





Learning a self-supervised tone mapping operator via feature contrast masking loss

C. Wang¹  B. Chen¹  HP. Seidel¹ K. Myszkowski¹  and A. Serrano² 

¹Max-Planck-Institut für Informatik, Germany

²Universidad de Zaragoza, I3A, Spain

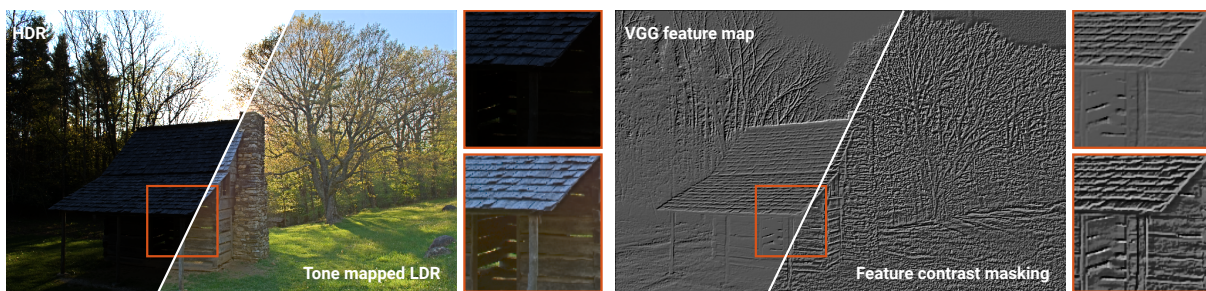


Figure 1: Our self-supervised tone mapping operator is directly guided by the input HDR image and a novel feature contrast masking loss that takes into account masking effects present in the Human Visual System. We minimize the difference between the HDR image and the tone mapped image in feature space, after applying our contrast masking model. Left: The input HDR image and our tone mapped result. Right: VGG feature map (1st layer, 12th channel) and our corresponding feature contrast masking response that effectively enhances low contrast image details while compressing high contrasts.

Abstract

High Dynamic Range (HDR) content is becoming ubiquitous due to the rapid development of capture technologies. Nevertheless, the dynamic range of common display devices is still limited, therefore tone mapping (TM) remains a key challenge for image visualization. Recent work has demonstrated that neural networks can achieve remarkable performance in this task when compared to traditional methods, however, the quality of the results of these learning-based methods is limited by the training data. Most existing works use as training set a curated selection of best-performing results from existing traditional tone mapping operators (often guided by a quality metric), therefore, the quality of newly generated results is fundamentally limited by the performance of such operators. This quality might be even further limited by the pool of HDR content that is used for training. In this work we propose a learning-based self-supervised tone mapping operator that is trained at test time specifically for each HDR image and does not need any data labeling. The key novelty of our approach is a carefully designed loss function built upon fundamental knowledge on contrast perception that allows for directly comparing the content in the HDR and tone mapped images. We achieve this goal by reformulating classic VGG feature maps into feature contrast maps that normalize local feature differences by their average magnitude in a local neighborhood, allowing our loss to account for contrast masking effects. We perform extensive ablation studies and exploration of parameters and demonstrate that our solution outperforms existing approaches with a single set of fixed parameters, as confirmed by both objective and subjective metrics.

CCS Concepts

• **Computing methodologies** → Computational photography; Neural networks; Image processing;

1. Introduction

High Dynamic Range (HDR) images can reproduce real-world appearance by encoding wide luminance ranges. With the fast-

paced developments in capturing devices, access to HDR image and video is becoming commonplace. Nevertheless, the majority of widespread displays are still limited in the dynamic range they can

reproduce, preventing direct reproduction of HDR content. Therefore, the use of tone mapping techniques is yet needed in order to adapt such content to current display capabilities.

Different tone mapping techniques have been developed for decades [RHD*10, BADC17], however the performance of even the most prominent techniques strongly depends on the HDR image content and specific parameter settings [ZWZW19, GJ21, PKO*21]. Subjective evaluations of these different techniques indicate that both specific algorithms as well as default parameter settings, as often proposed by their respective authors, do not generalize well across scenes [LCTS05, YBMS05, ČWNA08]. In these cases, manual selection of a suitable algorithm and experience in fine-tuning its parameters might be required for high-quality results. This hinders smooth adoption of HDR technology and is the key obstacle in developing machine learning solutions, as we discuss next. Recently, an increasing number of learning-based tone mapping methods have been proposed that show huge potential in terms of generality and quality with respect to their traditional algorithm-based predecessors. However, there are still some shortcomings. First, most learning-based models regard tone mapping as an image restoration task and optimize the network in a fully supervised manner, which requires high-quality paired HDR and tone mapped training data. Given a set of HDR scenes as training set, the Tone Mapping Quality Index (TMQI) [YW12] is typically used to select the best tone mapped image for each scene from a preexisting set of tone mapped results, limiting the quality of newly generated results to that of preexisting methods [ZWZW19, RSV*19, PKO*21]. And second, since such methods treat tone mapping as an image restoration task instead of an information reduction process, they usually involve large-scale networks. Motivated by these observations, in this work we seek an alternative solution that allows for reducing the network size and optimizes information reduction for a given HDR image content.

Since image contrast is arguably one of the key cues in the Human Visual System (HVS) while seeing and interpreting images [Pal99], we aim at reproducing perceived contrast in HDR scenes while also ensuring structural fidelity by reproducing visible image details. To this end, we propose a simple image-specific, self-supervised tone mapping network that is trained at test time and does not require any data labeling. The only training data is the input HDR image, and the key novelty in our approach is the loss function that directly compares the content in the HDR and tone mapped images. Since the compared signals present different luminance and contrast ranges, a direct computation of the loss in the feature space, as in e.g., perceptual VGG loss [SZ14, JAFF16], leads to sub-optimal results as we demonstrate in Sec. 4. To mitigate this problem, we first propose an adaptive μ -law compression that accounts for the scene brightness and brings HDR image histograms closer to those of low dynamic range images (Fig. 4). Then, motivated by contrast perception literature [LF80, Fo194, WS97, DZLL00], we introduce in our loss function a non-linearity considering both the HVS response for stronger contrasts and visual neighborhood masking, and we model it in the network's feature space. To this end, we first formulate a local contrast measure in the feature space, normalizing local feature differences by their average magnitude in a local neighborhood. This also allows for further abstraction from the magnitude difference

between HDR and LDR signals. Then, we introduce a compressive non-linearity as a function of feature contrast magnitude for the HDR signal, so that for higher magnitudes any departs in feature contrast are less strongly penalized in the loss function. This directly translates into compressing higher contrasts while preserving image details in the tone mapped images generated by our network. Finally, we also introduce feature contrast neighborhood masking, so that the loss function penalizes more weakly changes in feature contrast when similar features are present in the spatial neighborhood of the image.

We perform extensive ablation studies and exploration of parameters and demonstrate that our solution outperforms existing approaches for a single set of parameters, as confirmed by both objective and subjective metrics. Our contributions are as follows:

- We present a self-supervised tone mapping network that takes as an input multiple exposures derived from an HDR image and produces state-of-the-art tone mapping results, overcoming the need for annotated HDR-tone mapped image pairs for training.
- We propose a perceptually-inspired feature contrast masking loss function which is derived in feature space. This function effectively enables a direct supervision of the tone mapping process by the HDR image content so that perceived luminance contrast losses in the tone mapped results are minimized.
- We introduce an adaptive version of μ -law compression that brings HDR histograms closer to those typically observed in low dynamic range tone-mapped images. This facilitates self-supervision by the input HDR image.

Our code is available at: <https://self-supervisedTMO.mpi-inf.mpg.de>.

2. Related work

In this section we first summarize related research on contrast perception modeling, and then we discuss tone mapping techniques with emphasis on those that explicitly process contrast as well as more recent learning-based methods.

2.1. Contrast perception modeling

Visual sensitivity is affected by a number of key image dimensions such as luminance level [AJP92], spatial and temporal frequency [Rob66], or local image contrast [Wat89], as well as their interactions. In particular, changes in sensitivity as a function of local image contrast are generally termed masking effects. There are two main masking effects related to spatial contrast perception that have been widely studied and applied to computer graphics applications: contrast self-masking and visual contrast masking. Contrast self-masking [DZLL00] is characterized by a strong non-linearity that allows for stronger absolute changes for higher supra-threshold contrasts than for lower near-threshold contrasts before such changes become noticeable [KW96]. Visual masking (also called neighborhood masking) [LF80, Fo194, WS97, DZLL00] is a phenomenon in which sensitivity is locally reduced with increases in image local contrast [Dal92]. When contrasts with similar spatial frequencies are present in a close neighborhood, the thresholds for detection of lower contrasts and for change discrimination of higher contrast rise. To model this effect, the input image contrast

is decomposed into frequency bands using a filter bank such as a Laplacian pyramid [Pel90, MDC*21], a cortex transform [Dal92], wavelets [DZLL00], or discrete cosine transforms (DCTs) [Wat93], and then the visual masking is modeled for each frequency band.

Modeling contrast self-masking and neighborhood masking has been proven to be beneficial for several applications including image [DZLL00] or video [YJS*21] compression, image quality evaluation [Lub95, MDC*21], rendering [BM98, RPG99], and foveated rendering [TAKW*19]. Existing works apply such contrast perception models in the primary image contrast domain (or disparity domain [DER*10]), and employ predefined filter banks. Instead, we use a neural network and compute per-channel contrast signals over feature maps, where optimal filters are learned for the task at hand, and we formulate a novel loss function that models contrast self-masking and neighborhood masking in the feature contrast domain.

2.2. Tone mapping techniques

We first summarize traditional global and local tone mapping operators and then discuss learning-based tone mapping techniques.

