Towards Efficient SDRTV-to-HDRTV by Learning From Image Formation

Xiangyu Chen, *Member, IEEE*, Zheyuan Li, Zhengwen Zhang, Jimmy S. Ren, Yihao Liu, Jingwen He, Yu Qiao, *Senior Member, IEEE*, Jiantao Zhou, *Senior Member, IEEE*, and Chao Dong

***Abstract*—Contemporary display enables video content rendering with high dynamic range (HDR) and wide color gamut (WCG). However, the majority of existing content remains in standard dynamic range (SDR) format. Therefore, the conversion of SDR content to HDRTV standards holds significant value. This paper delineates and analyzes the SDRTV-to-HDRTV conversion by modeling the formation of SDRTV/HDRTV content. The findings reveal that a naive end-to-end supervised training pipeline suffers from severe gamut transition errors. To address this, we propose a new three-step solution called HDRTVNet++, which includes adaptive global color mapping, local enhancement, and highlight refinement. The adaptive global color mapping step utilizes global statistics for image-adaptive color adjustments, followed by a local enhancement network for detail improvement. These two components are integrated as a generator, with GAN-based joint training ensuring highlight consistency. Our method, tailored for ultra-high-definition TV content, offers both effectiveness and computational efficiency in processing 4 K resolution images. We also construct HDRTV1K, a dataset comprising HDR videos adhering to the HDR10 standard, featuring 1235 training and 117 testing images at 4 K resolution. Furthermore, we employ five metrics to assess SDRTV-to-HDRTV performance. Our results demonstrate state-of-the-art performance both quantitatively and visually.**

***Index Terms*—Ultra high definition, high dynamic range, gamut mapping, image enhancement.**

Received 11 October 2024; revised 19 January 2025; accepted 22 February 2025. Date of publication 1 September 2025; date of current version 12 November 2025. This work was supported in part by Macau Science and Technology Development Fund under Grant SKLIOTSC-2021-2023, Grant 0022/2022/A1, and Grant 0014/2022/AFJ, in part by the Research Committee at University of Macau under Grant MYRG-GRG2023-00058-FST-UMDF and Grant MYRG2022-00152-FST, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515012536, and in part by the National Natural Science Foundation of China under Grant 62276251. The associate editor coordinating the review of this article and approving it for publication was Prof. Susanto Rahardja. *(Xiangyu Chen and Zheyuan Li are co-first authors.) (Corresponding authors: Jiantao Zhou; Chao Dong.)*

Xiangyu Chen is with the State Key Laboratory of Internet of Things for Smart City, the Department of Computer and Information Science, University of Macau, Macau 999078, China, and also with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 100045, China.

Zheyuan Li and Jiantao Zhou are with the State Key Laboratory of Internet of Things for Smart City, the Department of Computer and Information Science, University of Macau, Macau 999078, China (e-mail: jtzhou@um.edu.mo).

Zhengwen Zhang, Yihao Liu, Jingwen He, Yu Qiao, and Chao Dong are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 100045, China (e-mail: chao.dong@siat.ac.cn).

Jimmy S. Ren is with Hong Kong Metropolitan University, Hong Kong, China.

The codes and models are available at https://github.com/xiaom233/HDRTVNet-plus.

Digital Object Identifier 10.1109/TMM.2025.3604961

I. Introduction

THE evolution of television and film content resolution has progressed from standard definition to full high definition, and subsequently to ultra-high definition (UHD). A key feature of UHDTV is high dynamic range (HDR), which offers a wider color gamut and higher dynamic range than standard dynamic range (SDR) content, allowing viewers to experience images and videos closer to real life. Although HDR display devices are now common, most available resources remain in the SDR format, necessitating algorithms to convert SDR content to HDR. This task, referred to SDRTV-to-HDRTV, presents a valuable yet under-researched area of study. The primary reasons are twofold: first, HDRTV standards (e.g., HDR10 and HLG) have only recently become well-defined; second, there is a lack of comprehensive datasets for algorithm development and evaluation. To advance this emerging field, this paper conducts an in-depth study and analysis of the SDRTV-to-HDRTV problem. This task is complex due to the inherent disparities between the two content types in dynamic range, color gamut, and bit depth. While both SDRTV and HDRTV are derived from the same raw files, they adhere to different processing standards. It is crucial to distinguish the SDRTV-to-HDRTV task from the low dynamic range (LDR)-to-HDR task, which involves predicting scene luminance in the linear domain. In Section II, we provide detailed explanations of SDRTV/HDRTV concepts and the SDRTV-to-HDRTV task. From an imaging formation perspective, SDRTV-to-HDRTV can be viewed as an image-to-image translation task. Due to substantial differences in color gamut, photo retouching methods can be applied to manage color transformations in this task. Although not exclusively focused on SDRTV-to-HDRTV, early works like Deep SR-ITM [5] and JSI-GAN [6] address this task by combining super-resolution with SDRTV-to-HDRTV. We illustrate SDRTV-to-HDRTV and compare various methods in Fig. 1. As shown, previous methods struggle to effectively address this task. This study aims to address the SDRTV-to-HDRTV conversion through a thorough examination of its underlying challenges. We introduce a simplified formation pipeline for SDRTV/HDRTV content, comprising tone mapping, gamut mapping, transfer function, and quantization. Based on this, we propose HDRTVNet++, which incorporates adaptive global color mapping (AGCM), local enhancement (LE), and highlight refinement (HR). Specifically, AGCM incorporates a new module to capture global image characteristics and accommodate diverse inputs. By utilizing

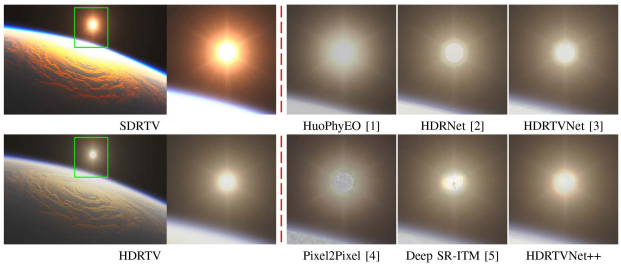


Fig. 1. Visual comparison of different methods to solve SDRTV-to-HDRTV.

only 1×1 convolutions, this approach achieves enhanced performance while minimizing parameter count, surpassing alternative photo retouching methods such as CSRNet [7], HDRNet [2] and Ada-3DLUT [8]. Building upon AGCM's foundation, we propose a U-shape architecture incorporating spatial conditioning for local enhancement. This solution pipeline enables further performance improvements while mitigating color transition artifacts commonly associated with end-to-end networks. After sequential training of AGCM and LE, we present that joint finetuning further improves results. Despite these advancements, highlight areas remain challenging due to severe information loss. To address this, we implement generative adversarial training tailored for highlight refinement. To advance research in this area, we construct a new dataset consisting of 4 K resolution SDR/HDRTV image pairs, called HDRTV1K, and select five evaluation metrics: PSNR, SSIM, SR-SIM [9], ΔE_{ITP} [10] and HDR-VDP3 [11].

In summary, our contributions are four-fold:

- We conduct a detailed analysis of the SDRTV-to-HDRTV task by modeling SDRTV/HDRTV content formation.
- We propose an efficient SDRTV-to-HDRTV method achieving state-of-the-art performance.
- We present a global color mapping network with outstanding accuracy and only 35 K parameters.
- We provide an HDRTV dataset and select five metrics for evaluating SDRTV-to-HDRTV algorithms.

A preliminary version of this work was presented at ICCV 2021 [3]. Since then, subsequent studies have further explored the SDRTV-to-HDRTV problem [12], [13], [14], [15], [16], [17]. This paper introduces HDRTVNet++, an enhanced version that significantly improves upon the initial method through a refined pipeline and more effective network design. Key advancements include: 1) **Enhanced Network Design:** We propose HDRTVNet++ with an improved pipeline and network architecture. Notably, we optimize the training of AGCM and further jointly finetune AGCM and LE, achieving better restoration accuracy. The heavy sub-network for highlight generation is also replaced with a joint adversarial training strategy, leading to a large performance gain of 0.99 dB on PSNR with reduced parameters. 2) **Detailed Analysis:** We provide detailed explanations of the motivation and rationale behind our solution pipeline. By formulating pixel-independent and region-dependent operations, we highlight the importance of the correct sequence of global color mapping and local enhancement. Additional experiments further illustrate this crucial point for SDRTV-to-HDRTV conversion. 3) **Extensive Experiments:** We conduct comprehensive ablation studies and detailed investigations of the network design. These experiments demonstrate the effectiveness of our proposed modules and the overall method.

II. Preliminary

In this section, we clarify the concepts of SDRTV/HDRTV and the distinctions between SDRTV-to-HDRTV and LDR-to-HDR, as these terms are often confused and underexplored in existing literature.

Concept: We use SDRTV/HDRTV to denote content (images and videos) adhering to the respective standards. SDRTV is defined by standards such as Rec.709 [18] and BT.1886 [19], while HDRTV is specified in Rec.2020 [20] and BT.2100 [21]. HDRTV is characterized by features such as a wide color gamut [20], PQ or HLG OETF [20], and a bit depth of 10-16 bits. Content not conforming to HDRTV is generally considered SDR, with clearer requirements outlined under the SDRTV standard, such as the Rec.709 color gamut and gamma-OETF. Both SDRTV and HDRTV can encode the same content, but they differ in information capacity, resulting in distinct visual experiences. The terms LDR and SDR both refer to low dynamic range content, but their usage varies. SDR typically pertains to display standards used in content production, often derived from high dynamic range RAW files. In contrast, LDR content refers to images captured at specific exposure levels, with dynamic range determined during imaging. Thus, their formation processes differ. For clarity, we use the terms **LDR-to-HDR** and **SDRTV-to-HDRTV** to denote the conventional image HDR reconstruction and the conversion of content from SDRTV to HDRTV standard.

