IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Deep Tone-mapping Operator Using Image Quality Assessment Inspired Semi-supervised Learning

Cheng Guo[1], Xiuhua Jiang[2]

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
[2]Laboratory of Digital Video Quality Assessment, Communication University of China, Beijing 100024, China

Corresponding author: Cheng Guo (e-mail: guocheng@cuc.edu.cn).

ABSTRACT Tone-mapping operator (TMO) is intended to convert high dynamic range (HDR) content into a lower dynamic range so that it can be displayed on a standard dynamic range (SDR) device. The tone-mapped result of HDR content is usually stored as SDR image. For different HDR scenes, traditional TMOs are able to obtain a satisfying SDR image only under manually fine-tuned parameters. In this paper, we address this problem by proposing a learning-based TMO using deep convolutional neural network (CNN). We explore different CNN structure and adopt multi-scale and multi-branch fully convolutional design. When training deep CNN, we introduce image quality assessments (IQA), specifically, tone-mapped image quality assessment, and implement it as semi-supervised loss terms. We discuss and prove the effectiveness of semi-supervised loss terms, CNN structure, data pre-processing, etc. by several experiments. Finally, we demonstrate that our approach can produce appealing results under diversified HDR scenes.

INDEX TERMS High Dynamic Range, Tone-mapping, Convolutional Neural Network, Semi-supervised Learning, Image Quality Assessment

I. INTRODUCTION

Dynamic range of scene is defined as the ratio of maximum luminance to the minimum. Real scenes have a wide range of luminance ranging from 10^{-4} to 10^{5} cd/m^2, thus the dynamic range of specific scene can be up to 10^{9}, which is far beyond the capture and display capability of standard dynamic range (SDR) devices. High dynamic range (HDR) image can record real-world luminance in a photometrically linear [1] and scene-referred manner, and store it in 32-bit float-point data encapsulated in .hdr or .exr format, in most cases. However, display devices capable of rendering HDR image are still costly. Thus, tone mapping operator (TMO) capable of approximate the appearance of HDR content in traditional SDR displays has become the prerequisite under most circumstances.

The aim of TMO is reproducing a perception that matches real-world scene as possible [1], in other words, selectively maintaining some features from the original HDR scene and producing a reduced-information version [3] of it. But most traditional TMOs are parametric dependent to yield a visually plausible results due to the diversity of HDR scenes. Artifacts like over-enhancement, over-stylization, halo effect and blurring are common in tone-mapped SDR images produced by traditional TMO with improper parameters.

This naturally raises the idea of scene-adaptive TMO which can generate high quality tone-mapped images under diversified HDR scenes. With the emergence of deep learning and its success on image transformation tasks, we are able to learn a deep convolutional neural network (CNN) based scene-adaptive TMO using easily available HDR data.

Unlike other tasks such as classification, object detection and style transfer etc., high-level semantic features undergo nearly no change during tone-mapping. Hence, for tone mapping, fully convolutional layers (where tensor's height and width undergo no change) is enough, and U-net [4] (encoder-decoder) architecture becomes improper especially when dealing high-resolution images [5]. Although fully convolutional architecture has an exclusive advantage in arbitrary input size, it does suffer from the shortcoming of insufficient global comprehension brought by limited receptive field. To overcome this, different approaches have been proposed by several related works. Based on insight and experiments, our fully convolutional network decomposes input into 2 components with different scales, and send them into separate task-specific CNN branches and assign another

CNN to polish the merged output.

Training, i.e., optimizing CNN's parameters, is another predominant aspect in deep learning. Accustomed to the routine of previous works on image transformation, almost every related work uses supervised training, i.e., calculating loss function between output and label images (both in SDR, in the case of tone-mapping). When it comes to image quality assessment (IQA), specifically, tone-mapped image quality assessment where objective score is calculated between output and input images (SDR vs. HDR), there comes a natural idea whether we can directly optimize the quality score. Inspired by this, 2 terms in our loss function are calculated in a IQA way, i.e., output vs. input (without label, unsupervised). Our training can be broadly termed semi-supervised because both supervised and unsupervised loss terms are involved. Since supervised losses require paired label images, we collect a training set containing high quality label SDR images in a unique and elaborate way.

We systematically study the CNN structure, training method, etc. of all HDR related deep CNNs before designing our method. Based on this, several other improvements such as multi-pass [8] or multi-group [9] convolution and instance normalization [10] were also applied.

In a nutshell, our works are:

1) Proposing a learning-based TMO using CNN.
2) To the best of our knowledge, we first introduce IQA inspired semi-supervised training in HDR related deep CNN. We made a small step bridging the gap between perceptual quality and HDR related CNN.
3) In semantic-free task, we explore a distinctive low-cost and flexible way to strengthen the global comprehension of CNN, i.e., multi-scale decomposing and multi-group convolution on fully convolutional layers.

The rest of this paper is organized as follows. Section II reviews related works. Section III details the network structure and training of proposed method. Ablation studies on semi-supervised loss terms etc., extra experiments, and comparison with other methods are presented in Section IV. Finally, Section V remarks the present and future work.

II. RELATED WORK

A. TRADITIONAL TONE-MAPPING OPETATOR

A considerable amount of traditional TMOs have been proposed in last 2 decades. They can be mathematically explicitly defined, and can be classified into 4 categories namely global, local, frequency/gradient and segmentation [2]. Global approach such as Ward94 [11], Larson97 [12], Pattanaik00. [13], Drago03 [14], Mantiuk06 [15] and iCAM06 [16] apply same operation to all pixels in HDR image. Local approach like Reinhard02 [17] process pixel value based on its neighbors. Frequency approach Durand02 [18] use bilateral filter to decompose input image into base and detail frequency components, and process them separately. Gradient approach Fattal02 [19] process pixels in gradient domain. Segmentation method Krawczyk05 [20] apply different operations on segmented image regions.

More diversified TMOs have sprung up in last decade. Li et al. [21] combine tone-mapping with visual saliency. Some TMOs are designed for application scenarios other than human perception. Yang et al. [22] is for object detection, and Rana et al. [23] is for image matching. Despite artifacts like over-enhancement, over-stylization, halo effect and blurring etc. brought by traditional TMOs, some of their ideas such as decomposing in [18] are still affecting deep CNN based TMOs.

B. DEEP CNN BASED TONE-MAPPING OPETATOR

So far, there are 5 deep CNN based TMOs. Patel et al. [24] proposed a generative adversarial network [28] (GAN) based TMO whose generator is a 14-layers encoder-decoder similar to U-net [4]. They trained their network with 957 HDR-SDR (label) image pairs. These labels were generated by the traditional TMO who gives the best TMQI [6] over others (TMQI is an objective tone-mapped images quality assessment method, see §II.D for details). This label generation is referred to as "best TMQI" in Table I.

CNN in Yang et al. [25]'s method only contains fully convolutional layers: 2 same 5-layers branches to process different component and a 10-layers CNN to polish the merged output. Single-channel luminance of HDR image was transferred into logarithm domain and then decomposed into base/detail component in different scales by Laplacian pyramid. Their training set was fine-tuned, evaluated and selected by photographers and volunteers (denoted as "manually fine-tuned" in Table I).

Zhang et al. [26] applied multi-scale 2-branch CNN similar to [25]. Their 9-layer large-scale branch with dilated convolution [55] is responsible for processing details, 5-layer small-scale encoder branch is for global information, and 2-layer "tail" is for merging the output of 2 branches. Their loss function contains variants of l_1 norm (on gradient magnitude map/Gaussian filtered image, measuring local/ global detail) and other customized terms for their binocular vision task i.e., producing 2 tone-mapped images with their own emphasis.

Rana et al. [27] proposed a method named DeepTMO, based on conditional generative adversarial network [29] (cGAN). They tried 4 combinations of generator and discriminator, and applied the best generator containing a small-scale branch with 15-layer U-net and a large-scale branch with 7 fully convolutional layer. Besides cGAN term, their loss function also contains l_1 norm and perceptual loss extracted form Siamese pre-trained 19-layer VGG-Net [30] (denoted "VGG" in Table I).

Zhang et al. [34]'s method converts HDR image into HSV color space. S and V channels are processed by CNN while H is preserved to avoid hue shift during tone-mapping. Their training was supervised by loss terms including SSIM [35], using photographer-fine-tuned label images.

There are 3 works where tone-mapping was implemented in part of their CNN. Sheth et al. [31] processed HDR images by a simple 4-layer convolution separately on 4 channels in Lab color space. Hou et al. [3] processed luminance channel

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

of HDR image using 4-layer convolution, note that this was the only HDR related CNN involving unsupervised training. Yang et al. [33] used a 12-layer U-net to transfer their intermediate-stage HDR image into enhanced tone-mapped image.

Comparison of deep CNN based TMOs ("deep TMOs" for short) is listed in Table I. Here, "conv5" represents 5 fully convolutional layers, and "Unet14" means encoder-decoder structure of totally 14 convolutional layers, etc. In 6th column, "reg" means regularization term to prevent over-fitting, "GAN" is the loss terms of specific GAN.

From the 3rd column we know that some deep TMOs ([25], [26], [27]) were influenced by traditional Durand02 TMO [18] in that they assigned separate CNN branches to handle different frequency components. Specifically, they use CNN with large receptive filed to handle global/low-frequency component, and CNN with small receptive filed to deal with local/high-frequency component. C=1 in 5th column mean that only luminance channel is processed by CNN (except [31]), and the color of output was reconstructed form the ratio of original HDR image using one of the methods in [32], which is a common practice in traditional TMOs.

C. OTHER HDR RELATED DEEP CNN

Other HDR related deep CNNs including reverse tone mapping operator (rTMO, single SDR to HDR) are listed in Table I as well, since there are innovations worth learning. Here, "simulated exposure" means shooting linear-light HDR image with simulated camera response functions (CRFs) to get non-linear SDR image, "EV0 in MEF stack" will be detailed in §III.B.2.

In HDR related CNN, decomposing was first applied by Eilertsen et al. [36]. However, their illuminance/reflectance (I/R) [37] decomposing was implemented in loss function (different weight for I/R component, same as [46]) rather than network structure. Later work Marnerides et al [39] first applied multi-branch structure by assigning different kernel size in each branch to focus on global/local features. Later, decomposing and multi-branch had become more popular, as they were used by [31], [25], [26], [27], [34], [40], [41], [42], [45], [5] and [47]. Moreover, Wang et al. [42] first tackled denoising, while Xu et al. [46] first introduced 3D convolution considering temporal information of HDR videos.

Latest works [5], [47] and [56] combined HDR (rTMO) with super resolution (SR). It was in multi-task exploration (rTMO + SR) that they found U-net structure no longer proper, the conclusion which is stated in §I. Several other mechanisms which are frequently used in other computer vision (CV) tasks were introduced to HDR related CNN (MEF) too. Yan et al. first introduced spatial attention (mask) in [50], and multi-pass convolution in [8], to gain a better comprehension of global information for CNN in both works.

TABLE I
COMPARISON OF HDR RELATED DEPP CNNS. HERE, B AND C REPRESENT THE NUMBER OF BRANCHES AND CHANNELS, RESPECTIVELY.

Type	Method	B	CNN structure	C	Domain	Loss function	Dataset size	SDR generating
TMO (our task)	Sheth et al. [31]	4	4×conv4	1	-	-	958	best TMQI
	Patel et al. [24]	1	Unet14	3	-	l-1reg, l-2, GAN	958	best TMQI
	Hou et al. [3]	1	conv4	1	logarithm	VGG19	unsupervised	-
	Yang et al. [33]	1	Unet12	3	logarithm	l-2	not in HDR	-
	Yang et al.[25]	2	2×conv5+conv10	1	logarithm	l-2, l-2reg, VGG16	2100	manually fine-tune
	Zhang et al. [26]	2	conv9+encoder5	1	logarithm	variants of l-1, etc.	3620	Durand02 [18] TMO
	Rana et al. [27]	2	conv7+Unet15+conv2	1	-	l-1reg, cGAN, VGG19	698	best TMQI
	Zhang et al. [34]	2	conv2+Unet12	2	logarithm	l-2, SSIM, wGAN	1000	manually fine-tune
rTMO	Eilertsen et al.[36]	1	Unet27	3	logarithm	weighted l-2	1211	simulated exposure
	Zhang et al. [38]	1	Unet10	3	nonlinear	l-1	50000	simultaneous shot
	Marnerides et al. [39]	3	conv+encoder	3	-	l-1, °cosine-similarity	1013	random traditional TMO
	Jang et al. [40]	2	2×Unet31	3&1	-	l-2, ΔE_{76}	8156	EV0 SDR image in MEF stack
	Kinoshita et al. [41]	2	Unet with 2 encoder	3	-	nonlinear l-1, cosine-similarity	336	simulated exposure
	Wang et al. [42]	2	2×conv8+Unet16	3	logarithm	l-1, VGG, weighted l-2, LS-GAN	3000	simulated exposure
	Santos et al. [43]	1	Unet14	3	logarithm	l-1, gram matrix [44], VGG	2000	simulated exposure
	Kim et al. [45]	2	conv12	3	-	l-2, l-2reg	7268	YouTube
	Xu et al. [46]	1	3D-Unet	3	logarithm	weighted l-2, VGG	360videos	traditional TMOs
rTMO+ SR	Kim et al. [5]	2	conv+spacial mask	3	-	l-2	59818	YouTube
	Kim et al. [47]	3	3×(dynamic)conv11	3	-	l-2, RaHingeGAN	59818	YouTube
	Zeng et al. [56]	1	(multi-group)conv17	1	YUV	l-1, ESR-GAN	23229	Reinhard02 [17] TMO

There were also several training innovations. Marnerides et al [39] first considered color information in training by introducing cosine-similarity loss term along different RGB channels. Jang et al. [40] introduced ΔE_{76} color difference formula in their loss function, which was the first human perception related loss term in HDR related CNN. Santos et al. [43] first introduced gram matrix [44] to measure style loss, and feature masking to emphasize global difference.

