2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3315-4364-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/CVPR52734.2025.00686

High Dynamic Range Video Compression: A Large-Scale Benchmark Dataset and A Learned Bit-depth Scalable Compression Algorithm

Zhaoyi Tian, Feifeng Wang*, Shiwei Wang, Zihao Zhou, Yao Zhu, Liquan Shen†
Shangha University, China
{kinda,wff0520,ieemia,yi_yuan,yaozhu,jsslq}@shu.edu.cn

Abstract

Recently, learned video compression (LVC) is undergoing a period of rapid development. However, due to absence of large and high-quality high dynamic range (HDR) video training data, LVC on HDR video is still unexplored. In this paper, we are the first to collect a large-scale HDR video benchmark dataset, named HDRVD2K, featuring huge quantity, diverse scenes and multiple motion types. HDRVD2K fills gaps of video training data and facilitate the development of LVC on HDR videos. Based on HDRVD2K, we further propose the first learned bit-depth scalable video compression (LBSVC) network for HDR videos by effectively exploiting bit-depth redundancy between videos of multiple dynamic ranges. To achieve this, we first propose a compression-friendly bit-depth enhancement module (BEM) to effectively predict original HDR videos based on compressed tone-mapped low dynamic range (LDR) videos and dynamic range prior, instead of reducing redundancy only through spatio-temporal predictions. Our method greatly improves the reconstruction quality and compression performance on HDR videos. Extensive experiments demonstrate the effectiveness of HDRVD2K on learned HDR video compression and great compression performance of our proposed LBSVC network. Code and dataset will be released in https://github.com/sdkinda/HDR-Learned-Video-Coding.

1. Introduction

In the past decade, high dynamic range (HDR) video capturing techniques and display devices have made remarkable progress and the demand for HDR video technology is increasing in various domains, such as photography, medical imaging, gaming. HDR videos can produce more realistic scenes on HDR displays with full visible light range compared with low dynamic range (LDR) videos [48]. However, LDR videos typically employ a 24-bit (three 8-bit integer) per pixel encoding format, instead HDR videos represent natural scenes as floating point values. Common uncompressed floating point HDR content formats include .hdr [56], .exr [9] and .tiff [27], in which .exr format [9] even require total 48 bits for three color channels in one pixel. More bits contributes to HDR videos inevitably occupy huge transmission bandwidth and storage space, which hinders its widespread application.

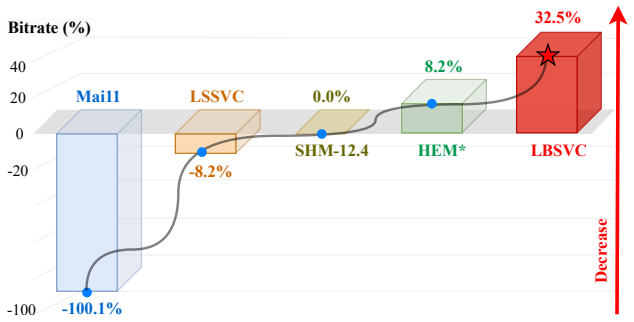


Figure 1. BD-Rate (PU-SSIM) comparisons with Mai11 [38], LSSVC [7], SHM [11], and HEM* [31]. The test dataset is HDM dataset [16]. Our proposed LBSVC has a large performance drop under this setting.

In order to compress HDR video efficiently without compromising perceptual quality, several traditional HDR video compression algorithms have been proposed. They can be classified into two categories 1) single-layer perception-based algorithms and 2) two-layer scalable algorithms. **Single-layer perception-based algorithms** [15, 17, 34, 35, 40, 46, 59, 61] typically apply a perceptual quantization transformation to map HDR data to the maximum bit depth supported by the encoder and then compress the input according to the selected traditional codec. For compatibility with both LDR and HDR monitors, **two-layer scalable algorithms** [25, 26, 28, 29, 38, 39, 41, 45, 50, 57, 58], also named backwards compatible algorithms, compress 8-bit

*Equal contribution

†Corresponding author. This work was partially supported by China NSFC grant (no.62271301 and no.61931022), Shanghai SEALP grant (23XD1401400), Shanghai SSTP grant (22511105200).

2575-7075/25/$31.00 ©2025 IEEE
DOI 10.1109/CVPR52734.2025.00686

tone-mapped LDR video and then reconstruct 16-bit floating point HDR video based on compressed LDR video.

However, existing traditional HDR video compression methods rely heavily on the statistical properties of videos to optimize each hand-crafted modules, whose development has been limited. Therefore, it is necessary to explore further improvements with a whole compression system optimization. Recently, researchers [32, 33, 36, 52] have been exploring learned video coding (LVC) based on deep neural networks to compress videos in a end-to-end way, which have achieved outstanding results and surpassed traditional codecs. Consequently, it is desirable to further improve HDR video compression performance with LVC methods by jointly optimizing the whole compression system.

To promote LVC methods on HDR videos, **the first issue** to be addressed urgently is the absence of training data. As far, there are some publicly accessible HDR video datasets [6, 16, 37, 54]. These datasets are proposed to evaluate performance of HDR video compression or quality assessment algorithms, whose quantity, numbers of distinct scenes and motion patterns are limited. If LVC methods are trained on these methods, problems including insufficient model generalisation ability and lacking of diversity are inevitable. Hence, it is urgent for a large-scale HDR video dataset to promote LVC on HDR videos. Besides the datasets, **another key issue** is the inapplicability of existing LVC methods on HDR videos. Existing LVC methods [32, 33, 36, 52] typically optimize their model on LDR videos and fail to consider the unique dynamic range characteristic of HDR videos, which achieve sub-optimal compression performance on HDR videos. Thus, novel LVC framework specifically designed for HDR videos are urgently needed.

Based on the above analysis, to facilitate the development of LVC on HDR videos, we build the first large-scale HDR video dataset, named HDRVD2K. To achieve high quality HDR video clips as training data, we collect 500 high-quality HDR videos and extract 2200 diverse video clips from these HDR videos. To verify the realism and dynamic range of these clips, careful subjective experiments on professional HDR display device are conducted. Meanwhile, we consider that selected clips should contain rich scenarios and different motion types, which are indispensable in video training datasets, so sufficient data analysis is conducted to verify the enrichment of HDRVD2K.