2.2.1. Traditional methods

Traditional tone mapping methods can be roughly categorized into global methods, that apply the same transfer function to the whole image [DMAC03, JH93, LRP97, MDK08], and local methods, in which the applied function varies for each pixel by taking into account the influence of neighborhood pixels [FLW02, RSSF02, DD02, MMS06]. An interested reader may refer to extensive surveys that discuss in length these methods [RHD*10, BADC17, MMS15]. Some existing works have proposed a number of techniques for selecting or generating various tone mapping operators tailored to different target images or applications. For instance, Mantiuk and Seidel [MS08] propose to use a generic model to approximate both global and local methods, which is useful for backward-compatible HDR image compression and comparisons of existing tone mapping algorithms. With the goal of displaying HDR images on LDR displays efficiently using the GPU, Banterle et al. [BAS*16] segment input HDR images according to luminance zones and then select the best performing tone mapping operator for each zone. DeBattista [Deb18] uses genetic programming to automatically generate tone mapping operators suitable for different applications (such as visualization, feature mapping or compression). Other techniques, such as exposure fusion [MKVR07, RC09] directly composite a high-quality low dynamic range image by fusing a set of bracketed exposures, bypassing the reconstruction of an HDR image. In our pipeline we also leverage several input exposures to obtain the final tone mapped image, however, the tone mapping process is actively guided by our contrast masking loss in feature space.

Recently, a number of model-based tone mapping algorithms achieving higher performance have been proposed. We introduce them here and include them in our comparisons in Sec. 4.2.1. Shan et al. [SJB09] introduce a method that operates in overlapping windows over the image in which dynamic range compression is optimized globally for the image while window-based local constraints

are also satisfied. This allows for both small details and large image structures to be preserved. Ma et al. [MYZW15] propose an iterative algorithm that directly optimizes the resulting tone mapped image to maximize structural fidelity and statistical naturalness following a new tone mapping quality metric based on TMQI. Liang et al. [LXZ*18] design a hybrid l_1 - l_0 multi-scale decomposition model that decomposes the image into a base layer, to which an l_1 sparsity term is imposed to enforce piecewise smoothness, and a detail layer, to which an l_0 sparsity term is imposed as structural prior, in order to avoid halos and over-enhancement of contrast. Li et al. [LJZ18] propose to decompose HDR images into color patches and cluster them according to different statistics. Then, for each cluster they apply principal component analysis to find a more compact domain for applying different tone mapping curves. In general, these traditional methods are model-based and need to introduce prior information, furthermore, they usually require careful parameter tuning which is not user-friendly for non-expert users.

The closest to our goals are gradient domain techniques [FLW02, MMS06, STO16] that effectively compress/enhance contrast. Fattal et al. [FLW02] consider gradients between neighboring pixels, while Shibata et al. [STO16] employ a base- and detail-layer decomposition, manually pre-selecting parameters to guide contrast manipulations. Mantiuk et al. [MMS06] proposes a multi-resolution contrast processing that is driven by a perceptual contrast self-masking model [KW96]. In contrast, our model additionally takes into account neighborhood masking effects, further, contrast masking effects are computed in feature domain instead of traditional image domain.

2.2.2. Learning-based methods

Due to the great success of deep learning in image processing tasks, new learning-based tone mapping operators have been proposed during the last years. Most of these works fall under the category of supervised methods, i.e., they need HDR-LDR (low dynamic range) image pairs in order to train their proposed models. Patel et al. [PSR17] propose to train a generative adversarial network (GAN) in order to learn a combination of traditional tone mapping operators that allows for better generalization across scenes. During training, in order to select the target tone mapped image, they select the best scoring tone mapped image (using the TMQI metric [YW12]) among a set of tone mapped results using different traditional methods. Rana et al. [RSV*19] instead propose using a multi-scale conditional generative adversarial network, and then follow the same procedure for selecting tone mapped images for training. Zhang et al. [ZWZW19] use a carefully designed loss function to push tone mapped images into the natural image manifold. The target tone mapped images for training are manually adjusted by photographers using the tone mapping operators available in Photomatix2 and HDRToolbox [BADC17]. Su et al. [SWL*21] propose an explorable tone mapping network based on BicycleGAN [ZZP*17] and use LuminanceHDR to generate suitable tone mapped target images, selecting the top-scoring ones using the TMQI metric. To mitigate the challenge of finding target tone mapped images suitable for training, Panetta et al. [PKO*21] use low-light images based on the insight that they have under-exposed regions that model well the distribution of HDR images while also having characteristics considered as under-exposed when viewed

in displays with limited dynamic range. The method proposed by Yang et al. [YXS*18] allows to recover image details by training an end-to-end network for reconstructing HDR images from LDR ones, and then performing tone mapping. They use Adobe Photoshop as a black box to empirically generate ground truth tone mapped images with human supervision. Recently, Zhang et al. [ZZWW21] propose a semi-supervised method by combining unsupervised losses and a supervised loss. In this manner, their method only requires a few HDR-LDR pairs with well tone mapped images. For the supervised loss term, they use LDR images from the previously discussed work of Zhang et al. [ZWZW19]. Inspired by image quality assessment metrics, Guo and Jiang [GJ21] also propose a semi-supervised method and obtain HDR-LDR pairs from fine-tuning raw bracketed exposures using Adobe Photoshop.

These learning methods are intrinsically limited by the input data they see during training. Therefore, using images tone mapped with existing methods as target, although allows for training new models with better generalization, fundamentally limits the quality that such new learned models can achieve. In contrast, we propose a self-supervised network that only takes as input the original HDR image for training, and learns a tone mapping operator relying on a carefully designed loss function that takes into account contrast masking effects present in the Human Visual System. To our knowledge, the only work that does not need carefully selected HDR-LDR image pairs is the work of Hou et al. [HDQ17], however, they rely on combining feature losses from different layers chosen empirically and only demonstrate their approach for two selected images.

3. Proposed Method

In this section we present our image-specific, self-supervised tone mapping network, whose structure is shown in Fig. 2. The input HDR image is first normalized, and then decomposed into three differently exposed LDR images (Sec. 3.1). The three exposures are first transformed into feature space by an encoder with shared weights. Then, they are fused in this feature space to produce a corrective residual that is then decoded, and finally added to the three input exposures to generate the output tone mapped image (Sec. 3.2). To compute the training loss, first the normalized HDR image is processed by an adaptive μ -law compression that brings its distribution closer to typical LDR image histogram distribution (Sec. 3.3). Then, the processed HDR image along with its tone mapped counterpart go through a VGG network to derive their respective feature spaces (we employ VGG19 [SZ14], which we denote VGG for brevity). Finally, we compute the L1 loss in feature space, however, instead of the standard perceptual loss between corresponding features [SZ14, JAFF16], we compute feature contrast and model contrast self-masking and neighborhood masking effects (Sec. 3.4).

3.1. Multiple Exposure Selection

HDR images are stored in linear intensity space and might feature extremely large dynamic range compared to LDR images. Moreover, the distribution of pixel intensities in the HDR image is also unbalanced, where pixels with very high intensity have large values

but correspond small image regions, while low-brightness pixels usually cover larger portions of the image [EKD*17, PKO*21]. In neural network applications, to align such HDR image characteristics to those of LDR images, the logarithm of HDR pixel intensities is often applied [EKD*17, ZWZW19, SWL*21]. This way a compressive response of the HVS to increasing luminance values (the Weber-Fechner law) is modeled as it is common in the tone mapping literature [RHD*10]. However, as we detail in Sec. 4.3, we found that our tone mapping network leads to better results when multiple differently exposed images with linear pixel intensity relation are used instead.

As HDR images are typically stored as relative positive values, before choosing the exposures, we first derive the normalized I_{HDR} image, where each pixel value is multiplied by 0.5, and divided by the mean of the original HDR image I_{SRC} intensity [EKM17]. To estimate the exposure range for each exposure we employ an automatic procedure originally proposed for HDR image quality evaluation [ANSAM21]. This way we obtain the low e_{low} and high e_{high} exposures[†], and additionally we derive an intermediate third exposure as $e_{\text{mid}} = (e_{\text{low}} + e_{\text{high}})/2$. As we show in detail in the supplemental (Sec. S1.3), we found that these three exposures lead to overall good results. Only two exposures are not sufficient for capturing all the dynamic range of challenging scenes, and four or more exposures do not improve the quality of the results, while increasing the computation time. An alternative for exposure selection is the work of Gallo et al. [GTM*12], which provides an optimization method for exposure metering based on the input HDR histogram. This approach guarantees that all pixels will be well represented (avoiding saturation), however, the resulting images are non-linear LDR images quantized in 8-bits. Note that in contrast to standard multi-exposure 8-bit LDR images, our three floating-point exposure selection is sufficient to represent high dynamic range information for tone mapping purposes. As we do not perform pixel quantization, the only information loss is due to clipping higher intensities into the range $[0,1]$ with the $\text{clip}()$ function:

$$I_{e-x} = \text{clip}(2^{e_x} I_{\text{HDR}}), \quad (1)$$

where I_{e-x} represents one of multi-exposure images with the exposure factor e_x .

3.2. Tone Mapping Network

The tone mapping network is composed of an encoder \mathcal{E} , a feature module \mathcal{F} , and a decoder \mathcal{D} . All details on the number of layers, as well as the per-layer kernel size, channel number, and stride extent are specified in the bottom-left corner of Fig. 2. All layers use Relu [NH10] as the activation function except for the last layer using sigmoid. The three input exposures are used as an input to the encoder network \mathcal{E} with shared weights between the exposures. The resulting feature maps are used to derive a corrective residual signal (module \mathcal{F}) that after decoding (\mathcal{D}) is added up with the three exposures. Finally, the resulting image is fitted into the range $[0,1]$ by

[†] Note that the goal of Andersson et al. [ANSAM21] is to get aesthetical results. We divide by two both e_{low} and e_{high} , resulting in less dark and less bright exposures, respectively. Our goal is that all relevant content is within reasonable pixel values (not too dark, not too bright).