Currently, the only simultaneously-shot HDR-SDR pair was proposed by Zhang et al [38]. However, its low-resolution (64*128) has limited their application in future works because only global feature is contained. Note that, different from linear HDR usually applied in photography, image-based rendering and medicine, HDR images/videos in [40], [45] , [46], [5], [47] and [56] were non-linear transformed by PQ [48] / HLG [49] optic-electronic transfer function (OETF) which is used in consumer grade HDR television and HDR films. Kim et al. found that traditional TMOs designed for linear-light HDR images perform poorly on OETF transferred non-linear HDR content, thus they decided to collect SDR counterpart using YouTube default method ([45], [5] and [47]).

D. TONE-MAPPED IMAGE QUALITY ASSESSMENT

As a branch of IQA, tone-mapped image quality assessment treats HDR as original image while SDR as distorted one. The quality of tone-mapped image can be measured in a full-reference (FR) way i.e., comparing HDR with SDR images, or in a non-reference (NR) way i.e., only assessing tone-mapped SDR image [51].

To be implemented in deep learning as loss function, IQA method need to be mathematically explicit and differentiable to suit the chain derivation rule in backpropagation. TMQI [6] (tone-mapped image quality index) assesses the quality of tone-mapped images from 2 aspects: FR structure fidelity between HDR and SDR image, and NR statistical naturalness of tone-mapped SDR image. The latter is non-differentiable, thus excluded from our work. While the former local structure fidelity term of the former is improved from SSIM [35]:

$$SF_{local}(x,y)=\frac{2\sigma_x'\sigma_y'+C_1}{\sigma_x'^2+\sigma_y'^2+C_1}\cdot\frac{\sigma_{xy}+C_2}{\sigma_x\sigma_y+C_2} \quad (1)$$

where σ_x, σ_y and σ_{xy} are the local standard deviations and cross correlation between corresponding patches in HDR and SDR images, respectively. Here, stabilizing constants are set to default $C_1=0.01$ and $C_2=10$. Superscript of σ_x', σ_y' represent a non-linear normalization using a cumulative distribution function (CDF) of normal distribution, σ_x' and σ_y' are to replace the original luminance term using mean value μ which will definitely change dramatically in tone-mapping. Structure fidelity $SF(X,Y)$ between 2 images is derived from averaging the $SF_{local}(x,y)$ of all 11*11 sliding patch. Finally, $SF(X,Y)$ of all different scales are calculated into structure fidelity part of TMQI using same coefficients as MS-SSIM [52].

Nafchi et al. proposed a FR method FSITM [7] (feature similarity index for tone-mapped images) measuring the phase congruency between HDR and SDR images via locally-weighted mean phase angle. Apart from above, latest NR methods like BTMQI [73] and BLIQUE-TMI [70] involving more comprehensive feature extraction and regression reached better performance (higher correlation with subjective score), but they're either non-differentiable or too complex to be implemented as loss function in CNN. More about tone-mapped image quality assessment can be found in survey [53].

III. PROPOSED METHOD

A. NETWORK STRUCTURE

As is illustrated in Fig. 1, CNN in our method consists of 3 sub-networks namely Full-scale Local Branch (N_L), Small-scale Global Branch (N_G) and Polishing Network (N_P). Input HDR image (H) in 3 RGB channels is first decomposed into full-scale detail component (H_D) and small-scale base component (H_B), and sent to N_L and N_G respectively to yield intermediate S_D and S_B. Then, S_D and up-scaled S_B (S_{B_F}) are pixel-wise added, thus recomposed as the input of N_P. Finally, output tone-mapped SDR image (S) is given by N_P.

The Full-scale Local Branch (N_L) consists of 5 fully convolutional layers (blue box in Fig. 1), it is responsible for processing detail information thus not responsible for handling global information. Under this guideline, we add 2 skip-connections in order to maintain image's structure, and apply only 3×3 convolutional kernels of relatively small receptive filed to focus on detail information.

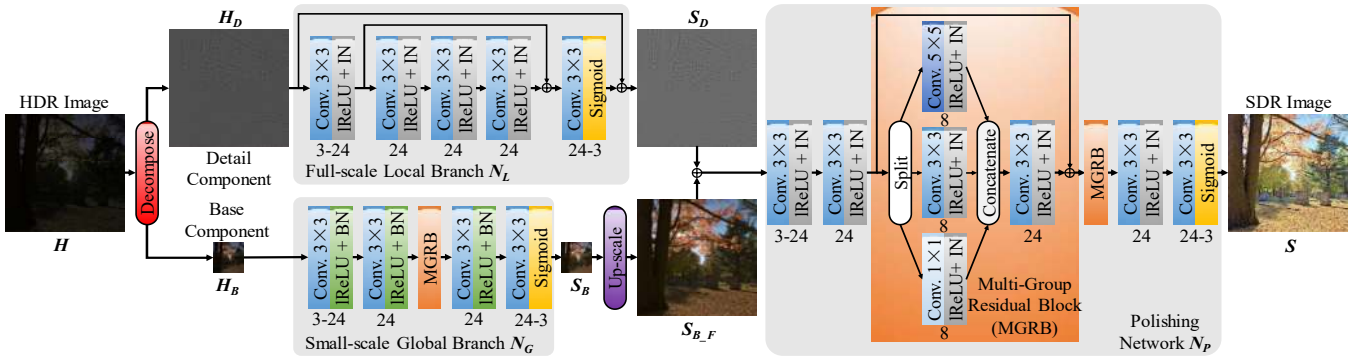


FIGURE 1. Overview of our multi-scale and multi-branch CNN structure. Here, HDR image (H) is linear mapped for display.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

The Small-scale Global Branch (N_G) contains a Multi-Group Residual Block (MGRB, orange box in Fig. 1) and 4 convolutional layers before and after. The MGRB aims for enhancing global comprehension by enlarging receptive filed, it will be detailed in §III.A.2.

The Polishing Network (N_P) consists of 2 MGRBs and 4 convolutional layers (totally 8-layers). Among related works where filter decomposing and branch network structure were adopted ([31], [25], [42], [45], [5] and [47]), polishing network is used by 4 of them ([25], [45], [5] and [47]). We took the same design, and an experiment was later conducted to prove its necessity.

In Fig. 2, the number below blue box (e.g., 24-3) represents "number of input-output channel" of current layer, a single number for short if the above two are same. While the number of channels changes, tensor/image's size keep unchanged because our CNN involves no encoder-decoder structure (no deconvolution layer), and strides for all fully convolutional layers are set to 1 (with symmetric padding).

All neurons excluding those in the last layers of N_L, N_G and N_P are activated by leaky rectified linear unit (lReLU) with a slope of 0.2 to accelerate computing and avoid vanishing gradient. Sigmoid activations are applied after 3 last layers to increase network's non-linearity. Front layers are followed by normalization, we applied batch normalization [58] (BN, green box in Fig. 1) for N_G where image size is small thus large batch-size can be applied when training. Since instance normalization [10] was proven helpful for small batch-size by [27], we used it (IN, gray box) for N_L and N_P where batch-size is limited by images size.

1) DECOMPOSING STRATEGY

Our decomposing strategy is designed based on the following considerations. As can be concluded from related works, there are 3 types of multi-branch strategies. First, resizing i.e., processing full-size input image with local/detail/large-scale branch while sending down-sampled image into global/small-scale branch ([26], [27], [39] and [41]). Second, filtering i.e., decomposing image into base and detail components using filter (usually edge-preserving filter like bilateral filter) and respectively sending them into global and local/detail branches ([31], [42], [45], [5] and [47]). Third, "decomposing" by color channel ([31], [34] and [40]). We considered the 2nd idea feasible because task-specific branches can focus on different image components using customized structure.

Distinctively, Yang et al. [25] combined resizing and filtering using Laplacian pyramid. Here, specific level is the residual between corresponding level in Gaussian pyramid and its up-sampled blurred next level. The lowest level of a 4-level Laplacian pyramid is chosen as the input of global branch, while rest levels are reformulated as the input of local/detail branch [54]. In this case, their global branch receives a 1/16 down-scaled condensed image, thus the computational cost is significantly reduced.

Hence, we decided to combine image pyramid with our idea of decomposing. First, we found that the 1/16 scale of the lowest level of 4-level pyramid in [25] is too small and thus too blurry when recomposing, therefore we use a 3-level pyramid $\{l_0, l_1, l_2, l_3\}$ thus the base component (H_B) is of 1/8 scale. Second, we replace the Gaussian filter on the first level of "Gaussian" pyramid $\{g_0\}$ with an edge-preserving bilateral filter, and directly subtract levels in "Gaussian" pyramid $\{g_0, g_1, g_2\}$ with its filtered image so that the highest level in our "Laplacian" pyramid $\{l_0\}$ is exactly the same as the detail component of bilateral filter decomposition (same as [31]). Our decomposing is illustrated in right(red) part in Fig. 2.

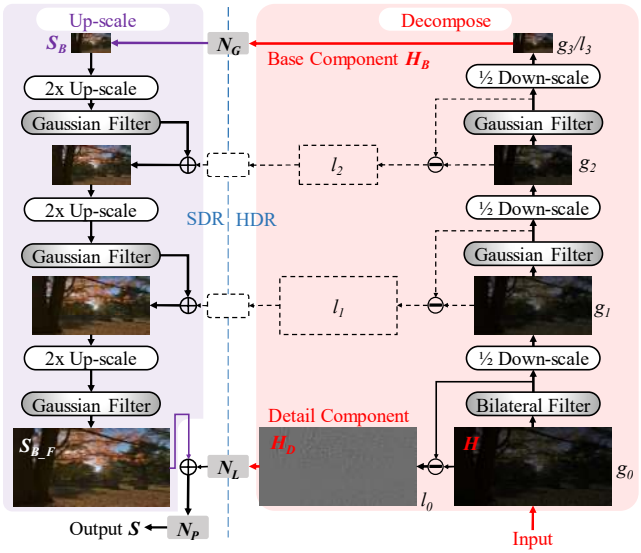


FIGURE 2. Our decomposing and recomposing method using a modified pyramid. l_i/g_i represent the i th level of "Gaussian"/"Laplacian" pyramid, dashed lines represent parts which are unused.

The prototype of our up-scale and recomposing method is also image pyramid, but the difference is that only the highest and lowest level $\{l_0, l_3\}$ (detail and base component) are processed by CNN and utilized in pyramid reconstruction. We bypass those rest middle layers $\{l_1, l_2\}$ since [54] found that the performance degradation caused by their missing is negligible compared with the reduction of network complexity.

2) MULTI-GROUP RESIDUAL BLOCK

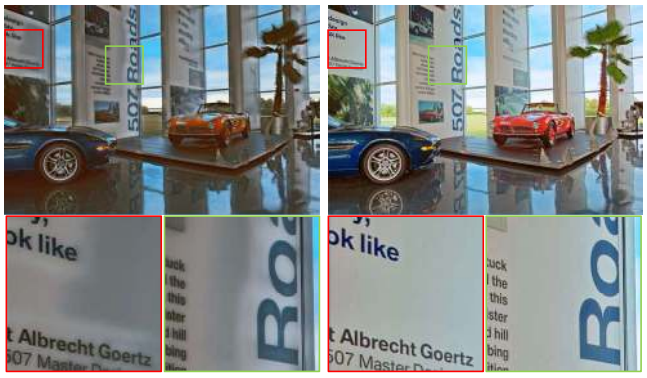
As is illustrated with red box in Fig. 1, our Multi-Group Residual Block (MGRB) first split input image/tensor into 3 groups with same channel number (24 to 3*8), and separately convolve then with 1×1, 3×3 and 5×5 kernel. Channel number of each group stay unchanged during convolution, so that they can be subsequently concatenated in the original order. Then, concatenated image/tensor is followed by a 3×3 convolution same as those outside MGRB. Finally, output is pixel-wise added with input residual.

