This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A New Journey from SDRTV to HDRTV

Xiangyu Chen[1*] Zhengwen Zhang[1*] Jimmy S. Ren[2,3] Lynhoo Tian[2] Yu Qiao[1,4] Chao Dong[1†]
[1]ShenZhen Key Lab of Computer Vision and Pattern Recognition, SIAT-SenseTime Joint Lab,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences [2]SenseTime Research
[3]Qing Yuan Research Institute, Shanghai Jiao Tong University [4]Shanghai AI Laboratory, Shanghai
{chxy95, zhengwen.zhang02, jimmy.sj.ren, lynhoo.tian}@gmail.com {yu.qiao, chao.dong}@siat.ac.cn

Abstract

Nowadays modern displays are capable to render video content with high dynamic range (HDR) and wide color gamut (WCG). However, most available resources are still in standard dynamic range (SDR). Therefore, there is an urgent demand to transform existing SDR-TV contents into their HDR-TV versions. In this paper, we conduct an analysis of SDRTV-to-HDRTV task by modeling the formation of SDRTV/HDRTV content. Base on the analysis, we propose a three-step solution pipeline including adaptive global color mapping, local enhancement and highlight generation. Moreover, the above analysis inspires us to present a lightweight network that utilizes global statistics as guidance to conduct image-adaptive color mapping. In addition, we construct a dataset using HDR videos in HDR10 standard, named HDRTV1K, and select five metrics to evaluate the results of SDRTV-to-HDRTV algorithms. Furthermore, our final results achieve state-of-the-art performance in quantitative comparisons and visual quality. The code and dataset are available at https://github.com/chxy95/HDRTVNet.

1. Introduction

The resolution of television (TV) content has increased from standard definition (SD) to high definition (HD) and most recently to ultra-high definition (UHD). High dynamic range (HDR) is one of the biggest features of the latest TV generation. HDRTV[1] content has much wider color space and higher dynamic range than SDR content, and HDRTV standard allows us to create images and videos that are closer to what we see in real life. While HDR display devices have become prevalent in daily life, most accessible resources are still in SDR format. Therefore, we need algorithms to convert SDRTV content to their HDRTV version. The task, denoted as SDRTV-to-HDRTV, is of great practical value, but received relatively less attention in the research community. The reason is mainly two-fold. First, existing HDRTV standards (e.g., HDR10 and HLG) have not been well defined until recent years. Second, there is a lack of large-scale datasets for training and testing. This work aims at promoting the development of this emerging area, by presenting the analysis of this problem, basic methods, evaluation metrics and a new dataset.

SDRTV-to-HDRTV is a highly ill-posed problem. In actual production, contents of SDRTV and HDRTV are derived from the same Raw file but are processed under different standards. Thus, they have different dynamic ranges, color gamuts and bit-depths. To some extent, SDRTV-to-HDRTV is similar to image-to-image translation such as Pixel2Pixel [11] and CycleGAN [37]. On the contrary, the task of LDR-to-HDR, which is similar in terms of name, has completely different connotations. LDR-to-HDR methods [21, 26, 10, 24, 5] aim to predict the HDR scene luminance in the linear domain, which is closer to Raw file in essence. SDRTV-to-HDRTV has recently been touched in Deep SR-ITM [19] and JSI-GAN [20], which try to solve the problem of joint super-resolution and SDRTV-to-HDRTV. Although the above-mentioned works are all related to SDRTV-to-HDRTV, they are not dedicated to this problem. We will detail these comments in Sec. 2 and Sec. 3.

This paper aims to address SDRTV-to-HDRTV based on deep understanding of this problem. We first provide a simplified formation pipeline for SDRTV/HDRTV content, which consists of tone mapping, gamut mapping, transfer function and quantization, as in Fig. 1(a). Based on the formation pipeline, we propose a solution pipeline, including adaptive global color mapping (AGCM), local enhancement (LE) and highlight generation (HG). For AGCM, we propose a novel color condition block to extract global image prior and adapt different images. With only 1×1 filters, the network achieves the best performance with less parameters compared with other photo retouching methods such as

*indicates contribute equally. †Corresponding author.
[1]We add a suffix TV after HDR/SDR to indicate content in HDRTV/SDR-TV format and standard.

CSRNet [8], HDRNet [6] and Ada-3DLUT [35]. Besides, we adopt a commonly used ResNet-based network and a GAN-based model for LE and HG, respectively.

To promote the progress of this new research area, we collect a new large-scale dataset, named HDRTV1K. We also select five evaluation metrics – PSNR, SSIM, SR-SIM [36], ΔE_{ITP} [17] and HDR-VDP3 [25], to evaluate the mapping accuracy, structural similarity (SSIM and SR-SIM), color difference and visual quality, respectively.

Our contributions are four-fold: (1) We conduct a detailed analysis for SDRTV-to-HDRTV task by modeling the formation of SDRTV/HDRTV content. (2) We propose a three-step SDRTV-to-HDRTV solution pipeline and a corresponding method, which performs best in quantitative and qualitative comparisons. (3) We present a novel global color mapping network based on color condition blocks. With about only 35K parameters, it can still achieve state-of-the-art performance. (4) We provide a HDRTV dataset and select five metrics to evaluate SDRTV-to-HDRTV algorithms.