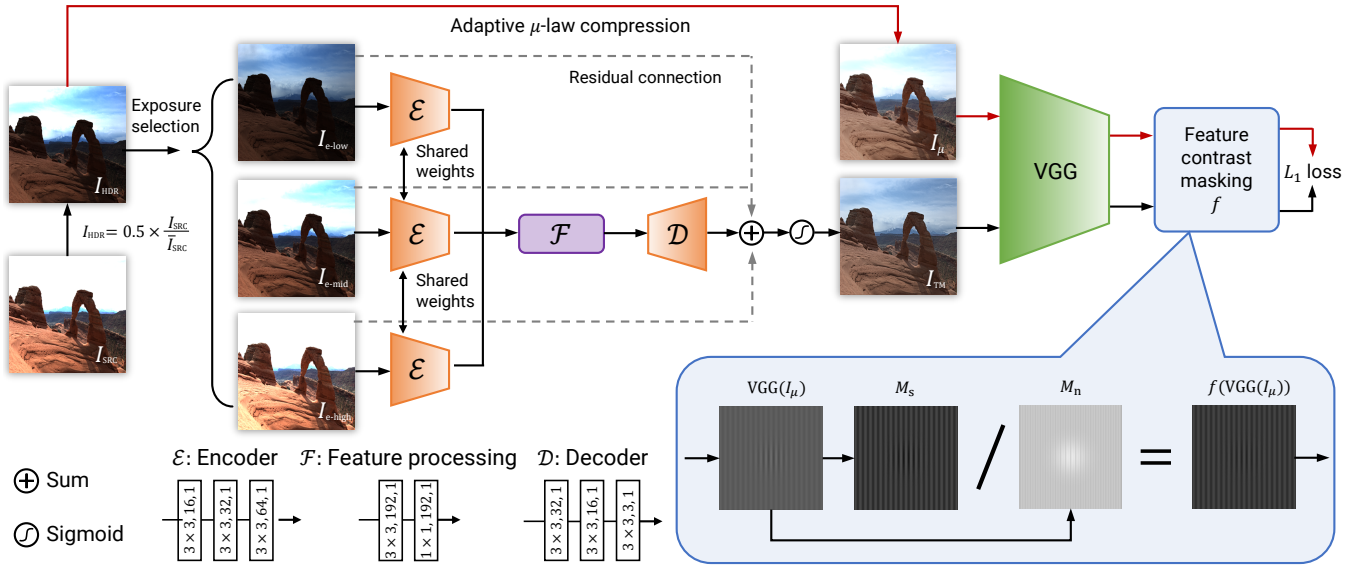


Figure 2: Overview of our method. The input HDR image I_{SRC} is first normalized into I_{HDR} , and then three exposures I_{e-low} , I_{e-mid} , and I_{e-high} are selected. A network with learnable HDR-image-specific weights is used to encode (\mathcal{E}) each exposure into feature space (\mathcal{F}), where a corrective residual is derived, that after decoding (\mathcal{D}) is summed up with all three exposures. The resulting image values are compressed into the range of $[0,1]$ by a sigmoid function, and the output tone mapped image I_{TM} is obtained. The appearance and contrast of I_{TM} is directly guided by the normalized HDR I_{HDR} that is further processed into I_{μ} using an adaptive μ -law compression. Both I_{TM} and I_{μ} are transformed into feature space (VGG) and compared taking into account our novel contrast masking model f . As shown in the bottom-right inset I_{μ} is transformed into the feature space $VGG(I_{\mu})$, where the ratio between feature contrast self-masking M_s and feature contrast neighborhood masking M_n define the feature contrast masking model $f(VGG(I_{\mu}))$. Finally, the L_1 loss is computed between $f(VGG(I_{\mu}))$ and $f(VGG(I_{TM}))$. The inset in the bottom-left shows the structure of decoder (\mathcal{E}), feature space residual correction (\mathcal{F}), and decoder (\mathcal{D}) networks.

a sigmoid function (Fig. 2) to produce the final tone mapped image I_{TM} . While similar residual corrections have been proposed in the past [HZRS16, XZR18], our residual has two goals specific to our application. First, I_{μ} (Sec. 3.3) is outside the range $[0,1]$ as shown in Fig. 5 (left) so it needs to be compressed. Through our perceptual loss, the corrective residual is trained such that deviations from the compressed HDR image I_{μ} in the tone mapped image I_{TM} are more strongly penalized for low contrast image details than for high contrast details, encouraging their compression. This way the desired appearance of the final tone mapped image I_{TM} is achieved, while an exact reconstruction of I_{μ} is avoided (Fig. 3 and the insets in Fig. 5 show the relatively poor quality of I_{μ}). Second, during our exposure selection, darker pixels in I_{HDR} (Fig. 2) are linearly scaled according to the three selected exposures and brighter pixels may be affected in some cases by clipping in I_{e-high} or I_{e-mid} (Eq. 1). The residual can be seen as a correction that, when added to the three exposures and after applying the sigmoid function, brings each pixel to the right intensity value while accounting for the clipping non-linearity and assuring perceptually plausible contrast processing (following the loss guided by I_{μ}). In Fig. 3 we show an example of the residual together with the three exposures that it corrects.

3.3. Adaptive μ -law compression

Our image-specific tone mapping network is self-supervised by the input HDR image I_{SRC} that is transformed by the VGG network into

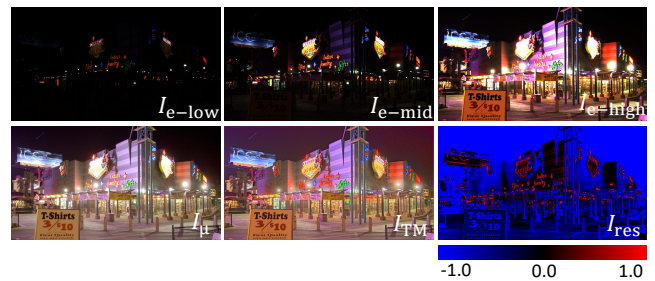


Figure 3: Visualization of three input exposures (I_{e-low} , I_{e-mid} , and I_{e-high}), corresponding compressed I_{μ} and resulting tone mapped image I_{TM} , together with the residual image I_{res} displayed in false color. For visualization, positive (red) and negative (blue) values in the residual are normalized to the range $[-1, 1]$. As the sum of the three exposures falls into the range $[0, 3]$, and the argument space of a sigmoid function spans an $[-\infty, +\infty]$ input range, large positive residual values are expected in bright regions to push output values in I_{TM} towards 1. Similarly, large negative residual values are expected in dark regions to push output values in I_{TM} towards 0.

feature space (Fig. 2). We first adapt the range of intensity values by converting I_{SRC} into I_{HDR} as discussed in Sec. 3.1. As shown in Fig. 4 (top row), I_{HDR} still has a strong skew of its histogram towards

low intensity values (typical for HDR images [EKD*17,PKO*21]), which strongly differs from LDR image characteristics. Since LDR images are typically used for training VGG, we resort to a μ -law compression to correct the histogram of the image I_{HDR} :

$$I_{\mu} = \frac{\log(1 + \mu I_{\text{HDR}})}{\log(1 + \mu)}, \quad (2)$$

This algorithm is widely used in HDR image coding [JKX*11] and inverse tone mapping [WXTT, STKK20] to re-arrange pixel intensity distribution. Fig. 4 (bottom row) shows an example of this transformation, where image details become more visible and the long tail in the histogram is corrected.

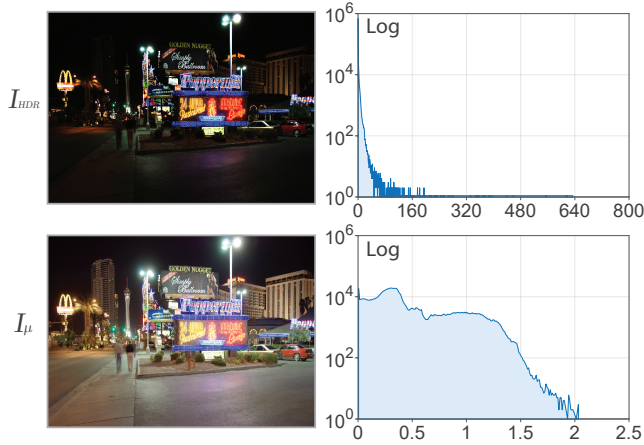


Figure 4: Image appearance and logarithmic histogram of pixel intensities for a transformed HDR image. Top Row: I_{HDR} with the mean normalized to 0.5. Bottom row: I_{μ} derived using the adaptive μ -law compression (Eqs. 2 and 3). Notice the dramatically different scales on the x-axis.

Typically a fixed μ is used to derive the transformed image I_{μ} , but as we show in the insets in Fig. 5 (left), the resulting image appearance strongly depends on the selected μ value. We observe that the choice of μ also affects greatly the performance of the VGG-based loss that drives our tone mapping network, where larger μ values are required for darker HDR images. We propose an adaptive μ -law compression, where the μ value changes with the median pixel intensity value i_{HDR} that is computed for the I_{HDR} image:

$$\mu = \lambda_1 (i_{\text{HDR}})^{\gamma_1} + \lambda_2 (i_{\text{HDR}})^{\gamma_2}, \quad (3)$$

with fitted parameter values $\lambda_1 = 8.759$, $\gamma_1 = 2.148$, $\lambda_2 = 0.1494$, and $\gamma_2 = -2.067$. We derive this function experimentally for a number of representative HDR images featuring different appearances as well as different i_{HDR} (i.e., brightness). We use the TMQI quality metric [YW12] to select the best performing μ values, and then by visual inspection we confirm this selection. Fig. 5 (right) shows the fitted curve based on this procedure.

3.4. Feature Contrast Masking Loss

In this section we propose the feature contrast masking (FCM) loss that guides our tone mapping network to reproduce image details

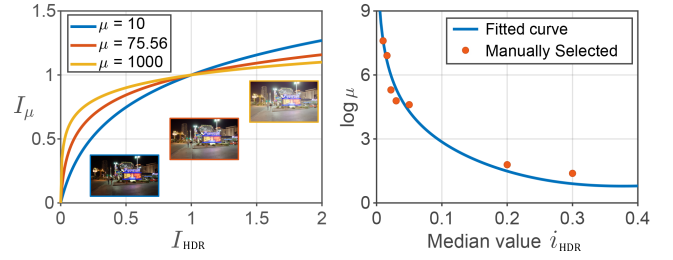


Figure 5: Left: Adaptive μ -law compression (Eq. 2) for different μ values and the corresponding compressed HDR images. In this case $\mu = 75.56$ has been selected for this image using Eq. 3. Right: Experimentally derived μ selection as a function of HDR image median intensity i_{HDR} .

and overall perceived contrast. To this end, we first model feature contrast, and then introduce self contrast masking and neighborhood masking for such feature contrast inspired by their analogues in image domain described in Sec. 2.1.