Based on HDRVD2K, we further propose the first learned bit-depth scalable video compression network (LBSVC) for HDR videos. In LBSVC, a dynamic range prior guided bit-depth enhancement module (BEM) is proposed to effectively predict HDR content based on compressed LDR videos and dynamic range prior. Specifically, BEM extracts dynamic range prior information from HDR videos and compresses them losslessly with few bits. Then BEM utilizes dynamic range prior to enhance the dynamic range of compressed tone-mapped LDR video and predicts higher bit-depth HDR video. Hence, reconstruction quality of HDR video can be better and code stream to represent HDR video can be smaller. Experimental results show that our LBSVC can achieve 32.5% bitrate saving over traditional scalable codec SHM [11] in PU-SSIM metric [5], as shown in Fig. 1. In summary, our contributions are as follows:

- We are the first to propose a large-scale HDR video dataset named HDRVD2K, whose main features contain huge quantity, diverse scenes and multiple motion types. And proposed dataset fills the gap of training data on learned HDR video compression methods.
- We propose a LBSVC network for HDR videos, in which a BEM module is designed to effectively predict HDR video content based on compressed LDR video with the guidance of dynamic range prior, greatly improving the reconstruction quality of compressed HDR videos.
- Extensive analysis and experiments demonstrate the superiority of our dataset and our method. Our work can become a new platform for researchers to explore LVC methods on HDR videos.

2. Related Work

2.1. Traditional HDR Video Compression

Single-layer perception based method Considering input data is typically integer in digital image and video processing, high bit-depth floating-point data of HDR data should be transformed into low bit-depth integer data. Mantiuk et al. [40] first propose a perception quantization scheme, which only requires only 10-11 bits to encode 12 orders of magnitude of visible luminance range, and then compresses HDR sequence with perception quantization transformation in MPEG-4. After that, a prominent perceptual quantizer (PQ) transfer function is presented in [44], which has the best perceptual uniformity in the considered luminance range [47] and has been selected as Anchor for HDR/WCG video coding [15, 37]. After that, some works to further promote single-layer perception based HDR video compression by optimizing PQ transfer function [34, 35, 59, 61]. However, these methods fail to be compatible with LDR display device, limiting their spread to a certain extent.

Two-layer scalable method In order to be compatible with LDR technology and display devices, several backwards-compatible methods are proposed, which can compress both HDR content and its tone-mapped LDR version. They first utilize tone mapping operator to map linear floating HDR content down to LDR bit-depths. Then tone-mapped LDR contents are compressed with a legacy encoder, referred to base layer (BL). Subsequently, these algo-

Table 1. Metrics to assess the diversity of different video datasets.

Metrics on the extent of HDR	
FHLP	Fraction of HighLight Pixel: defined in [18]
EHL	Extent of HighLight: defined in [18]
Metrics on the overall-style	
ALL	Average Luminance Level: defined in [18]
DR	Dynamic Range [20]: calculated as the log10 differences between the highest 2% luminance and the lowest 2% luminance.
Metrics on intra-frame diversity	
SI	Spatial Information: defined in [1]
CF	Colorfulness: defined in [19]
stdL	standard deviation of Luminance: defined in [18]
Metrics on the temporal motion	
TI	Temporal Information: defined in [1]

rithms reconstruct HDR content (referred to enhancement layer, EL) with the information of BL. Mantiuk et al. [41] propose a backward compatible HDR video compression method, which utilizes MPEG-4 to compress tone-mapped LDR version of HDR sequence and the residual between compressed tone-mapped LDR version and original HDR sequence. Subsequently, some researches focus on optimizing inter-layer prediction module to reduce the redundancy between compressed tone-mapped LDR and HDR video and improve reconstruction quality of HDR videos. However, these methods typically rely on the statistical properties of videos to optimize each hand-crafted modules and it is necessary to explore further improvements with a whole system optimization.

2.2. Learned Video Compression

Recently, LVC has garnered significant interest from researchers, which optimize the whole compression sysytem in a end-to-end way. Existing single-layer video compression methods [21, 22, 24, 30–33, 36, 52, 60] have seen rapid advancements in compression performance, even surpassing traditional video compression algorithms like H.266/VVC [10]. However, these methods focus on optimizing the compression efficiency of a single video signal, which is inadequate for meeting the demands of HDR video playback across different display devices. Scalable video coding schemes, on the other hand, offer greater flexibility and adaptability, supporting multiple resolutions [7] and tasks [12, 13, 23] within their frameworks. This paper falls into this category and introduces the first learned bit-depth scalable scheme for HDR videos, effectively eliminating content redundancy between videos of varying bit depths (8-bit tone-mapped LDR and 16-bit HDR video).

2.3. HDR video dataset

To evaluate various HDR image and video quality metrics, Banitalebi-Dehkordi et al. [6] propose a DML-HDR dataset. Froehlich et al. [16] present a cinematic wide gamut HDR-video test set named HDM, designed for the evaluation of temporal tone mapping operators and HDR-displays. Luthra el al. [37] propose a HDR video dataset as common test condition for HDR video compression experiments and assessment. Considering the majority of available datasets focusing HDR content are HD resolutions Song et al. [54] propose a UHD HDR video dataset named SJTU. However, these datasets are typically presented for testing and evaluation of HDR video compression methods, whose quantities and scenes are insufficient for LVC network training.

3. Proposed Dataset

Construction of Our Dataset To favor the development of blooming LVC methods on HDR videos, we construct a large-scale HDR video dataset, named HDRVD2K. Actually, it is extremely challenging to collect large-scale and useful HDR video sequences, which is time-consuming and laborintensive. Former researchers have explained that the process of HDR video shooting can be very hard, resulting in HDR video datasets that are small in size and limited in terms of content. Alternatively, we resort to video platforms where many videos are taken with professional cameras.

Nowadays, there exists many high-quality HDR videos in website and we collect 500 videos in a professional HDR video format, which is SMPTE ST 2086, HDR10 compatible and whose color primaries are mainly BT.2020. In order to save storage space and accelerate the process of video editing, these videos are downloaded in a 3840×2160 resolution with 60 fps, whose source resolution can be 4K or even 8K. After that, we utilize professional video editing software DaVinci Resolve Studio to edit downloaded HDR videos and generate useful HDR video clips. Specifically, during the 'EDIT' stage, we manually select clips whose duration are about 3 seconds and ensure the scenes and motions are distinctive in a video. Then we set the output gamma in color space transformation to Linear during the 'COLOR' stage in DaVinci Resolve Studio. After that, during the 'DELIVER' stage, output format of each clip is selected 'EXR' [9] and the resolution is set to 1920×1080, which aims to reduce the storage space of dataset. Hence, thousands of video clips are generated and each clip contains 180 frames.