2. Preliminary

Background. In this paper, we use SDRTV/HDRTV to represent the content (including image and video) under SDR-TV/HDR-TV standard. The two standards are specified in [14, 12] and [15, 16], respectively. The basic elements of the HDR-TV standard generally include wide color gamut [15], PQ or HLG OETF [16] and 10-16 bits depth. In terms of name, "LDR" is often used in academia, and "SDR" is generally used in industry. Essentially, both of them represent content with low dynamic range but generated from TV production and camera ISP, respectively. For the convenience of distinction and reference, we uniformly use "LDR-to-HDR" and "SDRTV-to-HDRTV" to represent the conventional image HDR reconstruction and conversion of content from SDR-TV to HDR-TV standard.

Explanation. Before introducing our method, we first explain that SDRTV-to-HDRTV is functionally different from the previous LDR-to-HDR (i.e., inverse tone mapping) problem. Although the concept of "HDR" is involved in these issues, it is undeniable that the connotations of HDR are not the same. It is non-trivial to explain the concepts and differences exhaustively due to the overwhelm of data-level details. In general, the previous LDR-to-HDR methods aim to predict the luminance of images in the linear domain, which is the physical brightness of the scene. In contrast, our goal is to predict HDR images with the HDR *display format* in the pixel domain, which are encoded in HDR-TV standards, such as HDR10, HLG and Dolby Vision. Essentially, content in HDR-TV standard can also be generated from HDR scene radiance. However, the process is an engineering problem and still requires a lot of operations, such as tone mapping and gamut mapping. Therefore, the methods of these two different tasks are not generalizable.

3. Analysis

In this section, we first present a simplified SDRTV/HDRTV formation pipeline which contains the most critical steps in actual production. Then, based on the analysis of the formation pipeline, we propose a novel three-step solution based on the idea of "divide-and-conquer". Finally, we compare different solution pipelines of previous methods.

3.1. SDRTV/HDRTV Formation Pipeline

To further understand the task of SDRTV-to-HDRTV, we introduce a simplified formation pipeline of SDRTV and HDRTV based on camera ISP and HDR-TV content production [18] as depicted in Fig. 1(a). Although there are some operations we do not mention here, such as denoising and white balance in camera pipeline and color grading in HDR content production, we retain the four key operations that lead to the difference between SDRTV and HDRTV, which are tone mapping, gamut mapping, opto-electronic transfer function and quantization. In the following equations, we use the subscript "S" to represent SDRTV and "H" to represent HDRTV. More details about the pipelines can be found in the supplementary material.

Tone mapping. Tone mapping is used to transform the high dynamic range signals to low dynamic range signals for adapting different display devices. The process includes global tone mapping [3, 28, 33] and local tone mapping [22, 23]. Global tone mapping processes all pixels equally with the same function, while parameters are generally related to global image statistics (e.g., average luminance). Local tone mapping can be adaptive to local content and human preference but often brings high computational cost and artifacts. Thus, global tone mapping is mainly used in the HDRTV/SDRTV image formation. The function of global tone mapping can be denoted as:

$$I_{tS} = T_S(I|\theta_S),\ I_{tH} = T_H(I|\theta_H), \tag{1}$$

where T_S and T_H represent the specific tone mapping functions, θ_S and θ_H are coefficients related to image statistics. It is noteworthy that S-shape curves are commonly used for global tone mapping and clipping operations often exist in actual process of tone mapping.

Gamut mapping. Gamut mapping is to convert colors from source gamut to target gamut while still preserving the overall look of the scene. According to the standards of ITU-R [14] and [15], the transformations from the original XYZ space to SDRTV (rec.709) and HDRTV (rec.2020) can be represented as :

$$\begin{bmatrix} R_{709} \\ G_{709} \\ B_{709} \end{bmatrix} = M_S \begin{bmatrix} X \\ Y \\ Z \end{bmatrix},\ \begin{bmatrix} R_{2020} \\ G_{2020} \\ B_{2020} \end{bmatrix} = M_H \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{2}$$

where M_S and M_H are 3×3 constant matrices.

Opto-electronic transfer function. Opto-electronic transfer function abbreviates as OETF. It is used to convert linear optical signals to non-linear electronic signals in the image formation pipeline. For SDRTV, it approximates a gamma exponential function as $I_{fS} = f_S(I) = I^{1/2.2}$. For HDRTV, there are several kinds of OETFs for different standards such as PQ-OETF [13] for HDR10 standard and HLG-OETF [16] for HLG standard (HLG stands for Hybrid Log-Gamma). We take the PQ-OETF as an example:

$$I_{fH} = f_H(I) = (\frac{a_1 + a_2 I^{b_1}}{1 + a_3 I^{b_1}})^{b_2}, \tag{3}$$

where a_1, a_2, a_3, b_1, b_2 are constants.

Quantization. After the above operations, the encoded pixel values are quantized with the function:

$$I_q = Q(I, n) = \frac{\lfloor (2^n - 1) \times I + 0.5 \rfloor}{2^n - 1}, \tag{4}$$

where n is 8 for SDRTV and 10-16 for HDRTV.

In summary, the HDRTV and SDRTV content formation pipelines are given by:

$$I_S = Q_S \circ f_S \circ M_S \circ T_S(I_{raw}), \tag{5}$$

$$I_H = Q_H \circ f_H \circ M_H \circ T_H(I_{raw}), \tag{6}$$

where $\circ$ represents the connection between two operations.

