Towards Real-World HDRTV Reconstruction: A Data Synthesis-based Approach

Zhen Cheng[1*], Tao Wang[2*], Yong Li[2], Fenglong Song[2✉], Chang Chen[2], and Zhiwei Xiong[1✉]

[1] University of Science and Technology of China
mywander@mail.ustc.edu.cn,zwxiong@ustc.edu.cn
[2] Huawei Noah's Ark Lab
{wangtao10,liyong156,songfenglong,chenchang25}@huawei.com

Abstract. Existing deep learning based HDRTV reconstruction methods assume one kind of tone mapping operators (TMOs) as the degradation procedure to synthesize SDRTV-HDRTV pairs for supervised training. In this paper, we argue that, although traditional TMOs exploit efficient dynamic range compression priors, they have several drawbacks on modeling the realistic degradation: information over-preservation, color bias and possible artifacts, making the trained reconstruction networks hard to generalize well to real-world cases. To solve this problem, we propose a learning-based data synthesis approach to learn the properties of real-world SDRTVs by integrating several tone mapping priors into both network structures and loss functions. In specific, we design a conditioned two-stream network with prior tone mapping results as a guidance to synthesize SDRTVs by both global and local transformations. To train the data synthesis network, we form a novel self-supervised content loss to constraint different aspects of the synthesized SDRTVs at regions with different brightness distributions and an adversarial loss to emphasize the details to be more realistic. To validate the effectiveness of our approach, we synthesize SDRTV-HDRTV pairs with our method and use them to train several HDRTV reconstruction networks. Then we collect two inference datasets containing both labeled and unlabeled real-world SDRTVs, respectively. Experimental results demonstrate that, the networks trained with our synthesized data generalize significantly better to these two real-world datasets than existing solutions.

Keywords: Real-world HDRTV reconstruction, Data synthesis, Tone mapping operators

1 Introduction

Recent years have seen the huge progress on ultra high-definition (UHD) display devices such as OLED [14], which can display high dynamic range television

*Equal contribution. This work was done when Zhen Cheng was an intern in Huawei Noah's Ark Lab.

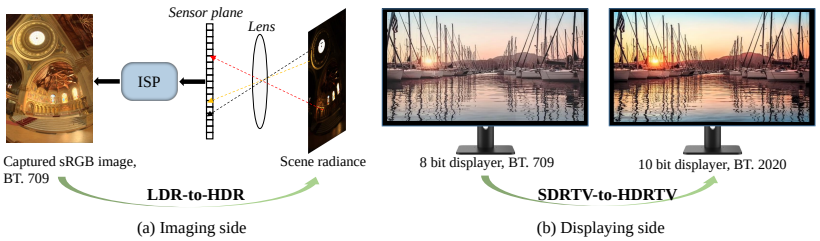

Fig. 1. Illustration of the difference between the tasks LDR-to-HDR (at the imaging side) and SDRTV-to-HDRTV (at the displaying side).

sources (HDRTVs) with high dynamic range (HDR, *e.g.*, 10 bit quantization) and wide color gamut (WCG, *e.g.*, BT.2020 [19]). However, while such HDR display devices (named HDR-TVs) become more popular, most available images/videos are still standard dynamic range television sources (SDRTVs).

To this end, previous researches [4,43,26,13,12,31,34,44] focus on recovering the linear and scene radiance maps from the captured sRGB sources, forming the LDR-to-HDR problem defined at the imaging side, as shown in Fig. 1(a). Then the scene radiance maps are transformed to HDRTVs via complicated post-processing [24,25,9]. However, such post-processing has been not well-defined for the standards of HDRTVs, resulting in severe color bias and artifacts [25,9]. Recently, researchers introduced deep learning techniques to straightforwardly reconstruct HDRTVs from their corresponding SDRTVs [24,54,25,9], forming the problem SDRTV-to-HDRTV at the dispalying side (Fig. 1(b)). Such solutions need to train convolutional neural networks (CNNs) relying on SDRTV-HDRTV pairs. Hence, the acquisition of such paired data becomes a vital problem.

There exists two possible ways to get SDRTV-HDRTV pairs: acquisition by cameras and synthesis by algorithms. The former acquires SDRTV-HDRTV pairs via asynchronous camera shots like those in super-resolution [7,5]. However, such approach faces difficulties to get large datasets for network training due to its high sensitivity to motion and light condition changes. The latter solution can be further divided into two categories: camera pipeline based and tone mapping operator (TMO) based. Camera pipeline based approaches get the scene radiance map first and then process it to SDRTV and HDRTV via different processing pipelines. However, mostly the processing from light radiance to HDRTV is unknown, which makes the solution unavailable [9]. In consequence, existing SDRTV-to-HDRTV methods rely on TMOs [33,10,16,27,40,42,29] that compress the dynamic range via global or local transformations as the degradation procedure to synthesize the SDRTV data.

However, through detailed analysis, we observe that, because TMOs aim at preserving the information from HDRTVs as much as possible, they may inherit too much information such as extreme-light details from HDRTVs, which often do not appear in real-world SDRTVs. Such information over-preservation, as shown in Fig. 2(a), is the main drawback of TMOs as SDRTV data synthesis

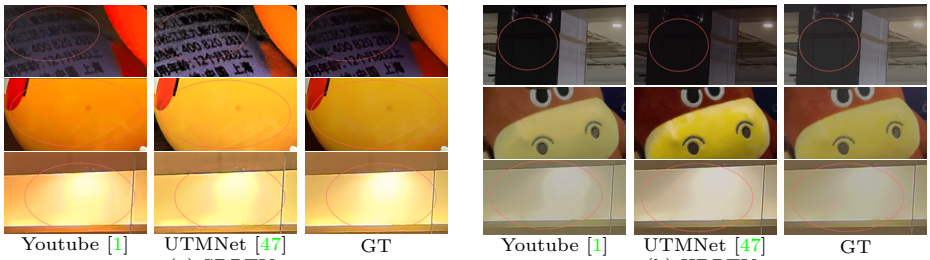

Fig. 2. (a) Drawbacks on SDRTV data synthesis of two representative TMOs [1,47]. From top to bottom: information over-preservation, color bias and artifacts. (b) Reconstruction artifacts on real-world HDRTVs of the networks (HDRTVNet [9]) trained with data synthesized by these two TMOs.

solutions. Moreover, most TMOs will also introduce color bias due to inaccurate gamut mapping and obvious artifacts such as wrong structures. Accordingly, the HDRTV reconstruction networks trained by TMO-synthesized SDRTV-HDRTV pairs are hard to generalize well to real-world cases as shown in Fig. 2(b).