Our MGRB is designed based on following considerations. Similar to other methods, global branch in our multi-branch structure is responsible for understanding global luminance

distribution, thus is supposed to have a global comprehension. Global or small-scale branch of [26], [39], [41] and [42] contain encoder structures to extract global features. Dilated convolution [55] was further used in [26] and [39] to enlarge their receptive field, thus strengthen the global comprehension of branches other than global branch. While encoder structure was regarded by above methods as capable of capturing global features, it has several intrinsic shortcomings. First, encoder-decoder (U-net) structure relies badly on skip-connections to keep structural consistency and avoid checkerboard artifacts. Second, encoder-decoder structure requires a fixed size input which is usually obtained by extra resize operation.

To overcome the second shortcoming, our whole network including MGRB only contains fully convolutional layers which have no limit on input size. To overcome the first shortcoming, we decided to deprecate U-net structure. Hence, an extra task of strengthening the global comprehension of Full-scale Local Branch $\boldsymbol{N_L}$ was posed. Other methods tackled this by introducing multi-group (in [56]) or multi-pass (in [8]) convolution where different kernel-size are assigned to each pass, using spatial attention mechanism (in [43], [50] and [5]), using 1-D and 2-D dynamic convolution (in [47]), and improving encoder-decoder structure (U-net) [57].

Different from methods above, we decided to strike a trade-off between performance and complexity. Multi-group conv. [9] split tensor along channel dimension and send them into separate groups, while multi-pass conv. [8] just copy tensor into different passes. To reduce memory cost and the number of parameters, we took multi-group convolution as prototype of MGRB. Also, to enlarge the receptive field, we transplant the characteristic of different kernel-size of multi-pass convolution onto MGRB.



(a) Without MGRB (b) With MGRB

FIGURE 3. The effect of multi-group residual block (MGRB). As seen in (b), MGRB can remove halo artifacts around edges.

Fig. 3 reveals the immediate effect of MGRB, i.e., the effective removal of ripple/halo artifacts at edges which are brought by limited receptive filed. Impact of MGRB will be further quantitatively evaluated in §IV.B.2.

B. TRAINING

Our CNN can be formulated as:

$$\boldsymbol{S}=N(\boldsymbol{H},\theta_L,\theta_G,\theta_P) \tag{2}$$

where N is whole network, θ_L, θ_G and θ_P represent model parameters in $\boldsymbol{N_L}$, $\boldsymbol{N_G}$ and $\boldsymbol{N_P}$, respectively. Then, training is to find the θ_L, θ_G and θ_P which minimize the loss function.

We adopted 2-step training strategy. Step 1 is the pre-training of $\boldsymbol{N_L}$ and $\boldsymbol{N_G}$, i.e., θ_L and θ_G were first optimized and θ_P were frozen. As shown in Fig. 4, in Step 1, label SDR image ($\boldsymbol{SL}$) were decomposed into detail component ($\boldsymbol{SL_D}$) and base component ($\boldsymbol{SL_B}$) using the same method applied to input HDR image. Then, supervised loss terms were separately calculated on label vs. output i.e., $\boldsymbol{SL_B}$ vs. $\boldsymbol{S_B}$, and $\boldsymbol{SL_D}$ vs. $\boldsymbol{S_D}$. Meanwhile, unsupervised loss terms were separately calculated on inputs vs. outputs i.e., $\boldsymbol{H_B}$ vs. $\boldsymbol{S_B}$, and $\boldsymbol{H_D}$ vs. $\boldsymbol{S_D}$. Step 2 is the end-to-end synchronous training of whole network, i.e., both θ_L, θ_G and θ_P were optimized. Here, supervised loss terms were calculated on $\boldsymbol{SL}$ vs. $\boldsymbol{S}$, while unsupervised loss terms were calculated on $\boldsymbol{H}$ vs. $\boldsymbol{S}$.

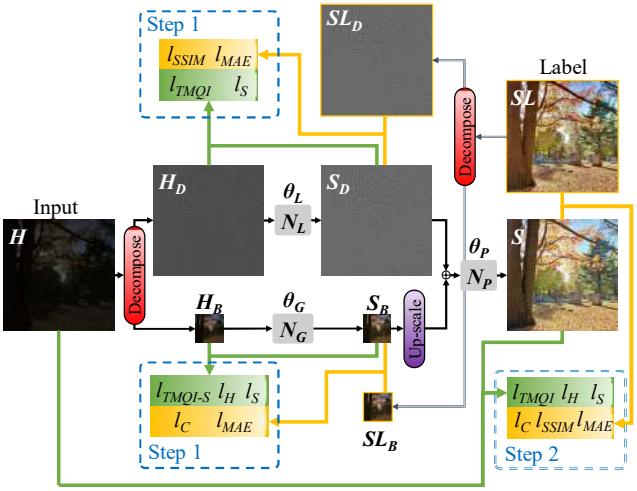


FIGURE 4. Semi-supervised two-step training. Yellow line and box denote the calculation of supervised loss terms, while green box and line denote unsupervised ones.

Similar to [25], [42], [45], [5] and [47] where decomposing, multi-branch and multi-step training were also applied, our 2-step training share the same intention of simplifying training and making network more interpretable. The motivation and implementation details of supervised and unsupervised loss terms are introduced below.

1) SEMI-SUPERVISED LOSS FUNCTION

Rather than using both labeled and unlabeled data, semi-supervised training here means simultaneously applying both supervised and unsupervised loss term on labeled data. There are 2 IQA inspired unsupervised terms ($\boldsymbol{l_{TMQI}}$ and $\boldsymbol{l_H}$), 2 IQA inspired supervised terms ($\boldsymbol{l_C}$ and $\boldsymbol{l_{SSIM}}$), and 2 conventional losses ($\boldsymbol{l_P}$ and $\boldsymbol{l_{MAE}}$):

TMQI loss (unsupervised). The TMQI (upper bound by 1) of output tone-mapped SDR image was optimized (maximized) by minimizing TMQI loss (l_{TMQI}) which consists of the differentiable structural fidelity part of TMQI [6]. Take Step 2 for example:

$$l_{TMQI}=1-\prod_{l=1}^{5} SF(S,H)^{\beta_l} \quad (3)$$

where $\beta_l=\{0.0448, 0.2856, 0.3001, 0.2363, 0.1333\}$ are weights of different scales, same as MS-SSIM [52]. Other implementation details have been introduced in eqn. (1) and §II.D Note that, it's inappropriate to apply multi-scale-implemented l_{TMQI} in small-scale global branch N_G (of 32×32 training patch), thus l_{TMQI} here was calculated in single-scale fashion (denoted $l_{TMQI\text{-}S}$):

$$l_{TMQI\text{-}S}=1-SF(S_B,H_B) \quad (4)$$

Hue shift loss (unsupervised). For most TMOs where color gamut mapping is not involved, color appearance management has long been an unsolved issue. Mantiuk et al [32] explored several color correction methods for tone-mapping and found that chroma change is more acceptable compared with hue shift. We didn't adopt their method because it's designed for TMO processing single luminance channel while our method directly handles 3 channel RGB image. But inspired by their finding, we started to limit the hue shift by minimizing hue shift loss (l_H).

Since CIE 1976 $L^*a^*b^*$ color space and its derivative $L^*C^*h^*$ have cross-contamination around blue color [62], i.e., chroma (C^*) around blue will change even if the hue (h^*) is restricted during tone-mapping (L^* decreasing), we turn to defined l_H in IPT color space. To be converted into IPT color space, pixels in RGB value need to be converted to XYZ tristimulus value based on the chromaticity coordinates of its color gamut. We assume that the destination color gamut of all output SDR images is sRGB, meanwhile, the source color gamut of input HDR images in training set [59] and [60] (see §III.B.2) is also sRGB. For single pixel p in sRGB color gamut:

$$\begin{bmatrix}X\\Y\\Z\end{bmatrix}=\begin{bmatrix}0.4124 & 0.3576 & 0.1805\\0.2126 & 0.7152 & 0.0722\\0.0193 & 0.1192 & 0.9505\end{bmatrix}\times\begin{bmatrix}R\\G\\B\end{bmatrix} \quad (5)$$

for Fairchild [1] HDR dataset where the color gamut of source HDR capturing devise was measured:

$$\begin{bmatrix}X\\Y\\Z\end{bmatrix}=\begin{bmatrix}0.4024 & 0.4610 & 0.0871\\0.1904 & 0.7646 & 0.0450\\-0.0249 & 0.1264 & 0.9873\end{bmatrix}\times\begin{bmatrix}R\\G\\B\end{bmatrix} \quad (6)$$

Then, XYZ tristimulus were converted into intermediate LMS color space:

$$\begin{bmatrix}L\\M\\S\end{bmatrix}=\begin{bmatrix}0.4002 & 0.7076 & -0.0808\\-0.2263 & 1.1653 & 0.0457\\0 & 0 & 0.9182\end{bmatrix}\times\begin{bmatrix}R\\G\\B\end{bmatrix} \quad (7)$$

The IPT was derived from non-linear L'M'S' value:

$$\begin{bmatrix}I\\P\\T\end{bmatrix}=\begin{bmatrix}0.4 & 0.4 & 0.2\\4.4550 & -4.8510 & 0.3960\\0.8056 & 0.3572 & -1.1628\end{bmatrix}\times\begin{bmatrix}L^{0.43}\\M^{0.43}\\S^{0.43}\end{bmatrix} \quad (8)$$

Finally, the hue of specific pixel p was defined as:

$$hue(p)=\tan^{-1}(P/T) \quad (9)$$

Let $s(i)$ denote a pixel in S, S_D or S_B, $h(i)$ is a pixel in H, H_D or H_B, and hw is the total pixel number of an image. Hue shift loss l_H is defined as the average of hue difference of all pixels

$$l_H=\frac{1}{hw}\sum_{i=1}^{hw}|hue[s(i)]-hue[h(i)]| \quad (10)$$

Color difference loss (supervised). According to Human Visual System (HVS) theory, human color perception changes accordingly with luminance, which means there will definitely be certain amount of color difference between corresponding HDR and tone-mapped SDR image. Therefore, unsupervised minimize of color difference between HDR and SDR images is meaningless and impossible. Hence, we turned to minimize the color difference loss (l_C) between output and label SDR images in a supervised way.

Traditional color difference formulas defined in CIE 1976 $L^*a^*b^*$ color space such as ΔE_{2000} or ΔE_{76} are designed for SDR scenario where luminance is under 100nit [48]. Hence, we defined color difference loss (l_C) in IC_tC_p color space which is suitable for HDR luminance up to 1000nit. RGB values of output and label SDR images were first converted to XYZ tristimulus using eqn. (5) and (6), then converted to LMS color space using a cross-talked matrix different to eqn. (7):

$$\begin{bmatrix}L\\M\\S\end{bmatrix}=\begin{bmatrix}0.3592 & 0.6976 & -0.0358\\-0.1922 & 1.1004 & 0.0755\\0.0070 & 0.0749 & 0.8434\end{bmatrix}\times\begin{bmatrix}R\\G\\B\end{bmatrix} \quad (11)$$

Then, IC_tC_p was derived from non-linear L'M'S' value:

$$\begin{bmatrix}I\\C_t\\C_p\end{bmatrix}=\begin{bmatrix}0.5 & 0.5 & 0\\1.6137 & -3.3234 & 1.7097\\4.3780 & -4.2455 & -0.1325\end{bmatrix}\times\begin{bmatrix}L^{0.43}\\M^{0.43}\\S^{0.43}\end{bmatrix} \quad (12)$$

Suppose $sl(i)$ is a specific pixel in SL, SL_D or SL_B, the color difference loss was defined as the average of ΔE_{IPT} [63] color difference value of all pixels:

$$l_C=\frac{1}{hw}\sum_{i=1}^{hw}\sqrt{\Delta I(i)^2+[0.5\Delta C_t(i)]^2+\Delta C_p(i)^2} \quad (13)$$

where the color difference Δ is calculated between $s(i)$ and $sl(i)$.

SSIM loss (supervised). While unsupervised l_{TMQI} ensures that tone-mapped SDR image will maintain abundant details and structure from original HDR image, it's also worth paying attention to the structure consistency between output and label SDR images. Therefore, we decided to use loss terms SSIM [35] which had been proven (by [64]) effective in "perceptual-motivated" (see §V for its definition) CNN. SSIM loss (l_{SSIM})

was calculated in single-scale fashion same as eqn. (4), but its local structure fidelity is different from eqn. (1):

$$SF_{local}(x,y)=\frac{2\mu_x\mu_y+C_1}{\mu_x^2+\mu_y^2+C_1}\cdot\frac{2\sigma_{xy}+C_2}{\sigma_x\sigma_y+C_2} \quad (14)$$

where μ_x, σ_x and σ_{xy} are mean, standard deviation and cross correlation of 2 patches. Here, $C_1=0.0001$, $C_2=0.0009$. Note that, both $\boldsymbol{l_{SSIM}}$ and $\boldsymbol{l_{TMQI}}$ were calculated on luminance channel Y (row 2 of eqn. (5) or (6), depending on color gamut).