Feature contrast. While the HDR image representation I_{μ} , which results from the adaptive μ -law compression (Sec. 3.3), greatly facilitates its meaningful processing by the VGG network, still significant intensity differences might exist with respect to its tone mapped version I_{TM} . Such intensity differences translate into corresponding differences in the feature magnitude when I_{μ} and I_{TM} are transformed by the VGG network (Fig. 2). To further reduce such feature magnitude differences we compute a per-channel local feature contrast:

$$C_p = \frac{f_p - \bar{f}_p}{|\bar{f}_p| + \epsilon}, \quad (4)$$

where f_p denotes the feature magnitude at pixel p , \bar{f}_p is the Gaussian-filtered feature value computed for the patch \mathcal{P} (centered at p), and ϵ is a small constant to avoid division by zero. We experimentally set the patch size \mathcal{P} to 13×13 pixels (Sec. S1.2 in the supplemental). To compute this contrast, the feature difference with respect to its local neighborhood is first computed and then normalized. This normalization enables further reduction of the impact of the differences in absolute feature magnitudes between I_{μ} and I_{TM} . Note that Eq. 4 is aligned with contrast definitions for images that also use Laplacian and Gaussian filter responses in the numerator and denominator, respectively [Pel90]. More complex filter banks such as Laplacian pyramids [MDC*21], cortex transforms [Dal92], wavelets [DZLL00], or discrete cosine transforms [Wat93] are often used for advanced contrast processing operations.

Feature contrast self-masking. An important property of contrast perception is a higher sensitivity to contrast discrimination for lower contrasts than for supra-threshold contrasts [KW96], this is termed contrast self-masking [DZLL00]. In the context of tone mapping this means that even small contrast changes can be perceived in low-contrast image regions, which are often neglected in global tone mapping operators [MMS15]. Conversely, for larger image contrast, even strong contrast compression might remain undetected. Many tone mapping operators take advantage of this effect [DD02, FLW02, MMS06]. Different from conventional meth-



Figure 6: Feature contrast masking visualization of VGG feature maps (1st layer, 18th channel). From left to right: tone mapped image I_{TM} , original VGG feature map $VGG(I_\mu)$, feature contrast self-masking M_s , neighborhood masking M_n , and final $f(VGG(I_\mu))$ masking terms.

ods, we do not model contrast self-masking in the image contrast domain, instead we model it in the feature contrast domain C defined in Eq. 4 (we skip the pixel index p for brevity):

$$M_s = \text{sign}(C)|C|^\alpha, \quad (5)$$

where M_s denotes a non-linear response to the feature contrast magnitude controlled by the compressive power factor α . The function $\text{sign}(x) = \frac{x}{|x|+\epsilon}$ preserves the feature contrast polarity in M_s . We experimentally set $\alpha_\mu = 0.5$ when processing I_μ while we keep $\alpha_{I_{TM}} = 1.0$ for I_{TM} , allowing for visually pleasant low-contrast detail enhancement (refer to Sec. S1.2 in the supplemental). Fig. 6 shows a feature map $VGG(I_\mu)$ for a selected VGG channel, as well as the response M_s to feature contrast C . As shown in this figure, M_s vividly responds for low-contrast details in the sky and rocks that are instead strongly suppressed in $VGG(I_\mu)$. Note that when $VGG(I_\mu)$ is directly used in the perceptual VGG loss driving the tone mapping operation $\mathcal{L}_{VGG} = \|VGG(I_\mu) - VGG(I_{TM})\|_1$ [JAFF16], such details are likely to be neglected in the resulting I_{TM} due to the low penalty in the loss.

Feature contrast neighborhood masking. Inspired by successful applications of neighborhood masking, as discussed in Sec. 2.1, we also model feature contrast neighborhood masking. Image contrast neighborhood masking is performed selectively for different spatial frequency bands; this requires image decomposition by a filter bank [DZLL00, Lub95]. We approximate this process by modeling feature contrast neighborhood masking per channel, where features with similar frequency characteristics are naturally isolated. Our goal is to suppress the magnitude of M_s when there is a high variation of feature magnitudes f_p in the local neighboring of pixel p that we measure as:

$$M_n = \frac{\sigma_b}{|\mu_b| + \epsilon}, \quad (6)$$

where μ_b and σ_b denote the mean and standard deviation of feature magnitude f_p in the patch \mathcal{P} that is centered at pixel p . Again, we experimentally set the patch \mathcal{P} size to 13×13 pixels. Finally, our feature contrast masking is calculated as the ratio of self- and neighborhood masking:

$$f(VGG(I)) = \frac{M_s}{1 + M_n} \quad (7)$$

As can be seen in Fig. 6, M_n vividly responds in the regions with high local feature variation as seen in $VGG(I_\mu)$, so that the final fea-

ture contrast masking measure $f(VGG(I_\mu))$ is strongly suppressed in such regions. In particular, this means that in regions with strong image contrast, such as the horizon line, $f(VGG(I_\mu))$ is relatively much smaller with respect to the original $VGG(I_\mu)$ feature magnitudes. Consequently, when including $f(VGG(I_\mu))$ into the loss computation that drives the tone mapping operation, the penalty for any distortion of such high contrast is much smaller than in the perceptual VGG loss \mathcal{L}_{VGG} that directly employs $VGG(I_\mu)$. Effectively, this gives the tone mapping network more freedom for compressing image contrast in such regions.

We compute our feature contrast masking (FCM) loss \mathcal{L}_{FCM} as the L_1 loss between the masked feature maps $f()$ for the transformed input HDR image I_μ and the output tone mapped image I_{TM} :

$$\mathcal{L}_{FCM} = \|f(VGG(I_\mu)) - f(VGG(I_{TM}))\|_1 \quad (8)$$

To further illustrate the behavior of our loss, in Fig. 7 we consider simple sinusoid patterns with three different contrasts (c_1 , c_2 , and c_3). The corresponding feature maps of the VGG network are shown in the top-left image row. We distort each sinusoid by increasing their respective amplitudes by the same factor δ and their corresponding feature maps are shown in the bottom-left row.

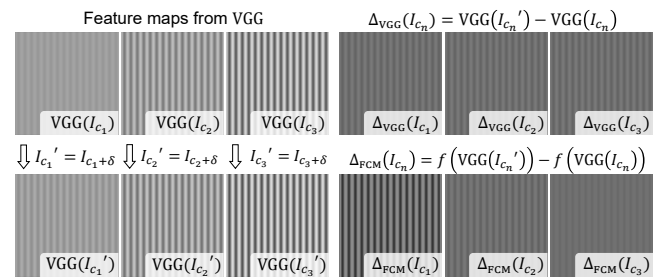


Figure 7: Left: VGG feature maps (1st layer, 2nd channel) for three input sinusoids of increasing image contrast ($c_1 < c_2 < c_3$) (upper row) that are distorted by further increasing their amplitudes by the same factor δ (bottom row). Right: the differences Δ_{VGG} between the VGG feature maps, which are computed for each sinusoid and its distorted version, show a weak dependence to the input sinusoid contrast (upper row), while the corresponding differences Δ_{FCM} resulting from our feature contrast masking model penalize more strongly the distortion for smaller contrast sinusoids.

We compare the feature map difference Δ_{VGG} used in the VGG loss \mathcal{L}_{VGG} [JAFF16] (top-right row) and the corresponding Δ_{FCM} used in our FCM loss \mathcal{L}_{FCM} (bottom-right row). As can be seen, Δ_{FCM} yields a higher penalty when the distortion δ is added to the lowest contrast pattern. This forces our tone mapping, driven by \mathcal{L}_{FCM} , to reproduce image details in low-contrast areas. The VGG loss \mathcal{L}_{VGG} remains similar irrespectively on the input sinusoid contrast. This puts equal pressure on the tone mapping network to reproduce image details for large contrast regions that cannot be perceived, and low-contrast regions where they are clearly visible.

Finally, we perform color correction in I_{TM} following Tumblin and Turk [TT99]:

$$C_{out}^{\{R,G,B\}} = L_{out} \left(\frac{C_{in}^{\{R,G,B\}}}{L_{in}} \right)^s \quad (9)$$

$$L_{in/out} = 0.2126 * R + 0.7152 * G + 0.0722 * B \quad (10)$$

where L_{in} and L_{out} are the luminance of the input HDR image I_{HDR} and the output tone mapped image I_{TM} respectively, and s is the saturation control parameter (we use $s = 0.6$ in all examples). Since we apply color correction as a post-processing operation, other algorithms [MMTH09, APBA18] could be also applied.

4. Results and Ablation Study

In this section we first describe the implementation details of our approach. Then, we provide objective and subjective comparisons including both traditional methods and state-of-the-art learning-based approaches. Finally, we perform an ablation study showing how each of the components of our approach contributes to achieving the final quality of our results.

4.1. Implementation

We adopt an online training strategy and train a model for each HDR image at test time. Our model is implemented on TensorFlow 1.10 and the results reported in the paper are computed with a RTX 8000 GPU. We use the Adam optimizer with an initial learning rate of 2×10^{-4} and an exponential decay factor of 0.9 every ten epochs. The training converges after 400 epochs, which translates into around 583 ± 6.62 seconds, for an image resolution of 768×384 . We fix a single set of parameters for all our experiments and results. As discussed in the previous section, we set $\mathcal{P} = 13 \times 13$, $\alpha_{\mu} = 0.5$, and $\alpha_{I_{\text{TM}}} = 1.0$. We compute our loss function based on the first three layers of VGG. Please refer to the supplemental for experimental exploration of these parameters.