Explanation: SDRTV-to-HDRTV and LDR-to-HDR differ in function. While both involve HDR, the meaning of HDR varies. LDR-to-HDR methods predict luminance in the linear domain, representing the physical brightness of a scene. Here, HDR refers to dynamic range information beyond the capture capabilities of LDR imaging. Conversely, SDRTV-to-HDRTV involves generating HDR images in the pixel domain using HDR display formats such as HDR10, HLG, and Dolby Vision. Both SDRTV and HDRTV content originate from the same HDR scene radiance. While HDRTV content can derive from linear domain HDR content, this requires additional operations such as tone mapping and gamut mapping. As a result, methods for these tasks are not interchangeable.

III. Related Work

A. Sdrtv-to-Hdrtv

The SDRTV-to-HDRTV task, central to this paper, is initially presented in [22], where SDRTV/HDRTV content is referred to as LDR/HDR. Notable advancements, such as Deep SR-ITM [5] and JSI-GAN [6], have successfully combined image super-resolution with SDRTV-to-HDRTV, significantly drawing research interest. Our conference version, HDRTVNet [3], provides an in-depth analysis, a foundational solution, and a dataset for this task. Building on that, several works have emerged.

For instance, He et al. [13] introduce HDCFM, employing hierarchical global and local feature modulation. Xu et al. [12] propose FMNet, utilizing frequency information for modulation to minimize structural distortions and artifacts. Cheng et al. [14] develop a learning-based approach for synthesizing realistic SDRTV-HDRTV pairs, enhancing method generalization. Guo et al. [16] contribute a new dataset and degradation models for practical conversion. Zhang et al. [17] design a efficient framework with reconstruction and enhancement models for this task. In this work, we enhance HDRTVNet [3] by introducing more effective networks, along with more comprehensive analysis and experiments.

B. Ldr-to-Hdr

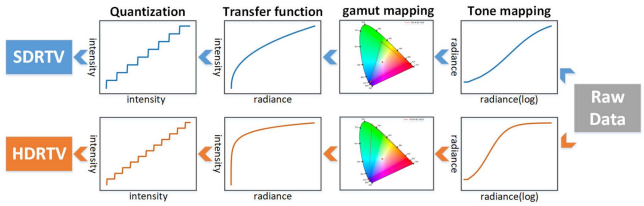
In general, LDR-to-HDR, also known as HDR reconstruction, seeks to generate HDR images from LDR photographs. Common methods include fusing multi-exposure LDR images [23], [24], [25], [26] and reconstructing HDR images from a single image [27], [28], [29], [30], [31]. The latter is more relevant to this work. Traditional methods for single image LDR-to-HDR leverage intrinsic image features to estimate scene luminance. For example, [27] estimates light source density to extend the dynamic range, and [1] employs a cross-bilateral filter to enhance LDR inputs. Additionally, deep learning-based methods have also emerged. Ref. [28] introduces HDRCNN to recover missing details in over-exposed areas, and [29] learns the LDR-to-HDR mapping by reversing the camera pipeline. Nonetheless, these techniques primarily focus on predicting linear HDR luminance and are not designed for content conversion under two display standards. Consequently, they are either hard to use for SDRTV-to-HDRTV or perform poorly when applied.

C. Gamut Extension

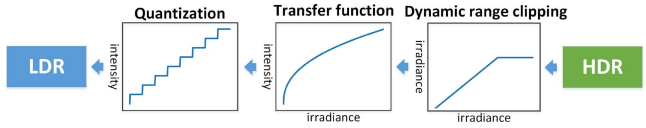
Gamut extension, a key concept in color science, involves converting content to a wider color gamut, essential in transitioning from Rec.709 to Rec.2020 during SDRTV-to-HDRTV conversion. Despite ITU-R [32] offering a color conversion matrix, it cannot consider multiple mappings involving color (i.e., tone mapping) used in production, limiting its effectiveness. Several gamut extension algorithms have been proposed [33], [34], [35], [36], [37], [38]. For example, [33] introduces a perceptually-based variational framework for spatial gamut mapping, and [38] presented a PDE-based optimization procedure considering hue, chroma, and saturation. However, these algorithms primarily address color mapping, lacking the capability for complex detail enhancement needed in SDRTV-to-HDRTV.

IV. Analysis and Method

In this section, we present a streamlined process for SDRTV/HDRTV formation, highlighting critical steps in actual production. We then analyze this pipeline and propose a new solution using a divide-and-conquer approach.



(a) SDRTV-to-HDRTV formation pipeline.



(b) LDR image formation pipeline in SingleHDR [29].

Fig. 2. Analysis of SDRTV-to-HDRTV and LDR-to-HDR formations.

A. SDRTV/HDRTV Formation Pipeline

We present a simplified pipeline for forming SDRTV and HDRTV content, grounded in camera ISP and HDRTV production [39], depicted in Fig. 2(a). While operations like denoising, white balance, and color grading are not covered, we focus on the key differences: tone mapping, gamut mapping, opto-electronic transfer function, and quantization. In the equations below, "S" represents SDRTV and "H" is HDRTV.

Tone mapping: This process converts high dynamic range signals to low dynamic range for display compatibility. It includes global tone mapping [40], [41], [42] and local tone mapping [43], [44]. Global tone mapping applies a uniform function to all pixels, based on global image statistics like average luminance, whereas local tone mapping adapts to content but is computationally intensive. Thus, global tone mapping is preferred in SDRTV/HDRTV. This can be expressed as:

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

where T_S and T_H are the tone mapping functions, and θ_S and θ_H are coefficients related to image statistics. S-shaped curves are frequently used for global tone mapping, with clipping operations commonly occurring during production. We take Hable tone mapping [45] as an example and show its curves when processing SDRTV (0–100 cd/m^2) and HDRTV (0–10000 cd/m^2) in Fig. 2(a).

Gamut mapping: This process adjusts colors from the source to the target gamut while maintaining the scene's appearance. According to ITU-R standards [18], [20], transformations from XYZ space to SDRTV (Rec.709) and HDRTV (Rec.2020) are:

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

where M_S and M_H are constant 3×3 matrices. Fig. 2(a) illustrates the target gamuts of SDRTV and HDRTV on the CIE chromaticity diagram.

Opto-electronic transfer function (OETF): This function converts linear optical signals into non-linear electronic signals. The OETF for SDRTV often approximates an exponential gamma

function as $I_{fS} = f_S(I) = I^{1/2.2}$. For HDRTV, there are several OETFs on different standards, such as PQ-OETF [46] on the HDR10 standard and HLG-OETF [21] on the HLG (Hybrid Log-Gamma) standard. The PQ-OETF is:

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

where a_1, a_2, a_3, b_1, b_2 are constants. The curves of gamma-OETF for SDRTV (0–100 cd/m^2) and PQ-OETF for HDRTV (0–10000 cd/m^2) are depicted in Fig. 2(a).

Quantization: This involves encoding pixel values using:

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

where n is 8 for SDRTV and 10-16 for HDRTV. Fig. 2(a) also presents the two quantization curves.

In summary, the SDRTV and HDRTV content formation pipelines are expressed as:

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

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

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

Comparison with LDR formation pipeline: Taking Single-HDR [29] as an example, the LDR image is derived from its HDR counterpart through a series of process including dynamic range clipping, non-linear mapping, and quantization. Unlike LDR-to-HDR, SDRTV and HDRTV originate from the same raw data with different operations, as shown in (5) and (6). Gamut extension is critical in SDRTV-to-HDRTV, illustrating key production differences.

B. SDRTV-to-HDRTV Solution Pipeline

According to the above pipeline, the SDRTV-to-HDRTV process 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 $T_S^{-1}, M_S^{-1}, f_S^{-1}, Q_S^{-1}$ are the inversions of corresponding operations. We propose a new solution pipeline as shown in Fig. 3(a)), based on two observations: the one is that many critical operations, such as global tone mapping, OETF, and gamut mapping, are pixel-independent; the other one is that some operations, like local tone mapping and dequantization, depend on regional information. To understand these operations, we define the mapping $f(\cdot)$ from the input I_{in} to the output I_{out} at (x, y) as:

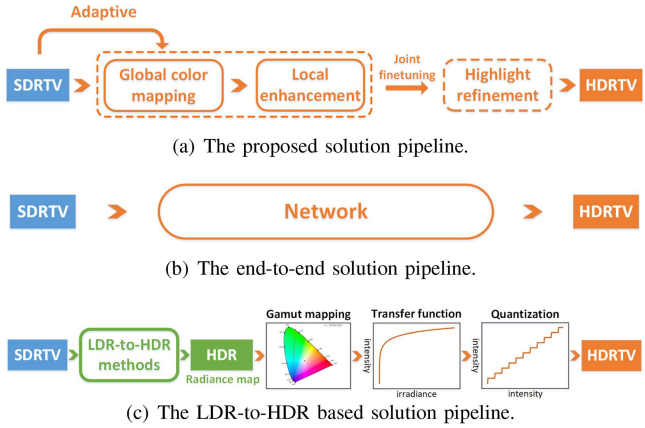
$$I_{out}(x, y) = f(\Omega(I_{in}(x, y), \delta)), \tag{8}$$

where $\Omega(I_{in}(x, y), \delta)$ denotes a local region. It composed of pixels whose distance from the center point $I_{in}(x, y)$ is no more than δ. Specially, $\delta = 1$ means that $\Omega(I_{in}(x, y), \delta)$ is equivalent to $I_{in}(x, y)$. For pixel-independent operations:

$$I_{out}(x, y) = f(\Omega(I_{in}(x, y), 1)) = f(I_{in}(x, y)), \tag{9}$$

and for region-dependent operations:

$$I_{out}(x, y) = f(\Omega(I_{in}(x, y), \delta)), \text{where } \delta > 1. \tag{10}$$



(a) The proposed solution pipeline.