3.2. Proposed Solution Pipeline

According to the above formation pipeline, the process of SDRTV-to-HDRTV can be formulated as:

$$I_H = Q_H \circ f_H \circ M_H \circ T_H \circ T_S^{-1} \circ M_S^{-1} \circ f_S^{-1} \circ Q_S^{-1}(I_S), \tag{7}$$

where $T_S^{-1}, M_S^{-1}, f_S^{-1}, Q_S^{-1}$ denote the inversions of the corresponding operations.

We obtain the following observations and insights based on the formation pipeline. Firstly, many critical operations in the formation pipeline are global operations, such as global tone mapping, OETF and gamut mapping. Moreover, the inversions of these operations are also global operations or can be approximately equal to global operations. These operations can be processed using global operators. Secondly, some operations such as local tone mapping and dequantization rely on local spatial information, which can be processed by local operators. Thirdly, there is severe information compression/loss. For example, highlight areas are generally processed through a shoulder operation or simply clipped by tone mapping.

Inspired by the observations, we propose a new SDRTV-to-HDRTV solution pipeline using the idea of "divide and conquer". Our method includes three steps, as shown in Fig. 1(d). The first step is adaptive global color mapping, which aims at dealing with global operations. This step roughly and adaptively converts the input from SDRTV domain to HDRTV domain. Then, we perform local enhancement, which utilizes local information to enhance the result of the first step. Finally, the highlight generation step is to restore the lost information in the overexposed regions.

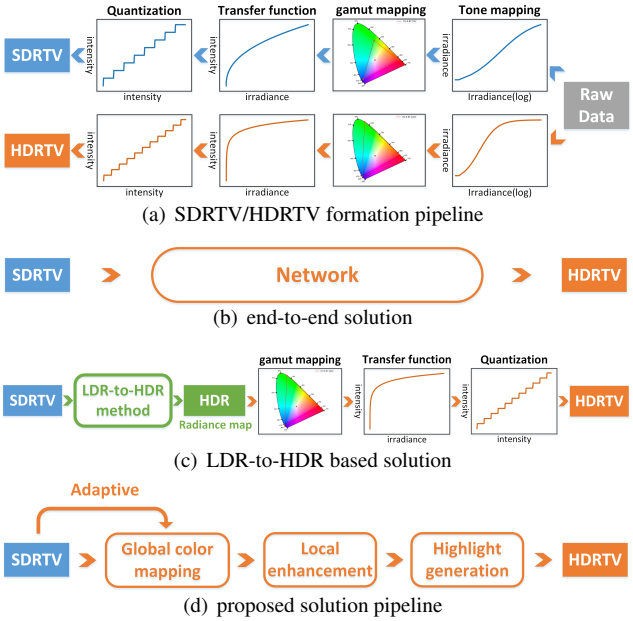

Figure 1. Analysis of SDRTV-to-HDRTV. (a) Simplified SDRTV /HDRTV content formation pipeline. (b) Existing end-to-end solution pipeline. (c) Solution pipeline based on LDR-to-HDR methods. (d) Proposed solution pipeline. Please refer to supplementary file for detailed explanations of the above diagrams and curves.

3.3. Comparison with Existing Solutions

As an actual industrial problem, SDRTV-to-HDRTV is rarely discussed in academia. In this part, we elaborate on the two groups of existing solutions.

End-to-end solution. Image-to-Image translation methods [11, 37] and joint SDRTV-to-HDRTV and super-resolution methods [19, 20] learn a direct mapping with an end-to-end model to solve the problem, as shown in Fig. 1(b). However, their methods do not consider the imaging mechanism of SDRTV and HDRTV, thus the results contain some obvious local artifacts and unnatural colors.

LDR-to-HDR based solution. LDR-to-HDR is discussed a lot in academia [29, 2, 27]. Although these methods are dedicated to predicting HDR scene radiance, it is still necessary to compare them with ours. As in Fig.1(c), the HDR radiance map generated by LDR-to-HDR methods needs to go through color gamut mapping to rec.2020, applying PQ or HLG OETF and quantization. For HDR10 standard, the results are obtained by setting the maximum brightness to 1000 cd/m^2. Since these steps need to be adjusted according to different data in actual processing, it

is hard to get a fair comparison. In this paper, we use the same processing pipeline as [19, 20] to compare with LDR-to-HDR methods.

4. Method

We propose HDRTVNet, a cascaded method consisting of three deep networks for SDRTV-to-HDRTV. Each network corresponds to each step of the solution pipeline.

4.1. Adaptive Global Color Mapping

As the first step, adaptive global color mapping (AGCM) aims to achieve image-adaptive color mapping from SDRTV to HDRTV domain. As shown in Fig. 2, the proposed model consists of a base network and a condition network.