To solve this problem, we propose an learning-based SDRTV data synthesis approach to synthesize realistic SDRTV-HDRTV pairs. Inspired by real-world degradation learning with the help of predefined degradations in super-resolution [52,35,8], we exploit the tone mapping priors in our method for both network structures and loss functions.

In specific, we model the SDRTV data synthesis with two streams, *i.e.*, a global mapping stream and a local adjustment one and use some representative tone mapping results to generate global guidance information for better HDRTV-to-SDRTV conversion. To train the network, we utilize different tone mapping results as the supervisor for regions with different light conditions, forming a novel unsupervised content loss to constraint different aspects of the synthesized SDRTVs. We also introduce an adversarial loss to emphasize the synthesized SDRTVs to be more realistic.

To validate the effectiveness of our approach, we synthesize SDRTV-HDRTV pairs using our method and use them to train several HDRTV reconstruction networks. For inference, we collect two inference datasets containing labeled SDRTVs captured by a smartphone and unlabeled SDRTVs from public datasets [25]. Quantitative and qualitative experimental results on these two inference datasets demonstrate that, the networks trained with our synthesized data can achieve significantly better performance than those with other data synthesis approaches.

2 Related Work

SDRTV-to-HDRTV methods. SDRTV-to-HDRTV is a highly ill-posed problem since the complicated degradation from HDRTVs to SDRTVs. While early

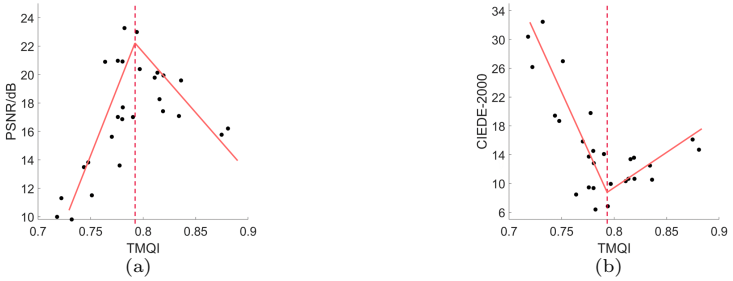

Fig. 3. The scatters showing statistical relationships between PSNR (a) and CIEDE-2000 [45] (b) *w.r.t* TMQI [53], respectively. We use solid lines and dotted lines to represent the trend and the turning point of trend changes, respectively. These metrics are evaluated and averaged on our collected RealHDRTV dataset.

researches aim at restoring HDR radiance map from a low dynamic range (LDR) input, which is called inverse tone mapping [4,43,26,13,12,31,34,44], they only consider HDR reconstruction at the imaging side and ignore the color gamut transform. Recently, SDRTV-to-HDRTV with deep learning techniques relying on synthesized SDRTV data becomes popular [24,25,9]. In this paper, we focus on the solution of data synthesis for real-world HDRTV reconstruction.

Tone mapping operators. TMOs aim at compressing the dynamic range of HDR sources but preserve image details as much as possible. Traditional TMOs always involve some useful tone mapping priors such as the Weber-Fechner law [11] and the Retinex Theory [28] to make either global mappings [42,16,1] or local mappings [10,27,33,40,29]. Recently, learning-based TMOs become popular due to their remarkable performance. They rely on ranking traditional TMOs [41,6,39,38,56] as labels for fully supervision or unpaired datasets for adversarial learning [47]. In this paper, we argue that TMOs have several drawbacks for realistic HDRTV data synthesis. Accordingly, we propose a learning-based method integrating tone mapping priors to solve these drawbacks.

3 Motivation

As we all know, the core target of TMOs is to preserve as much information as possible from the HDR sources. However, the essential of the degradation from HDRTVs to SDRTVs is to lose information selectively, *i.e.*, drop out details at extreme-light regions. Thus sometimes a contradiction will occur when we use TMOs to model the degradation. To get a deep-in understanding of this problem, we make an evaluation on 31 TMOs (detailed in the supplementary material) with our RealHDRTV dataset (detailed in Sec. 5.1).

Specifically, we use TMQI [53] (higher is better), which is mostly used for the evaluations of TMOs, to evaluate the amount of information an SDRTV preserves from the corresponding HDRTV. Meanwhile, we use PSNR (higher is

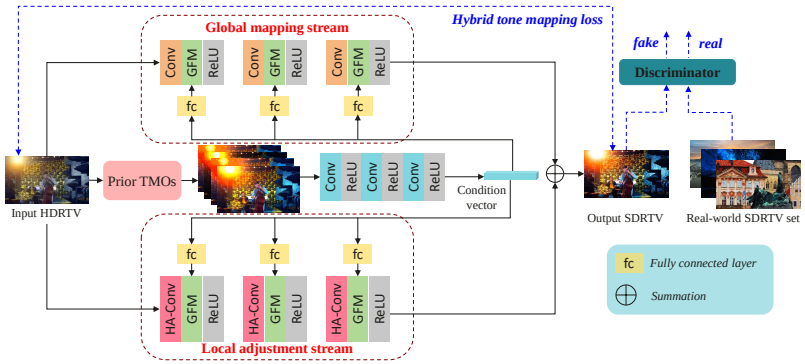

Fig. 4. Our proposed SDRTV data synthesis approach. We integrate several tone mapping priors into this framework, resulting a two-stream data synthesis network conditioned by prior tone mapping results and a novel content loss function formulated by tone mapping priors.

better) and CIEDE-2000 [45] (lower is better) to evaluate the distance between a synthesized SDRTV and the ground truth real-world one. We draw the evaluation results averaged over the RealHDRTV dataset to two scatters in Fig. 3 where each point represents a TMO.

Interestingly, we can see that, on our RealHDRTV dataset, when the TMQI of a TMO exceeds a threshold at about 0.8, the distance between synthesized and real-world data turns to increase. It indicates that the information preserved by this TMO may be too much compared with realistic SDRTVs. We can also observe such information over-preservation in Fig. 2(a). Such drawback may lead the trained HDRTV reconstruction networks fail to hallucinate the extreme-light details in real-world cases as shown in Fig. 2(b).

