Beyond Feature Mapping GAP: Integrating Real HDRTV Priors for Superior SDRTV-to-HDRTV Conversion

Gang He[1] , **Kepeng Xu**[1*] , **Li Xu**[1] , **Wenxin Yu**[23] , **Xianyun Wu**[1]
[1]Xidian University
[2]Southwest University of Science and Technology
[3]Fujiang Laboratory
kepengxu11@gmail.com

Abstract

The rise of HDR-WCG display devices has highlighted the need to convert SDRTV to HDRTV, as most video sources are still in SDR. Existing methods primarily focus on designing neural networks to learn a single-style mapping from SDRTV to HDRTV. However, the limited information in SDRTV and the diversity of styles in real-world conversions render this process an ill-posed problem, thereby constraining the performance and generalization of these methods. Inspired by generative approaches, we propose a novel method for SDRTV to HDRTV conversion guided by real HDRTV priors. Despite the limited information in SDRTV, introducing real HDRTV as reference priors significantly constrains the solution space of the originally high-dimensional ill-posed problem. This shift transforms the task from solving an unreferenced prediction problem to making a referenced selection, thereby markedly enhancing the accuracy and reliability of the conversion process. Specifically, our approach comprises two stages: the first stage employs a Vector Quantized Generative Adversarial Network to capture HDRTV priors, while the second stage matches these priors to the input SDRTV content to recover realistic HDRTV outputs. We evaluate our method on public datasets, demonstrating its effectiveness with significant improvements in both objective and subjective metrics across real and synthetic datasets.

1 Introduction

The dynamic range of a video, defined by the difference between its maximum and minimum luminance, enables High Dynamic Range (HDRTV) to deliver superior visuals. Advances in Electro-Optical Transfer Functions (EOTF), such as PQ/HLG, and Wide Color Gamut (WCG) RGB primaries (as per BT.2020), enhance HDR technology's potential.

Despite the rise of WCG-HDR displays, the production complexities result in limited WCG-HDR content. Consequently, many films remain in Standard Dynamic Range (SDRTV), driving the demand for SDRTV-to-HDRTV conversions. HDRTV offers a wider color gamut (Rec. 2020 vs Rec. 709), higher brightness range (0.01-1000 nits vs 0.1-100 nits), advanced EOTF curves (PQ/HLG vs Gamma), and greater color depth (10-bit vs 8-bit). However, the scarcity of HDRTV content compared to SDRTV makes SDRTV-to-HDRTV conversion essential, despite the inherent challenges due to the limitations of existing imaging systems and transmission protocols.

Traditional methods Huo *et al.* [2013]; Kovaleski and de Oliveira Neto [2014]; Ma *et al.* [2023] suffer from color inaccuracies and abnormal brightness restoration due to limitations in estimating curve parameters for SDRTV to HDRTV conversion. Meanwhile, recent neural network-based approaches Kim *et al.* [2019, 2020]; Chen *et al.* [2021b]; Cao *et al.* [2022]; Xu *et al.* [2022]; Shao *et al.* [2022] employ the strategy of encoding SDRTV content into a latent space and subsequently reconstructing it as HDRTV content. Models designed by these previous methods are trained and tested on a single data set, as shown in Figure 1 (a). However, SDRTV to HDRTV conversion models trained on a single dataset are difficult to adapt to the content diversity of the real world. This challenge arises because these models learn a fixed mapping that is inherently tied to the specific characteristics of the dataset on which they are trained. Actually, these characteristics can include, but are not limited to, lighting conditions, types of scenes, and tone mapping schemes.

Insights. In the process of video production, a single HDRTV content might correlate with various SDRTV-style versions. This scenario underscores the complexity faced when training neural networks to understand not just a single linear relationship. Due to the ill-posed of multiple mapping relationships and the lack of deterministic mapping rules, it is difficult for neural networks to learn such chaotic mapping functions. This complexity highlights the substantial challenge in developing neural networks that can directly learn the diverse SDRTV-to-HDRTV conversions reflective of real-world scenarios. Therefore, it is very difficult to build a neural network to directly learn real-world SDRTV-to-HDRTV conversion.

Previous Solution. HDRTVDMGuo *et al.* [2023] emphasizes the importance of dataset diversity for training neural networks that more closely match real-world content. Although the complexity of these datasets is close to actual con-