4.1.1 Base Network

The base network is designed to handle global operations, which works on each pixel independently. This global mapping can be denoted as:

$$I_B(x, y) = f(I_S(x, y)), \forall (x, y) \in I_S, \tag{8}$$

where (x, y) represents the coordinate of the pixel in the image I and I_B represents the output of base network. As demonstrated in CSRNet [8], a fully convolution network with only 1×1 convolutions and activation functions can achieve this kind of global mapping. Thus, our base network is composed of N_l convolutional layers with 1×1 filters and N_l-1 ReLU activations, which can be denoted as:

$$I_B = Conv_{1\times1} \circ (ReLU \circ Conv_{1\times1})^{N_l-1}(I_S). \tag{9}$$

The proposed base network takes an 8-bits SDRTV image as input and generates an HDRTV image encoded with 10-16 bits. Although the base network can only learn a one-to-one color mapping, it also achieves considerable performances, as shown in Tab. 1. It is noteworthy that the base network can perform like a 3D lookup table (3D LUT) with fewer parameters rather than learning 3D LUT directly, and please refer to supplementary material for more results.

4.1.2 Condition Network

The global priors are indispensable for adaptive global color mapping. For example, global tone mapping requires global image statistics. To achieve image-adaptive mapping, we add a condition network to modulate the base network. Previous works [34, 8] usually adopt the prior of image content, where the condition network can extract spatial and local information from the input image by $N_k \times N_k$ ($N_k > 1$) filters. However, for SDRTV-to-HDRTV problem, we find that global mapping is conditioned on global image statistics or color distribution. This kind of color condition is independent of spatial information, thus it is inappropriate to adopt previous structures of condition in this problem. Our proposed condition network focuses on extracting information related to color to realize adjustable mapping. As shown in Fig. 2, the proposed condition network consists of several color condition blocks (CCB), convolution layers, feature dropout and global average pooling.

Color condition block. A color condition block contains a convolution layer with 1×1 filters, an average pooling layer, a LeakyReLU activation and an instance normalization layer [4], which can be written as follows:

$$CCB(\cdot) = IN \circ LReLU \circ avgpool \circ Conv_{1\times1}(\cdot), \tag{10}$$

where $\cdot$ denotes the input of CCB. The condition network takes a down-sampled SDRTV image as input and outputs a condition vector V. Our condition network is denoted by:

$$V = GAP \circ Conv_{1\times1} \circ Dropout \circ CCB^{N_c}(I_S). \tag{11}$$

Since the convolutional layers only contain 1×1 filters, the condition network can not extract local features. With the help of pooling layers, the network can extract global priors based on image statistics. To avoid overfitting, we add a dropout before the last convolutional layer and global average pooling. It performs like adding a multiplicative Bernoulli noise on features which has an effect similar to data augmentation. It is worth noting that even if we take the image by shuffling the pixels randomly as input, we can also obtain comparable performance with the correct arrangement of pixels. It suggests that the effective prior is not related to spatial information in global color mapping.

4.1.3 Global Feature Modulation

To utilize the extracted global priors, we introduce the global feature modulation (GFM) [8] strategy to modulate the base network, which has been successfully applied in photo retouching tasks. Through GFM, the intermediate features of the base network can be modulated by scaling and shifting operations. It can be described as:

$$GFM(x_i) = \alpha_1 * x_i + \alpha_2, \tag{12}$$

where x_i denotes the feature map. α_1 and α_1 represent the scale and shift factor, respectively.

Overall, AGCM network can be formulated as:

$$\begin{aligned} I_{AGCM} = GFM \circ Conv_{1\times1} \circ (ReLU \circ \\ GFM \circ Conv_{1\times1})^{N_l-1}(I_S), \end{aligned} \tag{13}$$

where I_{AGCM} denotes the output of adaptive global color mapping. To optimize adaptive global color mapping, we minimize the distance of the output and the ground truth HDRTV image using L2 loss function.

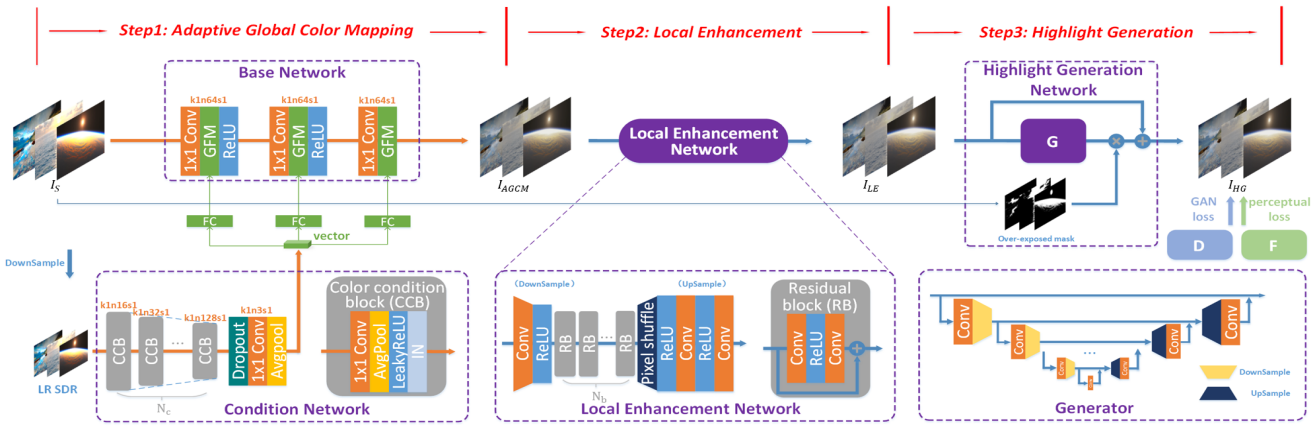

Figure 2. The architecture of the proposed three-step SDRTV-to-HDRTV method. Each step has a corresponding network.