To verify the dynamic range and realism of generated HDR video clips, some subjective tests are conducted. Following the ITU-R BT.500-13 Recommendation [51], the subjective tests have been conducted in a dark, quiet room, with the ambient illumination of the room at 2.154 lux and the luminance of the screen when turned off at 0.03

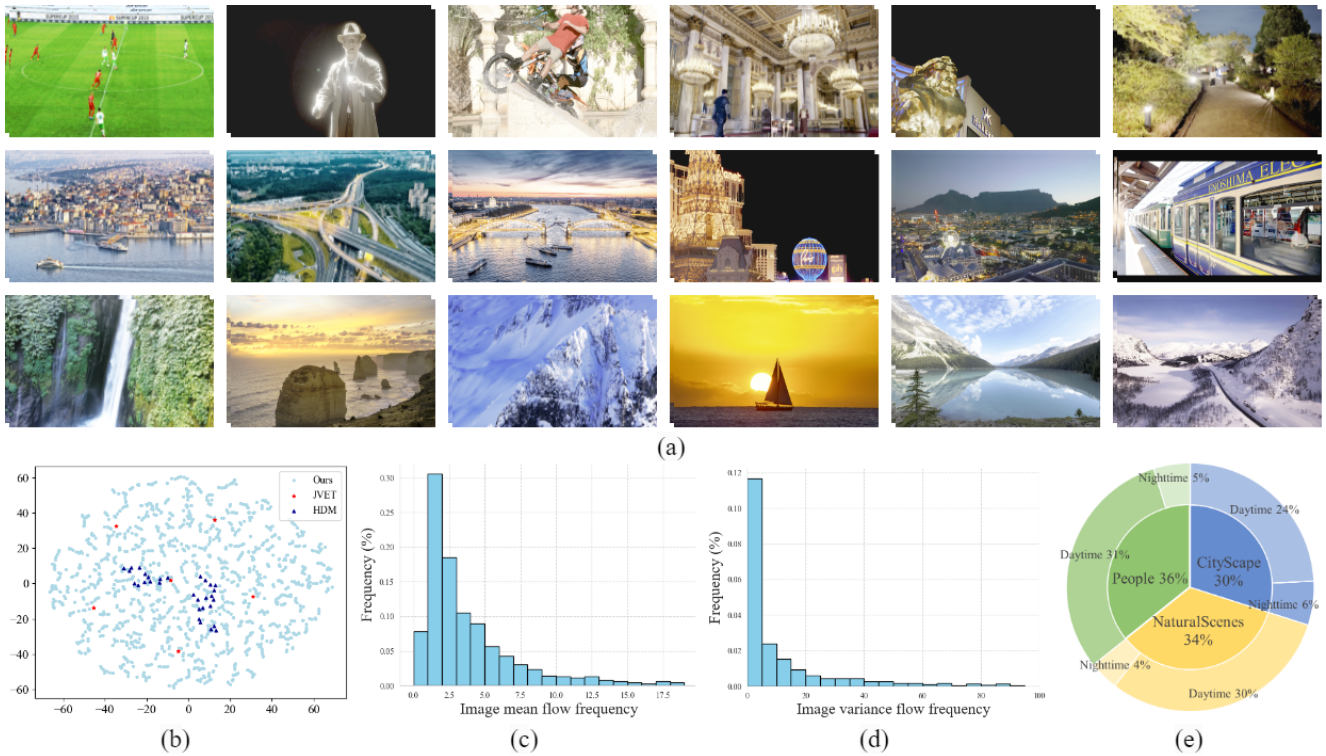


Figure 2. (a) Sampled video clips from HDRVD2K, which are all tome-mapped for visualization. (b) Diversity comparison among our dataset, JVET [37] and HDM [16]. (c) The histogram of flow mean magnitude of all pixels in the dataset. (d) The histogram of flow variance magnitude of all pixels in the dataset. (e) Scenes and illumination type percentage of HDRVD2K.

Table 2. Statistics of different datasets.

Metric / Dataset	Extent of HDR		Intra-frame Diversity			Overall-style		Motion
	FHLP↑	EHL↑	SI↑	CF↑	stdL↑	ALL↑	DR↑	TI↑
JVET	1.10	8.15	52.09	16.96	6.24	2.73	**4.55**	11.24
HDM	0.56	4.69	12.94	5.14	1.81	0.97	2.82	11.57
HDRVD2K	**9.08**	**29.7**	**60.17**	**25.89**	**16.09**	**12.32**	4.30	**14.58**

cd/m^2. The stimuli were presented on a calibrated HDR SIM2 HDR47ES4MB 47″ display [14] with 1920×1080 resolution, peak brightness of 4000 cd/m^2, used in its native HDR mode. The distance from the screen was fixed to three heights of the display, with the eyes of observers positioned zero degrees horizontally and vertically from the center of the display. Repeat and low-quality clips are discard during these tests. This way, the realism and dynamic range of generated video clips can be guaranteed.

As a result, we collect a new HDR video dataset, consisting of 500 videos with 2200 independent clips that are different from each other in content. To standard the input for training and testing, the resolution of all frames are fixed to 1920×1080 and each clip randomly select 15 successive frames from generated 180 frames. Fig. 2 (a) show some clips of our HDRVD2K, in which we can multiple contents including people, cityscapes and natural sceneries. For different scenes, we can observe scenes at different times including daytime and night time, whose distribution can be found in Fig. 2 (e).

Analysis of Our Dataset To quantitatively evaluate the superiority of our dataset, we analyze the diversity of the HDM dataset [16], JVET dataset [37] and our dataset. Following [18, 53], we utilize 8 metrics to assess the diversity of different HDR video datasets from the dimensions of intra-frame diversity, extent, overall style and temporal motion. For each video clip, 8 different metrics are calculated according to Table 1. Then we utilize the t-SNE [55] to project 8-D vector from Table 1 of each video clip to the corresponding 2D-coordinate for plotting the dataset distribution of our dataset and comparison datasets.

