. 1.

In order to successfully cope with images of different dynamic range our method adapts the compression level it applies. Moreover, it estimates this level from the input histogram and avoids inconsistencies in the luminance scales of images curated from different sources. Using a fixed $\lambda = 1000$ lowered the TMQI score considerably to 0.89,

and omitting the luminance normalization undermined both the pixFID to 1.096 and TMQI to 0.75. We provide visual demonstrations at the Supplemental Material.

Square-Root Transformations. As described above, we augment our U-Net architecture with \sqrt{x} transformations to better span smooth tone-mapping operators. By omitting these transformations the pixFID score increases to 1.219 and the TMQI decreases to 0.869. The visual impact of these transformations is substantial as we demonstrate in the Supplemental Material.

Discriminator Ensemble. We also evaluated the importance of using an ensemble of shallow discriminator networks applied at multiple image resolutions to two alternatives: (i) the use of a single shallow discriminator at the finest image resolution, and (ii) the use of a single deeper DCGAN architecture [37]. The first option results in inferior scores of pixFID 1.096 and TMQI 0.912, as well as a noticeable compromised ability to restore contrasts at coarse scales of the images as shown in the Supplemental Material. Figure 3 shows the artifacts resulting from the instabilities involved in training a deep DCGAN discriminator network, despite the fact that this option did not have a significant impact on the visual scores, namely, pixFID of 1.043 and TMQI of 0.92.

Adversarial Loss. Without the GAN loss, the network is trained only to minimize the structural loss and results in an identity mapping that outputs the input tone-mapped images unchanged, the TMQI and pixFID scores under these settings were 0.8 and 1.131 correspondingly.

Structural Preservation. Finally, we report that the removal of our structural-preservation loss did not allow our training to converge.

5. Conclusions

We presented a new training approach that, unlike existing DNN-based methods, does not require paired training examples by setting the goal of producing naturally looking LDR renditions. In order to allow this minimal setting to succeed, we presented a range-normalizing pre-process which estimates and applies a different level of curve-based compression, a loss that restricts the action of the network to local changes in brightness and contrast, and a more concise discriminator network, designed to promote the reproduction of low-level attributes native LDR poses. We evaluated our method on several image fidelity metrics and reported its superior performance. A visual inspection shows its ability to produce naturally-looking tone-mapping with balanced levels of contrast.

References

- [1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The*

- IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 5, 6
- [2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 3
 - [3] X. Cao, K. Lai, S.N. Yanushkevich, and M. R. Smith. Adversarial and adaptive tone mapping operator for high dynamic range images. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1814–1821, 2020. 2, 6
 - [4] Jeffrey DiCarlo and Brian Wandell. Rendering high dynamic range images. *Proceedings of SPIE - The International Society for Optical Engineering*, 3965:392–401, 05 2000. 2
 - [5] Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. *Comput. Graph. Forum*, 22:419–426, 2003. 1, 2, 4
 - [6] Frédo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, page 257–266, 2002. 1, 2, 8
 - [7] M.D. Fairchild. *The HDR photographic survey*, pages 233–238. 01 2007. 6
 - [8] Zeev Farbman, Raanan Fattal, Dani Lischinski, and Richard Szeliski. Edge-preserving decompositions for multi-scale tone and detail manipulation. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2008)*, 27(3):to appear, Aug. 2008. 1, 2, 6, 8
 - [9] Raanan Fattal, Dani Lischinski, and Michael Werman. Gradient domain high dynamic range compression. *ACM Trans. Graph.*, 21:249–256, 2002. 1, 2, 3, 5, 8
 - [10] Sira Ferradans, Marcelo Bertalmio, Edoardo Provenzi, and Vincent Caselles. An analysis of visual adaptation and contrast perception for tone mapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2002–2012, 2011. 6, 8
 - [11] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):118, 2017. 2
 - [12] B. Gu, W. Li, M. Zhu, and M. Wang. Local edge-preserving multiscale decomposition for high dynamic range image tone mapping. *IEEE Transactions on Image Processing*, 22(1):70–79, 2013. 2, 6, 8
 - [13] Ke Gu, Shiqi Wang, Guangtao Zhai, Siwei Ma, Xiaokang Yang, Weisi Lin, Wenjun Zhang, and Wen Gao. Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure. *IEEE Transactions on Multimedia*, 18(3):432–443, Mar. 2016. 6
 - [14] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3
 - [15] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 6
 - [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017. 6
 - [17] Xianxu Hou, Jiang Duan, and Guoping Qiu. Deep feature consistent deep image transformations: Downscaling, decolorization and HDR tone mapping. *CoRR*, abs/1707.09482, 2017. 2
 - [18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2
 - [19] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlighten: Deep light enhancement without paired supervision. *IEEE Trans. Image Process.*, 30:2340–2349, 2021. 3
 - [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 1, 2
 - [21] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017)*, 36(4), 2017. 2, 8
 - [22] Ishtiaq Rasool Khan, Wajid Aziz, and Seong-O. Shim. Tone-mapping using perceptual-quantizer and image histogram. *IEEE Access*, 8:31350–31358, 2020. 6
 - [23] Edwin Land and John McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61:1–11, 02 1971. 2
 - [24] Gregory Ward Larson, Holly Rushmeier, and Christine Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, pages 291–306, 1997. 1
 - [25] Zhetong Liang, Jun Xu, David Zhang, Zisheng Cao, and Lei Zhang. A hybrid l1-l0 layer decomposition model for tone mapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6, 8
 - [26] Kede Ma, Hojatollah Yeganeh, Kai Zeng, and Zhou Wang. High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Transactions on Image Processing*, 24(10):3086–3097, 2015. 6
 - [27] Zicong Mai, Hassan Mansour, Rafal Mantiuk, Panos Nasiopoulos, Rabab Ward, and Wolfgang Heidrich. Optimizing a tone curve for backward-compatible high dynamic range image and video compression. *IEEE transactions on image processing*, 20(6):1558–1571, 2010. 6, 8
 - [28] Rafal Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. Appl. Percept.*, 3(3):286–308, 2006. 2