Moreover, most TMOs transform the color gamut by simple transformation matrix [20] or color channel rescaling [47], resulting obvious color bias, let alone possible artifacts such as halo, wrong structures and color banding occur for most TMOs [47]. The data synthesized by TMOs will lead the trained reconstruction network to generate artifacts in real-world cases as shown in Fig. 2(b).

Motivated by these drawbacks of TMOs on realistic SDRTV data synthesis, we propose a learning-based approach to synthesize training data for better HDRTV reconstruction in real-world cases.

4 Learning-based SDRTV Data Synthesis

Fig. 4 illustrates the framework of our data synthesis method. Inspired by learning real-world degradation with the help of predefined downsampling methods in the field of image super-resolution [35,8,52], we involve the prior knowledge for designing TMOs to our framework. Although these priors themselves cannot be used for straightforward degradation modeling, some of them can provide re-

gional constraints or global guidance to benefit our learning. Thus, we integrate several tone mapping priors into both network structures and loss functions.

4.1 Conditioned two-stream network

Given an input HDRTV $H \in \mathbb{R}^{X \times Y \times 3}$ where X and Y denote the image size, our network N aims to convert it into an SDRTV $S \in \mathbb{R}^{X \times Y \times 3}$ whose properties are similar as the real-world SDRTVs. Considering that we need both global transformations such as color gamut mapping and local adjustments such as selective detail preservation at extreme-light regions, our network N includes a global mapping stream N_g and a local adjustment stream N_l as shown in Fig. 4.

The global stream N_g is composed of three 1×1 convolutions which performs similarly as global TMOs with 3DLUTs [1] or S-curves [42,16]. Such network has been validated effective for global color [9] and style [17] transformations. The other stream N_l is composed of three highlight-aware convolution blocks (HA-conv, detailed in the supplementary material), which shows superior performance on the task sensitive to extreme-light regions such as SVBRDF estimation [15]. For simplicity of the data synthesis network, we straightforwardly add the results of global and the local stream together to get the final synthesized SDRTVs.

Moreover, to benefit the learning, we involve the prior knowledge of existing TMOs into these two streams. For each input HDRTV H, we obtain a number of tone mapped versions $\{S_i | i = 1, 2, \cdots, K\}$ as the condition to guide the data synthesis. Specifically, we concatenate these condition images and feed them into a condition network N_c. The condition network is composed of three convolution layers with large kernel sizes and strides followed by a global average pooling layer. The pooling layer will output a 1D condition vector $v_c \in \mathbb{R}^{B \times C_{cond}}$ where B and C_{cond} denote the batch size and the channel number, respectively.

Because the condition vector embeds sufficient global information of the prior tone mapping results, it is then used to modulate the main branch of the two stream network N. For the output feature maps $F \in \mathbb{R}^{B \times C_{feat} \times X \times Y}$ of each layer/block in the global/local stream where C_{feat} denotes the channel number, we use a fully connected layer to transform v_c to scale factors $\omega_1 \in \mathbb{R}^{B \times C_{feat}}$ and shift factors $\omega_2 \in \mathbb{R}^{B \times C_{feat}}$ and modulate the feature maps F via global feature modulation (GFM) [17], which can be described as:

$$F_{mod} = F * \omega_1 + \omega_2. \tag{1}$$

Note that we do not share the fully connected layers used for N_g and N_l, they can provide different guidances for different transformation granularities.

4.2 Hybrid tone mapping prior loss

As analyzed in Sec. 3, the synthesized SDRTVs should have several aspects: globally compressed dynamic range, accurate color gamut and lost details at extreme-light regions. However, there are no paired HDRTV-SDRTV datasets and the acquisition of large-scale and high-quality datasets for training with

imaging devices is also difficult. Therefore, we follow these region-aware aspects and divide the whole image into several regions according to their brightness distributions. After that, we transform the input HDRTVs with existing TMOs to get weak supervisors for different regions, forming a novel content loss function, namely hybrid tone mapping prior (HTMP) loss ($\mathcal{L}_{htmp}$).

Region division. At the very first, we divide the input HDRTV H into three regions, *i.e.*, the high-, mid- and low-light regions. Specifically, we get the light radiance L by linearizing H with a PQ EOTF [21] and segment the radiance map into three regions by two truncation points α and β, which are the a-th and b-th percentiles of the radiance map's histogram, respectively. The resulting region division masks are calculated as:

$$M_{high} = I(L > a), M_{low} = I(L < b), M_{mid} = \mathbf{1} - M_{high} - M_{low}, \tag{2}$$

where $I(\cdot)$ denotes the indicative function and $\mathbf{1}$ is a all-one map.

High-light loss. For the high-light regions, the output SDRTV should be saturated. Thus we use a all-one map as the supervisor at this region as:

$$\mathcal{L}_{high} = \|M_{high} \odot (\mathbf{1} - N(H))\|_1, \tag{3}$$

where $\odot$ means element-wise production. Note that, although the supervisor at the high-light regions is a all-one map, due to the fact that CNNs have denoising and smoothing effects [46], the resultant SDRTVs will become smooth here.

Low-light loss. For the low-light regions, the output SDRTV should linearly compress the radiance due to its lower bit width. Thus we use the results of a simple TMO Linear [51] $l.$ as the supervisor:

$$\mathcal{L}_{low} = \|M_{low} \odot (l(H) - N(H))\|_1. \tag{4}$$

Mid-light loss. For the mid-light regions, we need to consider both global dynamic range compression and accurate color gamut. However, there is no proper TMO for both properties. Thus we combine two TMOs to achieve this goal. In specific, we firstly use a μ-law function [23] $\mu(\cdot)$ after global color gamut mapping [20] $CGM(\cdot)$. Since the μ-law function is a logarithm curve, which is similar to the compressive response to light in the human visual system, *i.e.*, the Weber-Fechner law [11], it can provide a visually pleasant global transformation for dynamic range compression and preserve low-light details by stretching the brightness. Meanwhile, such stretching will lead to under-saturated color, so we then introduce another TMO Youtube [1] $y(\cdot)$, which uses 3D lookup tables predefined by Youtube tools for online film showcase. Youtube can provide vivid but sometimes over-saturated color. Moreover, due to its point-wise processing nature, Youtube will generate discontinuous textures near the high-light regions. Because the μ-law function and Youtube are complementary to each other, we use an invert μ-law function, *i.e.*, $\mu^{-1}(\cdot)$ with the normalized linear radiance as input to generate a weighting matrix $W = \mu^{-1}(\frac{L-\beta}{\alpha-\beta})$. So the loss function at the mid-light regions can be described as:

$$\mathcal{L}_{mid} = \|M_{mid} \odot (W \odot \mu(CGM(H)) + (\mathbf{1} - W) \odot y(H)) - N(H)\|_1. \tag{5}$$

Finally, we add the above three loss functions, forming our HTMP loss via $\mathcal{L}_{htmp} = \mathcal{L}_{high} + \mathcal{L}_{mid} + \mathcal{L}_{low}$. We also illustrate a flowchart of our HTMP loss for a more intuitive understanding in the supplementary material.