As shown in Fig. 2 (b), our dataset contains wider frame distribution than JVET and HDM dataset, indicating that

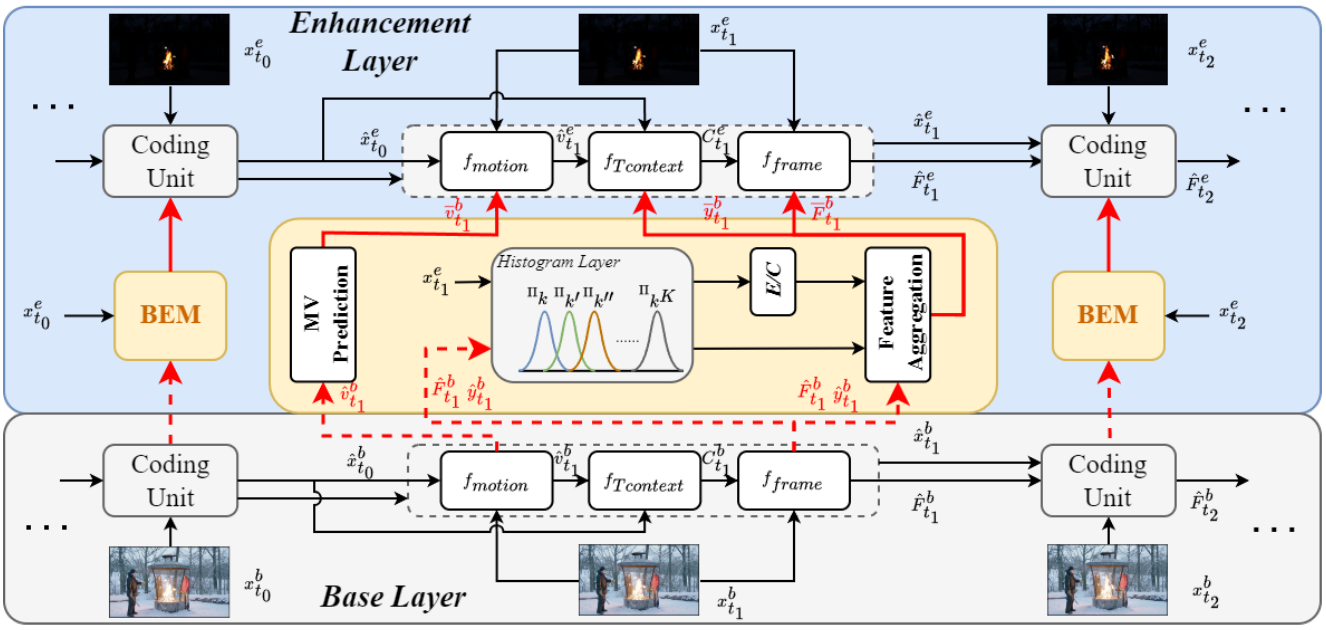


Figure 3. **Overview of our proposed LBSVC framework for HDR videos.** f_{motion}, $f_{Tcontext}$ and f_{frame} in coding unit denote motion information compression, contextual information extraction and contextual information compression individually. $x^e_{t_i}$ and $x^b_{t_i}$ indicate the i-th HDR frame and its tone-mapped LDR frame, which is the input of EL and BL. Code information in BL contains decoded LDR frame $\hat{x}^b_{t_i}$, decoded motion vector $\hat{v}^b_{t_i}$, decoded features $F^b_{t_i}$, decoded latent representation $\hat{y}^b_{t_i}$ and contextual information $C^b_{t_i}$. Code information in EL is similar with BL, with some external reference information from BEM module. E/C denotes dynamic range prior extraction and compression.

the networks trained with our dataset can be better generalized to different scenarios. Furthermore, we present the flow mean and variance magnitude in Fig. 1 (c)(d), which demonstrate the motion diversity of our dataset. Detailed statistics of different datasets are shown in Table 1 and we find our HDRVD2K outperforms other datasets on different assessment metrics. The diversity in both scenes and motion patterns makes that HDRVD2K can be used for training LVC methods and assessing the generalization capability of the networks across different scenes. More details about the dataset can be found in supplementary materials.

4. Proposed Method

Overview The coding structure of our proposed method is illustrated in Fig. 3, whose input contains original 16-bit floating HDR video data X^e=$\{x^e_t|t$ =1,...,n$\}$ and its tone-mapped LDR data X^b=$\{x^b_t|t$ =1,...,n$\}$. Firstly, LDR frames will be compressed in BL, which we use in LBSVC is a LVC method named DCVC-HEM [31]. And in EL, original HDR frames will be compressed based on BL information and a bit-depth enhancement module (BEM). BEM utilizes dynamic range prior and code information from BL to predict HDR information effectively, which can improve the reconstruction quality of HDR video. More details of proposed LBSVC and specific framework can be found in supplementary materials.

Base Layer Coding BL coding pipeline contains three core steps: f_{motion}, $f_{Tcontext}$ and f_{frame}. f_{motion} uses optical flow network to estimate the motion vector (MV) v^b_t, then v^b_t is encoded and decoded as $\hat{v}^b_t$. Based on $\hat{v}^b_t$ and the propagated feature $\hat{F}^b_{t-1}$ from the previous frame, $f_{Tcontext}$ extracts the motion-aligned temporal context feature C^b_t. Finally, f_{frame} encodes x^b_t into quantized latent representation $\hat{y}^b_t$ and the output frame $\hat{x}^b_t$ is reconstructed via the decoder and frame generator after entropy coding. So far, tone-mapped LDR video is compressed with BL and then EL will compress original HDR video based on compressed LDR video.

Enhancement Layer Coding During the BL coding stage, tone-mapped LDR video frames are compressed and some code information are retained. Then EL compresses original HDR frame $x^e_{t_0}$ based on these code information and temporal previous compressed HDR frame information. Specifically, as shown in Fig. 4 in EL, a motion estimation module is used to generate motion vectors information $v^e_{t_0}$ between $x^e_{t_0}$ and previous HDR frame $x^e_{t_0-1}$. Then a motion compression module compresses $v^e_{t_0}$ into bits based on $\overline{v}^b_t$ and decompress the bits into compressed motion vector information $\hat{v}^e_{t_0}$. After that, a hybrid temporal-layer context mining module in [7] is utilized to mine reconstructed temporal information ($\hat{v}^e_{t_0}$) and enhanced texture information ($\overline{F}^b_t$) as soon as possible to produce context hy-

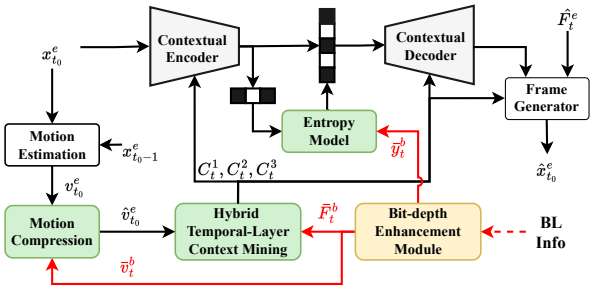


Figure 4. Architecture of EL coding unit.

brid C_t^1, C_t^2, C_t^3. The hybrid contexts C_t^l $(l = 1, 2, 3)$ are then refilled into contextual encoder/decoder for the compression of HDR video frames in the EL. In contextual encoder/decoder, $\overline{y}_t^b$ will be feed into entropy model to estimate the latent code y_t^e of HDR frame more effectively. Finally, reconstructed HDR frame $\hat{x}_{t_0}^e$ is generated with the frame generator module. Code information $\overline{v}_t^b$, $\overline{F}_t^b$ and $\overline{y}_t^b$ are the output of BEM module.