4.2. Local Enhancement

Local enhancement (LE) is performed followed by AGCM. Although AGCM can obtain considerable performance, local enhancement is indispensable for SDRTV-to-HDRTV. It is a remarkable fact that if we directly use local operations to learn end-to-end mapping before adaptive global color mapping, the output results often have obvious artifacts. Details can be founded in the supplementary material.

To achieve local enhancement, we directly adopt a classical ResNet structure [9]. More advanced architectures can be used here, but this is not the focus of this work. Details of the enhancement network are described as follows. This step takes the output of adaptive color mapping as input. During this process, the input is down-sampled by the first convolutional layer, then goes through several residual blocks (RB), and finally is up-sampled to the original size as output. The overall operation can be represented as:

$$\begin{aligned} I_{LE} = Conv \circ ReLU \circ Conv \circ ReLU \circ \\ PS \circ RB^{N_b} \circ ReLU \circ Conv(I_{AGCM}), \end{aligned} \tag{14}$$

where M is the number of residual blocks. L_{LE} is the HDR image generated by the proposed local enhancement network. L2 loss function is adopted for local enhancement.

4.3. Highlight Generation

The third step of our solution pipeline is highlight generation (HG), which aims at hallucinating some details that were lost due to extreme compression. To achieve this goal, we adopt a generative adversarial network (GAN) [7] that has the potential to generate details. This network consists of a generator and a discriminator. We adopt an encoder-decoder architecture with skip connections [30] as generator and an commonly used VGG architecture [32] as discriminator. The formulation of this step can be represented as:

$$I_o = I_{mask} \odot G(I_i) + I_i, \tag{15}$$

where G denotes the generator and $\odot$ denotes element-wise multiplication. The mask I_{mask} can be computed by $p = max(I_i - \gamma, 0)/1 - \gamma$ as [24]. For learning highlight generation network, we joint optimize three types of losses, including L1 loss, perceptual loss and GAN loss, which can be formulated as:

$$L_{HG} = \alpha L_1 + \beta L_P + \gamma L_{GAN}, \tag{16}$$

where α, β, γ are loss weights. The perceptual loss is to measures the similarity in feature space, as $L_P = \|\psi(I_{HG}) - \psi(I_{LE})\|_2^2$, where $\psi(I)$ represents the feature maps of image I. The GAN loss is denoted as $L_G = -\log D(G(I_{LE}))$, where D denotes the discriminator.

5. Experiments

5.1. Experimental Setup.

Dataset. We collect 22 videos under HDR10 standard and their counterpart SDR version from YouTube as [19]. All of these HDR videos are encoded by PQ-OETF and rec.2020 gamut. 18 video pairs are used for training and the left for testing. To avoid high coherence among extracted frames, we sample a frame every two seconds of each video and generate a training set with 1235 images. Besides, 117 images without repeated scenes are selected as the test set.

Training details. For the proposed AGCM, the base network consists of 3 convolution layers with 1×1 kernel size and 64 channels, and the condition network contains 4 CCBs. For LE, all convolution filters are of size 3×3 with 64 output channels, except for the convolution layer in the pixel shuffle module[31] with 256 channels, and the output layer with 3 channels. The number of the RBs is 16. Strides of all layers are set to 1 except for the first convolution layer with a stride of 2. For the part of HG, there are five convolution layers followed by max-pooling operation and also five convolution layers with pixel shuffle. Each filter has kernel size of 3×3. The number of channels is increased from

64 to 1024 in the downsampling process and reverses in the process of upsampling. More implementation details can be found in the supplementary material.

5.2. Evaluation of SDRTV-to-HDRTV

Evaluation metrics. We employ five evaluation metrics for comprehensive comparisons, including PSNR, SSIM, SR-SIM [36], HDR-VDP3 [25] and ΔE_{ITP} [17]. SSIM and SR-SIM are commonly used to measure image similarity. Although they are designed to evaluate SDR image, [1] shows that SR-SIM has a favorable performance of evaluation for HDR standard. Besides, we introduce ΔE_{ITP} to measure the color difference, which is designed for HDRTV. HDR-VDP3 is a new improved version of HDR-VDP2 that supports the rec.2020 gamut. To employ HDR-VDP3, results are compared by setting "side-by-side" task, "rgb-bt.2020" color encoding, 50 pixel per degree and option of "rgb-display" with "led-lcd-wcg".

Visualization. We directly show the HDRTV images encoded in 16-bits "PNG" format without extra processing. Since HDRTV images are decoded by gamma EOTF on SDR screens, they may look relatively darker than on HDR screens. Previous work [19, 20] shows HDRTV images by software for visualization. However, it introduces an unfair comparison since the video player may reduce the unnatural artifacts of original HDRTV images. In contrast, our visualization method preserves the details in highlight areas and compares all methods in the same conditions. Tone mapped results can be founded in the supplementary file.

5.3. Comparison with Other Methods

Compared methods. We compare our results with four types of methods including joint SR with SDRTV-to-HDRTV, image-to-image translation, photo retouching and LDR-to-HDR. Since these methods are not completely designed for this task, we have done the necessary processing of these methods. For joint SR with SDRTV-to-HDRTV methods, we change the stride of the first convolutional layer to 2 for downsampling to match the size of input and output[1]. For LDR-to-HDR methods, we process the results as illustrated in Sec. 3.3. Note that we adopt the same processing steps as [19, 20]. All data-driven methods are retrained on the proposed dataset.