[*]Corresponding author

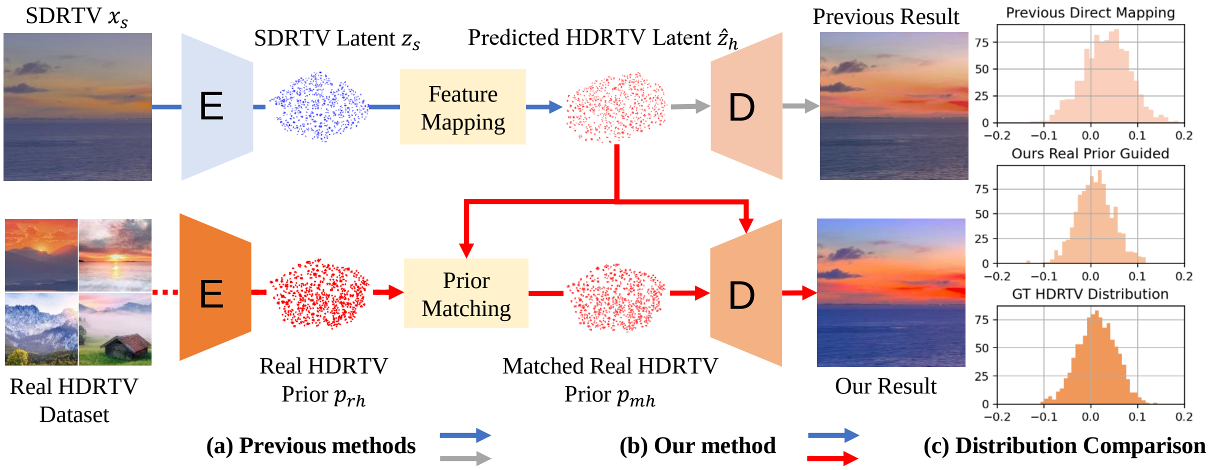

Figure 1: (a) Previous methods learn single-style SDRTV-to-HDRTV conversion on a single dataset. However, the SDRTV-to-HDRTV conversion distribution in real-world scenarios is complex and diverse, which makes it difficult for previous methods to effectively convert SDRTV-to-HDRTV conversion in the real-world. (b) Our method embeds rich and realistic HDRTV into the converted neural network, thereby greatly improving the conversion performance in real scenes. (c) The latent variable distribution of our method is closer to GT due to the incorporation of real HDRTV prior guidance.

ditions, existing neural networks face difficulties in fully capturing and learning the diverse mapping relationships present within the dataset. This situation calls for improvements in neural network architectures to better process and understand the rich, complex data reflective of real-world scenarios.

Our Solution. In contrast, our proposed RealHDRTVNet framework enhances the quality of SDRTV to HDRTV conversion by directly embedding HDRTV priors into the transformation process, as illustrated in Fig.1 (b). This approach effectively transforms the ill-posed restoration problem into a prior selection problem, significantly reducing the solution space size. By leveraging rich and diverse HDRTV priors, our method overcomes previous limitations, achieving more accurate, generalized, and reliable SDRTV to HDRTV mapping.

Moreover, a significant challenge in SDRTV-to-HDRTV conversion is the assessment of the perceptual quality of neural network-generated HDRTV content. Common metrics like LPIPS Zhang *et al.* [2018], NIQE Mittal *et al.* [2012], and FID Heusel *et al.* [2017], which are typically used for SDRTV quality evaluation, struggle to capture HDRTV's unique features within the PQ EOTF curve and Rec.2020 color gamut.

Inspired by this, our work extend tailored metrics for precise HDRTV quality assessment, including Learned Perceptual HDRTV Patch Similarity (LPHPS), Natural HDRTV Quality Evaluator (NHQE), and Fréchet Initial Distance (FHAD). These metrics are specifically designed to evaluate the subjective quality of HDRTV content directly. With these innovative metrics, both researchers and practitioners have the tools to conduct reliable subjective quality evaluations of HDRTV content.

This paper's contributions are in follow:

- We propose an a priori selected SDRTV to HDRTV conversion method, which significantly limits the solution space of the original high-dimensional ill-posed problem, thereby enabling efficient learning of real-world SDRTV to HDRTV conversion and improving the quality of the converted HDRTV.
- We quantitatively and qualitatively demonstrate that our proposed method outperforms previous methods.

2 Related Work

2.1 SDRTV-to-HDRTV Methods

Our Research Scope

Recent advancements in HDR imaging have seen the advent of various learning-based methods, as noted by Wang and Yoon Wang and Yoon [2022]. These methods have found different applications in the real world. HDR enhancement primarily involves the use of neural networks to convert SDR images to HDR Kovaleski and de Oliveira Neto [2014]. On the other hand, Multi-Exposure HDR Imaging employs exposure bracketing to create HDR images from a sequence of SDR images taken at different exposure levels Chaudhari *et al.* [2019]; Le *et al.* [2022]; Xu *et al.* [2021]; Chen *et al.* [2021a]. This paper concentrates on the transformation of SDRTV to HDRTV, aiming to achieve an enhanced dynamic range and wide color gamut in videos.

DNN-based SDRTV-to-HDRTV Methods

Recent studies primarily focus on devising feature modulation strategies for robust SDRTV-to-HDRTV conversion Kim *et al.* [2019, 2020]; Zeng *et al.* [2020]; Chen *et al.* [2021b]; Cao *et al.* [2022]; Xu *et al.* [2022]; Shao *et al.* [2022]. These strategies entail the utilization of diverse color-prior techniques to facilitate effective feature modulation. The HDRTVDM approach Guo *et al.* [2023] notably enhances HDRTV conversion quality by refining the dataset. In terms of integrated frameworks, SR-ITM Kim *et al.* [2019] presents a network design that concurrently addresses SDRTV-to-HDRTV transformation and super-resolution. Similarly,