(b) The end-to-end solution pipeline.

(c) The LDR-to-HDR based solution pipeline.

Fig. 3. SDRTV-to-HDRTV solution pipelines.

Color conversion is crucial in SDRTV-to-HDRTV production [32], so we implement pixel-independent and region-dependent operations separately, i.e., global color mapping and local enhancement in Fig. 3(a). Global color mapping should be image-adaptive (e.g., the average brightness and peak brightness are often used in tone mapping):

$$I_{out}(x, y) = f(I_{in}(x, y)|I_{in}). \tag{11}$$

Experiments in Section V-B illustrate the effectiveness and efficiency of this design. Performing pixel-independent operations first reduces color transition artifacts compared to end-to-end solutions, as shown in Fig. 8. This is likely due to the difficulty of convolution operators processing both pixel-independent low-frequency and region-dependent high-frequency transformations. For better performance, we optimize these operations jointly. Due to severe information loss in SDRTV, particularly in highlights, previous methods [3] struggle to recover missing information. However, generative adversarial training enhances visual quality by improving color transitions in highlights, aligning predictions closer to HDRTV distribution. We compare different SDRTV-to-HDRTV pipelines in Fig. 3. Existing methods [5], [6], [12] employ end-to-end networks in Fig. 3(b). Since LDR-to-HDR has been extensively discussed, we also demonstrate a method pipeline based on LDR-to-HDR principles in Fig. 3(c). Specifically, the HDR radiance map is first generated. Then, gamut mapping is applied to convert the radiance map to the Rec.2020 gamut. The PQ OETF is subsequently used to compress the dynamic range. Finally, quantization is performed to produce the HDRTV output. The details of these operations may vary in actual processing, and thus we follow the pipeline used in [5], [6]. Our solution uses a divide-and-conquer strategy, developing HDRTVNet++ with adaptive global color mapping, local enhancement, and highlight refinement.

C. Adaptive Global Color Mapping

Adaptive global color mapping (AGCM) is designed to facilitate image-specific color conversion from SDRTV domains to HDRTV domains. We employs an network analogous to the conference version. As depicted in Fig. 4, the proposed model comprises a base network and a condition network.

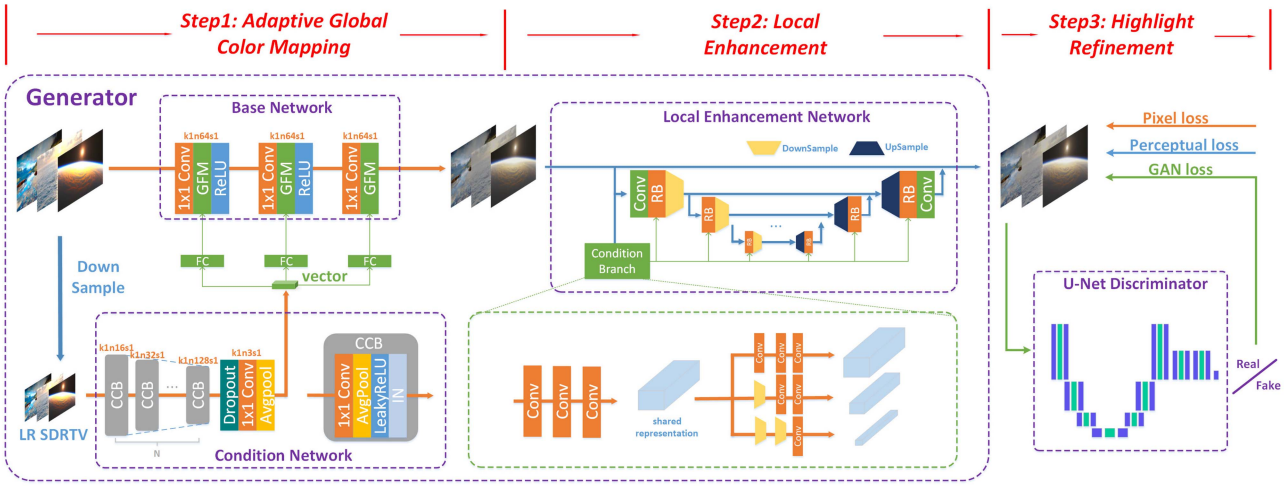


Fig. 4. The overall architecture of our proposed SDRTV-to-HDRTV method.

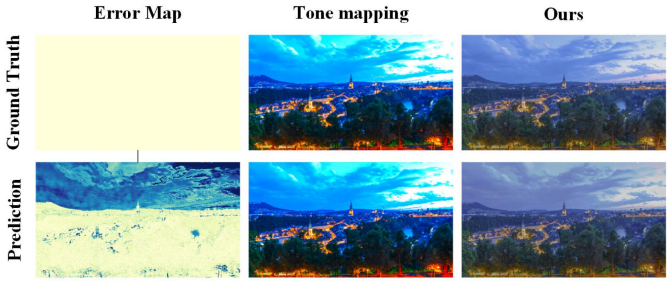


Fig. 5. Comparison of different visualization methods.

1) Base Network: The base network handles pixel-independent operations. Given an input SDRTV image I_S, the transformation is denoted as:

$$I_B(x, y) = f(I_S(x, y)), \forall (x, y) \in I_S, \tag{12}$$

where I_B represents the base network's output. As presented in CSRNet [47], a fully convolutional network utilizing exclusively 1×1 convolutions and activations can realize pixel-independent mapping. Therefore, we design the base network using N_l convolutional layers with 1×1 filters and N_l-1 ReLU activation functions, expressed as:

$$I_B = Conv_{1\times1} \circ (ReLU \circ Conv_{1\times1})^{N_l-1}(I_S). \tag{13}$$

This network processes an SDRTV image encoded with 8 bits and outputs a 10-16 bits HDRTV image. Despite learning a one-to-one color mapping, it performs well (see Table I). It is worth noting that this network can function similarly to a 3D lookup table, but more efficiently (see Section V-H).

2) Condition Network: Global priors play a crucial role in adaptive color mapping. To achieve image-specific transformation, we incorporate a condition network that modulates the base network. While previous studies [7], [48] have focused on extracting spatial and local information as conditioning factors, the SDRTV-to-HDRTV conversion primarily depends on global image statistics or color distribution, which are typically independent of spatial details. Our condition network extracts color-related information for adjustable mapping, as shown in Fig. 4. It comprises color condition blocks, convolution layers, a dropout layer, and global average pooling.

A color condition block (CCB) consists of a 1×1 convolution, average pooling, LeakyReLU activation, and instance normalization [49]:

$$CCB(I_f) = IN \circ LReLU \circ avgpool \circ Conv_{1\times1}(I_f), \tag{14}$$

where I_f is the input feature. The condition network processes a down-sampled SDRTV image $I_{S\downarrow}$ and outputs a condition vector V:

$$V = GAP \circ Conv_{1\times1} \circ Dropout \circ CCB^{N_c}(I_{S\downarrow}). \tag{15}$$

Without local feature extraction, the overall condition network focus on deriving global priors. Dropout before the convolution and pooling layers, which acts like adding a multiplicative Bernoulli noise to features, is used to prevent overfitting.

3) Global Feature Modulation: We introduce global feature modulation (GFM) to utilize global priors. GFM modulates intermediate features of the base network via scaling and shifting based on the condition vector:

$$GFM(x_i) = \alpha * x_i + \beta, \tag{16}$$

where x_i is the intermediate feature to be modulated, and α, β are scaling and shifting factors. Overall, the AGCM network is formulated as:

$$I_{AGCM} = GFM \circ Conv_{1\times1} \circ (ReLU \circ GFM \circ Conv_{1\times1})^{N_l-1}(I_S). \tag{17}$$

We optimize AGCM by minimizing the L_1 loss between the prediction and the ground-truth HDRTV image. In the conference version, the L_2 loss function is utilized to optimize the AGCM network, as it is commonly adopted in existing literature involving HDRTV conversion [5] or color mapping [8]. This paper demonstrates that the L_1 loss function yields better results for the SDRTV-to-HDRTV problem (see Section V-G).

Fig. 6. Visual comparison on HDRTV1K.