Quantitative comparison. As shown in Tab. 1, our method HDRTVNet outperforms other methods by a large margin on all evaluation metrics. It is worth noticing that our first step AGCM could already achieve comparable performance to Ada-3DLUT, but with only 1/17 of its parameters. The LDR-to-HDR based solutions have poor results as their pipeline is different from ours. It is hard to compare with them on a completely fair platform.

Qualitative comparison. The results of qualitative comparisons are shown in Fig.3. LDR-to-HDR based methods and image-to-image translation methods tend to generate low-contrast images. All approaches of LDR-to-HDR based, image-to-image translation and SDRTV-to-HDRTV produce unnatural colors and obvious artifacts except for HuoPhyEO [10]. Outputs generated by photo retouching methods are relatively better but suffer from the color cast. Our method HDRTVNet could produce natural colors and high contrast as referred ground truth and do not introduce any artifacts. Further, the visual quality improves with more processing steps, i.e., AGCM<AGCM+LE<AGCM+LE+HG. More results can be found in the supplementary material.

5.4. Color Transition Test

We observe that many previous methods perform poorly in the highlight regions, especially where the color changes. To reveal this phenomenon, we conduct a color transition test using a man-made color card as input image shown in Fig. 4. We obtain the following three observations. First, the unnatural transition and color blending problem can be easily observed in the outputs of several methods, which rely on learning local information (e.g., Deep SR-ITM, JSI-GAN, Pixel2Pixel, CycleGAN) or based on local conditions (e.g., 3D-LUT). Second, our method performs smooth transition even learning local information (e.g., AGCM+LE, AGCM+LE+HG), which shows the superiority of our cascaded solution pipeline. Third, blue regions suffer the most server unnatural color transition. A reasonable explanation is that blue colors are harder to recover than other colors, since more information is missing in the blue area in the process of extreme compression.

5.5. Ablation Study

Adaptive global color mapping. The process of adaptive color mapping can be observed by LUT cubes shown in Fig. 5. The color of each point in the cube corresponds to the input SDRTV color, and its coordinates in the cube correspond to values of HDRTV pixels after the current mapping. Note that if an SDRTV color is mapped to multiple HDRTV colors, the color will also appear in multiple positions of the cube. Since the basic network can only learn a one-to-one color mapping throughout the dataset, the color transition of the LUT manifold in the highlight areas is not smooth. It can be easily seen that the output of the base network suffers severe posterization artifacts. In contrast, the color condition network helps the base network learn image-adaptive color mapping. Then the artifacts disappear, and the LUT becomes concentrated. In Fig. 5, our method also performs better in the color transition test.

[1]We have also conducted experiments to remove the pixel shuffle layer instead of downsampling at the beginning, but the results show that it could not bring improvements but increase the computational cost significantly.

Method		Params↓	PSNR↓	SSIM↓	SR-SIM↓	ΔE_{ITP} ↓	HDR-VDP3↓
LDR-to-HDR	HuoPhyEO [10]	-	25.90	0.9296	0.9881	38.06	7.893
	KovaleskiEO [21]	-	27.89	0.9273	0.9809	28.00	7.431
image-to-image traslation	ResNet [9]	1.37M	**37.32**	0.9720	0.9950	**9.02**	8.391
	Pixel2Pixel [11]	11.38M	25.80	0.8777	0.9871	44.25	7.136
	CycleGAN [37]	11.38M	21.33	0.8496	0.9595	77.74	6.941
photo retouching	HDRNet [6]	482K	35.73	0.9664	0.9957	11.52	8.462
	CSRNet [8]	**36K**	35.04	0.9625	0.9955	14.28	8.400
	Ada-3DLUT [35]	594K	36.22	0.9658	**0.9967**	10.89	8.423
SDRTV-to-HDRTV	Deep SR-ITM [19]	2.87M	37.10	0.9686	0.9950	9.24	8.233
	JSI-GAN [20]	1.06M	37.01	**0.9694**	0.9928	9.36	8.169
HDRTVNet (ours)	Base Network	**5K**	36.14	0.9643	0.9961	10.43	8.305
	AGCM	**35K**	36.88	0.9655	0.9964	9.78	**8.464**
	AGCM+LE	1.41M	**37.61**	**0.9726**	**0.9967**	**8.89**	**8.613**
	AGCM+LE+HG	37.20M	**37.21**	**0.9699**	**0.9968**	**9.11**	**8.569**

Table 1. Quantitative comparisons. Red text indicates the best, blue text indicates the second and green text indicates the third.