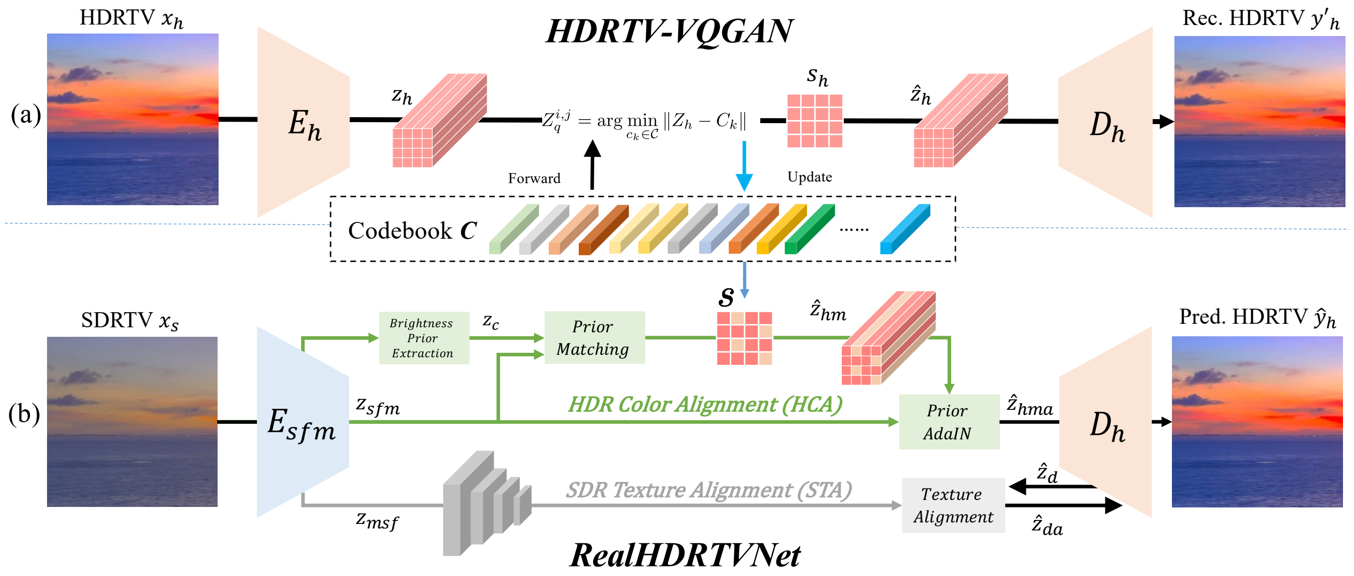

Figure 2: RealHDRTVNet framework. (a) HDRTV-VQGAN. We first pre-train an HDRTV-VQGAN to learn to store HDRTV priors through self-reconstruction. (b)RealHDRTVNet. The learning modulation encoder E_{sfm} obtains "nearly high-quality HDRTV features". Next, the HDR Color Alignment HCA module aligns the input features with HDRTV in the color dynamic range dimension. In addition, the SDR Texture Alignment STA module is used to align the texture with the input SDRTV. This makes the dynamic range information of the conversion result consistent with HDRTV, and the texture details consistent with SDRTV.

DIDnet Xu *et al.* [2023] introduces a network model that combines SDRTV-to-HDRTV conversion with the restoration of coding artifacts.

SDRTV-to-HDRTV Datasets

Within the current research landscape, a mere quartet of open-source HDRTV datasetsKim *et al.* [2019]; Chen *et al.* [2021b]; Guo *et al.* [2023] exists. Each dataset adheres to the BT.2020 RGB color gamut. Further, they conform to the PQ EOTF specifications with a luminance zenith of 1000 nits.

2.2 Generative Adversarial Network

The introduction of Generative Adversarial Networks (GANs), as demonstrated in Goodfellow et al.'s 2014 work Goodfellow *et al.* [2014], has revolutionized tasks like image restoration Wang *et al.* [2021]. GAN priors, known for capturing complex image distributions, contrast traditional methods by integrating natural image characteristics into degraded visuals through adversarial training. This approach has advanced subfields like denoising, inpainting, super-resolution Wang *et al.* [2018], and artifact reduction Wan *et al.* [2020]. The strength of GAN priors lies in their restoration ability and in producing outputs closely aligned with original images. Additionally, models like VQGAN have been extensively applied in areas like multi-modal conversion Esser *et al.* [2021].

3 Methodology

3.1 Preliminary

The conversion from SDRTV to HDRTV can be mathematically represented by the Maximum A Posteriori (MAP) estimate $p(h \mid s)$, which relies on the distribution of the SDRTV input $p(s)$. Traditional methods typically utilize inverse tone mapping functions to achieve this conversion. This process is formally defined in Equation 1:

$$\hat{h} = f(s; \theta_c) + \epsilon, \tag{1}$$

where $\hat{h}$ represents the reconstructed HDRTV, θ_c denotes the estimated curve parameters, and ϵ accounts for the error due to parameter estimation inaccuracies. This method often leads to color distortions and anomalies in brightness restoration in the converted HDRTV content.

Modern neural network-based methods Kim *et al.* [2019, 2020]; Zeng *et al.* [2020]; Chen *et al.* [2021b]; Cao *et al.* [2022]; Xu *et al.* [2022]; Shao *et al.* [2022] utilize a three-stage process to complete SDRTV-to-HDRTV conversion. Initially, an encoder module is employed to map the SDRTV content into a latent representation. Subsequently, the SDRTV latent features are input to the feature transformation module to obtain HDRTV latent features. The final stage involves a decoder module, which reconstructs the latent representation back into HDRTV. By minimizing the difference with real samples, the distribution of reconstructed HDRTV is close to the distribution of real data. This dependency can be modeled as a conditional probability distribution $P(h|s)$, which is only influenced by the distribution of the input SDRTV $P(s)$:

$$p_\theta(h|s) = \int p_{\theta_d}(h|\hat{z_h}) p_{\theta_\tau}(\hat{z_h}|z_s) p_{\theta_e}(z_s|s) ds \tag{2}$$

where:

- θ_e, θ_τ, and θ_d represent the parameters of the learned encoder module, feature transformation module, and decoder module of the neural network.