4.2. Results and comparisons

We include in our comparisons fourteen tone mapping operators including ten traditional methods which for simplicity we refer to as: Mantiuk [MMS06], Shan [SJB09], Durand [DD02], Drago [DMAC03], Mertens [MKVR07], Reinhard [RSSF02], Ma [MYZW15], Liang [LXZ*18], Shibata [STO16], and Li [LJZ18]; and four recent learning-based methods: Guo [GJ21], Zhang [ZZWW21], DeepTMO [RSV*19] and TMO-Net [PKO*21]. We use the publicly available implementations of these methods or if not available, their implementation

in HDRToolBox [BADC17]. For the case of DeepTMO [RSV*19] and TMO-Net [PKO*21], we were not able to access their implementations, therefore we directly use the results provided in their works for comparisons.

We use a large test set of 275 images gathered from the Fairchild dataset [Fai07], Poly Haven[‡], the Laval Indoor HDR Database [GSY*17], and the LVZ-HDR benchmark dataset [PKO*21], which cover various indoors, outdoors, bright and dark scenes.

4.2.1. Objective evaluation

For objective comparisons we adopt as metrics the Tone Mapped Image Quality Index (TMQI) [YW12], the Blind Tone Mapped Quality Index (BTMQI) [GWZ*16] and the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [MMB12]. The former two metrics are widely used for evaluating tone mapping operators [GJ21, ZZWW21, LXZ*18], while the latter is typically used as a blind metric for evaluating contrast [JYL19, LHLK17, SWH*20]. For completeness, we briefly discuss here these metrics. Additionally, we compute contrast distortion maps (loss of visible contrast and amplification of invisible contrast) between the tone mapped results and the original HDRs using the Dynamic Range Image Quality Assessment (DRIM) [AMMS08] metric (please, refer to Sec. S4.1 in the supplemental).

TMQI is a full-reference tone mapping image quality metric, which consists of two main terms assessing the structural fidelity (TMQI_S) and the naturalness (TMQI_N) of the tone mapped image. BTMQI is a no-reference tone mapping metric, which is composed of three terms accounting for entropy (richness of information), naturalness, and presence of structural details. BRISQUE is a well-known no-reference image quality metric based on natural scene statistics that quantifies the naturalness of an image and considers distortions such as noise, ringing, blur, or blocking artifacts.

We show in Table 1 the results of these objective metrics together with computation times for eleven of the tested methods with our testing set, while Table 2 shows the results for DeepTMO and TMO-Net, for which we use their provided test sets and results. We include examples of our results compared to the seven best performing methods in Fig. 8, and compared to DeepTMO and TMO-Net in Fig. 9. We include more results and comparisons in our supplemental. Additionally, we show in Table 3 running times for our method with Standard Definition and High Definition images. Since we adopt an online training strategy, our network has to be optimized for every different HDR image, resulting in moderately high running times. We have explored two different techniques for speeding-up computation: depth-wise convolutions [HZC*17] and progressive training (please, see Sec. S2.1 in the supplemental for details). This allows to heavily decrease our running times with a slight quality loss, still outperforming previous methods.

Our proposed approach outperforms previous methods in different aspects. We discuss now more in detail results for the best performing approaches. In general, all approaches except DeepTMO

[‡] <https://polyhaven.com/hdris>

Table 1: Mean and standard deviation for TMQI (including TMQI_S and TMQI_N), BTMQI and BRISQUE computed for the 275 images in our test set, and average computing time for an image with 768 × 384 pixels resolution.

Methods	TMQI (↑)	TMQI _S (↑)	TMQI _N (↑)	BTMQI (↓)	BRISQUE (↓)	time (s)
Ours	0.9248 ± 0.0432	0.8938 ± 0.0582	0.7062 ± 0.2369	2.9065 ± 1.0481	21.2208 ± 8.5981	583.26
Guo [GJ21]	0.8883 ± 0.0354	0.8166 ± 0.0762	0.5975 ± 0.2002	3.7110 ± 1.0239	23.9528 ± 8.2566	0.0283
Zhang [ZZWW21]	0.8767 ± 0.0600	0.8343 ± 0.0872	0.5118 ± 0.2686	3.7440 ± 1.4209	22.0520 ± 8.9580	0.0021
Liang [LXZ*18]	0.8964 ± 0.0490	0.8534 ± 0.0737	0.5194 ± 0.2605	3.5333 ± 1.0915	25.1433 ± 8.5510	1.1266
Shibata [STO16]	0.7689 ± 0.0506	0.7498 ± 0.0858	0.1203 ± 0.1719	4.3470 ± 0.7469	32.2273 ± 8.9383	11.655
Li [LJZ18]	0.8480 ± 0.0612	0.8301 ± 0.0743	0.3716 ± 0.2939	4.3670 ± 1.1621	24.0348 ± 8.5411	3.5321
Shan [SJB09]	0.8301 ± 0.0732	0.7458 ± 0.1448	0.4174 ± 0.2792	4.0685 ± 0.9839	22.8230 ± 8.5913	50.287
Durand [DD02]	0.8719 ± 0.0669	0.8375 ± 0.0989	0.4824 ± 0.2685	3.6537 ± 1.1016	22.0285 ± 8.4576	0.0665
Drago [DMAC03]	0.8794 ± 0.0537	0.8600 ± 0.0803	0.4840 ± 0.2558	4.0213 ± 1.2733	22.6600 ± 8.2405	0.0974
Mertens [MKVR07]	0.8425 ± 0.0717	0.8403 ± 0.0923	0.3373 ± 0.2848	4.8981 ± 1.6026	23.2085 ± 8.8589	1.4282
Reinhard [RSSF02]	0.8506 ± 0.0533	0.8176 ± 0.0864	0.3903 ± 0.2347	4.2774 ± 1.4580	25.5317 ± 7.7396	0.2393
Mantiuk [MMS06]	0.8529 ± 0.0753	0.8903 ± 0.0849	0.3238 ± 0.3050	4.5339 ± 1.4086	21.2943 ± 8.8302	1.8733
Ma [MYZW15]	0.8782 ± 0.0698	0.8644 ± 0.1043	0.4782 ± 0.2458	3.6444 ± 1.6857	24.9172 ± 8.0956	3225.4

Table 2: Mean and standard deviation for TMQI (including TMQI_S and TMQI_N), BTMQI and BRISQUE computed for the DeepTMO [RSV*19] and TMO-net [PKO*21] test sets. The former contains 100 images from the Fairchild dataset [Fai07] while the latter contains 457 captured images from their own dataset.

Methods	TMQI _Q (↑)	TMQI _S (↑)	TMQI _N (↑)	BTMQI (↓)	BRISQUE (↓)
Ours	0.9106 ± 0.0511	0.8987 ± 0.0664	0.6052 ± 0.2807	3.3420 ± 1.0686	19.5406 ± 9.1347
DeepTMO [RSV*19]	0.9052 ± 0.0619	0.8810 ± 0.0717	0.6015 ± 0.2679	3.4230 ± 1.1502	27.5489 ± 7.6865
Ours	0.9073 ± 0.0541	0.8939 ± 0.0551	0.6020 ± 0.3126	3.3069 ± 1.2266	23.1010 ± 8.5232
TMO-Net [PKO*21]	0.8609 ± 0.0594	0.8066 ± 0.0825	0.4723 ± 0.2871	3.9633 ± 1.2477	26.6078 ± 8.1895

Table 3: Computation times for our method with Standard Definition (SD, 720 × 480) and High Definition (HD, 1080 × 720) images for our baseline method and our two proposed techniques for speeding-up the computation: depth-wise convolutions [HZC*17] and progressive training.

Methods	Time-SD (s)	Time-HD (s)
Ours	654.72	1245.7
depth-conv	223.74	486.57
depth-conv & pro-training	126.28	251.36

yield low TMQI_N values, indicating that they fall short in preserving the naturalness of the image. For the case of DeepTMO, the low performance in the BRISQUE metric indicates that the tone mapped images do not preserve natural image statistics (Fig. 9, first three columns). We can also see that TMO-Net additionally produces saturated results in the brightest regions of the image (Fig. 9, last three columns). Looking into Fig. 8 we can observe that Guo over-enhances dark regions, sometimes resulting in heavy artifacts (*garage* scene), while Zhang tends to produce overly dark results in regions with low brightness. This is in agreement with a relatively low score in structural fidelity TMQI_S, indicating that the tone mapped images differ from the HDR in terms of conveyed structural information. While Liang and Ma perform well in terms of TMQI_S, they tend to produce under-saturated results with lower contrast (e.g., *garage* and *store* scenes), which is in agreement with

a relatively low performance in the BRISQUE metric. Drago and Durand also perform well in terms of TMQI_S. However, we can see that Drago produces blurry results and fails to reproduce fine details, such as the floor tiling in the *station* scene or the highlights of the bottles in the *store* scene. Durand presents good scores in terms of BRISQUE score which means the tone mapped images do align with natural image statistics, however in some cases it over-enhances contrast, producing artifacts such as those around the car windows in the *garage* scene or those in the luminous numbers of the *station* display sign. Mantiuk has the lowest TMQI_N of the methods included in Fig. 8. This operator tends to produce very dark images with low contrast.

Our method outperforms existing approaches for all tested metrics, exhibiting a good contrast reproduction while preserving the details present in the HDR images. Our results also produce natural images without visible artifacts.