4.3 Adversarial loss

With the content loss $\mathcal{L}_{htmp}$, the network has had the ability to model the region-aware properties of realistic SDRTVs. To further emphasize the synthesized SDRTVs to be more realistic, we introduce an additional adversarial loss with a discriminator following the GAN-based low-level researches [30]. Specifically, we collect a large real-world SDRTV dataset $\mathcal{S}$ containing 3603 4K SDRTVs from public datasets [25]. We split the dataset into train and inference subsets $\mathcal{S}_{train}$ and $\mathcal{S}_{test}$ while the latter contains 25 SDRTVs. The dataset $\mathcal{S}$ contains SDRTVs captured in different environments and with different devices.

During the adversarial training, we utilize the least square GAN approach [37] with a 70×70 PatchGAN [18,30,32,57] and the overall loss function for the generator network N is $\mathcal{L}_N = \mathcal{L}_{htmp} + \lambda \mathcal{L}_{adv}$, where λ is a weighting factor. More implementation details can be found in the supplementary material.

5 Experimental results

5.1 Experimental settings

For the training of our SDRTV data synthesis network N, we collect a dataset $\mathcal{H}$ containing 3679 HDRTVs (BT.2020 with PQ EOTF [20]) from public datasets [25] as the input of network N. To validate the effectiveness of our the trained data synthesis network, we firstly train several HDRTV reconstruction networks using the SDRTV-HDRTV pairs synthesized by our well-trained N. Then we inference these networks on two real-world SDRTV datasets to see the generalization ability of trained networks.

Datasets. With the unlabeled inference dataset $\mathcal{S}_{test}$ introduced in Sec. 4.3, we can only make visual comparisons and user study to validate the quality of reconstructed HDRTVs. In order to make full-reference evaluations, we also capture a dataset, named RealHDRTV, containing SDRTV-HDRTV pairs. Specifically, we capture 93 SDRTV-HDRTV pairs with 8K resolutions using a smartphone camera with the "SDR" and "HDR10" modes. To avoid possible misalignment, we use a professional steady tripod and only capture indoor or controlled static scenes. After the acquisition, we cut out regions with obvious motions (10+ pixels) and light condition changes, crop them into 4K image pairs and use a global 2D translation to align the cropped image pairs [7]. Finally, we remove the pairs which are still with obvious misalignment and get 97 4K SDRTV-HDRTV pairs with misalignment no more than 1 pixel as our labeled inference dataset. We've release the RealHDRTV dataset in https://github.com/huawei-noah/benchmark. More details about the dataset acquisition and post-processing can be found in the supplementary material.

Data synthesis baselines. As for baseline SDRTV synthesis methods, we use three traditional TMOs, *i.e.*, Youtube [1], Hable [16] and Raman [40] because they are often used for film showcase in different online video platforms. We then collect other 27 traditional TMOs (detailed in the supplementary material) and rank the 30 TMOs using TMQI [53] and choose the best one as a new baseline named Rank following [41,6,39,38]. In addition, the state-of-the-art learning-based TMO, named UTMNet [47] is also involved here for SDRTV synthesis.
HDRTV reconstruction networks. We use the public HDRTV dataset HDRTV1K [9] as the input of both our well-trained network N and other five baselines to synthesize SDRTV-HDRTV pairs. As a result, we get 6 datasets named after their synthesis methods to train HDRTV reconstruction networks. Specifically, we choose four state-of-the-art networks (JSI-Net [25], CSRNet [17], SpatialA3DLUT [48], and HDRTVNet-AGCM [9]). To compare with existing unpaired learning-based reconstruction methods, we also involve CycleGAN [57] as another reconstruction network. Note that because CycleGAN has no explicit modeling of the unique relationships between SDRTVs and HDRTVs, we do not involve it as a data synthesis baseline. The implementation details of these networks can be found in the supplementary material.
Evaluation metrics. With the labeled dataset, *i.e.*, our RealHDRTV dataset, we evaluate the reconstructed HDRTVs using several metrics for fidelity, perceptual quality and color difference. For fidelity, we use PSNR, mPSNR [3], SSIM [49], and MS-SSIM [50]. For perceptual quality, we use HDR-VDP-3 [36] and SR-SIM [55] because they are highly correlated to the human perceptions for HDRTVs [2]. For color difference, we utilize $\triangle E_{ITP}$ [22] which is designed for the color gamut BT.2020. For visualization, we visualize HDRTVs without any post-processing following [9] to keep the details in extreme-light regions.

5.2 Generalize to labeled real-world SDRTVs

Quantitative results. Quantitative results on the generalization to our RealHDRTV dataset are shown in Table 1. As we can see, for each network, the version trained by paired data synthesized by our method works the best in terms of every evaluation metric and achieves significant gains over the baseline methods. Taking HDRTVNet-AGCM [9], the state-of-the-art HDRTV reconstruction network, as an example, compared with the best-performed TMO Hable [16], our method gains 2.60dB, 0.014 and 6.7 in terms of PSNR, SR-SIM and $\triangle E_{ITP}$, respectively. Such results validate that, with our synthesized training data, the networks can generalize well to the real-world SDRTVs. Note that there are still small misalignment between SDRTVs and HDRTVs within this dataset, the absolute full-reference metrics will be not as high as those well-aligned ones, but the metric difference can still reflect the superiority of our method.
Qualitative results. We also show some visual examples in Fig. 5, we can see that, with CycleGAN, the reconstructed HDRTVs suffer from severe color bias and lose details at extreme light regions, which is consistent with the results shown in Table 1. Although the cycle consistency has been proved useful for style transfer [57], the real-world HDRTV reconstruction does not work well