- z_s is the initial latent representation into which the SDRTV image s is encoded.
- $\hat{z_h}$ is the intermediate latent representation obtained after encoding and feature transformation of the SDRTV image s.

These methods aim to learn a straightforward and static mapping relationship from SDRTV to HDRTV, inadequately capturing the intricate and multifaceted conversion process between SDRTV and HDRTV. HDRTVDMGuo *et al.* [2023] believes that increasing the complexity of the dataset can force the neural network to learn an SDRTV-to-HDRTV mapping function that is more in line with the real world. Although increasing the dataset is closer to the real world at the sample level, it is difficult for the neural network to directly learn such a complex mapping process.

Transitioning from SDRTV to HDRTV inherently involves conditional probability transition. On a singular dataset, these conditional transitions adhere to a stationary distribution, making it feasible for neural networks to learn such probabilistic mappings with relative ease. However, in the real-world scenario, the SDRTV-to-HDRTV conditional probability transitions do not conform to a single, fixed distribution, posing a challenge for neural architectures to learn this intricate probability transition.

To address this challenge, we propose a framework that can directly harness the prior knowledge encapsulated within HDRTV data to facilitate a superior-quality transformation from SDRTV to HDRTV. The conversion can be described by the following equation.

$$
\begin{aligned}
p_\theta(h|s,\pi) = \int & p_{\theta_d}(h|z_m) p_{\theta_m}(z_m|\hat{z_h},\pi) \\
& p_{\theta_\tau}(\hat{z_h}|z_s) p_{\theta_e}(z_s|s)\, ds
\end{aligned}
\tag{3}
$$

where:

- θ is the set of network parameters.
- z_m is the best matching HDRTV prior representation selected from the HDRTV prior distribution π.
- $\hat{z_h}$ is the intermediate latent representation obtained after encoding and mapping the SDRTV image s.
- π is the HDRTV prior set .
- z_s is the initial latent representation into which the SDRTV image s is encoded.

To circumvent the traditional reliance solely on SDRTV inputs, our approach ingeniously integrates HDRTV priors into the conversion workflow, thereby facilitating efficient and complex HDRTV reconstruction. Our method enables adaptive conditional probability transitions based on the embedded HDRTV priors, moving beyond fixed-type conditional transitions to achieve high-quality conversion from SDRTV inputs.

3.2 Overall Framework

Following the instantiation of our motivation, we propose RealHDRTVNet, a novel architecture designed to learn the complex and variable transformations from SDRTV to HDRTV, demonstrating superior generalization capabilities in authentic scenarios. Our methodology unfolds in three phases: initially, HDRTV-VQGAN is trained to embed real HDRTV priors. The subsequent two phase focuses on the SDRTV-to-HDRTV transformation, leveraging the pre-embedded HDRTV priors to augment the quality of HDRTV. The embedded HDRTV prior can serve as a powerful guide to ensure high-quality HDRTV restoration. The proposed method no longer learns a simple function mapping that only relies on SDRTV, but learns a transformation process guided by a real HDRTV prior. Guided by real HDRTV, our method can learn complex and diverse SDRTV-to-HDRTV conversion relationships. Based on this, our method adopts a three-phase strategy:

- Phase I: We train a VQGAN model on HDRTV domain, embedding real HDRTV priors.
- Phase II: We craft a preliminary modulation encoder for SDRTV to HDRTV transformation. Through feature modulation techniques, this model refines the SDRTV latent distribution towards the anticipated HDRTV latent space, ensuring that the next stage can more accurately match the HDRTV prior.
- Stage III: We propose the HDR Color Alignment module HCA and the SDR Texture Alignment module STA. HCA identifies the best HDRTV prior from the pre-trained VQGAN codebook and uses the identified HDRTV to assist the conversion process. The SDR Texture Alignment module STA aligns the transformed features with SDRTV in texture to ensure texture fidelity.

With this three-stage approach, we provide a novel solution for high-quality SDRTV-to-HDRTV conversion.

3.3 Phase I: HDRTV Vector Quantized AutoEncoder - HDR Prior Representation Learning

To reduce the uncertainty of SDRTV-to-HDRTV mapping and complement high-quality HDRTV color information, we first train a vector quantized autoencoder to learn a context-rich codebook that improves the quality of the converted HDRTV.

The specific structure is shown in Fig. 2 (a). First, the HDRTV SDRTV $x_h \in R^{H \times W \times 3}$ is input into the encoder E_h to get the latent feature $z_h \in R^{h \times w \times c}$. Next, find the closest codebook feature $\hat{z}_h$ to z_h in codebook $C \in R^{n \times k}$ (n is the number of vectors in the codebook) and the corresponding codebook index S. Then $\hat{z}_h$ is fed into the decoder D_h to obtain the reconstructed HDRTV y'_h.

We describe the details of HDRTV-VQGAN. First use the encoder E_h to encode the input HDRTV x_h into a latent representation z_h.

$$z_h = E_h(x_h), \tag{4}$$

where $x_h \in \mathbb{R}^{H \times W \times 3}$ represents the input HDRTV, and $z_h \in \mathbb{R}^{h \times w \times c}$ denotes the latent feature obtained from the encoder E_h.