- [29] Xudong Mao, Qing Li, Haoran Xie, Raymond Lau, Wang Zhen, and Stephen Smolley. Least squares generative adversarial networks. pages 2813–2821, 10 2017. **4**
- [30] Rico Montulet and A. Briassouli. Deep learning for robust end-to-end tone mapping. In *BMVC*, 2019. **2, 3, 4**
- [31] KI Naka and WA Rushton. S-potentials from colour units in the retina of fish (cyprinidae). *The Journal of physiology*, 185(3):536–555, August 1966. **2**
- [32] Karen Panetta, Landry Kezebou, Victor Oludare, Sos Agaian, and Zehua Xia. Tmo-net: A parameter-free tone mapping operator using generative adversarial network, and performance benchmarking on large scale hdr dataset. *IEEE Access*, 9:39500–39517, 2021. **2, 6**
- [33] Sylvain Paris, Samuel W. Hasinoff, and Jan Kautz. Local laplacian filters: Edge-aware image processing with a laplacian pyramid. *ACM Trans. Graph.*, 30(4), 2011. **1, 2, 6, 8**
- [34] Vaibhav Patel, Purvik Shah, and Shanmuganathan Raman. A generative adversarial network for tone mapping hdr images. *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics. NCVPRIPG*, pages 220–231, 2018. **1, 2, 4**
- [35] Sumanta N. Pattanaik, James A. Ferwerda, Mark D. Fairchild, and Donald P. Greenberg. A multiscale model of adaptation and spatial vision for realistic image display. *SIGGRAPH '98*, page 287–298. Association for Computing Machinery, 1998. **1, 2**
- [36] Sumanta N Pattanaik, Jack Tumblin, Hector Yee, and Donald P Greenberg. Time-dependent visual adaptation for fast realistic image display. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 47–54, 2000. **2**
- [37] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. **4, 8**
- [38] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic. Deep tone mapping operator for high dynamic range images. *IEEE Transactions on Image Processing*, 29:1285–1298, 2020. **1, 2, 4, 6, 8**
- [39] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Trans. Graph.*, 21(3):267–276, 2002. **1, 2**
- [40] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]). **5**
- [41] Christophe Schlick. Quantization techniques for visualization of high dynamic range pictures. pages 7–20. Springer-Verlag, 1994. **2, 3, 5**
- [42] Qi Shan, Jiaya Jia, and Michael S Brown. Globally optimized linear windowed tone mapping. *IEEE transactions on visualization and computer graphics*, 16(4):663–675, 2009. **6, 8**
- [43] Ziyi Shen, Wei Sheng Lai, Tingfa Xu, Jan Kautz, and Ming Hsuan Yang. Deep semantic face deblurring. In *Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 8260–8269. IEEE Computer Society, Dec. 2018. **1**
- [44] Takashi Shibata, Tanaka Masayuki, and Masatoshi Okutomi. Gradient-domain image reconstruction framework with intensity-range and base-structure constraints. pages 2745–2753, 06 2016. **2, 6, 8**
- [45] Rainer Storn and Kenneth Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. of Global Optimization*, 11(4):341–359, Dec. 1997. **5**
- [46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. **6**
- [47] Jack Tumblin and Holly Rushmeier. Tone reproduction for realistic images. *Computer Graphics and Applications, IEEE*, 13:42–48, 1993. **1, 2**
- [48] Jack Tumblin and Greg Turk. Lcis: A boundary hierarchy for detail-preserving contrast reduction. *SIGGRAPH '99*, page 83–90. ACM Press/Addison-Wesley Publishing Co., 1999. **1, 2, 3, 5**
- [49] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. **3**
- [50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. **2, 4**
- [51] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, Apr. 2004. **5**
- [52] Greg Ward. *A Contrast-Based Scalefactor for Luminance Display*, page 415–421. Academic Press Professional, Inc., USA, 1994. **2**
- [53] H. Yeganeh and Z. Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing*, 22(2):657–667, 2013. **2, 6**
- [54] Greg Zaal. Hdri haven dataset. <https://hdrihaven.com/>. **6**
- [55] N. Zhang, C. Wang, Y. Zhao, and R. Wang. Deep tone mapping network in hsv color space. In *2019 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2019. **2, 3, 4, 6**
- [56] Xian-Shi Zhang, Kai fu Yang, Jun Zhou, and Yong-Jie Li. Retina inspired tone mapping method for high dynamic range images. *Opt. Express*, 28(5):5953–5964, Mar 2020. **6**