4.2.2. Subjective evaluation

To further validate the performance of our approach we additionally performed a subjective study. We included the six most commonly used, best-performing methods in terms of average TMQI score according to our previous objective evaluation, in particular: Guo [GJ21], Zhang [ZZWW21], Liang [LXZ*18], Drago [DMAC03], Durand [DD02], and ours. The study was approved by the Ethical Review Board of the Computer Sciences department at Saarland University, and participants provided written consent for participating in the study. A total of twelve participants

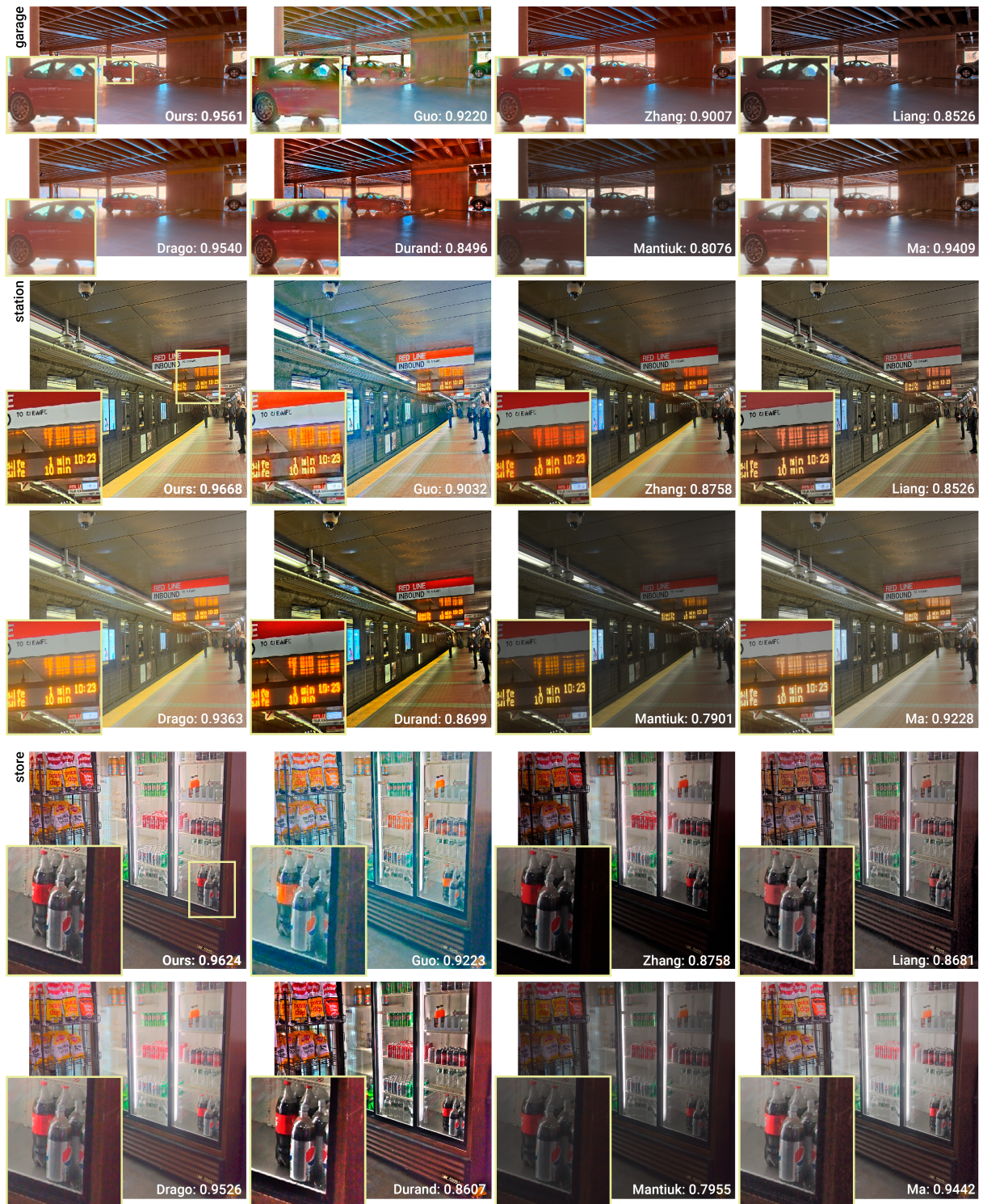


Figure 8: Visual comparisons for the best performing methods and their TMQI scores. Overall, our results achieve good contrast reproduction while preserving the details (highlights of the bottles in store, floor in station) and avoiding visual artifacts (display sign in station, car edges and windows in garage). Please refer to the text for an in-depth discussion of the observed differences.



Figure 9: Visual comparisons for DeepTMO [RSV*19] and TMO-Net [PKO*21] on images from their respective test set. DeepTMO tends to produce over-enhanced contrast and saturated results that do not preserve natural image statistics. TMO-Net can not handle some challenging scenarios, producing saturated results in very bright regions and overly dark pixels in dark regions.

(aged 25 to 34 years old with normal or corrected-to-normal vision) voluntarily took part in the study. We included fifteen scenes covering different scenarios and, for each scene, we showed the six tone mapped images at a random order. We asked the participants to rank them from 1 (preferred) to 6 (least preferred). Before the study, participants underwent a simple training with a few images showing common artifacts in tone mapped images such as over-saturated colors, over- or under-exposure, contrast loss and contrast over-enhancement. The study was conducted in a room with natural illumination and participants sat at a distance of 0.5 meters from the display. The images were displayed in a DELL UltraSharp U2421E monitor (1920 × 1200 resolution, 60 HZ refresh rate).

We show in Fig. 10 the preference rankings for each method, aggregated for all participants and scenes. We use Kruskal-Wallis (non-parametric extension of ANOVA) for analyzing the rankings, since these do not follow a normal distribution [JMB*14, RGSS10]. We then compute post-hocs using pairwise Kruskal-Wallis tests adjusted by Bonferroni correction for multiple test. Results reveal a statistically significant difference in the rankings for the different methods ($p < 0.001$), with our approach being ranked significantly higher than all others (refer to Sec. S3 in the supplemental for individual results for each scene and statistical tests for all pairwise comparisons).

4.3. Ablation Studies

In this section we evaluate the importance of each of the components in our method for achieving the final quality of the results. Table 4 shows the results of the objective metrics for different combinations of (i) input: linear HDR [RSV*19], log HDR [ZWZW19, SWL*21] or our multiple exposure inputs (MEI); (ii) HDR compression algorithm for computing the loss: linear, log or our adaptive μ -law compression (Ada μ); and (iii) loss function: \mathcal{L}_{VGG} or our \mathcal{L}_{FCM} . Fig. 11 shows the corresponding visual results. Please, refer to the supplemental for extended results on the ablation.

In general, we can see that our loss \mathcal{L}_{FCM} , which considers masking effects, plays an important role in emphasizing the local contrast, especially in large contrast regions, such as the clouds in the sky. Compared with our MEI, the logarithm input leads to overall darker results (brightness distortion), and the linear input can cause

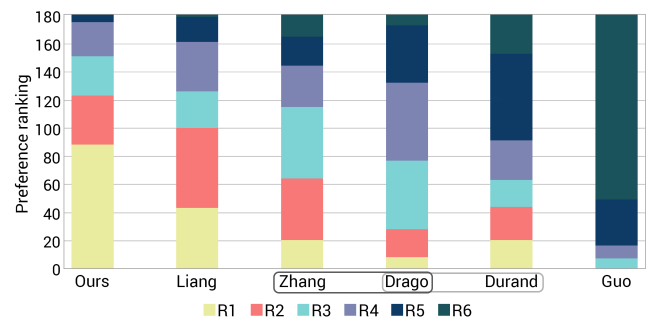


Figure 10: Preference rankings for each method aggregated across the twelve participants and fifteen scenes. Different colors indicate the received rankings (from R1 to R6). Pairwise comparisons between methods reveal that preference rankings are significantly different, except those methods marked in the same set (gray squares), which are statistically indistinguishable.

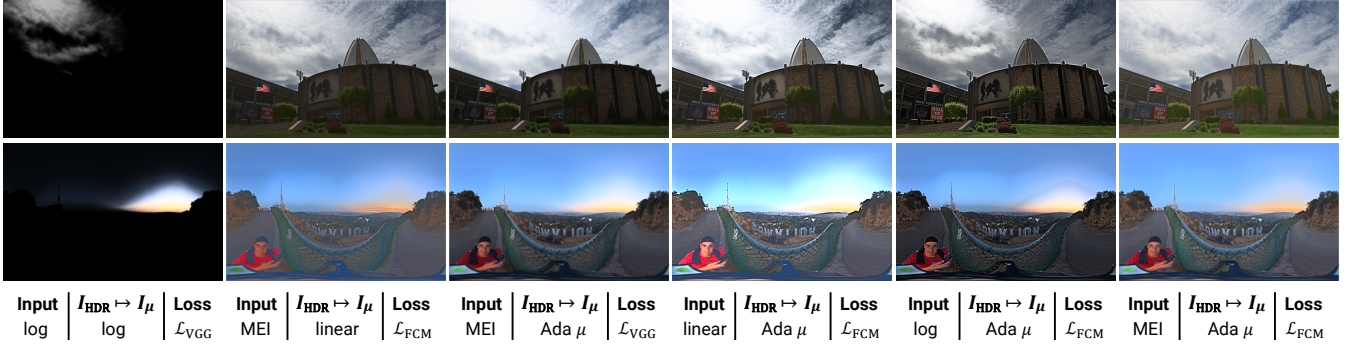
overexposure with missing information in the highlight regions. Finally, we can see that the Ada μ -law compression is important for overall image contrast, when linear or logarithmic transformations are applied instead, the resulting images are flatter in terms of contrast.

5. Conclusions

In this work we propose an image-specific self-supervised tone mapping approach that leads to consistent high-quality results for a large variety of HDR scenes. Previous learning-based approaches present two main limitations: (i) the variety of HDR content they can adequately tone map is limited by the images used during training, and (ii) most of these approaches are supervised, i.e., they need HDR-LDR image pairs for training. These LDR images are obtained either from tone mapped results from previous methods, or manually tone mapped images. Therefore, the quality of the learned tone mapper is limited to that of the selected training pairs. In contrast, our learned tone mapping operator is guided by a novel feature contrast masking loss that allows to represent in feature space

Table 4: Mean and standard deviation for TMQI (including $TMQI_S$ and $TMQI_N$), BTMQI and BRISQUE computed for different variations of components in our pipeline.