Next, the replaced features $\hat{z}_h$ and corresponding index s_h are obtained from the codebook C through nearest neighbor matching;

$$\hat{z}_h, s_h = \operatorname*{argmin}_{c_i \in C} \|z_h - c_i\|, \tag{5}$$

where $C \in \mathbb{R}^{n \times d}$ is the codebook containing n vectors, each of dimension d, and $\hat{z}_h$ is the codebook feature closest to z_h. The corresponding codebook index is obtained as $S =$ index of $\hat{z}_h$.

Then,$\hat{z}_h$ is processed via the decoder D_h get converted HDRTV:

$$y'_h = D_h(\hat{z}_h), \tag{6}$$

3.4 Phase II: SDRTV Modulation Encoder - Preliminary HDR Mapping

Given the substantial differences in dynamic range and color space between SDRTV and HDRTV, direct alignment of SDRTV features to HDRTV priors in latent space is inherently challenging. To mitigate this, we decompose the prior matching problem into a two-stage process.

First, we propose the SDR Feature Modulation Encoder (E_{sfm}) to transform SDRTV latent features into a space that is more congruent with HDRTV priors. The encoder E_{sfm} is composed of four sequential SDRTV Feature Modulation (SFM) blocks, which iteratively refine the latent representation z_{sfm} to approximate the HDRTV distribution.

Formally, the input SDRTV frame x_s (referred to as x_0 here) is processed as follows:

$$x_i = \text{Down}(\text{SFM}(\text{ResBlock}_c(x_{i-1}))), \tag{7}$$

where $i \in \{1, 2, 3, 4\}$, and $c \in \{64, 128, 256, 512\}$, representing the number of channels at each respective stage. In detail, the SFM is implemented by:

$$\hat{x}_i = \alpha_i \odot x_i + \beta_i; \;\; \alpha_i, \beta_i = \text{Conv}_\theta(x_i), \tag{8}$$

where $\odot$ denotes element-wise multiplication, α_i and β_i denote the modulation parameters derived via convolutional layers Conv_θ.

Subsequently, a Transformer module aggregates the modulated features into a compact latent representation:

$$z_{sfm} = \text{ResBlock}_{c512}(\text{Transformer}(x_{m_4})). \tag{9}$$

This process effectively aligns the SDRTV latent space with that of HDRTV, thereby facilitating optimal prior matching in subsequent processing stages.

3.5 Phase III: RealHDRTVNet - High-Quality Conversions with Pretrain HDR Prior

In phase I and II, E_{sfm}, the codebook C, and D_h are pretrained and then frozen in the subsequent phase. In phase III, to achieve efficient and nuanced HDRTV reconstruction, we propose RealHDRTVNet, which consists of two key components: **HDR Color Alignment Module(*HCA*)** and **SDR Texture Alignment(*STA*)**. These components ensure that the converted HDRTV retains the texture structure of the input SDRTV while aligning with the dynamic range and color of real HDRTV.

HDR Color Alignment Module(HCA)

To achieve accurate color alignment with real HDRTV priors, we introduce the HDR Color Alignment module(HCA), which integrates the functionalities of brightness prior extraction, prior matching, and prior adaptive instance normalization.

Given an SDRTV input x_s, the encoder E_{sfm} first extracts the multi-scale feature z_{msf} and the "basic" HDRTV feature z_{sfm}. Recognizing the necessity for adaptive processing in highlight regions, we employ a brightness prior extraction module to generate a luminance-aware position coding z_c using X_s and its highlight mask. This feature z_c is then fed into the Prior Matching (PM) module, where it is aligned with the optimal HDRTV prior $\hat{h}_{hm}$ from the codebook C, resulting in the matched feature $\hat{z}_{hm}$.

After obtaining $\hat{z}_{hm}$, Prior Adaptive Instance Normalization (AdaIN) adjusts the latent features of z_{sfm} and $\hat{z}_{hm}$, generating $\hat{z}_{hma}$. The feature $\hat{z}_{hma}$ is then fed into the pretrained D_h to produce the HDRTV result. This process ensures that the color and dynamic range of the generated HDRTV align with real HDRTV priors.

SDR Texture Alignment Module(STA)

. To maintain the texture structure of the input SDRTV in the converted HDRTV, we introduce the SDR Texture Alignment module. This module uses multi-scale features z_{msf} from the encoder E_{sfm} during the decoding process to align textures.

First, the multi-scale features z_{msf} from the encoder are concatenated with the decoder features $\hat{z}_d$ and fed into a deformable convolution to achieve alignment, resulting in aligned features $\hat{z}_n$. These aligned features are then used to modulate the decoder features $\hat{z}_d$, resulting in the aligned decoder feature $\hat{z}_{da}$, which is fed into decoder D_h for HDRTV reconstruction.

By performing SDR Texture Alignment on the decoder at four different resolutions, the decoder D_h is capable of recovering HDRTV with a realistic dynamic range while preserving high-quality texture details from the SDRTV.