Bit-depth Enhancement Module HDR videos and corresponding tone-mapped LDR videos are similar in content. Hence, information of compressed tone-mapped LDR content in BL can predict information of HDR content in EL with few bits instead re-encoding HDR videos and inter-layer prediction is utilized. On the other hand, dynamic range of tone-mapped videos is vastly different from HDR videos. Hence, bit-depth enhancement technique is utilized to transform tone-mapped LDR videos to HDR videos. Traditional scalable compression methods use simple linear scaling [11] or bit-consuming mapping [2] as bit-depth enhancement methods, which are inefficient to predict EL information with few bits. Instead, our BEM can predict EL information by combining BL information and dynamic range prior between HDR and tone-mapped LDR content. And dynamic range prior from HDR video can be represented with few bits. The detailed process of BEM is shown in Fig. 5.

Specifically, BEM first extracts histogram information as dynamic range prior from HDR frame with a differential histogram layer in [3, 20], in which differential threshold functions $t_j(x)$ are utilized for the luminance range slicing and binning of input content.

$$t_j(x) = exp\left(-\frac{(x - c_j)^2}{\sigma_j^2}\right) \tag{1}$$

where c_j is the center of the sliced luminance range, and σ_j represents the length of this range slice $(j = 1..k)$. The BEM aggregates dynamic range prior t_j^e extracted from HDR frames and threshold functions t_j^b extracted from BL information $(\hat{F}_t^b, \hat{y}_t^b)$, producing bit-depth enhanced output $\overline{F}_t^b, \overline{y}_t^b$. Dynamic range and content of output features are similar with corresponding features from original HDR content. Thus, fewer bits can be used to represent HDR contents with the assistance of enhanced features from BEM. Furthermore, dynamic range prior needs to be transmitted to decoder side, and we extract c_j^e and σ_j^e and compress them losslessly with additional few bits. The dynamic range prior compression of c_j^e and σ_j^e only costs 256 float-32 values while the prior can effectively represent the bin center and length of luminance range.

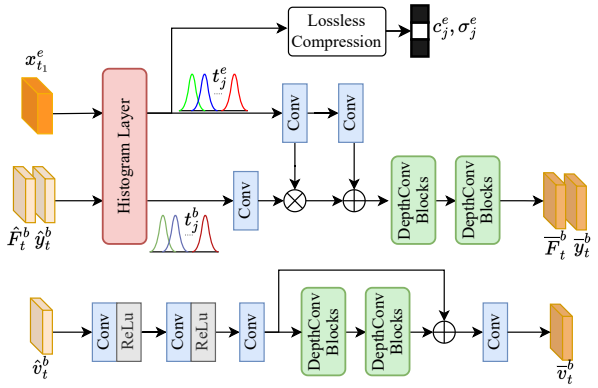


Figure 5. Architecture of the bit-depth enhancement module.

Code information $\hat{F}_t^b$ and $\hat{y}_t^b$ from BL have been enhanced with BEM for reducing the dynamic range difference between BL and EL. For another BL motion information $\hat{v}_t^b$, which represents compressed motion vector in BL, it can be similar with motion vector $\hat{v}_t^e$ in EL because motion in tone-mapped videos can be similar with original HDR videos. We utilize several depth convolution blocks, which have strong feature extraction performance and lightweight parameters, to minimize the representation redundancy between enhanced feature $\overline{v}_t^b$ and $\hat{v}_t^e$.

Loss function In traditional bit-depth scalable video coding of SHM [11], a hierarchical quality structure is applied to different layers, in which layers are assigned different quantization parameters (QPs). For single BL and EL training stage in our method, loss function is similar with previous single-layer methods [31, 52]:

$$L_{single} = \lambda \cdot D\left(x_t^n, \hat{x}_t^n\right) + R_t^n \tag{2}$$

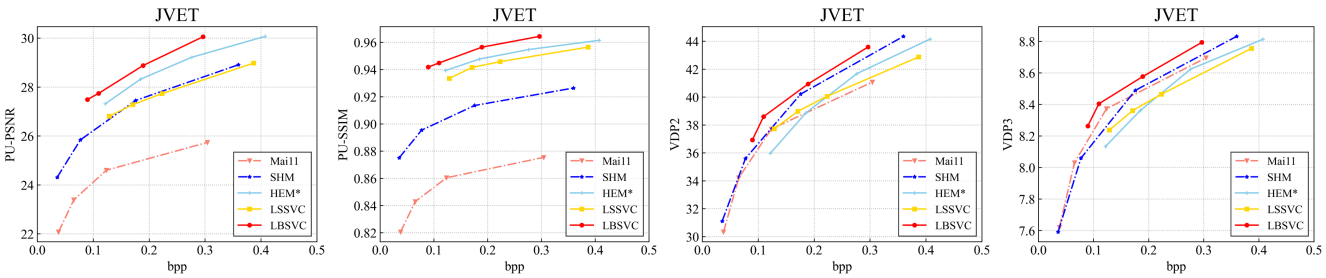
where $D\left(x_t^n, \hat{x}_t^n\right)$ denotes the distortion between original frame x_t^n and reconstructed frame $\hat{x}_t^n$ in the n-th layer. The $D\left(\cdot\right)$ in this paper is the common mean squared error (MSE) distortion. The R_t^n represents the bitrate estimated by the entropy model for encoding both MV and frame in the nth layer. For two-layer joint training stage, the loss function can be as follows:

$$\begin{aligned} L_{joint} = &\left(\omega_b \cdot \lambda_b \cdot D\left(x_t^b, \hat{x}_t^b\right) + R_t^b\right) \\ &+ \left(\lambda_e \cdot D\left(x_t^e, \hat{x}_t^e\right) + R_t^e\right), \end{aligned} \tag{3}$$

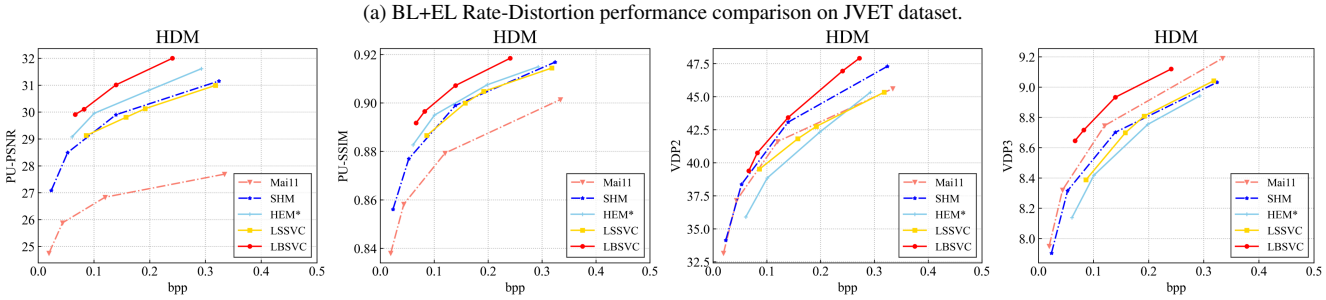
where ω_b we set 0.5 in this paper.