Input	$I_{HDR} \mapsto I_\mu$	Loss	TMQI (\uparrow)	$TMQI_S$ (\uparrow)	$TMQI_N$ (\uparrow)	BTMQI (\downarrow)	BRISQUE (\downarrow)
log	log	\mathcal{L}_{VGG}	0.5785 ± 0.1073	0.3759 ± 0.1965	0.0066 ± 0.0578	6.6324 ± 0.5731	42.5287 ± 7.4740
MEI	linear	\mathcal{L}_{FCM}	0.8776 ± 0.0610	0.8763 ± 0.0966	0.4555 ± 0.2642	3.9174 ± 1.4472	21.4104 ± 8.9001
MEI	Ada μ	\mathcal{L}_{VGG}	0.9178 ± 0.0462	0.8913 ± 0.0607	0.6596 ± 0.2470	3.1934 ± 1.1220	22.1773 ± 8.6294
linear	Ada μ	\mathcal{L}_{FCM}	0.8522 ± 0.0793	0.8166 ± 0.1206	0.4173 ± 0.2902	4.5342 ± 1.6754	30.2186 ± 13.7742
log	Ada μ	\mathcal{L}_{FCM}	0.8166 ± 0.0563	0.8921 ± 0.0774	0.0844 ± 0.0741	5.0949 ± 1.1779	22.4499 ± 8.3301
MEI	Ada μ	\mathcal{L}_{FCM}	0.9248 ± 0.0432	0.8938 ± 0.0582	0.7062 ± 0.2369	2.9065 ± 1.0481	21.2208 ± 8.5981

Figure 11: Example visualizations of our ablation study. Better contrast, brightness and detail reproduction is achieved with our full pipeline including multiple exposure inputs (MEI), adaptive μ -law compression (Ada μ), and \mathcal{L}_{FCM} loss.

important image contrast perception characteristics of the Human Visual System. The loss gives the network more freedom for compressing higher contrast while enhancing weak contrast, as it is often desirable for high quality HDR scene reproduction and an overall pleasant appearance, all in the context of local image content as modeled by neighborhood masking.

Limitations and future work In some rare cases our method may produce soft halos at high contrast edges as shown in Fig. 12. In future work we would like to experiment with edge stopping filters while deriving feature contrast and neighborhood masking for avoiding this issue. Nevertheless, these soft artifacts are not present in most of our results, and whenever present, they are not obviously visible as confirmed by both objective and subjective evaluations. As discussed in Trentacoste et al. [TMHD12], unsharp masking or weak counter-shading effects, similar to these soft halos, may be an effective way of enhancing perceived image contrast due to the Cornsweet illusion and are often employed for image enhancement. Our initial attempts at offline training (Sec. S2.3 in the supplemental) could not match the quality of our online training. This is somewhat expected since generalization is more challenging than dealing with a single image. This remains an interesting avenue for future work, since it would significantly decrease computation time. We would also like to investigate the utility of our adaptive μ -law compression for other learning-based applications that involve HDR content. Finally, another interesting avenue for future work would be employing our feature contrast masking model for other tasks such as image style transfer, where contrast characteristics in the source image should be conveyed to the target image.



Figure 12: Example exposure of the original HDR image (left) and our tone mapped result (right). The inset shows a failure case in which a soft halo appears around the edge of the mountain.

Acknowledgments

Ana Serrano acknowledges funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (project CHAMELEON, Grant no. 682080). We thank Thomas Leimkühler for his valuable suggestions and Aakanksha Rana for her help with DeepTMO.

References

- [AJP92] AHUMADA JR A. J., PETERSON H. A.: Luminance-model-based dct quantization for color image compression. In *Human Vision, Visual Processing, and Digital Display III* (1992), vol. 1666, Proc SPIE, pp. 365–374. 2
- [AMMS08] AYDIN T. O., MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: Dynamic range independent image quality assessment. *ACM Trans. Graph (SIGGRAPH 2008)* 27, 3 (2008), 1–10. 8

- [ANSAM21] ANDERSSON P., NILSSON J., SHIRLEY P., AKENINE-MÖLLER T.: Visualizing errors in rendered high dynamic range images. In *Eurographics Short Papers* (2021). 4
- [APBA18] ARTUSI A., POULI T., BANTERLE F., AKYÜZ A. O.: Automatic saturation correction for dynamic range management algorithms. *Signal Processing: Image Communication* 63 (2018), 100–112. 8
- [BADC17] BANTERLE F., ARTUSI A., DEBATTISTA K., CHALMERS A.: *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017. 2, 3, 8
- [BAS*16] BANTERLE F., ARTUSI A., SIKUDOVA E., LEDDA P., BASHFORD-ROGERS T., CHALMERS A., BLOJ M.: Mixing tone mapping operators on the GPU by differential zone mapping based on psychophysical experiments. *Signal Processing: Image Communication* 48 (2016), 50–62. 3
- [BM98] BOLIN M. R., MEYER G. W.: A perceptually based adaptive sampling algorithm. In *Proc ACM SIGGRAPH* (1998), pp. 299–309. 3
- [ČWNA08] ČADÍK M., WIMMER M., NEUMANN L., ARTUSI A.: Evaluation of hdr tone mapping methods using essential perceptual attributes. *Computers & Graphics* 32, 3 (2008), 330–349. 2
- [Dal92] DALY S. J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In *Human Vision, Visual Processing, and Digital Display III* (1992), vol. 1666, Proc SPIE, pp. 2–15. 2, 3, 6
- [DD02] DURAND F., DORSEY J.: Fast bilateral filtering for the display of high-dynamic-range images. In *ACM Trans. Graph* (2002), vol. 21, pp. 257–266. 3, 6, 8, 9
- [Deb18] DEBATTISTA K.: Application-specific tone mapping via genetic programming. In *Computer Graphics Forum* (2018), vol. 37, pp. 439–450. 3
- [DER*10] DIDYK P., EISEMANN E., RITSCHER T., MYSZKOWSKI K., SEIDEL H.-P.: Apparent display resolution enhancement for moving images. In *Proc ACM SIGGRAPH*. 2010, pp. 1–8. 3
- [DMAC03] DRAGO F., MYSZKOWSKI K., ANNEN T., CHIBA N.: Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer Graphics Forum* (2003), vol. 22, pp. 419–426. 3, 8, 9
- [DZLL00] DALY S. J., ZENG W., LI J., LEI S.: Visual masking in wavelet compression for JPEG-2000. In *Image and Video Communications and Processing* (2000), vol. 3974, Proc SPIE, pp. 66–80. 2, 3, 6, 7
- [EKD*17] EILERTSEN G., KRONANDER J., DENES G., MANTIUK R. K., UNGER J.: Hdr image reconstruction from a single exposure using deep cnns. *ACM Trans. Graph (SIGGRAPH Asia 2017)* 36, 6 (2017), 1–15. 4, 6
- [EKM17] ENDO Y., KANAMORI Y., MITANI J.: Deep reverse tone mapping. *ACM Trans. Graph (SIGGRAPH Asia 2017)* 36, 6 (2017). 4
- [Fai07] FAIRCHILD M. D.: The hdr photographic survey. In *Color and Imaging Conference* (2007), vol. 2007, pp. 233–238. 8, 9
- [FLW02] FATTAL R., LISCHINSKI D., WERMAN M.: Gradient domain high dynamic range compression. In *ACM Trans. Graph* (2002), vol. 21, pp. 249–256. 3, 6
- [Fol94] FOLEY J. M.: Human luminance pattern-vision mechanisms: masking experiments require a new model. *JOSA A* 11, 6 (1994), 1710–1719. 2
- [GJ21] GUO C., JIANG X.: Deep tone-mapping operator using image quality assessment inspired semi-supervised learning. *IEEE Access* 9 (2021), 73873–73889. 2, 4, 8, 9
- [GSY*17] GARDNER M.-A., SUNKAVALLI K., YUMER E., SHEN X., GAMBARETTO E., GAGNÉ C., LALONDE J.-F.: Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017). 8
- [GTM*12] GALLO O., TICO M., MANDUCHI R., GELFAND N., PULLI K.: Metering for exposure stacks. In *Computer Graphics Forum* (2012), vol. 31, pp. 479–488. 4
- [GWZ*16] GU K., WANG S., ZHAI G., MA S., YANG X., LIN W., ZHANG W., GAO W.: Blind quality assessment of tone-mapped images via analysis of information, naturalness, and structure. *IEEE Trans. on Multimedia* 18, 3 (2016), 432–443. 8
- [HDQ17] HOU X., DUAN J., QIU G.: Deep feature consistent deep image transformations: Downscaling, decolorization and hdr tone mapping. *arXiv preprint arXiv:1707.09482* (2017). 4
- [HZC*17] HOWARD A. G., ZHU M., CHEN B., KALENICHENKO D., WANG W., WEYAND T., ANDRETTA M., ADAM H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017). 8, 9
- [HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778. 5
- [JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *ECCV* (2016), pp. 694–711. 2, 4, 7, 8
- [JH93] JACK T., HOLLY R.: Tone reproduction for realistic images. *IEEE Computer Graphics and Applications* 13, 6 (1993), 42–48. 3
- [JKX*11] JINNO T., KAIDA H., XUE X., ADAMI N., OKUDA M.: μ -law based hdr coding and its error analysis. *IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences* 94, 3 (2011), 972–978. 6
- [JMB*14] JARABO A., MASIA B., BOUSSEAU A., PELLACINI F., GUTIERREZ D.: How do people edit light fields. *ACM Trans. Graph (SIGGRAPH 2014)* 33, 4 (2014), 4. 11
- [JYL19] JIANG X., YAO H., LIU D.: Nighttime image enhancement based on image decomposition. *Signal, Image and Video Processing* 13, 1 (2019), 189–197. 8
- [KW96] KINGDOM F. A., WHITTLE P.: Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing. *Vision research* 36, 6 (1996), 817–829. 2, 3, 6
- [LCTS05] LEDDA P., CHALMERS A., TROSCIANKO T., SEETZEN H.: Evaluation of tone mapping operators using a high dynamic range display. *ACM Trans. Graph* 24, 3 (2005), 640–648. 2
- [LF80] LEGGE G. E., FOLEY J. M.: Contrast masking in human vision. *Josa* 70, 12 (1980), 1458–1471. 2
- [LHLK17] LIM J., HEO M., LEE C., KIM C.-S.: Contrast enhancement of noisy low-light images based on structure-texture-noise decomposition. *Journal of Visual Communication and Image Representation* 45 (2017), 107–121. 8
- [LJZ18] LI H., JIA X., ZHANG L.: Clustering based content and color adaptive tone mapping. *CVIU* 168 (2018), 37–49. 3, 8, 9
- [LRP97] LARSON G. W., RUSHMEIER H., PIATKO C.: A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Trans Visual and Comp Graph* 3, 4 (1997), 291–306. 3
- [Lub95] LUBIN J.: A visual discrimination model for imaging system design and evaluation. In *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*. World Scientific, 1995, pp. 245–283. 3, 7
- [LXZ*18] LIANG Z., XU J., ZHANG D., CAO Z., ZHANG L.: A hybrid 11-10 layer decomposition model for tone mapping. In *CVPR* (2018), pp. 4758–4766. 3, 8, 9
- [MDC*21] MANTIUK R. K., DENES G., CHAPIRO A., KAPLANYAN A., RUFO G., BACHY R., LIAN T., PATNEY A.: Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Trans. Graph (SIGGRAPH 2021)* 40, 4 (2021), 1–19. 3, 6
- [MDK08] MANTIUK R., DALY S., KEROFKY L.: Display adaptive tone mapping. In *ACM Trans. Graph (SIGGRAPH 2008)*. 2008, pp. 1–10. 3
- [MKVR07] MERTENS T., KAUTZ J., VAN REETH F.: Exposure fusion. In *Pacific Graphics (PG)* (2007), pp. 382–390. 3, 8, 9