Overall Pipeline

The complete RealHDRTVNet pipeline is described as follows:

$$z_{sfm}, z_{msf} = E_{sfm}(x_s), \tag{10}$$

$$\hat{z}_{hma} = HCA(z_{sfm}, x_s, C), \tag{11}$$

$$\hat{y}_h = D_h(\hat{z}_{hma}, STA(z_{msf}, \hat{z}_d)). \tag{12}$$

4 Experiment

4.1 Evaluation Metrics

We use objective metrics, subjective metrics and user study to evaluate different methods. Objective metrics include PSNR, SSIM and HDRVDP3, which respectively evaluate the fidelity, structural similarity, color fidelity, and visual similarity of the converted HDRTV. Subjective metrics include LPHPS(Extended from LPIPS), FHAD(Extended from FID), and NHQE(Extended from NIQE), which can evaluate the perceptual similarity and distribution consistency of the converted HDRTV and the real HDRTV. These three subjective

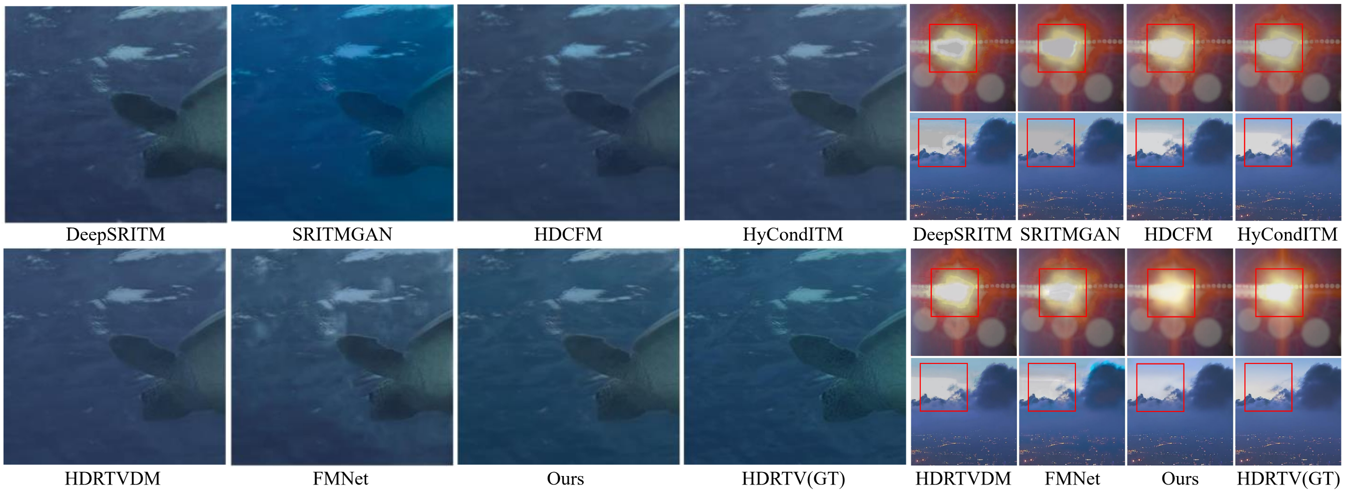

Figure 3: Qualitative results on synthetic datasets. Our RealHDRTVNet can recover realistic HDRTV color information through embedded real-world HDRTV priors. **(Zoom in for details)**

Methods	Test on HDRTV1K Dataset					Test on HDRTV4K Dataset					Test on SRITM Dataset				
	PSNR	SSIM	LPHPS	NHQE	FHAD	PSNR	SSIM	LPHPS	NHQE	FHAD	PSNR	SSIM	LPHPS	NHQE	FHAD
IRSDE	22.64	0.8683	0.3189	4.35	121.41	16.71	0.7663	0.3321	4.45	119.10	22.07	0.8721	0.2789	4.53	113.31
DeepSRITM	29.79	0.8945	0.3260	4.09	131.56	23.61	0.6925	0.4373	4.42	143.58	27.8	0.8767	0.3423	4.28	129.34
FMNet	34.18	0.9527	0.1812	3.94	103.78	27.43	0.8414	0.3077	5.95	124.06	29.26	0.9042	0.221	3.99	96.27
HDCFM	32.42	0.9414	0.1714	3.92	104.82	25.41	0.8404	0.2541	4.49	112.20	28.44	0.871	0.2157	**3.83**	96.22
HDRTVDM	34.09	0.9268	0.2189	4.22	95.73	27.47	0.8792	0.2286	5.38	102.27	31.07	0.9138	0.2102	4.14	85.73
HDRTVNet	**36.01**	0.9559	0.1593	3.98	100.42	28.06	0.8368	0.2836	5.64	121.20	29.47	0.8747	0.2327	3.94	92.84
HyConDITM	35.61	0.9566	0.1288	3.91	95.33	29.31	0.8702	0.2064	5.57	102.61	**31.16**	0.9176	0.1555	3.96	84.24
Ours	35.06	**0.9609**	**0.1166**	**3.88**	**91.03**	**30.31**	**0.8912**	**0.1804**	**4.03**	**96.15**	29.65	**0.9308**	**0.1392**	3.94	**81.56**

Table 1: Quantitative results on synthetic three SDRTV-to-HDRTV datasets.

metrics are obtained by expanding the previous subjective quality assessment methods in this paper.

4.2 Implementation Details

We validate our method's efficacy by training and testing on various datasets detailed in Section 2.1, and additionally evaluating real SDRTV datasets. Our approach employs the Adam optimizer with an initial learning rate of 2×10^{-5}, utilizing a cosine annealing schedule for learning rate decay. Training is divided into three phases, which means HDRTV-VQGAN, E_{sfm}, and RealHDRTVNet are trained respectively.

4.3 Comparisons with State-of-the-Art Methods

We compared our proposed RealHDRTVNet with state-of-the-art methods, including HDRTVNetChen *et al.* [2021b], HyCondITMShao *et al.* [2022], HDCFMHe *et al.* [2022], HDRTVDMGuo *et al.* [2023], FMNetXu *et al.* [2022], and DEEPSRITMKim *et al.* [2019]. Extensive experiments were conducted on both synthetic and real-world datasets.