Perceptual loss (supervised). Loss from another pre-trained network was referred to as "perceptual" loss in 6 of 20 related works ([3], [25], [27], [42], [43] and [46]), we decided to follow the same practice.

By separately feeding $\boldsymbol{S}$, $\boldsymbol{S_D}$ or $\boldsymbol{S_B}$ and $\boldsymbol{SL}$, $\boldsymbol{SL_D}$ or $\boldsymbol{SL_B}$ into pre-trained 19-layer VGG-Net [30], perceptual loss ($\boldsymbol{l_P}$) was calculated using the mean absolute error between same VGG layers with different inputs. Take step 2 for example:

$$l_S=\sum_{n=1}^{5}\frac{1}{h_nw_nc_n}\|\phi_n(S)-\phi_n(SL)\|_1 \quad (15)$$

where $\phi_n(x)$ represent specific layer in pre-trained VGG-Net receiving x as input, $\|\cdot\|_1$ is l_1 norm, $h_nw_nc_n$ denote the size of current layer, n from 1 to 5 means that we totally utilized 5 layers i.e., *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1* and *conv5_1*.

MAE loss (supervised). Mean absolute error (MAE) loss is one of the most widely-used pixel-wise loss terms in images transformation tasks:

$$l_{MAE}=\frac{1}{hwc}\sum_{i=1}^{hwc}|sl(i)-s(i)| \quad (16)$$

where *hwc* represent the total element number of tensors. We chose MAE (l_1 norm) rather than MSE (squared l_2 norm) loss because [38] found that l_2 loss will overestimate overexposed areas of HDR image even it has been converted into non-linear domain.

Totally 6 loss terms were calculated, 4 of them were inspired IQA ($\boldsymbol{l_{TMQI}}$, $\boldsymbol{l_H}$, $\boldsymbol{l_C}$, and $\boldsymbol{l_{SSIM}}$) and 3 of them were first introduced in HDR related CNN ($\boldsymbol{l_{TMQI}}$, $\boldsymbol{l_H}$ and $\boldsymbol{l_C}$). Finally, all loss terms were linearly added as total loss, their coefficients were empirically set as in Table II.

TABLE II
LINEAR COEFFICIENTS OF ALL LOSS TERMS

Loss terms	Unsupervised			Supervised		
	$\boldsymbol{l_{TMQI(-S)}}$	$\boldsymbol{l_H}$	$\boldsymbol{l_P}$	$\boldsymbol{l_C}$	$\boldsymbol{l_{SSIM}}$	$\boldsymbol{l_{MAE}}$
Step 1 ($\boldsymbol{N_G}$)	0.15	0.15	0.3	0.2	-	0.6
Step 1 ($\boldsymbol{N_L}$)	0.2	-	0.3	-	0.2	0.6
Step 2	0.1	0.15	0.3	0.2	0.1	0.6

2) TRAINING SET

A training set containing HDR-SDR (input-label) pairs is still indispensable for supervised loss terms. Our training set was obtained from Fairchild [1], Funt et al [59] and Waterloo IVC MEFI [60] datasets which totally contain 234 high-resolution diversified real-world HDR scenes, along with their source bracketing exposure SDR sequence. 200 of them were selected as training set while the rest 34 were used as part of our test set. For each HDR-SDR pair, 16 patches with 512×512 size were obtained: 15 were from random cropping while the last one was from resizing. Finally, we got 3200 pairs of patches. It is worth noting that target (label) SDR images in our training set were obtained in a distinctive way based on the following insights.

The common practice to obtain training pairs is to generate target SDR image from its HDR counterpart using traditional TMOs. This was adopted by almost all deep TMOs due to the wide accessibility of HDR content. However, we found that even when target SDR images are from parameter-fine-tuned TMO (in [25] and [34]) or selected according to best objective score (in [31], [24], [26] and [27]), they still contains artifacts brought by traditional TMOs e.g., over-enhancement or over-stylization. Therefore, to avoid those artifacts, we first turned to use simultaneously-shot real HDR-SDR pairs.

Since the only public-available real HDR-SDR pairs [38] was excluded due to the reason mentioned in §II.C, we turned to obtain simultaneously-shot SDR counterpart from bracketing exposure SDR sequence (available in [1], [59] and [60]). We started with treating exposure-value-0 (EV0) image as the SDR counterpart, but we found it ending up with unsatisfactory results, specifically, lack of details in both bright and dark areas. This was caused by target SDR images themselves: EV0 SDR images ((a) in Fig. 5) containing deficient details in bright and dark areas showed good result for rTMO because they taught CNN to recover lost details, however, they taught CNN to vanish those details when it comes to TMO.

(a) EV0 target SDR image (b) Fine-tuned target SDR image

FIGURE 5. An example of simultaneously-shot target SDR obtained from different ways. Fine-tuned one (b) is obviously more vivid in dark (green box) and bright areas (red box).

Therefore, we turned to generate better target SDR images by utilizing all raw information in bracketing exposure sequence. Some were accomplished by professional photographers' fine-tuning using Adobe Photoshop on a calibrated sRGB color gamut monitor, the rest were by tuning a pre-trained multi-exposure image enhancer SICE [61]. As is illustrated in (b) in Fig. 5, target SDR images acquired in this way have more details in bright and dark areas.

3) DATA PRE-PROCESSING

Since there is a huge difference between pixel value distribution of linear HDR and non-linear SDR images,

majority of related works followed the common practice of converting HDR images into logarithm (log) domain ([3], [33], [25], [26], [34], [36], [42], [43] and [46]) or other non-linear domain ([38]), which made pixel value more evenly-distributed and easier for CNN to process. While the rest of methods ([31], [24], [27], [39], [40], [41], [45], [5], [47] and [56]) just normalized HDR images without domain transfer.

Inspired by TV production process, we proposed a domain consistent strategy where HDR input and SDR output of our CNN are in same non-linear (gamma/display) domain. Recall that pixel value in HDR image is photometrically linear, thus HDR images were converted to non-linear after normalization. Let $h_o(i)$ and $s_o(i)$ denote original pixel value of HDR and SDR images, respectively. HDR pixel value was converted as:

$$h(i)=\left[\frac{h_o(i)-\min\{h(k)\}}{\max\{h(k)\}-\min\{h(k)\}}\right]^{0.4545},k\in\{1,\ldots,hwc\} \quad (17)$$

where power *0.4545* was derived from approximate sRGB [65] non-linear (gamma2.2) curve. Meanwhile, non-linear SDR images were normalized without curve conversion:

$$s(i)=s_o(i)/255 \quad (18)$$

We deprecated logarithm domain because it has no physical meaning, while non-linear/linear is display/scene-referred. We also deprecated unified linear domain (in this case, power *0.4545* in eqn. (17) was removed, and power *2.2* was added to eqn. (18)) because we found it producing unnatural color in dark areas of output SDR images. Experiment on data pre-processing was later conducted and demonstrated in §IV.C.2.

4) INPLEMENTATION DETAILS

Parameters in our CNN were initialized by truncated normal distribution, and optimized using adaptive moment estimation (ADAM) [66] optimizer in 0.0005 learning rate for both steps. Batch-size for step 1 and step 2 was set to 8($\boldsymbol{N_L}$), 32($\boldsymbol{N_G}$) and 4, respectively. TensorFlow implementation of our method is available at github.com/AndreGuo/IQATM/.

IV. EXPERIMENTS

Different from other tasks, tone-mapping is an information-reducing process, which means original HDR image is more informative even than elaborate label SDR image. Therefore, objective scores are calculated between output SDR and original HDR, rather than between output SDR and label SDR.

We selected TMQI [6] and FSITM [7] detailed in §II.D as the objective quality score. In experiments below, TMQI will be split into 2 parts namely FR structure fidelity (denoted as TMQI_S) and NR naturalness (denoted as TMQI_N). Since both TMQI and FSITM works in luminance channel Y (row 2 of eqn. (5) or (6), depends on color gamut), color information is ignored. Hence, we appended another FR tone-mapped image quality assessment method [67] previously proposed by our quality assessment laboratory, to measure the color preservation from HDR to SDR images. Its objective score "color difference matrix index" (CDMI) is given by calculating a modified color difference formula between each pixel in HDR and SDR images. All objective scores are upper-bound by 1 where higher means better.

A. TEST SET

Totally 87 HDR scenes were included in our test set: 34 from the rest part of training set ([1], [59] and [60], mentioned in §III.B.2), 15 from [6], and 38 from Laval Indoor HDR dataset [68]. Indoor HDR scenes from [68] were added to diversify our test set, thus to prove the scene-adaptability of our method. Since our training set is mainly composed of outdoor scenes, indoor scenes from [68] will also help us to further reveal the generalization of our trained model.

When testing, all input HDR images $\boldsymbol{H}$ followed the same pre-processing as training (eqn. (17)), no post-processing was applied since our CNN works in unified non-linear domain.

B. ABLATION STUDIES

Our ablation studies were done on 2 aspects namely MGRB and IQA inspired semi-supervised loss terms. As shown in Table III, 6 combinations of innovations (①-⑥) were tested using 5 abovementioned objective scores. Best performances on each score are highlighted in bold.

TABLE III
OBJECTIVE SCORES OF DIFFERENT COMBINATIONS OF INNOVATIONS

	①	②	③	④	⑤	⑥
MGRB	×	×	√	√	√	√
$\boldsymbol{l_{TMQI}}$,$\boldsymbol{l_{SSIM}}$	×	√	×	×	√	√
$\boldsymbol{l_H}$, $\boldsymbol{l_C}$	×	√	×	√	×	√
TMQI	0.7144	0.7315	0.8808	0.9059	0.8959	**0.9144**
TMQI_S	0.6290	0.6613	0.8680	0.8498	**0.8732**	0.8522
TMQI_N	0.1140	0.1237	0.4531	0.6389	0.5400	**0.6805**
FSITM	0.6718	0.6763	0.7745	0.7300	**0.8173**	0.8044
CDMI	0.7475	0.7584	0.7770	**0.8569**	0.8313	0.8457

1) ON IQA INSPIRED SEMI-SUPERVISED LOSS TERMS

When studying the effect of IQA inspired semi-supervised loss terms, $\boldsymbol{l_{MAE}}$ and $\boldsymbol{l_S}$ were chosen as the baseline (column ③). Note that 4 novel loss terms ($\boldsymbol{l_{TMQI}}$, $\boldsymbol{l_H}$, $\boldsymbol{l_C}$, and $\boldsymbol{l_{SSIM}}$) were divided into 2 groups based on their functionality (on structure or color) rather than mechanism (supervised or unsupervised). During ablation study, all loss terms shared same coefficients as Table II.

By comparing column ④ with ③, we can find that the introduction of $\boldsymbol{l_H}$ and $\boldsymbol{l_C}$ had improved the performance of color-related objective score CDMI. Form column ③ and ⑤ we know that the introduction of $\boldsymbol{l_{TMQI}}$ and $\boldsymbol{l_{SSIM}}$ had boosted structure-related score TMQI_S and FSITM. Meanwhile, some scores of final model ⑥ were sightly inferior to those where "expertise" loss terms were added along (④ and ⑤), but final model ⑥ reached a more balanced score.

The effect of color-related $\boldsymbol{l_H}$ and $\boldsymbol{l_C}$ is visualized in Fig. 6. Different from deep CNN based TMOs [25] and [27] handling only luminance channel, and [34] where only 2 of 3 channel are processed, our CNN directly handle 3 channels of RGB

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

image. Hence, severe color aberration may occur when only basic loss terms are used (case 2 and 3 in Fig. 6). This is caused by the intrinsic shortcoming of l_1 (MAE) loss that 2 output RGB values representing different colors may share the same MAE value with specific label RGB value.

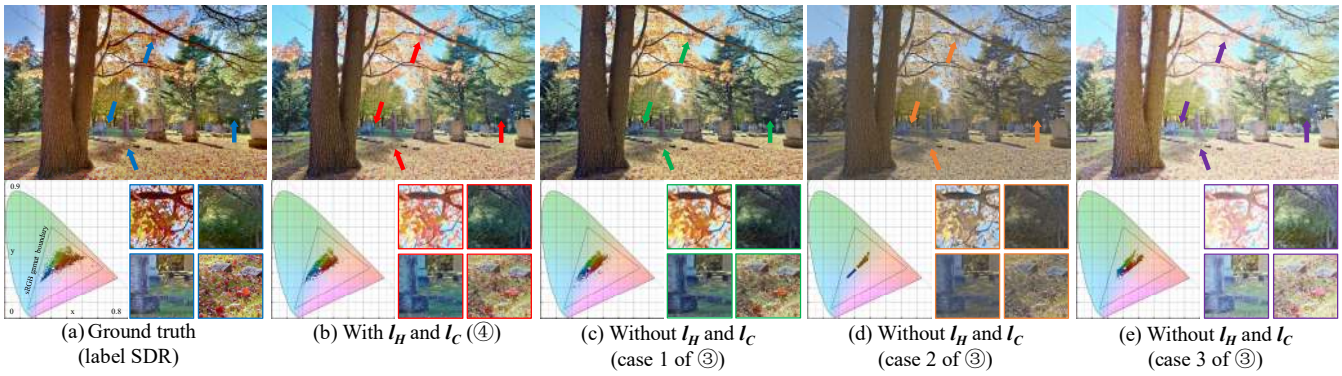


FIGURE 6. The effect of color-related semi-supervised loss terms. Ground truth label image along with 4 output images with/without color related loss are illustrated. Corresponding chromaticity diagrams are listed to visualized their color difference. As seen, the introduction of color-related losses had made the output's color appearance and pixel color distribution the closest to label.