- [MMB12] MITTAL A., MOORTHY A. K., BOVIK A. C.: No-reference image quality assessment in the spatial domain. *IEEE Trans. on Image Processing* 21, 12 (2012), 4695–4708. 8
- [MMS06] MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: A perceptual framework for contrast processing of high dynamic range images. *ACM Trans. on Applied Perception (TAP)* 3, 3 (2006), 286–308. 3, 6, 8, 9
- [MMS15] MANTIUK R., MYSZKOWSKI K., SEIDEL H.-P.: *High dynamic range imaging*. Wiley Encyclopedia of Electrical and Electronics Engineering, 2015. 3, 6
- [MMTH09] MANTIUK R., MANTIUK R., TOMASZEWSKA A., HEIDRICH W.: Color correction for tone mapping. In *Computer Graphics Forum* (2009), vol. 28, pp. 193–202. 8
- [MS08] MANTIUK R., SEIDEL H.-P.: Modeling a generic tone-mapping operator. In *Computer Graphics Forum* (2008), vol. 27, pp. 699–708. 3
- [MYZW15] MA K., YEGANEH H., ZENG K., WANG Z.: High dynamic range image compression by optimizing tone mapped image quality index. *IEEE Trans on Image Proc* 24, 10 (2015), 3086–97. 3, 8, 9
- [NH10] NAIR V., HINTON G. E.: Rectified linear units improve restricted boltzmann machines. In *Icml* (2010). 4
- [Pal99] PALMER S. E.: *Vision science: Photons to phenomenology*. MIT press, 1999. 2
- [Pel90] PELI E.: Contrast in complex images. *JOSA A* 7, 10 (1990), 2032–2040. 3, 6
- [PKO*21] PANETTA K., KEZEBOU L., OLUDARE V., AGAIAN S., XIA Z.: Tmo-net: A parameter-free tone mapping operator using generative adversarial network, and performance benchmarking on large scale hdr dataset. *IEEE Access* 9 (2021), 39500–39517. 2, 3, 4, 6, 8, 9, 11
- [PSR17] PATEL V. A., SHAH P., RAMAN S.: A generative adversarial network for tone mapping hdr images. In *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics* (2017), Springer, pp. 220–231. 3
- [RC09] RAMAN S., CHAUDHURI S.: Bilateral filter based compositing for variable exposure photography. In *Eurographics (short papers)* (2009), pp. 1–4. 3
- [RGSS10] RUBINSTEIN M., GUTIERREZ D., SORKINE O., SHAMIR A.: A comparative study of image retargeting. In *Proc ACM SIGGRAPH Asia*. 2010, pp. 1–10. 11
- [RHD*10] REINHARD E., HEIDRICH W., DEBEVEC P., PATTANAİK S., WARD G., MYSZKOWSKI K.: *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann, 2010. 2, 3, 4
- [Rob66] ROBSON J. G.: Spatial and temporal contrast-sensitivity functions of the visual system. *JOSA* 56, 8 (1966), 1141–1142. 2
- [RPG99] RAMASUBRAMANIAN M., PATTANAİK S. N., GREENBERG D. P.: A perceptually based physical error metric for realistic image synthesis. In *Proc ACM SIGGRAPH* (1999), pp. 73–82. 3
- [RSSF02] REINHARD E., STARK M., SHIRLEY P., FERWERDA J.: Photographic tone reproduction for digital images. In *Proc ACM SIGGRAPH* (2002), pp. 267–276. 3, 8, 9
- [RSV*19] RANA A., SINGH P., VALENZISE G., DUFAUX F., KOMODAKIS N., SMOLIC A.: Deep tone mapping operator for high dynamic range images. *IEEE Trans. on Image Processing* 29 (2019), 1285–1298. 2, 3, 8, 9, 11
- [SJB09] SHAN Q., JIA J., BROWN M. S.: Globally optimized linear windowed tone mapping. *IEEE Trans Visual and Comp Graph* 16, 4 (2009), 663–675. 3, 8, 9
- [STKK20] SANTOS M. S., TSANG R., KHADEMI KALANTARI N.: Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Trans. Graph (SIGGRAPH 2020)* 39, 4 (7 2020). doi:10.1145/3386569.3392403. 6
- [STO16] SHIBATA T., TANAKA M., OKUTOMI M.: Gradient-domain image reconstruction framework with intensity-range and base-structure constraints. In *CVPR* (2016), pp. 2745–2753. 3, 8, 9
- [SWH*20] SONG W., WANG Y., HUANG D., LIOTTA A., PERRA C.: Enhancement of underwater images with statistical model of background light and optimization of transmission map. *IEEE Trans. on Broadcasting* 66, 1 (2020), 153–169. 8
- [SWL*21] SU C.-C., WANG R., LIN H.-J., LIU Y.-L., CHEN C.-P., CHANG Y.-L., PEI S.-C.: Explorable tone mapping operators. In *ICPR* (2021), pp. 10320–10326. 3, 4, 11
- [SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 2, 4
- [TAKW*19] TURSUN O. T., ARABADZHIYSKA-KOLEVA E., WERNIKOWSKI M., MANTIUK R., SEIDEL H.-P., MYSZKOWSKI K., DIDYK P.: Luminance-contrast-aware foveated rendering. *ACM Trans. Graph (SIGGRAPH 2019)* 38, 4 (2019), 1–14. 3
- [TMHD12] TRENTACOSTE M., MANTIUK R., HEIDRICH W., DUFROT F.: Unsharp Masking, Countershading and Halos: Enhancements or Artifacts? In *Proc. Eurographics* (2012), p. to appear. 12
- [TT99] TUMBLIN J., TURK G.: LCIS: A boundary hierarchy for detail-preserving contrast reduction. In *Proc ACM SIGGRAPH* (1999), pp. 83–90. 8
- [Wat89] WATSON A. B.: Receptive fields and visual representations. In *Human Vision, Visual Processing, and Digital Display* (1989), vol. 1077, Proc SPIE, pp. 190–197. 2
- [Wat93] WATSON A. B.: Visually optimal dct quantization matrices for individual images. In *[Proceedings] DCC93: Data Compression Conference* (1993), IEEE, pp. 178–187. 3, 6
- [WS97] WATSON A. B., SOLOMON J. A.: Model of visual contrast gain control and pattern masking. *JOSA A* 14, 9 (1997), 2379–2391. 2
- [WXTT] WU S., XU J., TAI Y.-W., TANG C.-K.: Deep high dynamic range imaging with large foreground motions. In *ECCV, year = 2018*. 6
- [XZR18] XU K., ZHANG Z., REN F.: Lapran: A scalable laplacian pyramid reconstructive adversarial network for flexible compressive sensing reconstruction. In *ECCV* (2018), pp. 485–500. 5
- [YBMS05] YOSHIDA A., BLANZ V., MYSZKOWSKI K., SEIDEL H.-P.: Perceptual evaluation of tone mapping operators with real-world scenes. In *Human Vision and Electronic Imaging X* (2005), vol. 5666, Proc SPIE, pp. 192–203. 2
- [YJS*21] YI S., JEON D. S., SERRANO A., JEONG S.-Y., KIM H.-Y., GUTIERREZ D., KIM M. H.: Modeling surround-aware contrast sensitivity. In *Eurographics Symposium on Rendering* (2021). 3
- [YW12] YEGANEH H., WANG Z.: Objective quality assessment of tone-mapped images. *IEEE Trans. on Image Processing* 22, 2 (2012), 657–667. 2, 3, 6, 8
- [YXS*18] YANG X., XU K., SONG Y., ZHANG Q., WEI X., LAU R. W.: Image correction via deep reciprocating hdr transformation. In *CVPR* (2018), pp. 1798–1807. 4
- [ZWZW19] ZHANG N., WANG C., ZHAO Y., WANG R.: Deep tone mapping network in hsv color space. In *2019 IEEE Visual Communications and Image Processing (VCIP)* (2019), pp. 1–4. 2, 3, 4, 11
- [ZZP*17] ZHU J.-Y., ZHANG R., PATHAK D., DARRELL T., EFROS A. A., WANG O., SHECHTMAN E.: Multimodal image-to-image translation by enforcing bi-cycle consistency. In *Advances in neural information processing systems* (2017), pp. 465–476. 3
- [ZZWW21] ZHANG N., ZHAO Y., WANG C., WANG R.: A real-time semi-supervised deep tone mapping network. *IEEE Trans. on Multimedia* (2021), 1–1. 4, 8, 9