5. Experiments

Datasets For training LVC methods on HDR videos, our HDRVD2K dataset is selected. Original 16-bit floating



(a) BL+EL Rate-Distortion performance comparison on JVET dataset.



(b) BL+EL Rate-Distortion performance comparison on HDM dataset.

Figure 6. Rate and distortion curves for JVET and HDM datasets. Bpp is calculated by the sum of BL and EL. Four quality metrics are all calculated on HDR video frames. The intra period is set to 32. The metrics of HEM* tests on HDR video results and bpp is calculated based on two individual code streams.

Table 3. BD-Rate (%) on four metrics of different methods.

Methods	PU-PSNR		PU-SSIM		VDP2		VDP3	
	JVET	HDM	JVET	HDM	JVET	HDM	JVET	HDM
SHM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mai11	-	-	-	-	14.2	13.7	-6.6	-21.6
HEM*	-30.6	-27.4	-81.6	-9.4	29.1	66.6	28.0	39.2
LSSVC	8.7	14.7	-69.2	8.2	30.4	43.4	31.3	13.4
Ours	**-45.2**	**-49.8**	**-93.9**	**-32.5**	**-9.0**	**-7.9**	**-15.5**	**-44.2**

HDR video clips in dataset are used for EL stage and corresponding tone-mapped LDR version are used for BL stage. HDR and LDR video clips are all cropped into 256×256 patches. During the test stage, to evaluate the compression performance of different methods on HDR videos, we select 6 sequences from JVET [37] and 10 sequences from HDM [16] as our test datasets, whose resolutions are all 1920×1080.

Implementation Details We first train the BL model [31] for different bitrates using λ^b= (85,170,380,840) to compress the tone-mapped LDR videos. After that, we freeze the parameters of BL and train EL model for four different bitrates using λ^e= (85,170,380,840). Finally, the loss function defined in Eq. 3 is utilized to finetune the training on two-layer model jointly. The Adam optimizer is used with default settings. Batch size is range from 1 to 16 with the change of training stage. The learning rate starts at 1×10^{-4} and decays to 1×10^{-5} in the later training stage. Both the BL and the EL models are implemented by Pytorch and trained on 1 Nvidia RTX4090. The training takes about 240 hours to finish.

Settings We follow [7] to encode 96 frames for each sequence in test datasets. Intra period is set to 32. Our comparison algorithms include latest reference software SHM-12.4 of SHVC, a traditional backwards-compatible method Mai11 [38] combined with HEVC test model (HM-16.20), and two LVC methods including a modified LSSVC network [7] for bit-depth scalable video coding and a method HEM* that encodes two layers separately based on [31]. LSSVC is designed for spatial scalable coding and we modify some modules to fulfill the bit-depth scalable coding for HDR videos. In these four comparison algorithms, SHM, Mai11 and modified LSSVC are real bit-depth scalable methods, while HEM* method encode LDR and HDR videos separately for comparison, whose two code streams are unrelated. And follow existing LVC methods, our method and comparison algorithms are all tested on low-delay common test conditions. In SHM codec, we set the EL QP offsets ΔQP = 4 because of its higher compression performance. Stable tone-mapping operator Mai11 is selected to transfer HDR videos to LDR videos for BL coding and 12bit-PQ coding is utilized on 16-bit floating HDR videos as preprocessing for EL coding.

Evaluation Metrics To evaluate the performance of two-

layer HDR video coding, several quality assessment metrics are utilized. HDR-VDP-2 [42], HDR-VDP-3 [43], PU-PSNR and PU-SSIM are used for compressed HDR video results. PU-PSNR and PU-SSIM are computed after perceptually uniform encoding [5]. We abbreviate HDR-VDP-2 and HDR-VDP-3 as VDP2 and VDP3. When computing the VDP2 and VDP2, the diagonal display size is set to 21 inches and viewing distance in meters is set to 1. BD-Rate [8] calculates the bitrate difference between two methods under the same video quality and we utilize the same way to achieve four BD-Rates on above four metrics to evaluate overall compression quality of different methods.

Comparisons with SOTA Methods Backward-compatible HDR video code streams can be decoded and then displayed on both LDR and HDR monitors, thus BL+EL performance is compared in our paper as the main comparison setting. Furthermore, BL performance on tone-mapped HDR videos utilize PSNR and BD-Rate as metrics. In BL+EL performance comparison, the bpp is calculated by the sum of BL and EL, and the distortion is calculated on HDR video frames. More details can be found in supplementary materials.

Fig. 6 presents the two-layer bit-depth scalable video coding performance on JVET and HDM datasets. It can be observed that our scheme outperforms other methods by a large margin on under all testing conditions, which effectively demonstrates the high compression efficiency of our method on HDR videos. For PU-PSNR and PU-SSIM metric, LVC methods achieve great performance because 12-bit PQ encoding transfers HDR content into perceptual uniformly domain [44], which helps leverage the strengths of LVC methods. LVC methods have achieved great compression performance on LDR videos and gamma corrected pixels of LDR videos are considered in perceptual uniformly domain [4]. On the other hand, we can find that PU-PSNR and PU-SSIM metric of Mai11 [38] are very low, because Mai11 utilizes piece-wise function mapping to reconstruct HDR videos based on tone-mapped videos. However, other methods optimize the EL coding on 12-bit PQ coding HDR videos with MSE loss in perceptual uniformly domain, so their PU metrics are higher than Mai11.

LVC methods are no longer have the advantage on VDP2 and VDP3 metric because these metrics aim to evaluate images covering complete range of luminance the human eye can see, which are not the optimization direction of MSE Loss and PQ transfer function used in this paper. Our method utilize the high dynamic range prior in BEM to assist the reconstruction of HDR videos and thus we also achieve considerable performance on VDP2 and VDP3. To numerically visualize the overall rate-distortion performance of different methods, Table 3 is presented. It can be observed that our scheme surpasses all comparison algorithms and achieve the best compression performance under all test conditions, verifying the effectiveness of proposed method on HDR videos.

Table 4. Model complexity of different LVC methods.