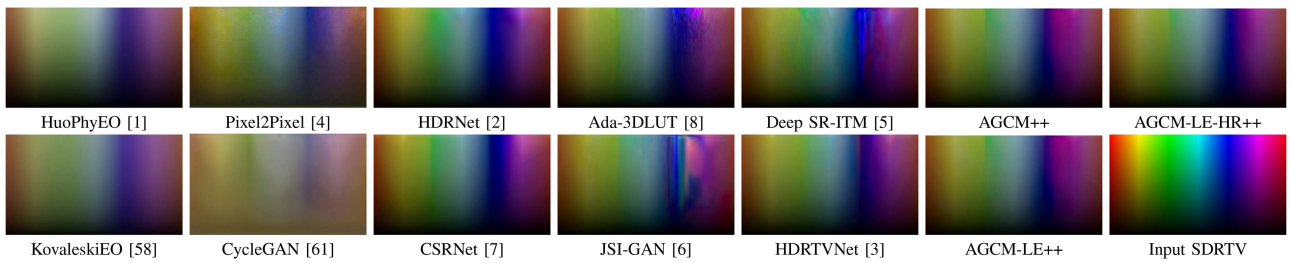


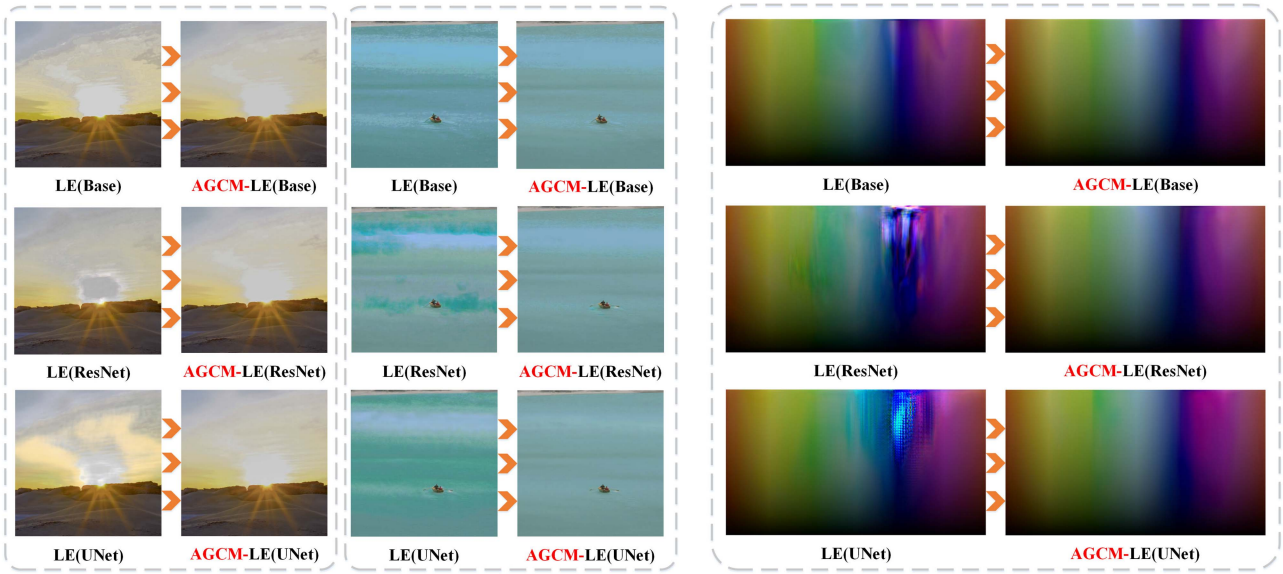
Fig. 7. Visual comparison of the color transition test.

D. Local Enhancement

Following AGCM, Local Enhancement (LE) is crucial for SDRTV-to-HDRTV conversion. While AGCM provides significant performance, LE addresses region-dependent mappings. Initially, a classic ResNet is used for LE [3], but it has limited performance and large computational load. Inspired by [50], we employ a UNet structure for LE, consisting of a main and a condition branch, as illustrated in Fig. 4. Specifically, The main branch is U-shape, and the condition branch generates vectors to modulate main branch features. The input $I_{AGCM} \in \mathbb{R}^{3\times H\times W}$ is transformed into high-dimensional features $F_0 \in \mathbb{R}^{C\times H\times W}$. A three-level encoder-decoder refines these features, using stride convolution and pixel-shuffle for downsampling and upsampling [51]. Skip connections assist feature recovery. In actual production, the resolution of SDRTV content is generally from 1 K to 4 K. The use of a U-shape structure can greatly reduce the computational burden required for processing. The condition branch processes inputs through three convolutions for shared representation, generating hierarchical conditions for spatial feature modulation using SFT layers [48]:

$$SFT(x_i) = m \odot x_i + n, \tag{18}$$

where $\odot$ denotes the element-wise multiplication. $x_i \in \mathbb{R}^{C\times H\times W}$ denotes the features to be modulated. $m \in \mathbb{R}^{C\times H\times W}$ and $n \in \mathbb{R}^{C\times H\times W}$ are both conditional maps predicted by the condition branch. It is noteworthy that without AGCM, employing local enhancement with region-dependent operations can cause artifacts (see Fig. 8). To achieve better optimization, we further jointly train the AGCM and LE networks. Experiments show that joint training can still bring a slight performance improvement. Benefiting from our AGCM and LE, the proposed method can significantly outperform existing approaches with high efficiency for SDRTV-to-HDRTV.



(a) Visual results. (b) Color transition test.

Fig. 8. Visual comparisons between methods w/ AGCM and w/o AGCM.

TABLE I
QUANTITATIVE COMPARISONS WITH EXISTING METHODS

Method		Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
LDR-to-HDR	HuoPhyEO [1]	-	25.90	0.9296	0.9881	38.06	7.893
	KovaleskiEO [58]	-	27.89	0.9273	0.9809	28.00	7.431
	ExpandNet [59]	2.7M	37.58	0.9743	0.9971	8.60	8.644
	HDRUNet [50]	1.65M	37.49	0.9742	0.9951	8.20	8.508
image-to-image translation	ResNet [60]	1.37M	37.32	0.9720	0.9950	9.02	8.391
	Pixel2Pixel [4]	11.38M	25.80	0.8777	0.9871	44.25	7.136
	CycleGAN [61]	11.38M	21.33	0.8496	0.9595	77.74	6.941
photo retouching	HDRNet [2]	482K	35.73	0.9664	0.9957	11.52	8.462
	CSRNet [7]	36K	35.04	0.9625	0.9955	14.28	8.400
	Ada-3DLUT [8]	594K	36.22	0.9658	0.9967	10.89	8.423
SDRTV-to-HDRTV	Deep SR-ITM [5]	2.87M	37.10	0.9686	0.9950	9.24	8.233
	JSI-GAN [6]	1.06M	37.01	0.9694	0.9928	9.36	8.169
	FMNet [12]	1.24M	37.94	0.9747	0.9957	8.10	8.510
	HDRTVDM [16]	325k	37.98	0.9707	0.9974	8.84	8.610
	DFT [62]	130K	38.15	0.9737	0.9967	8.25	8.586
	Chen et al. [63]	1.0M	38.48	**0.9751**	-	7.86	8.739
HDRTVNet [3]	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
HDRTVNet++ (ours)	AGCM++	35K	37.35	0.9666	0.9968	9.29	8.511
	AGCM-LE++	591K	38.45	0.9739	0.9970	7.90	8.666
	AGCM-LE++†	591K	**38.60**	0.9745	0.9973	**7.67**	8.696
	AGCM-LE-HR++	591K	38.36	0.9735	**0.9975**	8.28	**8.751**

[1] Results with the best and second-best performance are in **bold** and underline.
[2] † represents that the model is finetuned by joint training.

E. Highlight Refinement

Highlight Refinement (HR) aims to address color disharmony in highlight regions, which is often caused by dynamic range and color gamut clipping. MSE-based models struggle with this ill-posed problem, so we introduce generative adversarial training to alleviate the color disharmony. The initial version uses a separate UNet with a soft mask [3], but it has limited visual improvement and high computational costs. Instead, we leverage the pre-trained AGCM and LE networks as the generator, enhancing it with adversarial training to achieve better visual results, as shown in Fig. 4. This approach brings two advantages. First, it aligns output distribution closer to the ground-truth. Since well-exposed regions have already been effectively processed in previous stages, this training will focus on improving color transition in highlight regions. Second, it avoids extra

computational cost since there are no additional parameters to optimize. The generator is defined as:

$$I_H = Generator(I_S) = LE(AGCM(I_S)), \quad (19)$$

where $LE(\cdot)$ and $AGCM(\cdot)$ represent the pretrained LE and AGCM networks. We adopt the UNet-style network as the discriminator following [52] and optimize the GAN training based on Relativistic GAN [53]. The overall loss function is composed of L_1 loss, perceptual loss [54] and adversarial loss [55], [56], [57] as:

$$L_{HR} = \lambda_1 L_1 + \lambda_2 L_{Percep} + \lambda_3 L_{GAN}, \quad (20)$$

where $\lambda_1, \lambda_2, \lambda_3$ are constant, set to 0.01, 1 and 0.005.

V. Experiments

A. Experimental Setup

1) Dataset: To address the scarcity of SDRTV/HDRTV data pairs for training and testing, we construct the HDRTV1K dataset. This dataset comprises 22 HDR videos (compliant with the HDR10 standard) and their SDRTV counterparts, sourced from YouTube following the method [5]. All HDR videos are encoded using PQ-OETF within the Rec.2020 color gamut. We utilize 18 video pairs to generate image data pairs for training, reserving the remaining 4 for testing. To reduce content similarity, we sample one frame every two seconds, resulting in a training set of 1,235 images. The testing set consists of 117 unique images extracted from the videos.

2) Training Details: For AGCM, the base network comprises 3 convolutional layers with a 1×1 kernel size and 64 channels, while the condition network consists of 4 CCBs. Prior to training, images are cropped to 480×480 with a step of 240. During training, patches of size 480×480 are input into the base network, while full images downscaled by a factor of 4 are input into the condition network. The mini-batch size is set to 4 and we employ the L_1 loss function and Adam optimizer for 1×10^6 iterations. The initial learning rate is 4×10^{-4}, which decays by a factor of 2 at 5×10^5 and 8×10^5 iterations. For LE, the AGCM outputs serve as inputs. We set the mini-batch size as 8 and the patch size as 240×240. The initial learning rate is set to 1×10^{-4}, decaying by a factor of 2 every 2×10^5 iterations, over 1×10^6 iterations. The L_1 loss function and Adam optimizer are used for training. In joint training, the AGCM and LE networks are optimized simultaneously with the L_1 loss function and Adam optimizer, using a batch size of 4 and a patch size of 192. The initial learning rate is 1×10^{-4}, decaying by 2 every 1×10^5 iterations, over 5×10^5 iterations. Subsequently, the GAN model is trained with a batch size of 64 and a patch size of 128. The initial learning rate is 1×10^{-4}, with a total of 4×10^5 iterations. The learning rate decays by a factor of 0.5 at 5×10^4, 1×10^5, 2×10^5, and 3×10^5 iterations. We train all models using PyTorch and on NVIDIA 3090 GPUs.