We addressed this by restricting pixel color using $\boldsymbol{l_H}$ and $\boldsymbol{l_C}$. As seen in those CIE 1931 Yxy chromaticity diagrams in Fig. 6, the introduction of supervised loss $\boldsymbol{l_C}$ made the pixel color distribution of ④ the closest one to our elaborate label. Note that, there still exist several color differences between them, this is because unsupervised loss $\boldsymbol{l_H}$ will make output SDR image inheriting color distribution (especially hue) form input HDR rather than only from label SDR.

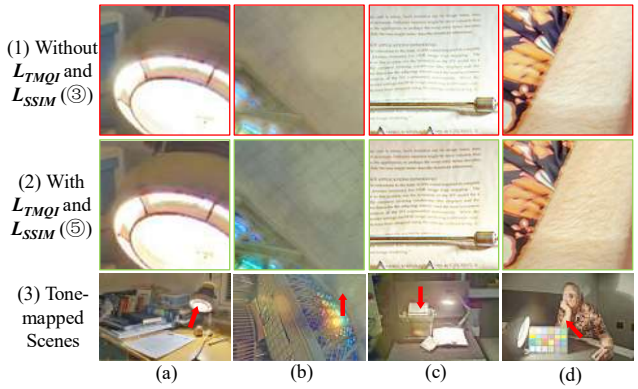


FIGURE 7. The effect of structure-related semi-supervised loss terms. As seen, the introduction of structure-related losses can make edges more obvious.

By avoiding up-sample, the deprecation of U-net's deconvolution layers had prevented our method from structure distortion like checkerboard artifact [71]. However, poorly-trained fully convolution layers may still vanish detail (high-frequency texture).

The effect of the introduction of structure-related $\boldsymbol{l_{TMQI}}$ and $\boldsymbol{l_{SSIM}}$ is illustrated in Fig. 7. As seen, there are more details in the output of ⑤compared with ③. Edges around bright area (lamp in (a) and light spot in (b)) are clearer, and high-frequency details (hair in (d) and text in (c)) become more obvious.

2) ON MULTI-GROUP RESIDUAL BLOCK

The immediate removal of ripple/halo artifacts brought by MGRB has been demonstrated in Fig. 3. The impact of MGRB can be quantified by comparing column ① vs. ③, and ② vs. ⑥ in Table III. As seen, MGRB's improvement on objective scores mainly lies on structure fidelity related TMQI_S and FSITM. We attribute this to the enlarged receptive field brought by MGRB, since it can eliminate structure discontinuity at the edge of ripple/halo.

C. OTHER EXPERIMRNTS

1) ON NETWORK STRCTURE

We designed 3 separate simplified networks to explore the effectiveness of decomposing and multi-branch strategy, and to prove the necessity of Polishing Network $\boldsymbol{N_P}$. As is illustrated in Fig. 8, simplified networks are full network without $\boldsymbol{N_L}$, without $\boldsymbol{N_P}$, or without both of them, separately. Note that, parameters of all simplified networks were individually trained to their best effort, rather than borrowed from the trained whole network.

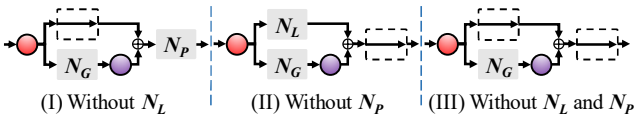


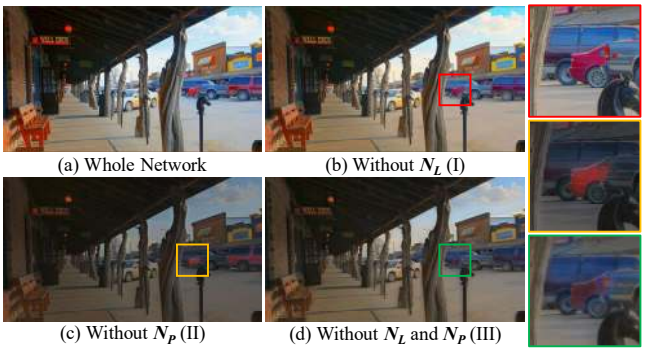
FIGURE 8. 3 different simplified networks. Red and purple circle represent same decomposing and up-scale in Fig. 1/2, respectively.

Performances of different simplified networks are list in the left side of Table IV. As seen, simplified network (I) reached the closest performance to whole network. Also, (II) decreased significantly on structure related score TMQI_S and FSITM, while decreased slightly on other scores. Last, (III) got poor result on all scores. Hence, it's acceptable to discard $\boldsymbol{N_L}$ to further reduce network complexity at the cost of slight performance degradation. Also, by comparing (I) with (III) we know that $\boldsymbol{N_P}$ is most indispensable among all sub-networks.

TABLE IV
OBJECTIVE SCORES OF OTHER EXPERIMENTS

Experiment Type	On network structure			On data pre-processing	
Network Structure	(I) without N_L	(II) without N_P	(III) without N_L & N_P	whole (linear)	whole (non-linear)
TMQI	0.8809	0.7660	0.7582	0.8763	0.9144
TMQI_S	0.8262	0.7584	0.7232	0.8290	0.8522
TMQI_N	0.5526	0.1127	0.1211	0.5075	0.6805
FSITM	0.7858	0.7356	0.6710	0.8103	0.8044
CDMI	0.8329	0.7350	0.7553	0.7622	0.8457

Fig. 9 can better support above conclusion. As seen, there is only slight difference between (b) and (a). While the (c) is overall dim, and has light-spot-artifact brought by limited network depth thus small receptive filed. At last, (d) is unacceptable since blur and halo artifact occur in whole image.



(a) Whole Network (b) Without N_L (I)
(c) Without N_P (II) (d) Without N_L and N_P (III)

FIGURE 9. Result comparison of different simplified networks.

2) ON DATA PRE-PROCESSING

Besides network structure, we also explored the effect of data pre-processing. Here, we compared the result of non-linear domain (detailed in §III.B.3) against linear domain. Their quantitative comparison is listed in last 2 columns of Table IV. As seen, non-linear model outperformed linear model in all objective scores except FSITM.

(a) Linear domain (b) Non-linear domain

FIGURE 10. The effect of domain transfer in data-preprocessing. As seen, (a) is under-saturated, and its pixel color distribution is limited.

The effect of data pre-processing is illustrated in Fig. 10. As seen, output SDR image from linear domain model (a) tends to be undersaturated, especially in dark areas. This is mainly due to the pixel value distribution of linear light HDR images. Take Fig. 10 for example, in HDR image, luminance of the outdoor aera is very low compared with building in high luminance, thus their pixel value become extremely low after normalization. In this case, it's hard for CNN to recover color information form tiny RGB ratio, as seen in the limited pixel color distribution in (a). However, when it is non-linear transferred, pixel values of dark areas become more notable, thus easier for CNN to recover color information.

D. COMPARISONS WITH OTHER METHODS

We compared our method with 8 TMOs including 5 traditional ones and 3 deep TMOs. Traditional TMOs namely Drago03 [14], Mantiuk06 [15], iCAM06 [16] and Reinhard02 [17] were implemented using official code/software with default parameters. In addition, we added a parameter-free traditional TMO Mai11 [72] which optimizes the global mapping-curve.

We reproduced deep-CNN-based Yang et al. [25]' method following the same data pre-processing with official model parameters (checkpoint). For deep TMOs Rana et al. [27] and Zhang et al. [34] where official checkpoint was not provided, we obtained their test set and corresponding official tone-mapped SDR images form authors. Hence, data in Table V was calculated on the test set intersection between ours and theirs. (In §IV.A, images in test set were selected in a way which will maximize this intersection.)

1) QUANTITATIVE EVALUATION

Mean value and standard deviation of 6 objective scores are listed in Table V. Higher mean value indicates better overall performance (except BTMQI where lower is better), while smaller standard deviation means better stability and scene-adaptability. Among deep TMOs, our method had got 8×1^{st}, 3×2^{ed}, 1×3^{rd} and no worst over all scores.

Comparing with all TMOs, ours reached the best TMQI and TMQI_N in both mean value and standard deviation, which means it's the most likely one to produce nature-looking results under various HDR scenes. Also, we got the best CDMI standard deviation and a CDMI mean value very close to the best one. This indicates that our method is able to scene-adaptively generate results with good color reproduction.

However, when it comes to TMQI_S and FSITM (both on structure), most deep TMOs including ours were not top-ranked, and didn't outperform all traditional ones. This is because while the output pixel value of global traditional TMOs (all except [17]) is not affected by its neighbors, that of deep TMOs may be undeservedly disturbed by its long-distance pixel dependency established by receptive filed.

In this section, we added an NR quality score BTMQI [73] assessing only tone-mapped SDR images, from its information, naturalness and structure. Our method got the best and second best BTMQI among deep and all TMOs, respectively. Since our results are best in naturalness and average in structure, a small BTMQI indicate that our results are informative.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

TABLE V
PERFORMANCE COMPARISON OF ALL TMOS. THE BEST AMONG ALL/DEEP TMOS ARE HIGHLIGHTED WITH BOLD/UNDERSCORE, RESPECTIVELY.

Type	Method	Scores (Mean Value)				Scores (Standard Deviation)			
		TMQI(TMQI_S/TMQI_N)	FSITM	CDMI	BTMQI	TMQI(TMQI_S/TMQI_N)	FSITM	CDMI	BTMQI
Traditional	Dargo03 [14]	0.8833(0.8703/0.4962)	0.8532	0.8681	3.4159	0.0550(**0.0382**/0.2844)	0.0392	0.0440	1.4357
	Mantiuk06 [15]	0.8660(**0.8984**/0.3546)	0.8575	0.8305	3.8534	0.0519(0.0548/0.2171)	**0.0354**	0.0631	**0.9054**
	iCAM06 [16]	0.8466(0.8601/0.3410)	0.7352	0.7687	4.1865	0.0679(0.0545/0.2685)	0.0476	0.0776	1.6121
	Reinhard02 [17]	0.8781(0.8756/0.4647)	0.8562	0.8702	3.4313	0.0645(0.0605/0.3354)	0.0413	0.0531	1.6220
	Mai11 [72]	0.9103(0.8776/0.6337)	0.8182	0.8956	**3.2975**	0.0500(0.0532/0.2648)	0.0506	0.0326	1.1349
Deep CNN Based	Yang et al. [25]	0.8728(0.8267/0.5204)	0.8494	0.8363	3.8057	0.1040(0.1456/0.3222)	0.0567	0.0561	1.9299
	Rana et al. [27]	0.8805(0.8658/0.4897)	**0.8624**	0.7968	3.3782	0.0725(0.0918/0.2859)	0.0412	0.0358	1.2227
	Zhang et al. [34]	0.8961(0.8389/0.6057)	0.8311	**0.8965**	3.5193	0.0552(0.0659/0.2607)	0.0643	0.0322	1.0159
	Proposed	**0.9189**(0.8492/**0.7221**)	0.8349	0.8945	3.3643	**0.0270**(0.0511/**0.1680**)	0.0456	**0.0316**	0.9114

2) VISUAL COMPARISON

Overall Performance. As seen in Fig. 11, [14], [17] and [72] got fair result, but they lack the ability to reveal details in dark areas (red arrow). [15] and [16] emphasized details, but their results appeared over-stylized and overall-dim. As for deep TMOs, results of [25], [27] and [34] got strange saturation and are thus less nature-looking. Besides, result of [27] didn't recover information in dark region (red box), while their bright area (green box) still tends to be over-exposed. In summary, our method can produce nature-looking result while maintaining information in both dark and bright areas.

Due to the extra resize operation (enlarging to the same size as others for display) applied to the results of deep TMOs [27] and [34], their details appear blurry, as seen in red and green box of (g) and (h). Due to their fixed-size U-net CNN structure, original output size of [27] is 2048×1024, while that of [34] is limited within 512×512. Hence, our method has a practical advantage of arbitrary input size, thanks to the design of fully convolutional layer.

In Fig. 12, we selected a hard-to-tackle indoor-outdoor-alternating HDR scene. As seen, our method did the best simultaneously revealing details in both bright reflectance (red arrow) and dim ceiling (green arrow), meanwhile, having a good overall-brightness. Besides, our method got the best TMQI (mainly from TMQI_N) and CDMI in Fig, 12 and Fig. 13, indicating that our method reached the best naturalness, and the most accurate color reproduction from HDR image. Similar conclusions are draw on the title of Fig. 13 and Fig. 14.