Network	TrainData	PSNR↑	mPSNR↑	SSIM↑	MS-SSIM↑	HDR-VDP3↑	SR-SIM↑	$\triangle E_{ITP}$↓
JSI-Net [25]	Raman [40]	18.91	13.23	0.708	0.719	3.54	0.736	74.2
	Rank	17.75	11.42	0.680	0.668	4.16	0.723	81.9
	UTMNet [47]	15.68	8.09	0.598	0.737	4.26	0.753	107.3
	Youtube [1]	25.47	18.56	0.842	0.923	6.32	0.942	33.6
	Hable [16]	25.45	19.60	0.851	0.918	5.71	0.926	33.8
	Ours	**27.80**	**22.92**	**0.878**	**0.933**	**6.38**	**0.943**	**27.2**
CSRNet [17]	Raman [40]	15.16	9.04	0.628	0.868	5.16	0.843	131.3
	Rank	19.41	13.43	0.749	0.912	6.28	0.929	84.0
	UTMNet [47]	12.37	5.40	0.433	0.829	4.63	0.815	172.2
	Youtube [1]	25.29	18.30	0.834	0.923	6.36	0.945	34.2
	Hable [16]	25.34	19.45	0.847	0.925	6.35	0.942	33.8
	Ours	**27.73**	**22.65**	**0.874**	**0.935**	**6.43**	**0.950**	**27.2**
Spatial-A3DLUT [48]	Raman [40]	15.35	10.77	0.726	0.882	5.61	0.852	117.4
	Rank	22.68	16.74	0.829	0.920	5.90	0.931	50.1
	UTMNet [47]	18.55	13.51	0.805	0.924	6.04	0.910	84.2
	Youtube [1]	25.27	18.23	0.832	0.921	6.34	0.943	34.2
	Hable [16]	25.48	19.40	0.846	0.924	6.35	0.942	33.5
	Ours	**27.56**	**22.44**	**0.871**	**0.933**	**6.37**	**0.945**	**27.7**
HDRTVNet-AGCM [9]	Raman [40]	19.35	13.61	0.749	0.902	5.90	0.904	88.6
	Rank	19.73	14.06	0.778	0.917	6.16	0.936	77.5
	UTMNet [47]	16.34	10.43	0.649	0.887	5.39	0.868	112.4
	Youtube [1]	25.26	18.29	0.833	0.922	6.36	0.945	34.1
	Hable [16]	25.44	19.48	0.847	0.925	6.36	0.943	33.6
	Ours	**28.04**	**22.82**	**0.876**	**0.938**	**6.47**	**0.957**	**26.9**
CycleGAN [57]	—	10.70	8.90	0.743	0.891	5.59	0.862	203.7

Table 1. Evaluation results of the HDRTV reconstruction results on the RealHDRTV dataset via various networks trained on datasets synthesized by different SDRTV data synthesis methods.

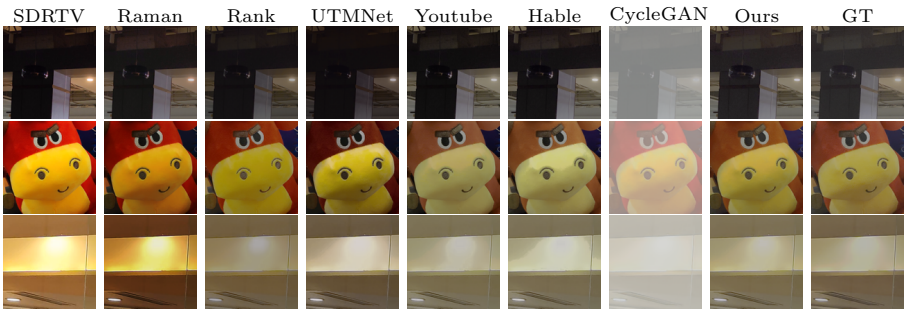

Fig. 5. Visual comparisons on real-world SDRTV-HDRTV pairs and the HDRTVs reconstructed by HDRTVNet-AGCM [9] trained with different data synthesis methods. The images are from our RealHDRTV dataset. Zoom in the figure for a better visual experience.

with such constraint. In contrast, by exploiting several tone mapping priors as both constraints and guidance, our method can perform pretty well in real-world cases. While the networks trained with data synthesized by other methods show weak ability to recover the low-light region and expand the accurate color gamut,

	Hable	Youtube	Ours	**Total**
Hable	–	125	70	**195**
Youtube	150	–	83	**233**
Ours	205	192	–	***397***

Table 2. The preference matrix from the user study on the unlabeled real-world dataset $\mathcal{S}_{test}$.

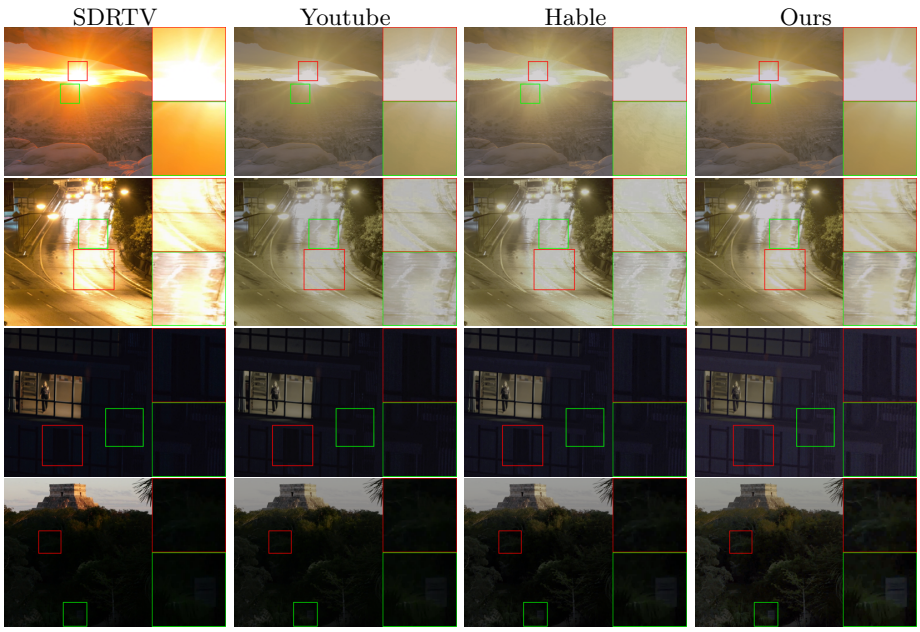

Fig. 6. Visual comparisons on the HDRTVs reconstructed by HDRTVNet-AGCM [9] trained with data synthesized by Youtube [1], Hable [16] and Ours. The input SDRTVs are from the dataset $\mathcal{S}_{test}$. Zoom in the figure for a better visual experience.

the network trained by our dataset show significant advantage over them and produce results much more close to the ground truth.