3) Evaluation: Five metrics are utilized for comprehensive evaluation: PSNR, SSIM, SR-SIM [9], HDR-VDP3 [11] and ΔE_{ITP} [10]. PSNR assesses SDRTV-to-HDRTV fidelity against ground truth HDRTV images. SSIM and SR-SIM are adopted to evaluate structural similarity of images; SR-SIM, although designed for SDR images, is effective for HDR standards as shown in [63]. ΔE_{ITP} measures color differences, tailored for HDRTV content. HDR-VDP3 is an image quality assessment metric supports the Rec.2020 color gamut. For HDR-VDP3, evaluations are implemented by setting the side-by-side" task, rgb-bt.2020" color encoding, 50pixels per degree, and led-lcd-wcg" for the rgb-display" option.

For visualization, we display 16-bit PNG HDRTV images directly without extra processing. They may appear lower in contrast due to gamma EOTF decoding on SDR devices, but the visual differences are still discernible. Previous work [5], [6] visualize HDRTV images using video players (i.e., MPC-HC player). However, this comparison may be unfair, because the software introduces unknown enhancement, particularly in highlight regions, leading to similar visual outcomes. Another approach is to use error maps to show the intensity difference between the produced result and the ground truth, while it may fail to reflect visual differences accurately. In contrast, our visualization preserves highlight details and aligns closely with human perception, as demonstrated in Fig. 5.

B. Comparison With Existing Methods

We compare our results with four types of methods: SDRTV-to-HDRTV, image-to-image translation, photo retouching, and LDR-to-HDR. Since these methods are not all specifically designed for this task, necessary adjustments are made. For joint SR with SDRTV-to-HDRTV methods, we modify the stride of the first convolutional layer to 2 for downsampling to match input and output sizes.[1] For LDR-to-HDR methods, we process results as illustrated in Fig. 3(c), following the same steps as previous works [5], [6]. Note that All data-driven methods are retrained on our dataset.

Quantitative comparison: As shown in Table I, our method significantly outperforms other methods on all metrics. Notably, our initial AGCM version achieves comparable performance to Ada-3DLUT with only 1/17 of its parameters. By further optimizing the training of AGCM, the improved version, AGCM++, surpasses all compared methods including recent works FMNet [12] and HDRTVDM [16]. When equipped with the LE network and joint training, our approach achieves 38.60 dB, surpassing all other approaches, including a 0.66 dB gain over FMNet, a 0.62 dB gain over HDRTVDM and about 1 dB over the initial version HDRTVNet. For the HR part, although generative adversarial training reduces PSNR performance, it achieves the best HDR-VDP3 perceptual quality scores. All the quantitative results show the superiority of our method, and it is noteworthy that HDRTVNet++ is efficient and has much fewer parameters than other methods.

Visual comparison: Fig. 6 presents the results of visual comparison. LDR-to-HDR and image-to-image translation methods often produce low-contrast images. Except for HuoPhyEO [1],

[1]We have also conducted experiments with removing the upsampling operation in these networks, but this do not improve performance and significantly increased memory and runtime costs.

TABLE II
QUANTITATIVE COMPARISONS BETWEEN METHODS W/ AND W/O AGCM

Method	Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
LE (Basic3x3)	40K	36.98	0.9706	**0.9989**	9.63	8.368
AGCM-LE (Basic3x3)	75K	**37.50**	**0.9721**	0.9988	**9.13**	**8.580**
LE (ResNet)	1.37M	37.32	0.9720	0.9950	9.02	8.391
AGCM-LE (ResNet)	1.41M	**37.61**	**0.9726**	**0.9967**	**8.89**	**8.613**
LE (UNet)	556K	37.08	0.9705	0.9956	9.27	8.315
AGCM-LE (UNet)	591K	**38.60**	**0.9745**	**0.9973**	**7.67**	**8.696**

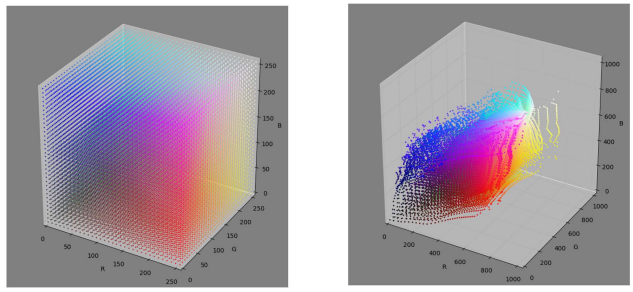
LDR-to-HDR-based, image-to-image translation, and SDRTV-to-HDRTV approaches all generate unnatural colors and noticeable artifacts. Photo retouching methods perform relatively better but suffer from color distortion. In contrast, our method produces natural colors and high contrast akin to the ground truth, without additional artifacts. Notably, the visual quality improves with processing steps: AGCM < AGCM-LE < AGCM-LE-HR, demonstrating the effectiveness of our proposed solution pipeline.

C. Color Transition Test

Previous methods often perform poorly in highlight regions, especially with color changes. We conduct a color transition test using a man-made color card as input, comparing outputs from different methods, as presented in Fig. 7. It can be observed that unnatural transitions and color blending appear in outputs from region-dependent methods (e.g., Deep SR-ITM, JSI-GAN, Pixel2Pixel, CycleGAN) and those based on region-dependent conditions (e.g., 3D-LUT). In contrast, our method achieves smooth transition with AGCM, based on pixel-independent operations. Even when learning region-dependent mapping (e.g., AGCM-LE, AGCM-LE-HR), Our method avoids color transition artifacts. This showcases the superiority of our AGCM to deal with the color gamut conversion, and the effectiveness of our entire solution pipeline to resolve the complex SDRTV-to-HDRTV mappings. Note that blue regions suffer the most severe unnatural color transition. This is due to the greater information loss appear in blue regions during color gamut compression in SDRTV production.

D. Significance of AGCM

To demonstrate the necessity of AGCM in the entire SDRTV-to-HDRTV solution pipeline, we conduct comprehensive experiments comparing methods with and without AGCM. In addition to using ResNet (used in the conference version) and UNet-based (in this paper) LE networks, we also implement a very small-scale LE network, denoted as Basic3x3. It has a very simple structure with only three layers of convolution with standard 3×3 filters. As presented in Table II, methods that perform AGCM before LE (i.e. AGCM-LE), with limited additional parameters, achieve significantly higher performance than methods that learn the LE network directly, regardless of the LE network's scale. For visual comparisons, Fig. 8(a) shows that outputs from methods without AGCM exhibit noticeable artifacts in over-exposed and saturated regions. Fig. 8(b) also demonstrates that methods without AGCM perform poorly in the color transition test. Additionally, we can observe an interesting phenomenon that when comparing the different LE networks (without AGCM), larger networks tend to produce more artifacts. The smallest LE network Basic3x3 produce the best visual results despite the lowest quantitative performance. All these results indicate that optimizing pixel-independent and region-dependent mapping together is challenging for an end-to-end LE network, and simply improve the LE network cannot address this challenge. Nevertheless, we find that performing AGCM prior to local enhancement is crucial for the final performance. This suggests that addressing color mapping before enhancement can effectively mitigates the optimization challenge and achieve considerable SDRTV-to-HDRTV results.



(a) The identity mapping of SDRTV colors.

(b) The SDRTV-to-HDRTV color mapping of the base network.

Fig. 9. Illustrations of 3D LUTs.

E. Analysis of Color Mapping Via LUT Manifold

In this section, we provide an analysis tool by visualizing the Look-Up Tables (LUTs) manifold to intuitively evaluate the model's function at various stages. We start by depicting the LUT manifold and its visualization. In a 3D LUT cube, each point possesses four fundamental attributes: color and three coordinate values, which determine the color's position within the current domain. Fig. 9(a) presents the 3D LUT cube of the identity mapping of SDRTV colors, where the coordinate values of each point align with the three-channel values of its color in the SDRTV domain. For instance, the color (R:128, G:128, B:128) is located at the position (128, 128, 128) in the space. Fig. 9(b) shows the LUT manifold of SDRTV-to-HDRTV color mapping using the base network. In this cube, the coordinate values of each point correspond to its HDRTV color. It can be observed that SDRTV colors (0-255) map to HDRTV colors (0-1023). To demonstrate the image-adaptive functionality, Fig. 10 shows LUT manifolds changing with different inputs,

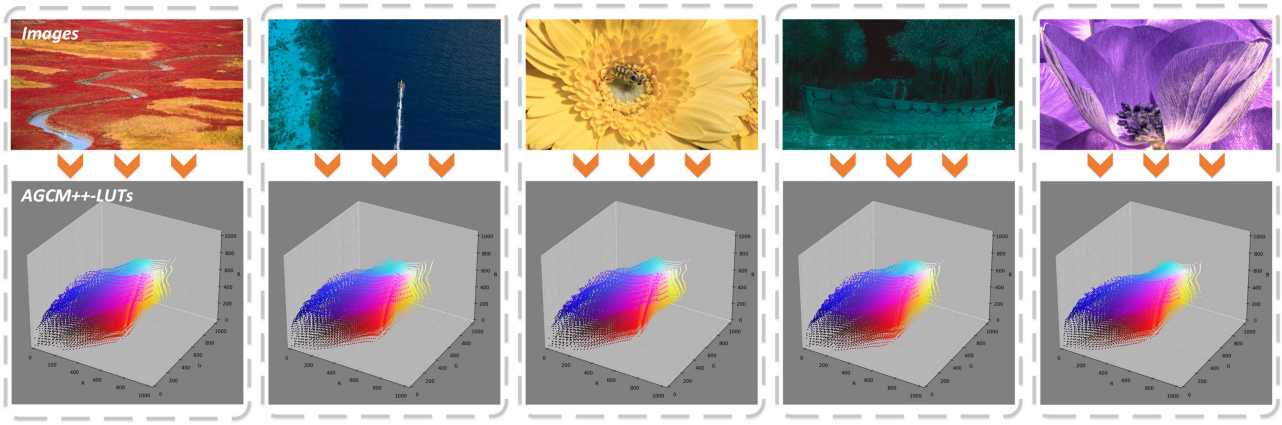


Fig. 10. 3D LUTs generated by using different images as input conditions to our AGCM network.