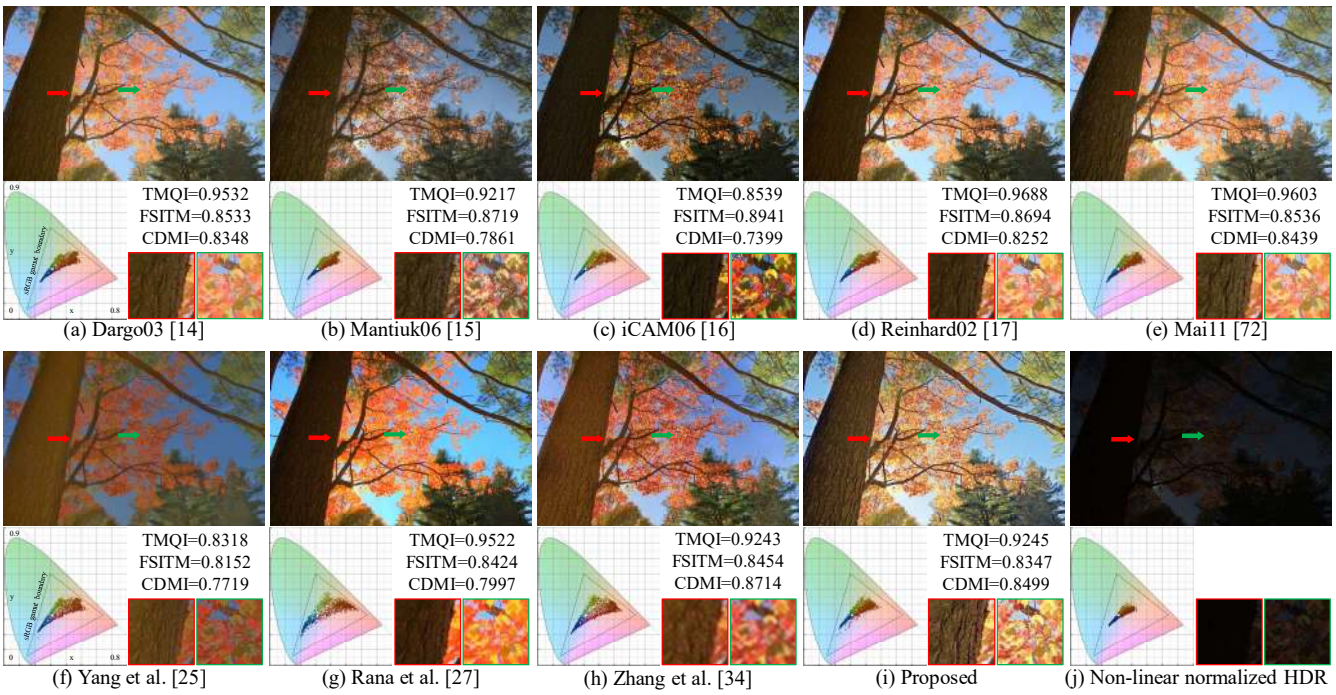


FIGURE 11. Result comparison on outdoor scene. Corresponding pixel color distribution is shown in the chromaticity diagram below the image. As seen, the proposed method can simultaneously reveal local details in both dark and bright areas. Note that, "(j) Non-linear normalized HDR" is visualized using Eqn. (17). In other words, it's the exact input of our CNN (*H*).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

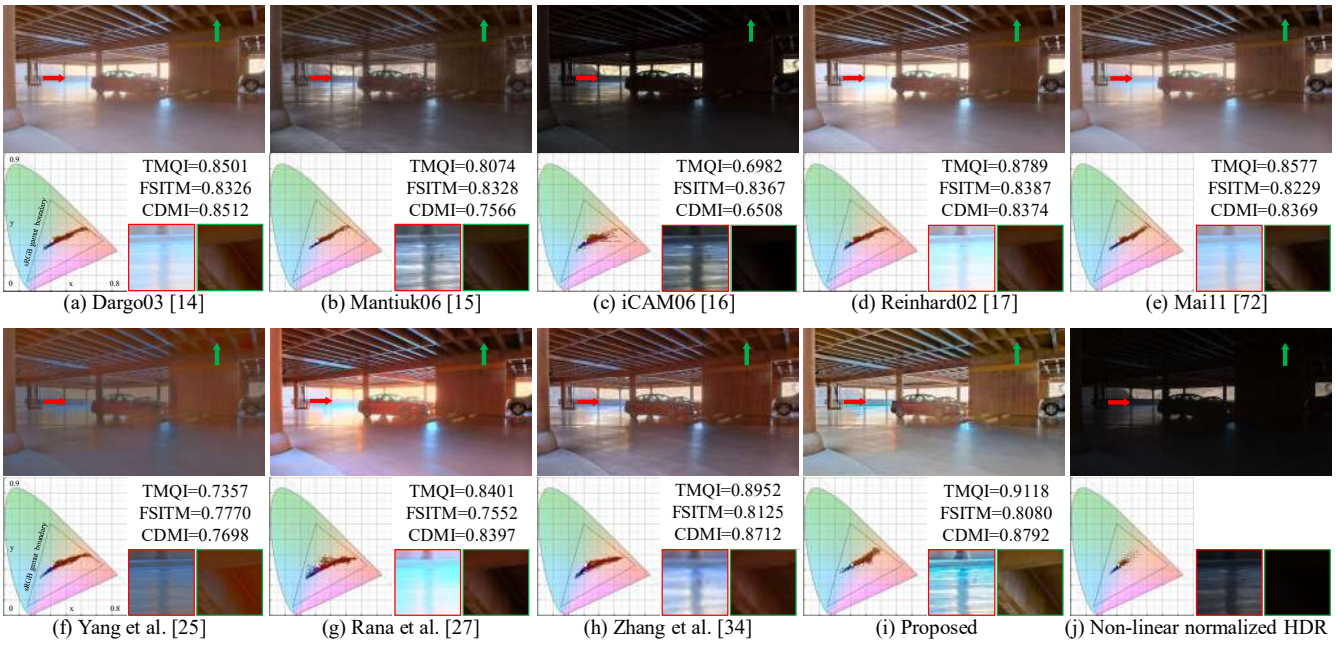


FIGURE 12. Visual comparison on indoor-outdoor scene. The proposed method is able to preserve information in both bright reflectance (red box) and dim areas (green box) while maintaining good color appearance and satisfying overall-brightness.

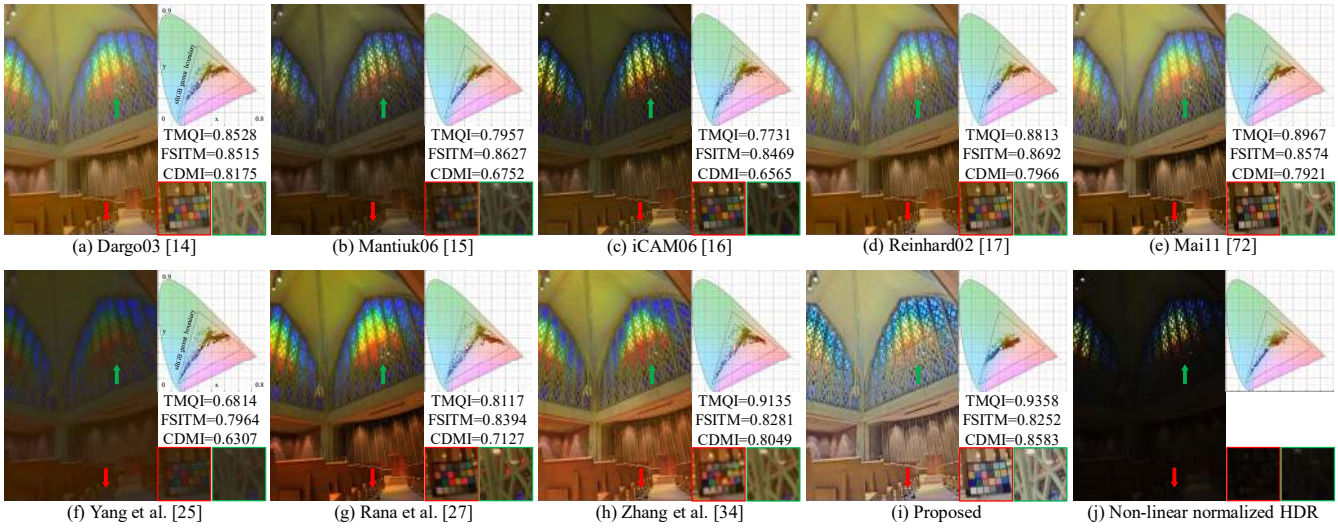


FIGURE 13. Visualizing results on indoor scene. Our result got highest TMQI and CMDI, i.e., the best naturalness and the most accurate color reconstruction form HDR image. Meanwhile, compared with others, details on the doom (green arrow) were better preserved by our method.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3080331, IEEE Access

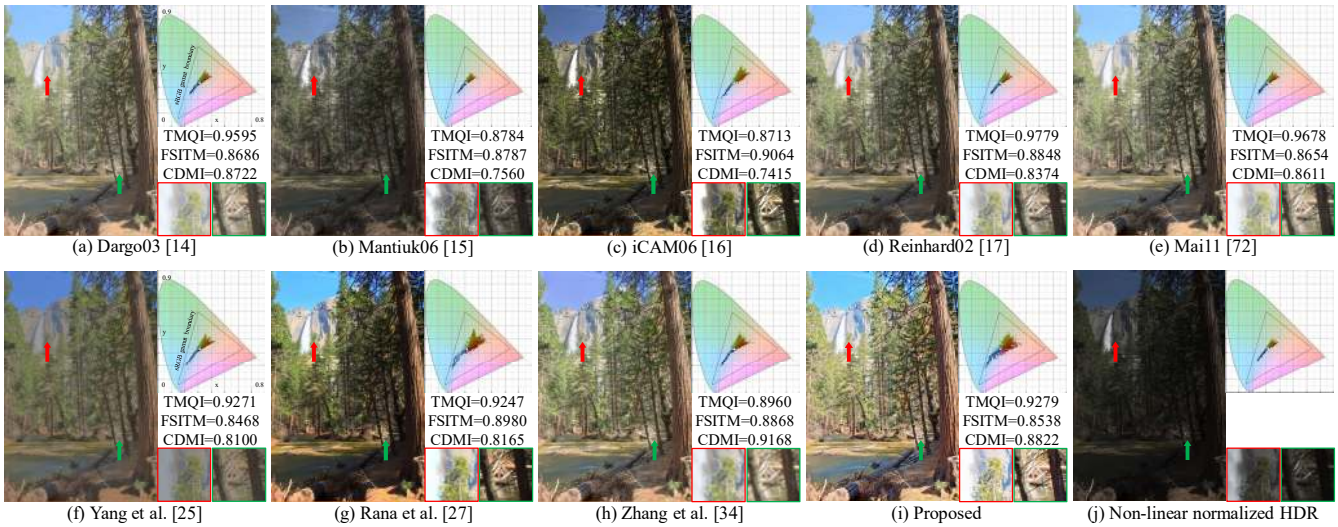


FIGURE 14. **Demonstrating results on outdoor scene. Our method didn't do the best preserving information in most-bright area (red arrow), but performed well revealing details in bright-dark-alternating area (green arrow), meanwhile, having good overall color appearance.**

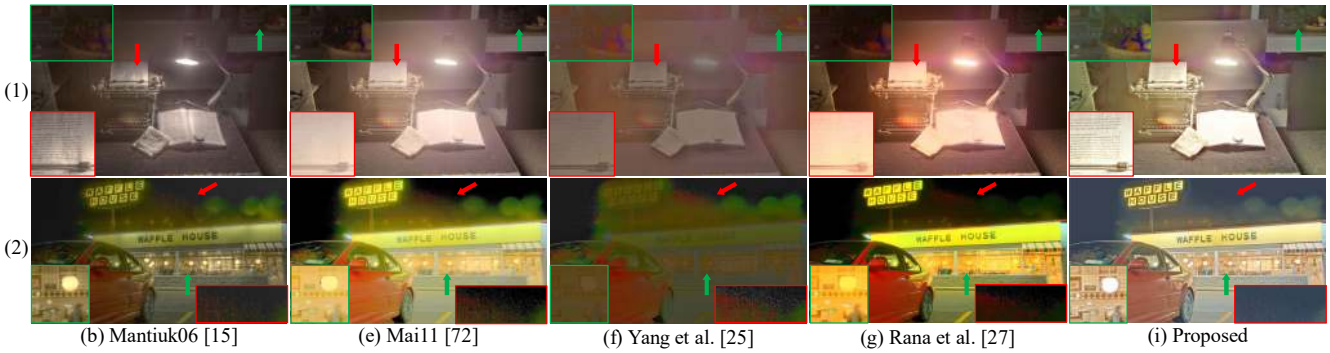


FIGURE 15. **Detailed comparison on structural preservation. As seen in (1), high-frequency details in bright region (red box) were well presented in both (b) and (i), while details in dark area (green box) were simultaneously preserved only by our method (i). As for scene (2), structure in bright area (green box) was better preserved in (b) and (i). Meanwhile, salt-and-pepper noise (red box) was amplified into a strange pattern by other deep TMOs in (f) and (g), while kept in (b) and (e), and suppressed by our method (i).**