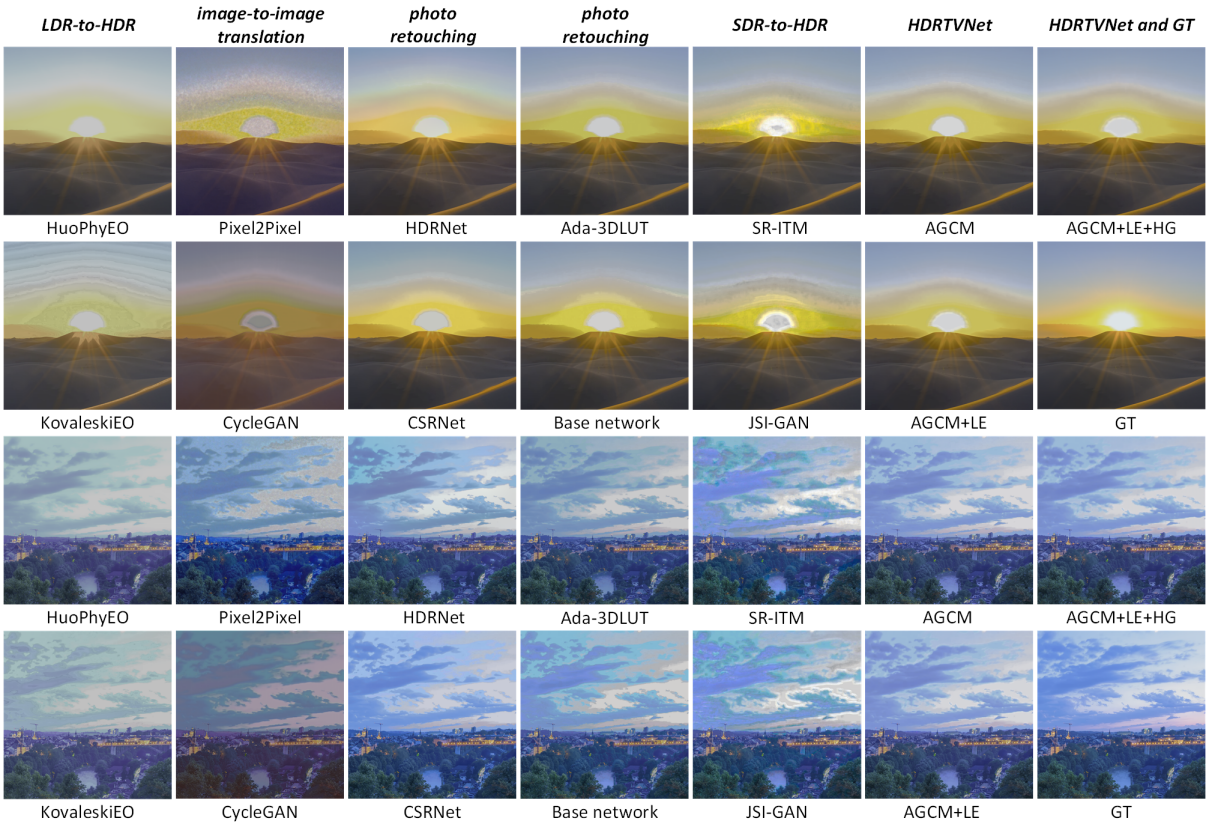

Figure 3. Qualitative comparisons. The top row describes the categories of algorithms.

Local enhancement. This part takes the adaptive-global-color-mapped image as input, aiming to handle the one-to-many color mapping and some local operations in the SDRTV-to-HDRTV process. The same color in SDRTV domain is mapped to multiple HDRTV colors, and the color distribution becomes more compact and smooth as in Fig. 5. Noting that adding local enhancement after adaptive global color mapping achieves the best performance in quantitative comparison (in Tab. 1) without introducing artifacts. If we apply the local enhancement as the first step, it will generate apparent local artifacts, similar to other end-to-end mapping methods. Please refer to the supplementary material.

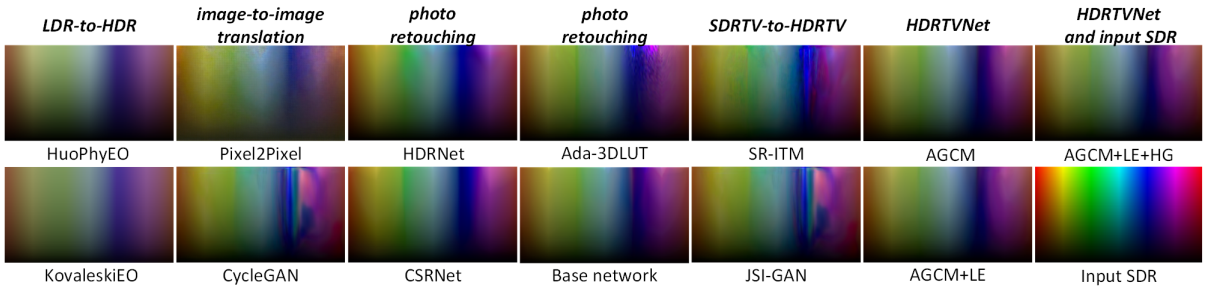

Figure 4. Color test visualization. The top row describes the categories of algorithms.

Figure 5. Visualization of LUTs. The LUT cube shows the position of SDR colors in the HDRTV domain at the coordinates determined by the pixel values of their corresponding HDRTV pixels.

Highlight generation. Highlight generation aims to hallucinate details in the over-exposed regions. We declare that this part is not designed for numerical improvement, but functional increase. Due to the inevitable loss of information in the production (e.g., dynamic range compression), we believe that it is necessary to use some generative methods to deal with these parts. Although our HG method has certain limitations for numerical evaluation, we can still observe that it makes colors in the LUT cube denser and makes the highlight regions look more natural.