5.3 Generalize to unlabeled real-world SDRTVs

We also reveal the generalization ability of the networks trained with our synthesized dataset in a more open situation, we compare the performance of three versions (Hable, Youtube, and Ours) of the network HDRTVNet-AGCM on the unlabeled inference dataset $\mathcal{S}_{test}$ collected from public datasets [25].

User study. We conduct a user study on the reconstructed HDRTVs with 11 professional photographers for subjective evaluation. Each participant is asked to make pairwise comparisons on 3 results of each image displayed on an HDR-TV (EIZO ColorEdge CG319X with a peak brightness of 1000 nits) in a darkroom. The detailed settings can be found in the supplementary material. We show the

	PSNR↑	SSIM ↑	CIEDE ↓	TMQI ↑
Clip	13.82	0.719	18.68	0.7477
Linear [51]	16.46	0.758	15.15	0.7353
Reinhard [42]	19.94	0.776	10.65	0.8194
Raman [40]	20.97	0.627	9.52	0.7759
Kuang [27]	20.92	0.717	9.35	0.7804
Youtube [1]	22.99	0.824	6.83	0.7940
Hable [16]	23.27	0.840	6.38	0.7822
Liang [33]	16.21	0.676	14.81	0.8807
Rank [41]	16.57	0.692	14.32	**0.8850**
UTMNet [47]	15.77	0.681	16.14	0.8747
Ours	**24.54**	**0.844**	**5.80**	0.7988

Table 3. Evaluation metrics on fidelity and color difference between the SDRTVs synthesized by several methods and the ground truth ones on our RealHDRTV dataset.

preference matrix in Table 2. We can see that, when comparing our method with the best-performed TMOs, *i.e.*, Hable and Youtube, 74.5% and 69.8% of users prefer our results, respectively.

Qualitative results. We also show some examples for visual comparison in Fig. 6. We can find that while the networks trained by Youtube's and Hable's data has less awareness of high-light (the top two) and low-light (the bottom two) regions, the network trained by our data can enrich the details as well as preserve continuous structures.

To sum up, while the training datasets for both our data synthesis network and the HDRTV reconstruction networks have no overlap with our RealHDRTV and $\mathcal{S}_{test}$ datasets, the networks trained by our data show notable performance gains in both numerical and visual comparisons as well as the user study. It indicates that, our approach can serve as a better solution for paired SDRTV-HDRTV synthesis towards real-world HDRTV reconstruction.

5.4 The quality of synthesized SDRTVs

In addition to the generalization evaluations of networks trained by our data, we also evaluate the quality of synthesized SDRTVs. Specifically, we feed the HDRTVs in our RealHDRTV dataset into our well-trained data synthesis network and evaluate the distance and difference between our synthesized SDRTVs and the real-world ones. We evaluate the distances in terms of fidelity metrics PSNR and SSIM [49] and color difference for color gamut BT.709, *i.e.*, CIEDE-2000 [45]. Following the experiment in Sec. 3, we also calculate TMQI [53] to evaluate the ability of information preservation from HDRTVs. Besides the baselines compared in Sec. 5.2 and Sec. 5.3, we involve more representative TMOs for the comparison as shown in Table 3.

We can observe that, although the state-of-the-art TMOs like Liang [33] and UTMNet [47] have significantly high TMQI values, the SDRTVs generated by them are far away from the ground truth SDRTVs. On the contrary, the SDRTVs generated by our method shows much better fidelity and color accuracy by 1.27dB gain of PSNR and 0.58 drop of CIEDE-2000 compared with the best performed TMO Hable [16]. Such results are consistent with what we observe in

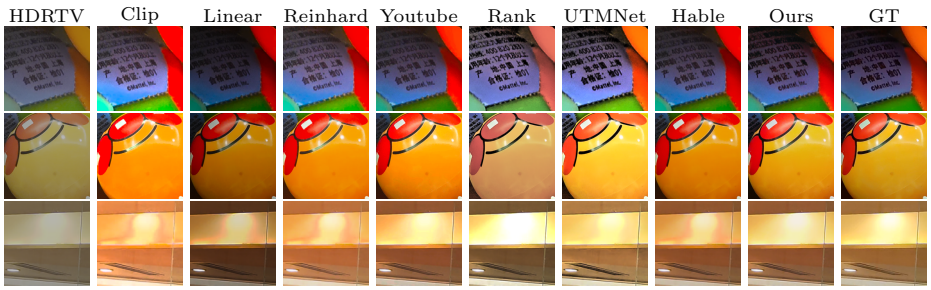

Fig. 7. Visual comparisons on SDRTVs synthesized by different representative synthesis methods together with the input HDRTVs and ground truth SDRTVs. The images are from our RealHDRTV dataset. Zoom in the figure for a better visual experience.

Fig. 3 and interestingly, we find that our average TMQI value is pretty close to the turning point in the scatters, *i.e.*, about 0.8 for this dataset. It reveals our success on avoiding information over-preservation.

We also show some visual examples in Fig. 7. We can see that, compared with the ground truth SDRTVs, the information over-preservation (*e.g.*, Clip and Rank for the top example), color bias (*e.g.*, Hable and UTMNet for the middle example) and artifacts (*e.g.*, Reinhard and Linear in the bottom example) are very obvious. Meanwhile, our method can selectively preserve the information from the HDRTVs, transform color gamut accurately and avoid the artifacts.

5.5 Ablation

With the evaluations on synthesized SDRTVs, we'd like to show some ablation studies about the network structures and loss functions, particularly the effects on the tone mapping priors we utilize in our framework. We compare the values of PSNR and CIEDE-2000 [45] calculated on the synthesized SDRTVs by different variants in Table 4 and show visual comparisons in the supplementary material.