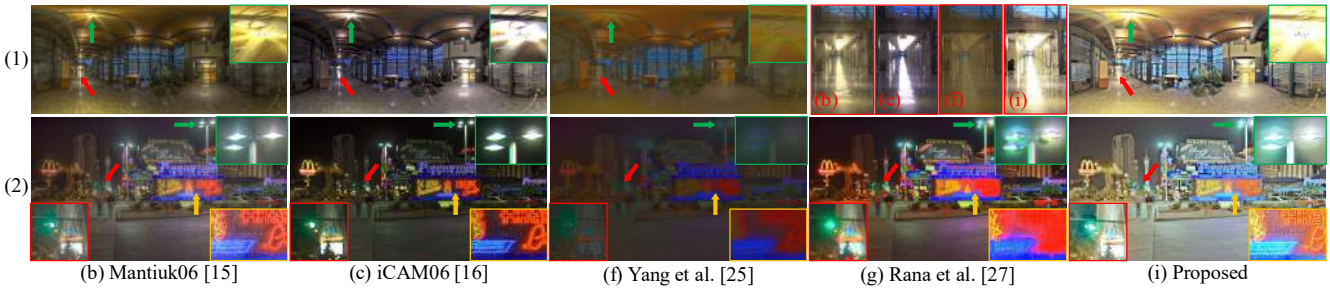


FIGURE 16. **Focusing on detail reconstruction around illuminant. All methods except (b) perform similarly on scene (1). As for scene (2), (g) got extra pattern around lamp (green box), and "dyeing" color on neon light (yellow box). Our method performs well revealing details in (1), but has halo artifact in some cases (red box in (2)).**

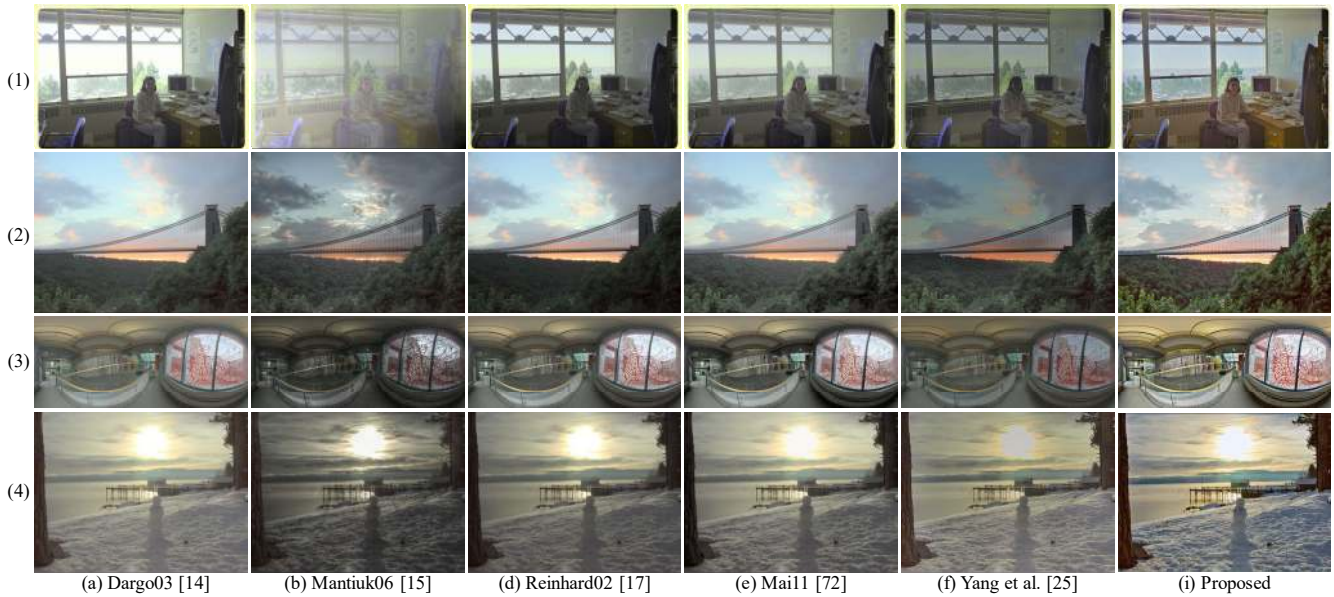


FIGURE 17. More HDR scenes are compared to show the scene-adaptability of our method. Among all scenes (1) high-contrast portrait, (2) outdoor, (3) high-contrast indoor and (4) high-contrast outdoor, our results (i) are obviously more colorful and vivid, and have sufficient details in both bright and dark region.

Color appearance and reproduction. In Fig. 11, 12, 13 and 14, pixel color distribution of different results is plotted within the assumed sRGB gamut boundary on CIE 1931 Yxy chromaticity diagrams. Here, color difference can be judged by distance (though it's not perceptually uniform), while hue is reflected in the angle from white point.

By comparing the chromaticity diagrams of others with (j), especially in Fig. 12, we know that our result most accurately reproduced the color appearance (especially hue) from HDR (j). We contribute this to our unsupervised hue loss $\boldsymbol{l_H}$.

The effect of supervised color difference loss $\boldsymbol{l_C}$ (color consistency from label) cannot be assessed here since there's no label for some image in test set. But from another angle, as seen in most cases except Fig. 14, other deep TMOs [25] (f), [27] (g) and [34] (h) tend to undeservedly extend color distribution, and distort the hue of their outputs. While the proposed method has learned a natural and traditional-TMO-like "conservative" color appearance.

Details and structure. More scenes (Fig. 15 and Fig. 16) were added to compare the structure preservation and detail revealing ability of different methods. Results of 5 methods are compared in each figure, including traditional Mantiuk06 [15] TMO with the most emphasis on structure, another traditional TMO, and 3 deep TMOs ([34] are excluded due to its 512×512 low-resolution output).

Description has been written on the title of each figure. In summary, in Fig. 15, our method did the best suppressing salt-and-pepper noise while avoiding structure distortion. Our result maintained as much structure information as structure-specialized Mantiuk06 [15] TMO (red box (1) and green box (2)), and overperformed others in simultaneously keeping structure in both dark (green box (1)) and bright region.

In Fig. 16, when preserving structure and detail around illuminants, our result (i) surpasses other deep TMOs (f) and (g). However, it has halo artifact which is unseen on the output of traditional TMOs (b) and (c). To this end, we must admit one of our limitations, that while most halo artifact was eliminated by the introduction of MGRB, it may occur in some extreme cases where the luminance around neighboring pixels varies dramatically.

Scene-adaptability. In §IV.D.1, the scene-adaptability of our method is reflected in the lower standard derivation on highly-diversified test set. Here, more HDR scenes are compared in Fig. 17 to prove this adaptability. As seen, our result is more vivid and detailed among all scenes.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a learning-based, scene-adaptive and size-adaptive TMO using deep CNN. During its design, we explored the effect of network structure, training, and data pre-processing. Most importantly, we introduced IQA to "perceptual-motivated" deep CNN (i.e., CNN whose output is to be evaluated by human perception). Due to the mechanism of IQA, specifically, tone-mapped image quality assessment, it's implemented by semi-supervised loss function.

Our work is just a small step bridging the gap between modern perceptual quality models and perceptual-motivated CNN. While our IQA losses were mathematically defined, there has been several recent "perceptual-motivated" deep CNN whose IQA scores were from a customized loss network. These loss networks were trained to mimic various quality scores to be used in loss function. For example, Chen et al. [69] trained their loss network to output objective VAMF between label and output, Talebi et al. [74] and Yang et al. [75] applied NIMA [76] as their loss network to get aesthetic subjective score on output image. Their loss networks shared the same motivation of mimicking a quality score which is unable to be directly implemented as loss function due to its complexity or non-differentiability (objective scores), and unquantifiability (subjective scores).

Recall that some objective scores (e.g., TMQI's naturalness term, BTMQI [73] and BLIQUE-TMI [70]) were excluded from our loss function due abovementioned reason. Hence, in further work, we are looking forward to use a loss network to learn those scores and act as loss function. We believe that compared with VGG-net-based loss network used in 6 of 20 HDR related CNN, a loss network with interpretable output can better represent the terminology "perceptual loss".

ACKNOWLEDGMENT

Author thanks his previous laboratory, Laboratory of Digital Video Quality Assessment, for the inspiration of combining IQA with perceptual motivated neural network.

REFERENCES

[1] M. Fairchild, "The HDR photographic survey," in *Proc. the 14th IS&T/SID Color Imaging Conference*, 2007, pp. 233-238.

[2] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers, *Advanced High Dynamic Range Imaging: Theory and Practice*, 2nd ed., Natick, MA, USA, A. K. Peters, Ltd., Feb. 2011, p. 1, 46, and 47.

[3] X. Hou, Y. Gong, B. Liu, K. Sun, J. Liu, B. Xu, J. Duan, and G. Qiu, "Learning based Image Transformation using Convolutional Neural Networks," *IEEE Access*, vol. 6, pp. 49779-49792, Sep. 2018, 10.1109/ACCESS.2018.2868733.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention*, Nov. 2015, pp. 234-241, 10.1007/978-3-319-24574-4_28.

[5] S. Y. Kim, J. Oh, and M. Kim, "Deep SR-ITM: Joint Learning of Super-Resolution and Inverse Tone-Mapping for 4K UHD HDR Applications," in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Nov. 2019, pp. 3116-3125, 10.1109/ICCV.2019.00321.

[6] H. Yeganeh and Z. Wang, "Objective Quality Assessment of Tone-Mapped Images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657-667, Feb. 2013, 10.1109/TIP.2012.2221725.

[7] H. Z. Nafchi, A. Shahkolaei, R. F. Moghaddam, and M. Cheriet, "FSITM: A feature similarity index for tone-mapped images," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1026–1029, Aug. 2015, 10.1109/LSP.2014.2381458.

[8] Q. Yan, L. Zhang, Y. Liu, Y. Zhu, J. Sun, Q. Shi, and Y. Zhang, "Deep HDR Imaging via A Non-Local Network,". *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 4308-4322, Feb. 2020, 10.1109/TIP.2020.2971346.

[9] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500, 10.1109/CVPR.2017.634.

[10] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," Jul. 2016, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1607.08022

[11] G. J. Ward, "The RADIANCE lighting simulation and rendering system," in *Proc. the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 94)*, Jul. 1994, pp. 459-472, 10.1145/192161.192286.

[12] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Trans. Visualization and Comput. Graph.*, vol. 3, no. 4, pp. 291-306, Oct. 1997, 10.1109/2945.646233.

[13] S. N. Pattanaik, J. Tumblin, H. Yee, and D. P. Greenberg, "Time-dependent visual adaptation for fast realistic image display," in *Proc. the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*, Jul. 2000, pp. 47-54, 10.1145/344779.344810.

[14] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, "Adaptive logarithmic mapping for displaying high contrast scenes," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 419–426, Nov. 2003, 10.1111/1467-8659.00689.

[15] R. Mantiuk, K. Myszkowski, and H. P. Seidel, "A perceptual framework for contrast processing of high dynamic range images," ACM Trans. Applied Percept., vol. 3, no. 3, pp. 286-308, Jul. 2006, 10.1145/1166087.1166095.

[16] J. Kuang, G. M. Johnson, and M. D. Fairchild, "iCAM06: a refined image appearance model for HDR image rendering," *Journal of Visual Comm. and Image Representation*, vol. 18, no. 5, pp. 406-414, Oct. 2007, 10.1016/j.jvcir.2007.06.003.

[17] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Digital Images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267-276, Jul. 2002, 10.1145/566654.566575.

[18] F. Durand and J. Dorsey, "Fast bilateral filtering for the display of high-dynamic-range images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 257-266, Jul. 2002, 10.1145/566654.566574.

[19] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 249-256, Jul. 2002, 10.1145/566654.566573.

[20] G. Krawczyk, K. Myszkowski, and H. Seidel, "Lightness perception in tone reproduction for high dynamic range images," *Comput. Graph. Forum*, vol. 24, no. 3, pp. 635–645, Oct. 2005, 10.1111/j.1467-8659.2005.00888.x.

[21] Z. Li and J. Zheng, "Visual-salience-based tone mapping for high dynamic range images," *IEEE Trans. Industrial Electronics*, vol. 61, no.12, pp. 7076-7082, Dec. 2014, 10.1109/TIE.2014.2314066.