Experimental Results on Synthetic Datasets

We first show quantitative results on three synthetic datasets in Table 1. On quality metrics PSNR, SSIM, LPHPS, FHAD, and NHQE, our RealHDRTVNet achieves the best scores than existing methods. This means that our method achieves the best performance and produces HDRTV results of higher subjective quality.

Our method achieves a lower value on the LPHPS metric, indicating that the perceptual difference between the proposed method and the real HDRTV is smaller, making it closer to the actual HDRTV. Additionally, the reduced FHAD and NHQE metrics suggest that the HDRTV produced by our method better matches the distribution of HDRTV captured from real-world scenes.

Furthermore, we performed a qualitative comparison in Fig.3. Previous methods failed to produce satisfactory conversion results, such as HDRTVDM and FMNet. Our method can convert HDRTV with more realistic and natural highlights and is particularly effective in reinstating natural and continuous color gradients in areas experiencing color transitions. This capability ensures that the images not only exhibit greater visual fidelity but also reflect a smoother and more authentic representation of real-world colors and shading, enhancing the overall viewing experience.

Visual Perceptual Quality Assessment via HDRVDP3 metrics

Table 3 presents a comparative performance analysis of various methods on the HDRTV1K, HDRTV4K, and SRITM

Datasets	BSD100		CBSD68-Noisy		CBSD68-Original		Set14		Set5	
Metrics	NHQE ↓	FHAD ↓	NHQE ↓	FHAD ↓	NHQE ↓	FHAD ↓	NHQE ↓	FHAD ↓	NHQE ↓	FHAD ↓
DEEPSRITM	3.73	147.06	3.72	144.58	3.67	148.92	3.68	176.66	3.80	193.45
FMNET	3.94	147.33	4.53	148.84	3.95	149.39	4.11	181.84	3.76	203.90
HDCFM	3.81	145.67	3.73	147.43	3.80	146.81	3.70	175.99	3.79	198.89
HDRTVDM	3.76	144.27	3.75	144.89	3.78	145.50	3.67	178.58	3.75	198.48
HDRTVNET	3.84	150.68	3.82	146.08	3.82	151.66	3.72	177.90	3.82	199.08
HyCondITM	3.74	141.63	3.68	142.72	3.73	144.49	3.66	174.74	3.85	198.41
Ours	**3.61**	**141.43**	**3.61**	**142.40**	**3.60**	**144.08**	**3.57**	**174.41**	**3.71**	**192.61**

Table 2: Quantitative results of Non-Reference metrics (FHAD and NHQE) on five real-world SDR datasets: BSD100 Martin *et al.* [2001], CBSD68-Noisy Martin *et al.* [2001], CBSD68-Original Martin *et al.* [2001], Set14 Zeyde *et al.* [2012], and Set5 Bevilacqua *et al.* [2012].

	HDRTV1K	HDRTV4K	SRITM
IRSDE	6.45	5.60	6.58
HDCFM	7.01	6.64	6.43
HDRTVDM	7.77	6.68	7.53
HDRTVNet	7.64	6.27	6.60
HyConDITM	8.22	7.58	7.71
Ours	8.28	7.91	7.80

Table 3: Results on HDRVDP3 metric(HDR Perceptual Quality).

datasets using the HDRVDP3 metric.

Our method achieves the highest HDRVDP3 scores on the HDRTV1K (8.28), HDRTV4K (7.91) and SRITM(7.8) datasets, indicating superior perceived quality of the HDRTV content.

Experimental Results on Real Datasets

To verify the generalization of our method in the real world, we evaluate it on multiple real-world datasets, including BSD100 Martin *et al.* [2001], CBSD68-Noisy Martin *et al.* [2001], CBSD68-Original Martin *et al.* [2001], Set5 Bevilacqua *et al.* [2012], and Set14 Zeyde *et al.* [2012]. As shown in Table 2, our RealHDRTVNet achieves the best FHAD score as well as NHQE perceptual quality on real-world SDR datasets.

E_{sfm}	HCA	STA	PSNR↑	LPHPS ↓	FHAD ↓
✓			34.92	0.1186	91.62
✓	✓		34.95	0.1183	91.58
	✓	✓	35.01	0.1173	91.53
✓	✓	✓	35.06	0.1166	91.03

Table 4: Ablation study results for feature modulation encoder E_{sfm}, HDR Color Alignment module HCA and the SDR Texture Alignment module STA.

4.4 Ablation Study

To assess the impact of the proposed components, we start with a baseline and systematically integrate each component one by one. In Table 4, we can observe that each component has brought improvement, and the improvement in the PSNR metric reached 0.14. Meanwhile, our visual ablation results, presented in Fig. 4, demonstrate that the modules we designed significantly enhance visual quality.

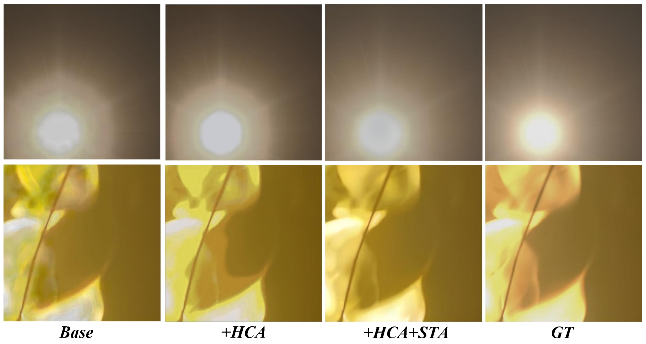

Figure 4: Visual ablation. The HDR Color Alignment HCA module and the SDR Texture Alignment STA module are added.