Network design. We conduct several experiments to validate the effectiveness of tone mapping priors used for network designs. Specifically, we remove the condition network or use the input HDRTV itself to replace the condition tone mapping results to keep the parameter numbers the same. We can see in Table 4 that the condition network as well as the condition tone mapping results make very important contributions to more accurate real-world data synthesis. As Fig. 7 shows that, the condition TMOs, *i.e.*, Clip, Linear, Reinhard and Youtube we use here shows different performance advantages at different regions. For example, Linear performs very well at losing low-light details. Meanwhile, our method apparently take merits of these conditions, which validates the importance of them again. In addition, the ablation results on only the global or local stream validate the effectiveness of combining them, the visual results in the supplementary material also validate the advantages of these two streams on global and local mappings, respectively.

Loss		Network			PSNR	CIEDE
$\mathcal{L}_{htmp}$	$\mathcal{L}_{adv}$	N_c	N_l	N_g		
✓	✓	✓	✓	✓	**24.54**	**5.80**
×	✓	✓	✓	✓	11.74	27.01
✓	×	✓	✓	✓	24.24	6.00
S-Linear	✓	✓	✓	✓	16.35	15.33
S-μ-law	✓	✓	✓	✓	18.18	12.42
S-Youtube	✓	✓	✓	✓	23.32	6.64
✓	✓	×	✓	✓	24.08	6.03
✓	✓	*Self*	✓	✓	24.15	5.97
✓	✓	✓	×	✓	24.33	5.99
✓	✓	✓	✓	×	24.44	5.83

Table 4. Ablation study on the RealHDRTV dataset.

Loss function. As we can see in the table, if we only use $\mathcal{L}_{adv}$ to train the network, the network will synthesize SDRTVs far away from the real-world ones due to the lack of content and structure constraints. However, it does not mean that $\mathcal{L}_{adv}$ is useless, we can see that with the help of $\mathcal{L}_{adv}$, the network with only $\mathcal{L}_{htmp}$ achieves a notable performance gain. In addition, to show the impacts on the involved TMOs for $\mathcal{L}_{htmp}$, we use simple L_1 loss function between the tone mapping results of each TMO as the content loss to replace $\mathcal{L}_{htmp}$. As we can see in the table, with either TMO as the supervisor, the network performance will be inferior than our $\mathcal{L}_{htmp}$. Such results validate the effectiveness of our region-aware content loss. We also show a visual example in the supplementary material to illustrate their complementarity.

6 Conclusion

In this paper, we propose a data synthesis approach to synthesize realistic SDRTV-HDRTV pairs for the training of HDRTV reconstruction networks to benefit their generalization ability on real-world cases. Through statistical and visual analysis, we observe that, existing TMOs suffer from several drawbacks on the modeling of HDRTV-to-SDRTV including information over-preservation, color bias and artifacts. To solve this problem, we propose a learning-based SDRTV data synthesis to learn the aspects of real-world SDRTVs. We integrate several tone mapping priors into both network structures and loss functions to benefit the learning. Experimental results on our collected labeled and unlabeled datasets validate that, the HDRTV reconstruction networks trained by our synthesized dataset can generalize significantly better than other methods. In addition, we believe that integrating degradation priors into degradation learning framework may also be promoted to benefit other low-level vision tasks.

Acknowledgments

We acknowledge funding from National Key R&D Program of China under Grant 2017YFA0700800, and National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

1. https://www.youtube.com. 3, 4, 6, 7, 9, 10, 11, 12
2. Athar, S., Costa, T., Zeng, K., Wang, Z.: Perceptual quality assessment of uhd-hdr-wcg videos. In: 2019 IEEE International Conference on Image Processing (ICIP). pp. 1740–1744. IEEE (2019) 9
3. Banterle, F., Artusi, A., Debattista, K., Chalmers, A.: Advanced high dynamic range imaging. AK Peters/CRC Press (2017) 9
4. Banterle, F., Ledda, P., Debattista, K., Chalmers, A.: Inverse tone mapping. In: Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia. pp. 349–356 (2006) 2, 4
5. Cai, J., Zeng, H., Yong, H., Cao, Z., Zhang, L.: Toward real-world single image super-resolution: A new benchmark and a new model. In: ICCV. pp. 3086–3095 (2019) 2
6. Cao, X., Lai, K., Yanushkevich, S.N., Smith, M.: Adversarial and adaptive tone mapping operator for high dynamic range images. In: 2020 IEEE Symposium Series on Computational Intelligence. pp. 1814–1821. IEEE (2020) 4, 9
7. Chen, C., Xiong, Z., Tian, X., Zha, Z.J., Wu, F.: Camera lens super-resolution. In: CVPR. pp. 1652–1660 (2019) 2, 8
8. Chen, S., Han, Z., Dai, E., Jia, X., Liu, Z., Xing, L., Zou, X., Xu, C., Liu, J., Tian, Q.: Unsupervised image super-resolution with an indirect supervised path. In: CVPRW (June 2020) 3, 5
9. Chen, X., Zhang, Z., Ren, J.S., Tian, L., Qiao, Y., Dong, C.: A new journey from sdrtv to hdrtv. In: ICCV. pp. 4500–4509 (2021) 2, 3, 4, 6, 9, 10, 11
10. Chiu, K., Herf, M., Shirley, P., Swamy, S., Wang, C., Zimmerman, K., et al.: Spatially nonuniform scaling functions for high contrast images. In: Graphics Interface. pp. 245–245. Canadian Information Processing Society (1993) 2, 4
11. Drago, F., Myszkowski, K., Annen, T., Chiba, N.: Adaptive logarithmic mapping for displaying high contrast scenes. In: Computer graphics forum. vol. 22, pp. 419–426. Wiley Online Library (2003) 4, 7
12. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. ACM Transactions on Graphics **36**(6), 1–15 (2017) 2, 4
13. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping **36**(6) (Nov 2017). https://doi.org/10.1145/3130800.3130834, https://doi.org/10.1145/3130800.3130834 2, 4
14. Geffroy, B., Le Roy, P., Prat, C.: Organic light-emitting diode (oled) technology: materials, devices and display technologies. Polymer international **55**(6), 572–582 (2006) 1
15. Guo, J., Lai, S., Tao, C., Cai, Y., Wang, L., Guo, Y., Yan, L.Q.: Highlight-aware two-stream network for single-image svbrdf acquisition. ACM Transactions on Graphics **40**(4), 1–14 (2021) 6
16. Hable, J.: Uncharted 2: Hdr lighting. In: Game Developers Conference. p. 56 (2010) 2, 4, 6, 9, 10, 11, 12
17. He, J., Liu, Y., Qiao, Y., Dong, C.: Conditional sequential modulation for efficient global image retouching. In: ECCV. pp. 679–695. Springer (2020) 6, 9, 10
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. pp. 1125–1134 (2017) 8
19. ITU-R: Parameter values for ultra-high definition television systems for production and international programme exchange. Recommendation ITU-R BT pp. 2020–2 (2015) 2