Methods	Parameters (M)	Flops (G)
HEM*	35.04	217.1
LSSVC	29.4	246.48
Ours	41.8	257.35

Table 5. Ablation study of BEM with BD-rate (%) on PU-PSNR and PU-SSIM.

BEM	PU-PSNR		PU-SSIM	
	JVET	HDM	JVET	HDM
✓	0.0	0.0	0.0	0.0
✗	8.51	7.49	8.05	7.11

Furthermore, in this paper, LVC methods [7, 31] are directly utilized to compress HDR videos with some modification and achieve considerable performance, which demonstrates the feasibility of LVC on HDR videos and further reflects the significance of proposed HDRVD2K. Our work bridges a platform for researchers to explore LVC methods on HDR videos.

Complexity Complexity of three scalable LVC methods are shown in Table 4. They are all calculated with BL+EL. HEM* are two individual HEM compression model so we double its complexity based on original HEM [31]. We can find the FLOPs and parameters of our LBSVC are similar to other LVC methods. However, our algorithm outperforms them greatly, which demonstrates the superiority of our method.

Ablation Study In our proposed LBSVC scheme, BEM is proposed to help reconstruct HDR videos with BL information and dynamic range prior. As shown in Table 5, for method without BEM, BD-Rate (PU-PSNR/PU-SSIM) will increase 8.51/7.49 and 8.05/7.11 on JVET and HDM dataset, which proves the effectiveness of BEM in HDR video reconstruction.

6. Conclusion

In this paper, we propose a large scale dataset named HDRVD2K and a HDR video compression framework LBSVC. HDRVD2K features huge quantity, diverse scenes and multiple motion types and fills gaps of HDR video training data. Based on HDRVD2K, LBSVC is proposed to compress HDR videos, which effectively exploits bit-depth redundancy between videos of multiple dynamic ranges. To achieve this, a compression-friendly BEM is designed to effectively predict original HDR videos based on compressed tone-mapped LDR videos and dynamic range prior, greatly improving the reconstruction quality and compression performance on HDR videos.

References

[1] ITU-R Recommendation BT. 1788. Methodology for the subjective assessment of video quality in multimedia applications, 2019. 3

[2] Alessandro Artusi, Rafał K Mantiuk, Thomas Richter, Philippe Hanhart, Pavel Korshunov, Massimiliano Agostinelli, Arkady Ten, and Touradj Ebrahimi. Overview and evaluation of the jpeg xt hdr image compression standard. *Journal of Real-Time Image Processing*, 16:413–428, 2019. 6

[3] Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv. Deephist: Differentiable joint and color histogram layers for image-to-image translation. *arXiv preprint arXiv:2005.03995*, 2020. 6

[4] Tunç O. Aydın, Rafal Mantiuk, and Hans-Peter Seidel. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging XIII*, page 68060B. International Society for Optics and Photonics, SPIE, 2008. 8

[5] Maryam Azimi et al. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 2, 8

[6] A Banitalebi-Dehkordi, M Azimi, Y Dong, MT Pourazad, and P Nasiopoulos. Quality assessment of high dynamic range (hdr) video content using existing full-reference metrics. *ISO/IEC JTC1/SC29/WG11, France*, 2014. 2, 3

[7] Yifan Bian, Xihua Sheng, Li Li, and Dong Liu. Lssvc: A learned spatially scalable video coding scheme. *IEEE Trans. Image Process.*, 33:3314–3327, 2024. 1, 3, 5, 7, 8

[8] G Bjøntega. Calculation of average psnr differences between rdcurves. document VCEG-M33, 2001. 8

[9] R Bogart, F Kainz, and D Hess. Openexr image file format. In *ACM SIGGRAPH*, page 28, 2003. 1, 3

[10] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Trans. Circuit Syst. Video Technol.*, 31(10):3736–3764, 2021. 3

[11] J Chen, J Boyce, Y Ye, and MM Hannuksela. Scalable hevc (shvc) test model 10 (shm 10). *JCT-VC of ITU-T SG16 WP 3and ISO/IEX JTC*, 1, 2015. 1, 2, 6

[12] Qiaoxi Chen, Changsheng Gao, and Dong Liu. End-to-end learned scalable multilayer feature compression for machine vision tasks. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 1781–1787, 2024. 3

[13] Hyomin Choi and Ivan V. Bajić. Scalable video coding for humans and machines. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2022. 3

[14] SIM2 Corporation. Sim2 hdr47 monitors, 2016. http://www.sim2.com/. 4

[15] Edouard François, Chad Fogg, Yuwen He, Xiang Li, Ajay Luthra, and Andrew Segall. High dynamic range and wide color gamut video coding in hevc: Status and potential future enhancements. *IEEE Trans. Circuit Syst. Video Technol.*, 26 (1):63–75, 2016. 1, 2

[16] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, pages 279–288. SPIE, 2014. 1, 2, 3, 4, 7

[17] Jens-Uwe Garbas and Herbert Thoma. Temporally coherent luminance-to-luma mapping for high dynamic range video coding with h.264/avc. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 829–832, 2011. 1

[18] Cheng Guo, Leidong Fan, Ziyu Xue, and Xiuhua Jiang. Learning a practical sdr-to-hdrtv up-conversion using new dataset and degradation models. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 22231–22241, 2023. 3, 4

[19] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95. SPIE, 2003. 3

[20] Xiangyu Hu, Liquan Shen, Mingxing Jiang, Ran Ma, and Ping An. La-hdr: Light adaptive hdr reconstruction framework for single ldr image considering varied light conditions. *IEEE Trans. Multimedia*, 25:4814–4829, 2022. 3, 6

[21] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 1502–1511, 2021. 3

[22] Zhihao Hu, Guo Lu, Jinyang Guo, Shan Liu, Wei Jiang, and Dong Xu. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5911–5920, 2022. 3

[23] Wei Jiang, Hyomin Choi, Fabien Racapé, Simon Feltman, and Fatih Kamisli. Face restoration-based scalable quality coding for video conferencing. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 206–211, 2023. 3

[24] Dengchao Jin, Jianjun Lei, Bo Peng, Zhaoqing Pan, Li Li, and Nam Ling. Learned video compression with efficient temporal context learning. *IEEE Trans. Image Process.*, 32: 3188–3198, 2023. 3

[25] Alper Koz and Frederic Dufau. Optimized tone mapping with perceptually uniform luminance values for backward-compatible high dynamic range video compression. In *IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, pages 1–6, 2012. 1

[26] Alper Koz and Frederic Dufaux. Methods for improving the tone mapping for backward compatible high dynamic range image and video coding. *Signal Processing: Image Communication*, 29(2):274–292, 2014. 1