Importance of SDRTV Feature Modulation Encoder E_{sfm}

We first study the effectiveness of feature modulation encoders. As shown in the third line of Table 4, deleting the modulation module in the encoder will cause the LPIPS and PSNR metric to deteriorate.

Effectiveness of HDR Color Alignment HCA and SDR Texture Alignment STA

We ablate the brightness prior extraction module HCA and the SDR Texture Alignment module STA. As can be shown in Table 4, after adding HCA and module STA in sequence, the performance of our model improved.

5 Conclusion

In this work, we introduce a novel paradigm for SDRTV-to-HDRTV conversion: HDRTV prior-guided high-quality SDRTV-to-HDRTV transformation. In contrast to traditional approaches that solely rely on SDRTV, our method achieves a more realistic and superior quality in HDRTV reconstruction. Additionally, we extend commonly used subjective quality evaluation metrics in SDRTV, such as FHAD, NHQE, and LPHPS, to assess the quality of HDRTV. Our proposed technique exhibits significant improvements in visual quality.

Acknowledgements

This paper is supported by the construction project of the Fujiang Laboratory Nuclear Medicine Artificial Intelligence Research Center (No. 2023ZYDF074) , supported by the Fundamental Research Funds for the Central Universities under Grant ZYTS25270. **Gang He and Kepeng Xu** contributed equally to this work. Please ask Dr. Kepeng Xu(kepengxu11@gmail.com) for correspondence.

References

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie-Line Alberi Morel. Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference (BMVC)*, Guildford, Surrey, United Kingdom, September 2012.

Gaofeng Cao, Fei Zhou, Han Yan, Anjie Wang, and Leidong Fan. Kpn-mfi: A kernel prediction network with multi-frame interaction for video inverse tone mapping. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 806–812. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

Prashant Chaudhari, Franziska Schirrmacher, Andreas Maier, Christian Riess, and Thomas Köhler. Merging-ISP: Multi-exposure high dynamic range image signal processing, 2019. arXiv preprint arXiv:1911.04762.

Jie Chen, Zaifeng Yang, Tsz Nam Chan, Hui Li, Junhui Hou, and Lap-Pui Chau. Attention-guided progressive neural texture fusion for high dynamic range image restoration, 2021. arXiv preprint arXiv:2107.06211.

Xiangyu Chen, Zhengwen Zhang, Jimmy S. Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from SDRTV to HDRTV. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4500–4509. IEEE, October 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

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.

Cheng Guo, Leidong Fan, Ziyu Xue, and Xiuhua Jiang. Learning a practical SDR-to-HDRTV up-conversion using new dataset and degradation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22231–22241. IEEE, June 2023.

Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai. Sdrtv-to-hdrtv via hierarchical dynamic context feature mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 2890–2898, New York, NY, USA, 2022. Association for Computing Machinery.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 6626–6637, 2017.

Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(4-5):507–517, 2013.

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/CVF International Conference on Computer Vision (ICCV)*, pages 3116–3125. IEEE, 2019.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11294–11301. AAAI Press, 2020.

Rafael Pacheco Kovaleski and Manuel Menezes de Oliveira Neto. High-quality reverse tone mapping for a wide range of exposures. In *Proceedings of the 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE, 2014.

Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-image HDR reconstruction by multi-exposure generation, 2022. arXiv preprint arXiv:2210.15897.

Siwei Ma, Junlong Gao, Ruofan Wang, Jianhui Chang, Qi Mao, Zhimeng Huang, and Chuanmin Jia. Overview of intelligent video coding: from model-based to learning-based approaches. *Visual Intelligence*, 1(1):15, 2023.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423 vol.2, 2001.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Tong Shao, Deming Zhai, Junjun Jiang, and Xianming Liu. Hybrid conditional deep inverse tone mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 1016–1024, New York, NY, USA, 2022. Association for Computing Machinery.

Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2747–2757, 2020.

Lin Wang and Kuk-Jin Yoon. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 8874–8895, Dec 2022.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

Tao Wang, Lingbo Xu, Jie Yang, Shengfeng Zhang, Enhua Wang, Xianhui Lin, and Zheng-Jun Zha. Towards real-world blind face restoration with generative facial prior. *arXiv preprint arXiv:2101.04061*, 2021.

Ke Xu, Qin Wang, Huangqing Xiao, and Kelin Liu. Multi-exposure image fusion algorithm based on improved weight function. *Frontiers in Computer Science*, 3:646339, 2021.

Gang Xu, Qibin Hou, Le Zhang, and Ming-Ming Cheng. Fmnet: Frequency-aware modulation network for sdr-to-hdr translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6425–6435, New York, NY, USA, 2022. Association for Computing Machinery.

Kepeng Xu, Gang He, Li Xu, Xingchao Yang, Ming Sun, Yuzhi Wang, Zijia Ma, Haoqiang Fan, and Xing Wen. Towards robust sdrtv-to-hdrtv via dual inverse degradation network. *arXiv preprint arXiv:2307.03394*, 2023.

H. Zeng, X. Zhang, Z. Yu, and Y. Wang. Sr-itm-gan: Learning 4k uhd hdr with a generative adversarial network. *IEEE Access*, 8:182815–182827, 2020.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker, editors, *Curves and Surfaces*, pages 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595. IEEE, 2018.