20. ITU-R: Colour gamut conversion from recommendation itu-r bt.2020 to recommendation itu-r bt.709. Recommendation ITU-R BT pp. 2407–0 (2017) 5, 7, 8
21. ITU-R: Image parameter values for high dynamic range television for use in production and international programme exchange. Recommendation ITU-R BT pp. 2100–2 (2018) 7
22. ITU-R: Objective metric for the assessment of the potential visibility of colour differences in television. Recommendation ITU-R BT pp. 2124–0 (2019) 9
23. Kalantari, N.K., Ramamoorthi, R., et al.: Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics **36**(4), 144–1 (2017) 7
24. Kim, S.Y., Oh, J., Kim, M.: Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In: ICCV. pp. 3116–3125 (2019) 2, 4
25. Kim, S.Y., Oh, J., Kim, M.: Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In: AAAI. vol. 34, pp. 11287–11295 (2020) 2, 3, 4, 8, 9, 10, 11
26. Kovaleski, R.P., Oliveira, M.M.: High-quality brightness enhancement functions for real-time reverse tone mapping. The Visual Computer **25**(5), 539–547 (2009) 2, 4
27. Kuang, J., Johnson, G.M., Fairchild, M.D.: icam06: A refined image appearance model for hdr image rendering. Journal of Visual Communication and Image Representation **18**(5), 406–414 (2007) 2, 4, 12
28. Land, E.H., McCann, J.J.: Lightness and retinex theory. Josa **61**(1), 1–11 (1971) 4
29. Larson, G.W., Rushmeier, H., Piatko, C.: A visibility matching tone reproduction operator for high dynamic range scenes. IEEE Transactions on Visualization and Computer Graphics **3**(4), 291–306 (1997) 2, 4
30. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4681–4690 (2017) 8
31. Lee, S., An, G.H., Kang, S.J.: Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In: ECCV. pp. 596–611 (2018) 2, 4
32. Li, C., Wand, M.: Precomputed real-time texture synthesis with markovian generative adversarial networks. In: ECCV. pp. 702–716. Springer (2016) 8
33. Liang, Z., Xu, J., Zhang, D., Cao, Z., Zhang, L.: A hybrid l1-l0 layer decomposition model for tone mapping. In: CVPR. pp. 4758–4766 (2018). https://doi.org/10.1109/CVPR.2018.00500 2, 4, 12
34. Liu, Y.L., Lai, W.S., Chen, Y.S., Kao, Y.L., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Single-image hdr reconstruction by learning to reverse the camera pipeline. In: CVPR. pp. 1651–1660 (2020) 2, 4
35. Maeda, S.: Unpaired image super-resolution using pseudo-supervision. In: CVPR. pp. 291–300 (2020) 3, 5
36. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Transactions on Graphics **30**(4), 1–14 (2011) 9
37. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. pp. 2794–2802 (2017) 8
38. Montulet, R., Briassouli, A., Maastricht, N.: Deep learning for robust end-to-end tone mapping. In: BMVC. p. 194 (2019) 4, 9

39. 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. pp. 220–231. Springer (2017) 4, 9
40. Raman, S., Chaudhuri, S.: Bilateral filter based compositing for variable exposure photography. In: Eurographics. pp. 1–4 (2009) 2, 4, 9, 10, 12
41. Rana, A., Singh, P., Valenzise, G., Dufaux, F., Komodakis, N., Smolic, A.: Deep tone mapping operator for high dynamic range images. IEEE Transactions on Image Processing **29**, 1285–1298 (2019) 4, 9, 12
42. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques. pp. 267–276 (2002) 2, 4, 6, 12
43. Rempel, A.G., Trentacoste, M., Seetzen, H., Young, H.D., Heidrich, W., Whitehead, L., Ward, G.: Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. ACM Transactions on Graphics **26**(3), 39–es (2007) 2, 4
44. Santos, M.S., Ren, T.I., Kalantari, N.K.: Single image hdr reconstruction using a cnn with masked features and perceptual loss. ACM Transactions on Graphics **39**(4), 80–1 (2020) 2, 4
45. Sharma, G., Wu, W., Dalal, E.N.: The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. Color Research & Application **30**(1), 21–30 (2005) 4, 5, 12, 13
46. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9446–9454 (2018) 7
47. Vinker, Y., Huberman-Spiegelglas, I., Fattal, R.: Unpaired learning for high dynamic range image tone mapping. In: ICCV. pp. 14657–14666 (2021) 3, 4, 5, 9, 10, 12
48. Wang, T., Li, Y., Peng, J., Ma, Y., Wang, X., Song, F., Yan, Y.: Real-time image enhancer via learnable spatial-aware 3d lookup tables. In: ICCV. pp. 2471–2480 (2021) 9, 10
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004) 9, 12
50. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003) 9
51. Ward, G.: A contrast-based scalefactor for luminance display. Graphics Gems **4**, 415–21 (1994) 7, 12
52. Wei, Y., Gu, S., Li, Y., Timofte, R., Jin, L., Song, H.: Unsupervised real-world image super resolution via domain-distance aware training. In: CVPR. pp. 13385–13394 (2021) 3, 5
53. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. IEEE Transactions on Image Processing **22**(2), 657–667 (2012) 4, 9, 12
54. Zeng, H., Zhang, X., Yu, Z., Wang, Y.: Sr-itm-gan: Learning 4k uhd hdr with a generative adversarial network. IEEE Access **8**, 182815–182827 (2020) 2
55. Zhang, L., Li, H.: Sr-sim: A fast and high performance iqa index based on spectral residual. In: ICIP. pp. 1473–1476. IEEE (2012) 9
56. Zhang, N., Wang, C., Zhao, Y., Wang, R.: Deep tone mapping network in hsv color space. In: VCIP. pp. 1–4. IEEE (2019) 4
57. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV. pp. 2223–2232 (2017) 8, 9, 10