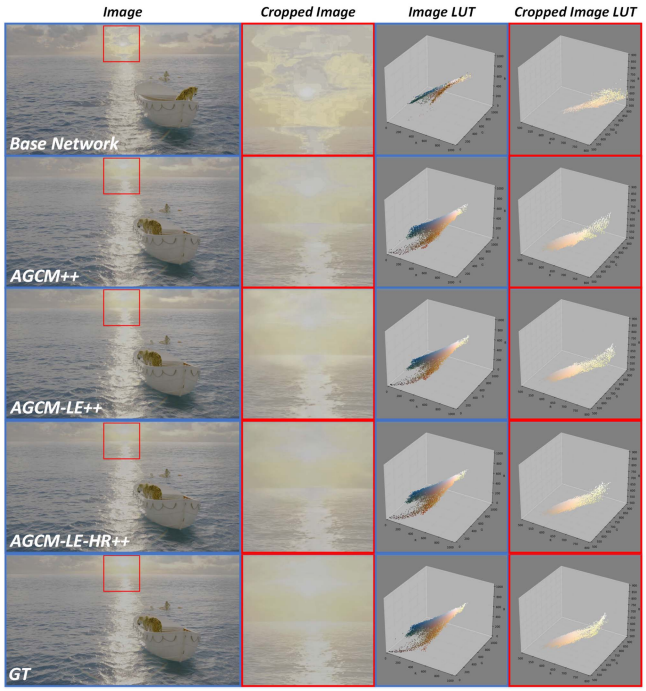


Fig. 11. Visual results and 3D LUT manifolds in different stages.

indicating effective color conditioning of our AGCM. We further demonstrate the functionality of different steps using LUT manifold visualization in Fig. 11. Firstly, the base network can only learn a one-to-one color mapping across the dataset, resulting in a non-smooth color transition of the LUT manifold in highlight areas and severe artifacts in the output. In contrast, the color condition network assists the base network in learning image-adaptive color mapping, removing artifacts in the results produced by AGCM and densifying the LUT manifold. Due to local enhancement, the LUT manifold becomes more compact and smooth, allowing an SDRTV color to map to multiple HDRTV colors through region-dependent operations (i.e., convolutions). It can handle one-to-many color mapping and greatly improve visual quality. Lastly, we can see that highlight refinement further compacts and densifies the LUT manifold, and the results also have natural color transition in highlight regions. The above LUT manifold analysis intuitively reflects the role of different stages in our SDRTV-to-HDRTV solution pipeline.

F. Network Investigation

In this section, we implement comprehensive experiments to investigate specific network design in the proposed method.

Adaptive Global Color Mapping: We first examine effects of different depths of the base network and the condition network in AGCM, as shown in Table III and Table IV. We vary the depth of the base network from 2 to 5 and set the depth of the condition network from 3 to 6. Experimental results show that a base network depth of 3 and a condition network depth of 5 achieves the best performance, thereby being set as the default setting in our method. We also perform an ablation study on the key components of AGCM. As presented in Table V, removing the Dropout layer slightly reduces performance, indicating the effectiveness of Dropout in the condition network. Notably, we can see that without instance normalization will result in a significant performance drop, performing only slightly better than the base network without the condition. This illustrates that this normalization layer is critical for the proposed color condition. We believe this is because instance normalization can help the network learn the key property (i.e., contrast) as the condition.

Local Enhancement: We improved upon the preliminary ResNet-style network by introducing a UNet-style network for local enhancement (See Section IV-D). For fair comparison, we adopt the same AGCM model, i.e., AGCM++, prior to the LE network. As shown in Table VI, our LE++ outperforms previous LE by over 0.8 dB with about half the parameters, indicating the superiority of LE++. There are two primary reasons for the success of our LE++ design: (1) its U-shape design enables stronger representation ability to learn useful features, particularly for high-resolution inputs; and (2) its ability to handle spatially varying mappings via conditional branches makes it

TABLE III
QUANTITATIVE COMPARISONS ON THE DEPTH OF THE BASE NETWORK IN AGCM

Method	Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
AGCM-C5B2[1]	30.2K	36.65	0.9664	0.9966	10.11	8.497
AGCM-C5B3	35.3K	**37.35**	**0.9666**	**0.9968**	9.29	**8.511**
AGCM-C5B4	40.3K	37.31	0.9662	0.9966	**9.16**	8.497
AGCM-C5B5	45.4K	37.31	0.9670	0.9969	9.35	8.504

[1] *B* means the depth of the base network, while *C* represents the depth of the condition network.

TABLE IV
QUANTITATIVE COMPARISONS ON THE DEPTH OF THE CONDITION NETWORK IN AGCM

Method	Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
AGCM-C3B3[1]	8.4K	36.42	0.9645	0.9966	10.25	8.492
AGCM-C4B3	13.9K	36.84	**0.9667**	0.9965	9.56	8.509
AGCM-C5B3	35.3K	**37.35**	0.9666	**0.9968**	**9.29**	**8.511**
AGCM-C6B3	118.8K	37.05	0.9663	0.9966	9.56	8.486

[1] *B* means the depth of the base network, while *C* represents the depth of the condition network.

TABLE V
QUANTITATIVE COMPARISONS ON THE CRITICAL LAYERS IN AGCM

Method	Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
BaseModel-B3	4.6K	36.37	0.9556	0.9963	10.22	8.397
AGCM-woDropout[1]	35.3K	37.00	0.9658	**0.9968**	9.39	8.497
AGCM-woIN[2]	34.8K	36.60	0.9645	0.9967	11.36	8.473
AGCM	35.3K	**37.35**	**0.9666**	**0.9968**	**9.29**	**8.511**

[1] *woDropout* means the Dropout layers are disabled.
[2] *woIN* indicates the Instance Normalization layers are disabled.

TABLE VI
QUANTITATIVE COMPARISONS ON DIFFERENT NETWORKS FOR LE

Method	Params↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
LE[1]	1368K	38.32	0.9736	0.9971	8.14	8.635
LE++	556K	38.45	0.9739	0.9970	7.90	8.666

[1] *LE* denotes the network in the initial version [3] trained based on the same AGCM++ outputs.

Fig. 12. Training curves based on two different loss functions.

particularly suitable for operations that vary across different regions, such as local tone mapping operators.

G. Loss Comparison

In previous works involving SDRTV-to-HDRTV [5], [6], the L_2 loss function is commonly employed to optimize networks, as well as in tasks dominated by color mapping [8]. Thus, we adopt this loss function to optimize AGCM in the initial work. Our experiments, however, indicate that the choice of loss function significantly impacts performance for SDRTV-to-HDRTV. As shown in Fig. 12, we compared training curves using L_1 and L_2 loss functions. The model optimized with L_1 loss exhibits notably better results than with L_2 loss. Previous studies, such as [64], have also demonstrated that L_1 loss often achieves better convergence and higher final performance in image restoration. Therefore, in this study, we utilize L_1 loss to optimize the AGCM network. As demonstrated in Table I, AGCM++ achieves significantly improved performance compared to the original AGCM.

H. Conversion From Base Network to 3D LUT

3D LUTs are widely used in real-world applications for image style and tone manipulation, especially in movie post-production, as part of the digital media process. Many tools directly use the costumed 3D LUT to edit images, thereby constructing available SDRTV-to-HDRTV 3D LUTs is of great value for practical applications. In this section, we show that the base network in our method can be converted into a SDRTV-to-HDRTV 3D LUT with acceptable performance loss. Concretely, we take a 3D grid composed of SDRTV colors as the input to the network and obtain the corresponding HDRTV colors. We can then build a lookup table based on these paired data. When performing color transformation, we can use lookup and trilinear interpolation operations as [8]. As shown in Table VII, we present the results of two settings for converting our base network to 3D LUT. Conversion to a small 3D LUT with 33 nodes (i.e., 3DLUT_s33) results in minimal performance drop (0.16 dB), while a large 3D LUT with 64 nodes (i.e., 3DLUT_s64) almost

TABLE VII
QUANTITATIVE COMPARISONS OF THE BASE NETWORK WITH ITS CONVERTED 3D LUTS

Method	Params ↓	PSNR(dB) ↑	SSIM ↑	SR-SIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑
Base Network	5k	36.14	0.9643	0.9961	10.43	8.305
3DLUT_s33[1]	108k	35.98	0.9645	0.9958	10.60	8.322
3DLUT_s64	786k	36.13	0.9643	0.9960	10.46	8.309

[1] s33 or s64 denote the size of LUT. These two sizes of LUTs are widely used in the actual production.