[27] Gregory Ward Larson. Logluv encoding for full-gamut, high-dynamic range images. *Journal of Graphics Tools*, 3 (1):15–31, 1998. 1

[28] Sébastien Lasserre, Fabrice Le Léannec, Tangi Poirier, and Franck Galpin. Backward compatible hdr video compression system. In *2016 Data Compression Conference (DCC)*, pages 309–318, 2016. 1

[29] Mikaël Le Pendu, Christine Guillemot, and Dominique Thoreau. Inter-layer prediction of color in high dynamic range image scalable compression. *IEEE Trans. Image Process.*, 25(8):3585–3596, 2016. 1

[30] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *Adv. Neural Inform. Process. Syst. (NIPS)*, 34: 18114–18125, 2021. 3

[31] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *ACM Int. Conf. Multimedia (ACMMM)*, pages 1503–1511, 2022. 1, 5, 6, 7, 8

[32] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 22616–22626, 2023. 2

[33] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 26099–26108, 2024. 2, 3

[34] Yi Liu, Naty Sidaty, Wassim Hamidouche, Olivier Déforges, Giuseppe Valenzise, and Emin Zerman. An adaptive quantizer for high dynamic range content: Application to video coding. *IEEE Trans. Circuit Syst. Video Technol.*, 29(2):531–545, 2019. 1, 2

[35] Yi Liu, Naty Sidaty, Wassim Hamidouche, Olivier Déforges, and Cheolkon Jung. Visual attention-aware high dynamic range quantization for hevc video coding. *IEEE Trans. Circuit Syst. Video Technol.*, 32(7):4296–4311, 2022. 1, 2

[36] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 10998–11007, 2019. 2, 3

[37] A Luthra, E Francois, and W Husak. Call for evidence (cfe) for hdr and wcg video coding, document n15083, iso. Technical report, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, 2015. 2, 3, 4, 7

[38] Zicong Mai, Hassan Mansour, Rafal Mantiuk, Panos Nasiopoulos, Rabab Ward, and Wolfgang Heidrich. Optimizing a tone curve for backward-compatible high dynamic range image and video compression. *IEEE Trans. Image Process.*, 20(6):1558–1571, 2010. 1, 7, 8

[39] Zicong Mai, Hassan Mansour, Panos Nasiopoulos, and Rabab Kreidieh Ward. Visually favorable tone-mapping with high compression performance in bit-depth scalable video coding. *IEEE Trans. Multimedia*, 15(7):1503–1518, 2013. 1

[40] Rafal Mantiuk, Grzegorz Krawczyk, Karol Myszkowski, and Hans-Peter Seidel. Perception-motivated high dynamic range video encoding. *ACM Trans. Graph.*, 23(3):733–741, 2004. 1, 2

[41] Rafał Mantiuk, Alexander Efremov, Karol Myszkowski, and Hans-Peter Seidel. Backward compatible high dynamic range mpeg video compression. In *ACM SIGGRAPH*, page 713–723, New York, NY, USA, 2006. Association for Computing Machinery. 1, 3

[42] Rafał Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4), 2011. 8

[43] Rafal K Mantiuk, Dounia Hammou, and Param Hanji. Hdr-vdp-3: A multi-metric for predicting image differences, quality and contrast distortions in high dynamic range and regular content. *arXiv preprint arXiv:2304.13625*, 2023. 8

[44] Scott Miller, Mahdi Nezamabadi, and Scott Daly. Perceptual signal coding for more efficient usage of bit codes. *SMPTE Motion Imaging Journal*, 122(4):52–59, 2013. 2, 8

[45] Junaid Mir, Dumidu S. Talagala, Hemantha Kodikara Arachchi, and Anil Fernando. Adaptive residual mapping for an efficient extension layer coding in two-layer hdr video coding. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 1394–1398, 2016. 1

[46] Ajit Motra and Herbert Thoma. An adaptive logluv transform for high dynamic range video compression. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2061–2064, 2010. 1

[47] Charles Poynton and Brian Funt. Perceptual uniformity in digital image representation and display. *Color Research & Application*, 39(1):6–15, 2014. 2

[48] Erik Reinhard, Greg Ward, Sumanta Pattanaik, and Paul Debevec. *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. 1

[49] E Reinhard, G Ward, S Pattanaik, and P Debevec. Scalable hevc (shvc) test model 9 (shm 9), 2010.

[50] Andrew Segall. Scalable coding of high dynamic range video. In *IEEE Int. Conf. Image Process. (ICIP)*, pages I – 1–I – 4, 2007. 1

[51] B Series. Methodology for the subjective assessment of the quality of television pictures. *Recommendation ITU-R BT*, 500(13), 2012. 3

[52] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Trans. Multimedia*, 25:7311–7322, 2023. 2, 3, 6

[53] Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world hdr video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 2879–2888, 2024. 4

[54] Li Song, Yankai Liu, Xiaokang Yang, Guangtao Zhai, Rong Xie, and Wenjun Zhang. The sjtu hdr video sequence dataset. In *Proceedings of International Conference on Quality of Multimedia Experience (QoMEX 2016)*, page 100, 2016. 2, 3

[55] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4

[56] Gregory J Ward. The radiance lighting simulation and rendering system. In *Proc. 21st Annu. Conf. Comput. Graph. Interact. Techn.*, pages 459–472, 1994. 1

[57] Zhe Wei, Changyun Wen, and Zhengguo Li. Local inverse tone mapping for scalable high dynamic range image coding. *IEEE Trans. Circuit Syst. Video Technol.*, 28(2):550–555, 2018. 1

[58] Martin Winken, Detlev Marpe, Heiko Schwarz, and Thomas Wiegand. Bit-depth scalable video coding. In *IEEE Int. Conf. Image Process. (ICIP)*, pages I – 5–I – 8, 2007. 1

[59] Shengtao Yu, Cheolkon Jung, and Peng Ke. Adaptive pq: Adaptive perceptual quantizer for hevc main 10 profile-based hdr video coding. In *IEEE Int. Conf. Vis. Commun. Image Process. (VCIP)*, pages 1–4, 2016. 1, 2

[60] M. Akın Yılmaz and A. Murat Tekalp. End-to-end rate-distortion optimized learned hierarchical bi-directional video compression. *IEEE Trans. Image Process.*, 31:974–983, 2022. 3

[61] Yang Zhang, Matteo Naccari, Dimitris Agrafiotis, Marta Mrak, and David R. Bull. High dynamic range video compression exploiting luminance masking. *IEEE Trans. Circuit Syst. Video Technol.*, 26(5):950–964, 2016. 1, 2