[22] J. Yang, A. Horé, U. Shahnovich, K. Lai, S. N. Yanushkevich, and O. Yadid-Pecht, "Multi-Scale histogram tone mapping algorithm enables better object detection in wide dynamic range images," in *Proc. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2017, 10.1109/AVSS.2017.8078533.

[23] A. Rana, G. Valenzise , and F. Dufaux, "Learning-Based Tone Mapping Operator for Image Matching," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 256-268, May 2018, 10.1109/TMM.2018.2839885.

[24] V. A. Patel, P. Shah, and S. Raman, "A Generative Adversarial Network for Tone mapping HDR images," in *Proc. the 6th National Conference on Computer Vis., Pattern Recog., Image Processing and Graphics (NCVPRIPG 17)*, Apr. 2018, pp. 220-231, 10.1007/978-981-13-0020-2_20.

[25] J. Yang, Z. Liu, M. Lin, S. Yanushkevich, O. Yadid-Pecht, "Deep Reformulated Laplacian Tone Mapping," Feb. 2021, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/2102.00348

[26] Z. Zhang, C. Han, S. He, X. Liu, and T. T. Wong, "Deep binocular tone mapping," *The Visual Comput.*, vol. 35, no. 6-8, pp. 997-1011, Jun. 2019, 10.1007/s00371-019-01669-8.

[27] A. Rana, P. Singh, G. Valenzise, F. Dufaux, N. Komodakis, and A. Smolic, "Deep tone mapping operator for high dynamic range images," *IEEE Trans. Image Process.*, vol. 29, pp. 1285-1298, Sep. 2019, 10.1109/TIP.2019.2936649.

[28] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. the 27th Int. Conf. Neural Info. Process. Syst. (NIPS 14)*, Dec. 2014, pp. 2672-2680.

[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," Nov. 2014, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1411.1784

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," Sep. 2014, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1409.1556

[31] K. Sheth, "Deep neural networks for HDR imaging," Nov. 2016, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1611.00591

[32] R. Mantiuk, R. Mantiuk, A. Tomaszewska, and W. Heidrich, "Color correction for tone mapping," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 193-202, Mar. 2009, 10.1111/j.1467-8659.2009.01358.x.

[33] X. Yang, K. Xu, Y. Song, Q. Zhang, X. Wei, and R. W.H. Lau, "Image Correction via Deep Reciprocating HDR Transformation," in *Proc. 2018 IEEE/CVF Conf. Comput. Vis. and Pattern Recog. (CVPR)*, Jun. 2018, pp. 1798-1807, 10.1109/CVPR.2018.00193.

[34] N. Zhang, C. Wang, Y. Zhao, and R. Wang, "Deep tone mapping network in HSV color space," in *Proc. 2019 IEEE Visual Comm. and Image Process. (VCIP)*, Dec. 2019, 10.1109/VCIP47243.2019.8965992.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E.P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process*., vol. 13, no. 4, pp. 600-612, Apr. 2004, 10.1109/TIP.2003.819861.

[36] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1-15, Nov 2017, 10.1145/3130800.3130816.

[37] H. G. Barrow, J. M. Tenenbaum, A Hanson, and E Riseman, "Recovering intrinsic scene characteristics," *Comput. Vis. Syst.*, vol. 2, pp. 3-26, 1978, 10.1.1.34.7382.

[38] J. Zhang and J. Lalonde, "Learning High Dynamic Range from Outdoor Panoramas," in *Proc. 2017 IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4529-4538, 10.1109/ICCV.2017.484.

[39] D. Marnerides, T. Bford-Rogers, J. Hatchett, and K. Debattista, "ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content," *Comupt. Graph. Forum*, vol. 37, no. 2, pp. 37-49, May 2018, 10.1111/cgf.13340.

[40] H. Jang, K. Bang, J. Jang, and D. Hwang, "Inverse Tone Mapping Operator Using Sequential Deep Neural Networks Based on the Human Visual System," *IEEE Access*, vol. 6, pp. 52058-52072, Sep. 2018, 10.1109/ACCESS.2018.2870295.

[41] Y. Kinoshita and H. Kiya, "iTM-Net: Deep Inverse Tone Mapping Using Novel Loss Function Considering Tone Mapping Operator," *IEEE Access*, vol. 7, pp. 73555-73563, May. 2019, 10.1109/ACCESS.2019.2919296.

[42] C. Wang, Y. Zhao, and R. Wang, "Deep Inverse Tone Mapping for Compressed Images," *IEEE Access*, vol. 7, pp. 74558-74569, Jun. 2019, 10.1109/ACCESS.2019. 2920951.

[43] M. S. Santos, T. I. Ren, and N. K. Kalantari, "Single image HDR reconstruction using a CNN with masked features and perceptual loss," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 80:1-80:10, Jul. 2020, 10.1145/3386569.3392403.

[44] L. A Gatys, A. S Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. the 2016 IEEE Conf. Comput. Vis. and Pattern Recog. (CVPR)*, Jun. 2016, pp. 2414–2423, 10.1109/CVPR.2016.265.

[45] S. Y. Kim, D. E. Kim, and M. Kim, "ITM-CNN: Learning the Inverse Tone Mapping from Low Dynamic Range Video to High Dynamic Range Displays Using Convolutional Neural Networks," in *Proc. 2018 Asian Conf. Comput. Vis. (ACCV)*, May 2019, pp. 395-409, 10.1007/978-3-030-20893-6_25.

[46] Y. Xu, L. Song, R. Xie, and W. Zhang, "Deep Video Inverse Tone Mapping," in *Proc. 2019 IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 142-147, 10.1109/BigMM.2019.00-32.

[47] S. Y. Kim, J. Oh, and M. Kim, "JSI-GAN: GAN-Based Joint Super-Resolution and Inverse Tone-Mapping with Pixel-Wise Task-

Specific Filters for UHD HDR Video," Sep. 2019, arXiv preprint. [Online]. Available: https://arxiv.org/abs/1909.04391.

[48] *SMPTE ST-2084: High dynamic range electro-optical transfer function of mastering reference displays*, Society of Motion Picture and Television Engineers (SMPTE), White Plains, NY, USA, 2014, pp. 1-14, 10.5594/SMPTE.ST2084.2014.

[49] *Recommendation ITU-R BT.2100-1: Image parameter values for high dynamic range television for use in production and international programme exchange*, International Telecommunication Union (ITU), Geneva, Swiss Confederation, Jul. 2017, pp. 5-13. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2100-1-201706-I!!PDF-E.pdf

[50] Q. Yan, "Attention-Guided Network for Ghost-Free High Dynamic Range Imaging," in *Proc. 2019 IEEE/CVF Conf. Comput. Vis. and Pattern Recog. (CVPR)*, Jun. 2019, pp. 1751-1760, 10.1109/CVPR.2019.00185.

[51] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*, 1st ed., Williston, VT, USA, Morgan & Claypool, Jan. 2006, pp. 12-13.

[52] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. the 37th Asilomar Conf. Signals, Syst. and Comput.*, Nov. 2003, pp. 1398-1402, 10.1109/ACSSC.2003.1292216.

[53] S. L. Tade, V. Vyas, "Tone Mapped High Dynamic Range Image Quality Assessment Techniques: Survey and Analysis," *Arch. of Computat. Methods in Eng.*, vol. 28, no. 3, Apr. 2020, 10.1007/s11831-020-09428-y.

[54] M. Lin, J. Yang, and O. Yadid-Pecht, "Deep Single Image Enhancer," in *Proc. 16th IEEE Int. Conf. Advanced Video and Signal Based Surveillance*, Sep. 2019, pp. 1-6. 10.1109/AVSS.2019.8909891.

[55] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," Nov. 2015, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1511.07122

[56] H. Zeng, X. Zhang, Z. Yu and Y. Wang, "SR-ITM-GAN: Learning 4K UHD HDR with a Generative Adversarial Network," *IEEE Access*, vol. 8, pp. 182815-182827, Oct. 2020, 10.1109/ACCESS.2020.3028584.

[57] D. Marnerides, T. Bashford-Rogers, and K. Debattista, "Spectrally Consistent UNet for High Fidelity Image Transformations," Apr. 2020, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/2004.10696

[58] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Feb. 2015, *arXiv preprint*. [Online]. Available: https://arxiv.org/abs/1502.03167

[59] B. Funt, "Funt et al. HDR Dataset," 2010, *School of Computer Science, Simon Fraser University*. [Online]. Available: https://www2.cs.sfu.ca/~colour/data/funt_hdr/

[60] K. Ma, K. Zeng and Z. Wang, "Perceptual Quality Assessment for Multi-Exposure Image Fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345-3356, Jun. 2015, 10.1109/TIP.2015.2442920.

[61] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049-2062, Apr. 2018, 10.1109/TIP.2018.2794218.

[62] J. Morovic, *Color Gamut Mapping*, 1st ed., Barcelona, Spain, Hewlett-Packard Company, Jun. 2008, p. 112.

[63] *Recommendation ITU-R BT.2124-0: Objective metric for the assessment of the potential visibility of colour differences in television*, International Telecommunication Union (ITU), Geneva, Swiss Confederation, Jan. 2019, pp. 2-3. [Online]. Available: https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2124-0-201901-I!!PDF-E.pdf

[64] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imaging*, vol. 3, no. 1, pp. 47-57, Dec. 2016, 10.1109/TCI.2016.2644865.

[65] *IEC 61966-2-1:1999 - Multimedia systems and equipment - Colour measurement and management - Part 2-1: Colour management - Default RGB colour space - sRGB*, International Electrotechnical Commission (ICE), Geneva, Swiss Confederation, 1999, pp. 17-18. [Online]. Available: https://webstore.iec.ch/publication/6169

[66] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. on Learning Representations (ICLR 2015)*, May 2015, pp. 1-15.

[67] C. Guo and X. Jiang, "Color Difference Matrix Index for Tone-mapped Images Quality Assessment," in *Proc the 3rd Int. Conf. Comm., Info. Management and Network Security (CIMNS)*, Oct 2018, pp. 75-78, 10.2991/cimns-18.2018.17.

[68] M. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J. Lalonde, "Learning to predict indoor illumination from a single image," *ACM Trans. Graph.*, vol. 36, no. 6, pp. 1-14, Nov. 2017, 10.1145/3130800.3130891.

[69] L. H. Chen, C. G. Bampis, Z. Li, A. Norkin and A. C. Bovik, "ProxIQA: A Proxy Approach to Perceptual Optimization of Learned Image Compression," *IEEE Trans. Image Process.*, vol. 30, pp. 360-373, Nov. 2020, 10.1109/TIP.2020.3036752.

[70] Q. Jiang, F. Shao, W. Lin and G. Jiang, "BLIQUE-TMI: Blind Quality Evaluator for Tone-Mapped Images Based on Local and Global Feature Analyses," *IEEE Trans. Circuits & Sys. Video Tech.*, vol. 29, no. 2, pp. 323-335, Feb. 2019, 10.1109/TCSVT.2017.2783938.

[71] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," Oct. 2016, *Distill*. [Online]. Available: https://distill.pub/2016/deconv-checkerboard

[72] Z. Mai, H. Mansour, R. Mantiuk, P. Nasiopoulos, R. Ward and W. Heidrich, "Optimizing a Tone Curve for Backward-Compatible High Dynamic Range Image and Video Compression," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1558-1571, Jun. 2011, 10.1109/TIP.2010.2095866.

[73] K. Gu et al., "Blind Quality Assessment of Tone-Mapped Images Via Analysis of Information, Naturalness, and Structure," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 432-443, Mar. 2016, 10.1109/TMM.2016.2518868.

[74] H. Talebi and P. Milanfar, "Learned perceptual image enhancement," 2018 IEEE International Conference on Computational Photography (ICCP), Pittsburgh, PA, USA, 2018, pp. 1-13, 10.1109/ICCPHOT.2018.8368474.

[75] W. Yang, S. Wang, Y. Fang, Y. Wang and J. Liu, "Band Representation-Based Semi-Supervised Low-Light Image Enhancement: Bridging the Gap Between Signal Fidelity and Perceptual Quality," IEEE Trans. Image Process., vol. 30, pp. 3461-3473, Mar. 2021, 10.1109/TIP.2021.3062184.

[76] H. Talebi and P. Milanfar, "NIMA: Neural Image Assessment," IEEE Trans. Image Process., vol. 27, no. 8, pp. 3998-4011, Aug. 2018, 10.1109/TIP.2020.3036752.

CHENG GUO received his B.E. degree in communication engineering from Shandong University, Weihai, China, in 2017, and the M.E. degree in communication and information system from Communication University of China, Beijing, China, in 2020, where he is currently pursuing the Ph.D. degree in communication and information system. His research interests include high dynamic range image/video, image processing, deep learning, and image quality assessment.

XIUHUA JIANG received her M.S. degree from Shandong University, Jinan, China, in 1982. She is a professor and a Ph.D. tutor in School of Information and Communication Engineering, Communication University of China. Her research interests include image quality assessment, image processing, and video compression, where she has filed 4 patents, authored 6 books, and published 13 journal papers.