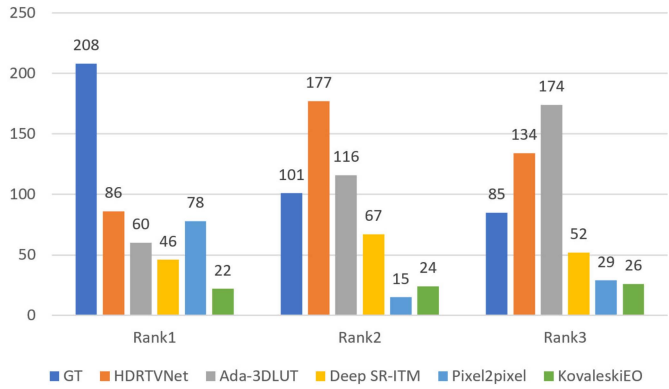


Fig. 13. User study rankings for different methods. Rank 1 means the best subjective feeling.

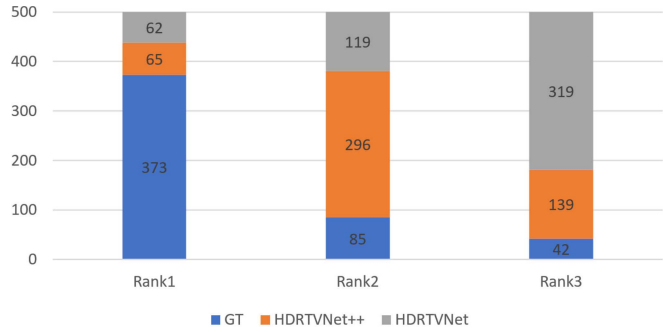


Fig. 14. User study rankings for HDRTVNet and HDRTVNet++. Rank 1 means the best subjective feeling.

maintains performance. This demonstrates the flexibility of our base network to be converted an available SDRTV-to-HDRTV 3D LUT. Furthermore, our base network' efficiency (with only 5 k parameters) allows for efficient training, and it allows modulation according to the condition network to generate customized 3D LUTs.

I. User Study

In the conference version, a user study is conducted with 20 participants to evaluate HDRTVNet's visual quality compared to four top-performing methods. For the experimental setup, 25 images are randomly selected from the test set and display in a darkroom on an HDR TV (Sony X9500 G, peak brightness 1300 nits) configured to the Rec.2020 color gamut and HDR10 standard. We then instruct the participants to consider three main factors when evaluating the images: (1) the presence of artifacts and unnatural colors, (2) the naturalness and comfort of the overall color, brightness, and contrast, and (3) the perception of contrast and highlight details. Using these criteria, participants rank the images in each scenario.

The results of five approaches including KovaleskiEO [58], Pixel2pixel [4], Deep SR-ITM [5], Ada-3DLUT [8] and HDRTVNet [3], along with the ground-truth (GT) images are compared. When ranking the images for a specific scene, participants can view six images from different methods simultaneously or compare any two images until they determined the ranking. The counts of results in the top three ranks are presented in Fig. 13. The GT images and HDRTVNet account for 41.6% (208 counts) and 17.2% (86 counts) of the results considered to have the top-ranked visual quality, respectively. Similarly, HDRTVNet accounts for 35.4% of the results considered to generate the second-best results. In conclusion, HDRTVNet's results are only inferior to the GT images in terms of visual experience in subjective evaluation. We further implement a user study to compare the results by our proposed HDRTVNet++ with the previous HDRTVNet [3]. Participants are asked to rank each set of images in this experiment using the same settings as above. As shown in Fig. 14, the ground-truth still achieve the best visual quality, while HDRTVNet++ shows a better ranking over HDRTVNet. HDRTVNet++ accounts for 59.2% (296 counts) of the results considered to have the second-best visual quality, which greatly outperforms HDRTVNet 23.8% (119 counts). These results demonstrate the superiority of our method in terms of objective visual quality.

VI. CONCLUSION

This paper introduces a new SDRTV-to-HDRTV solution pipeline, leveraging a divide-and-conquer strategy based on the SDRTV/HDRTV formation process. Additionally, we develop HDRTVNet++ to address this challenge effectively. Our approach distinguishes between pixel-independent and region-dependent operations in the formation pipeline, allowing us to implement adaptive global color mapping and local enhancement separately. We design a new color condition network, which offers improved performance with fewer parameters compared to existing methods, to facilitate SDRTV-to-HDRTV color mapping. For enhanced visual results, we employ generative adversarial training to refine highlights. Furthermore, we construct a new HDRTV dataset for rigorous training and testing. Extensive experiments show the superiority of our solution, demonstrating significant improvements on both quantitative metrics and visual quality.

REFERENCES

[1] Y. Huo, F. Yang, L. Dong, and V. Brost, "Physiological inverse tone mapping based on retina response," *Vis. Comput.*, vol. 30, no. 5, pp. 507–517, 2014.

[2] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017.

[3] X. Chen et al., "A new journey from SDRTV to HDRTV," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4500–4509.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[5] S. Y. Kim, J. Oh, and M. Kim, "Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4 K UHD HDR applications," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3116–3125.

[6] 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." in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11287–11295.

[7] J. He, Y. Liu, Y. Qiao, and C. Dong, "Conditional sequential modulation for efficient global image retouching," in *Proc. Comput. Vis.–ECCV 2020: 16th Eur. Conf.*, Glasgow, U.K., Springer, 2020, pp. 679–695.

[8] H. Zeng, J. Cai, L. Li, Z. Cao, and L. Zhang, "Learning image-adaptive 3D lookup tables for high performance photo enhancement in real-time," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 2058–2073, Apr. 2022.

[9] L. Zhang and H. Li, "SR-SIM: A fast and high performance IQA index based on spectral residual," in *Proc. 19th IEEE Int. Conf. Image Process.*, 2012, pp. 1473–1476.

[10] International Telecommunication Union Radiocommunication Sector (ITU-R), "Objective metric for the assessment of the potential visibility of colour differences in television," ITU-R, Geneva, Switzerland, Tech. Rep. BT.2124-0, 2019.

[11] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–14, 2011.

[12] G. Xu, Q. Hou, L. Zhang, and M.-M. Cheng, "FMNet: Frequency-aware modulation network for SDR-to-HDR translation," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 6425–6435.

[13] G. He et al., "SDRTV-to-HDRTV via hierarchical dynamic context feature mapping," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 2890–2898.

[14] Z. Cheng et al., "Towards real-world HDRTV reconstruction: A data synthesis-based approach," in *Proc. Comput. Vis.–ECCV 2022: 17th Eur. Conf.*, Tel Aviv, Israel, Springer, 2022, pp. 199–216.

[15] M. Yao, D. He, X. Li, Z. Pan, and Z. Xiong, "Bidirectional translation between UHD-HDR and HD-SDR videos," *IEEE Trans. Multimedia*, vol. 25, pp. 8672–8686, 2023.

[16] C. Guo, L. Fan, Z. Xue, and X. Jiang, "Learning a practical SDR-to-HDRTV up-conversion using new dataset and degradation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22231–22241.

[17] H. Zhang et al., "EffiHDR: An efficient framework for HDRTV reconstruction and enhancement in UHD systems," *IEEE Trans. Broadcast.*, vol. 70, no. 2, pp. 620–636, Jun. 2024.

[18] International Telecommunication Union Radiocommunication Sector (ITU-R), "Parameter values for the hdtv standards for production and international programme exchange," ITU-R, Geneva, Switzerland, Tech. Rep. BT.709-6, 2015.

[19] International Telecommunication Union Radiocommunication Sector (ITU R), "Reference electro-optical transfer function for flat panel displays used in hdtv studio production," ITU-R, Geneva, Switzerland, Tech. Rep. BT.1886, 2011.

[20] International Telecommunication Union Radiocommunication Sector (ITU R), "Parameter values for ultra-high definition television systems for production and international programme exchange," ITU-R, Geneva, Switzerland, Tech. Rep. BT.2020-2, 2015.

[21] International Telecommunication Union Radiocommunication Sector (ITU-R), "Image parameter values for high dynamic range television for use in production and international programme exchange," ITU-R, Geneva, Switzerland, Tech. Rep. BT.2100-2, 2018.

[22] 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. Comput. Vis.–ACCV 2018: 14th Asian Conf. Comput. Vis.*, Perth, Australia, Springer, 2019, pp. 395–409.

[23] F. Kou et al., "Intelligent detail enhancement for exposure fusion," *IEEE Trans. Multimedia*, vol. 20, pp. 484–495, Feb. 2018.

[24] S. Lee, S. Y. Jo, G. H. An, and S.-J. Kang, "Learning to generate multi-exposure stacks with cycle consistency for high dynamic range imaging," *IEEE Trans. Multimedia*, vol. 23, pp. 2561–2574, 2021.

[25] X. Tan, H. Chen, K. Xu, Y. Jin, and C. Zhu, "Deep SR-HDR: Joint learning of super-resolution and high dynamic range imaging for dynamic scenes," *IEEE Trans. Multimedia*, vol. 25, pp. 750–763, 2023.

[26] Q. Yan et al., "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1751–1760.

[27] F. Banterle et al., "High dynamic range imaging and low dynamic range expansion for generating HDR content," *Comput. Graph. Forum*, vol. 28, no. 8, pp. 2343–2367, 2009.

[28] 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, 2017.

[29] Y.-L. Liu et al., "Single-image HDR reconstruction by learning to reverse the camera pipeline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1651–1660.