5.6. User Study

We conduct a user study with 20 participants for subjective evaluation. Four methods with the best performance in each category are selected to compare with our method and ground truth. Participants are asked to rank them according to the visual quality. 25 images are randomly selected from the testing set and shown to participants on HDR-TV (Sony X9500G with a peak brightness of 1300 nits) in darkroom. More details of how we conduct experiment can be founded in the supplementary material. As suggested in Fig. 6, our method achieves a better visual ranking than other competitors except for the ground truth.

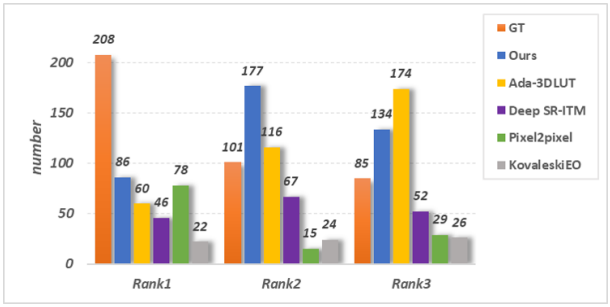

Figure 6. Ranking results of user study. Rank1 means the best subjective feeling.

6. Conclusion

We have provided a novel SDRTV-to-HDRTV solution pipeline based on the HDRTV/SDRTV formation pipeline using divide-and-conquer. Moreover, we have introduced a new method, HDRTVNet, for the problem. According to the three types of operations in HDRTV/SDRTV formation pipeline including global operation, local operation and shoulder operation, the whole method is divided into adaptive global color mapping, local enhancement and highlight generation correspondingly. Besides, a novel color condition network has been proposed with fewer parameters and better performance than other photo retouching approaches. Comprehensive experiments show the superiority of our solution in quantitative comparison and visual quality.

Acknowledgement. This work is partially supported by National Natural Science Foundation of China (61906184), the Science and Technology Service Network Initiative of Chinese Academy of Sciences (KFJ-STS-QYZX-092), the Shanghai Committee of Science and Technology, China (Grant No. 21DZ1100100).

References

[1] Shahrukh Athar, Thilan Costa, Kai Zeng, and Zhou Wang. Perceptual quality assessment of uhd-hdr-wcg videos. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1740–1744. IEEE, 2019. 6

[2] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006. 3

[3] Frédéric Drago, Karol Myszkowski, Thomas Annen, and Norishige Chiba. Adaptive logarithmic mapping for displaying high contrast scenes. In *Computer graphics forum*, volume 22, pages 419–426. Wiley Online Library, 2003. 2

[4] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 4

[5] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 1

[6] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2, 7

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 5

[8] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. *arXiv preprint arXiv:2009.10390*, 2020. 2, 4, 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 5, 7

[10] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5):507–517, 2014. 1, 6, 7

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 3, 7

[12] ITU-R. Reference electro-optical transfer function for flat panel displays used in hdtv studio production. Technical report, ITU-R Rec, BT.1886, 2011. 2

[13] ITU-R. High dynamic range electro-optical transfer function of mastering reference displays. Technical report, SMPTE, SMPTE ST2084:2014, 2014. 3

[14] ITU-R. Parameter values for the hdtv standards for production and international programme exchange. Technical report, ITU-R Rec, BT.709-6, 2015. 2

[15] ITU-R. Parameter values for ultra-high definition television systems for production and international programme exchange. Technical report, ITU-R Rec, BT.2020-2, 2015. 2

[16] ITU-R. Image parameter values for high dynamic range television for use in production and international programme exchange. Technical report, ITU-R Rec, BT.2100-2, 2018. 2, 3

[17] ITU-R. Objective metric for the assessment of the potential visibility of colour differences in television. Technical report, ITU-R Rec, BT.2124-0, 2019. 2, 6

[18] ITU-R. High dynamic range television for production and international programme exchange. Technical report, ITU-R Rec, BT.2390-8, 2020. 2

[19] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3116–3125, 2019. 1, 3, 4, 5, 6, 7

[20] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In *AAAI*, pages 11287–11295, 2020. 1, 3, 4, 6, 7

[21] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE, 2014. 1, 7

[22] Gregory Ward Larson, Holly Rushmeier, and Christine Piatko. A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics*, 3(4):291–306, 1997. 2

[23] Dani Lischinski, Zeev Farbman, Matt Uyttendaele, and Richard Szeliski. Interactive local adjustment of tonal values. *ACM Transactions on Graphics (TOG)*, 25(3):646–653, 2006. 2

[24] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1660, 2020. 1, 5

[25] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)*, 30(4):1–14, 2011. 2, 6

[26] Belen Masia, Sandra Agustin, Roland W Fleming, Olga Sorkine, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–8. 2009. 1

[27] Belen Masia and Diego Gutiérrez. Dynamic range expansion based on image statistics. *Multimedia Tools and Applications*, 76, 01 2017. 3

[28] Erik Reinhard. Parameter estimation for photographic tone reproduction. *Journal of graphics tools*, 7(1):45–51, 2002. 2

[29] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of

legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3):39–es, 2007. 3

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[33] Jack Tumblin and Holly Rushmeier. Tone reproduction for realistic images. *IEEE Computer graphics and Applications*, 13(6):42–48, 1993. 2

[34] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 4

[35] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 7

[36] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *2012 19th IEEE international conference on image processing*, pages 1473–1476. IEEE, 2012. 2, 6

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 3, 7