[30] Y. Nam, J. Kim, J.-H. Shim, and S.-J. Kang, "Deep conditional HDRI: Inverse tone mapping via dual encoder-decoder conditioning method," *IEEE Trans. Multimedia*, vol. 26, pp. 8504–8515, 2024.

[31] X. Hu, L. Shen, M. Jiang, R. Ma, and P. An, "La-HDR: Light adaptive hdr reconstruction framework for single ldr image considering varied light conditions," *IEEE Trans. Multimedia*, vol. 25, pp. 4814–4829, 2023.

[32] International Telecommunication Union Radiocommunication Sector (ITU-R), "Colour conversion from recommendation itu-r bt.709 to recommendation itu-r bt.2020," ITU-R, Geneva, Switzerland, Tech. Rep. BT.2087-0, 2015.

[33] S. W. Zamir et al., "Gamut mapping in cinematography through perceptually-based contrast modification," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 3, pp. 490–503, Jun. 2014.

[34] S. W. Zamir, J. Vazquez-Corral, and M. Bertalmıó, "Gamut extension for cinema: Psychophysical evaluation of the state of the art and a new algorithm," *Hum. Vis. Electron. Imag. XX*, vol. 9394, pp. 278–289, 2015.

[35] F. Schweiger, T. Borer, and M. Pindoria, "Luminance-preserving color conversion," *SMPTE Mot. Imag. J.*, vol. 126, no. 3, pp. 45–49, 2017.

[36] S. W. Zamir, J. Vazquez-Corral, and M. Bertalmıó, "Gamut extension for cinema," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1595–1606, Apr. 2017.

[37] L. Xu, B. Zhao, and M. R. Luo, "Color gamut mapping between small and large color gamuts: Part ii. gamut extension," *Opt. Exp.*, vol. 26, no. 13, pp. 17335–17349, 2018.

[38] S. W. Zamir, J. Vazquez-Corral, and M. Bertalmio, "Vision models for wide color gamut imaging in cinema," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1777–1790, May 2021.

[39] International Telecommunication Union Radiocommunication Sector (ITU-R), "High dynamic range television for production and international programme exchange," ITU-R, Geneva, Switzerland, Tech. Rep. BT.2390-8, 2020.

[40] 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, 2003.

[41] E. Reinhard, "Parameter estimation for photographic tone reproduction," *J. Graph. Tools*, vol. 7, no. 1, pp. 45–51, 2002.

[42] J. Tumblin and H. Rushmeier, "Tone reproduction for realistic images," *IEEE Comput. Graph. Appl.*, vol. 13, no. 6, pp. 42–48, Nov. 1993.

[43] G. W. Larson, H. Rushmeier, and C. Piatko, "A visibility matching tone reproduction operator for high dynamic range scenes," *IEEE Trans. Vis. Comput. Graphics*, vol. 3, no. 4, pp. 291–306, Oct.–Dec. 1997.

[44] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski, "Interactive local adjustment of tonal values," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 646–653, 2006.

[45] J. Hable, "Uncharted 2: HDR lighting," in *Proc. Game Developers Conf.*, 2010, p. 56.

[46] *High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, Standard ST2084:2014, Society of Motion Picture and Television Engineers, White Plains, NY, USA, 2014.

[47] Y. Liu et al., "Very lightweight photo retouching network with conditional sequential modulation," *IEEE Trans. Multimedia*, vol. 25, pp. 4638–4652, 2023.

[48] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 606–615.

[49] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *Proc. Int. Conf. Learn. Representations*, 2016.
[50] X. Chen, Y. Liu, Z. Zhang, Y. Qiao, and C. Dong, "HDRUNet: Single image HDR reconstruction with denoising and dequantization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 354–363.
[51] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.
[52] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1905–1914.
[53] A. Jolicoeur-Martineau, "The relativistic discriminator: A key element missing from standard GAN," in *Proc. Int. Conf. Learn. Representations*, 2018.
[54] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Comput. Vis.–ECCV 2016: 14th Eur. Conf.*, Amsterdam, The Netherlands, Springer, 2016, pp. 694–711.
[55] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[56] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6228–6237.
[57] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.
[58] R. P. Kovaleski and M. M. Oliveira, "High-quality reverse tone mapping for a wide range of exposures," in *Proc. 27th SIBGRAPI Conf. Graph.*, Patterns Images, 2014, pp. 49–56.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.
[60] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2223–2232.
[61] G. Xu, Q. Hou, and M.-M. Cheng, "Dual frequency transformer for efficient sdr-to-hdr translation," *Mach. Intell. Res.*, vol. 21, no. 3, pp. 538–548, 2024.
[62] P. Chen, W. Yang, and S. Wang, "Reviving standard-dynamic-range videos for high-dynamic-range devices: A learning paradigm with hybrid attention mechanisms," *IEEE MultiMedia*, vol. 30, no. 3, pp. 110–118, Jul.–Sep. 2023.
[63] S. Athar, T. Costa, K. Zeng, and Z. Wang, "Perceptual quality assessment of uhd-hdr-wcg videos," in *Proc. 2019 IEEE Int. Conf. Image Process.*, 2019, pp. 1740–1744.
[64] 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.

Xiangyu Chen (Member, IEEE) received the B.E. and M.E. degrees from Northwestern Polytechnical University, Xi'an, China, in 2017 and 2020, respectively. He was a Research Assistant in Multimedia Laboratory, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China, from 2019 to 2021. He is currently a joint Ph.D. Student with the University of Macau, Macao, China, and Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is also a research intern in Shanghai Artificial Intelligence Laboratory. His research interests include high dynamic range imaging, image restoration and enhancement, and general low-level vision.

Zheyuan Li received the bachelor's degree from Northwestern Polytechnical University, Xi'an, China, in 2022. He was a Research Assistant in Multimedia Laboratory with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Beijing, China. He is currently working toward the Ph.D. degree with the University of Macau, Macao, China. His research interests include computer vision and multi-modal model.

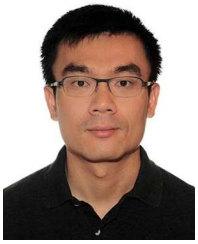
Zhengwen Zhang received the M.S. degree from the Beijing Institute of Technology, Beijing, China, in 2019. He is currently a Research Assistant in Multimedia Laboratory, Shenzhen institutes of Advanced Technology, Chinese Academy of Sciences, Beijing. He is also supervised by Prof. Yu Qiao and Prof. Chao Dong. His research interests include computer vision, image/video enhancement, and weakly supervised learning.

Jimmy S. Ren received the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2013. He was an Executive Research Director with SenseTime. He is currently an Assistant Professor in Hong Kong Metropolitan University, Hong Kong. The computational photography products he shipped were adopted by tens of millions smartphone users in the world. His research interests include computational photography, image processing, and computer vision.

Yihao Liu received the B.S. and Ph.D. degrees from the University of Chinese Academy of Sciences, Beijing, China, in 2018 and 2023, respectively. He was affiliated with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He holds a research position with the Shanghai Artificial Intelligence Laboratory. His research interests include computer vision and image processing, with a distinct focus on image, and video restoration and enhancement.

Jingwen He received the B.Eng. degree in computer science and technology from Sichuan University, Chengdu, China, in 2016, and the M.Phil. degree in electronic and information engineering from the University of Sydney, Camperdown, NSW, Australia, in 2019. She is currently working toward the Ph.D degree with the Informtion Engineering Department, Chinese University of Hong Kong, Hong Kong. Her current research interests include deep learning and computer vision.

Yu Qiao (Senior Member, IEEE) is currently a Professor with Shanghai AI Laboratory and the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Sciences, Beijing, China. He has authored or coauthored more than 600 articles in international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SIGNAL PROCESSING, CVPR, and ICCV. His research interests include computer vision, deep learning, and bioinformation. He was the recipient of the First Prize of the Guangdong Technological Invention Award, Jiaxi Lv Young Researcher Award from the Chinese Academy of Sciences, and Distinguished Paper Award in AAAI 2021. His group achieved the first runner-up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition, and the winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He was the Program Chair of IEEE ICIST 2014.



Jiantao Zhou (Senior Member, IEEE) received the B.Eng. degree from the Department of Electronic Engineering, Dalian University of Technology, Dalian, China, in 2002, the M.Phil. degree from the Department of Radio Engineering, Southeast University, in 2005, and the Ph.D. degree from the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, in 2009. He held various research positions with the University of Illinois at Urbana- Champaign, Champaign, IL, USA, Hong Kong University of Science and Technology, and McMaster University, Hamilton, ON, Canada. He is currently a Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macao, China. His research interests include multimedia security and forensics, multimedia signal processing, artificial intelligence, and Big Data. He holds four granted U.S. patents and two granted Chinese patents. He has coauthored two papers that received the Best Paper Award at the IEEE Pacific-Rim Conference on Multimedia in 2007 and the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2016. He is also an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, and IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING.

Chao Dong is currently a Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Science (SIAT), Beijing, China, and Shanghai AI Laboratory. He was in SenseTime from 2016 to 2018, as the Team Leader of Super-Resolution Group. In 2014, he first introduced deep learning method – SRCNN into the super-resolution field. This seminal work was chosen as one of the top ten "Most Popular Articles" of TPAMI in 2016. His current research interests include low-level vision problems, such as image/video super-resolution, denoising, and enhancement. His team has awarded several championships in international challenges –NTIRE2018, PIRM2018, NTIRE2019, NTIRE2020 AIM2020, and NTIRE2022. In 2021, he was chosen as one of the World's Top 2% Scientists. In 2022, he was recognized as the AI 2000 Most Influential Scholar Honorable Mention in computer vision.