HDRMovieformer: A Transformer Framework and Benchmark for Cinematic SDR-to-HDR Conversion

Xianwei Li, Huiyuan Fu*, Chuanming Wang, Huadong Ma

Beijing University of Posts and Telecommunications, China
{lixianwei,fhy,wcm,mhd}@bupt.edu.cn

Abstract

With the growing prevalence of HDR-capable cinema venues such as Cinity LED theaters, there is an increasing demand to convert existing Standard Dynamic Range (SDR) films into High Dynamic Range (HDR) formats for theatrical presentation. However, existing SDR-to-HDR conversion methods are primarily tailored for consumer-grade content such as television and therefore fall short of the stringent requirements of professional cinematic material. To bridge this gap, we present HDRMovie7K, the first large-scale, lossless dataset of cinematic SDR-HDR frame pairs sourced from professional Digital Cinema Distribution Master (DCDM) workflows. Based on this foundation, we introduce HDRMovieformer, a transformer-based framework featuring a Luminance Estimator module for luminance guidance, a Luminance-Guided Multi-Head Self-Attention to focus on critical fine-detail recovery, and a Chroma Refiner for color accuracy, optimized with a novel Wide Color Gamut Loss. To further evaluate our model in online streaming media scenarios, we introduce HDRMovie1K, a dataset curated from publicly available HDR film clips. Extensive experiments on both HDRMovie7K and HDRMovie1K demonstrate that our method achieves state-of-the-art performance.

Code —
https://github.com/XianWeiLee/HDRMovieformer

Introduction

With the advancement of the film industry and the rapid development of projection technologies, high-end cinema venues such as Cinity LED theaters (Yang and Khoo 2024) are becoming increasingly prevalent. A key feature of these theaters is their support for HDR film playback. These venues offer richer contrast and enhanced color depth, creating a more immersive visual experience for audiences. Despite these advances, the vast majority of existing films are still mastered and distributed in SDR formats, necessitating advanced conversion techniques to unlock the full potential of HDR displays in theatrical settings.

Recent years have seen considerable progress in SDR-to-HDR conversion. Learning-based methods for HDR imaging span several distinct tasks (Guo et al. 2023). HDR-style enhancement (Cai et al. 2024), which synthesize HDR views on SDR displays; multi-exposure HDR imaging (Zhang et al. 2024), which fuses multiple SDR images into a linear HDR image; single-image HDR reconstruction (Cui et al. 2024), which generates a linear HDR from a single SDR; and SDRTV-to-HDRTV up-conversion (Chen et al. 2021), which produces display-referred HDRTV frames, typically in specific electro-optical transfer function (EOTF) and wide color-gamut (WCG) RGB primaries, from a single SDR image. Our work aligns closely with this task, but targets the cinematic domain, aiming to adapt existing SDR content for modern HDR cinema presentation.

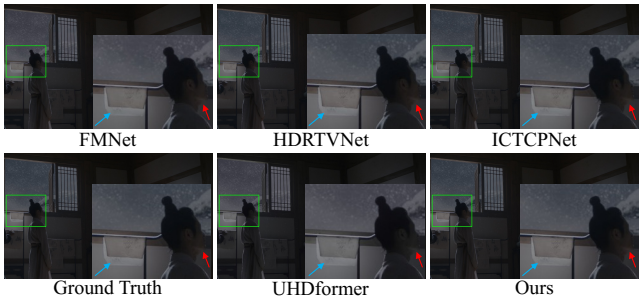

Figure 1: Our method outperforms others in preserving natural outdoor brightness (blue) and in avoiding unnatural skin tones or poor shadow detail (red). These subtle differences become more noticeable at larger image scales (cinema).

Unlike SDRTV-to-HDRTV tasks, cinematic SDR-to-HDR conversion presents unique challenges due to different standards (Digital Cinema Initiatives 2024), equipment, and viewing environments. As summarized in Table 1, television content is typically mastered using consumer-grade equipment within the Rec.709 (ITU-R 2015a) gamut and viewed under diverse ambient lighting. Consequently, many existing SDRTV-to-HDRTV methods emphasize global tone mapping (Chen et al. 2021), often neglecting subtle luminance and chroma details that are unlikely to be perceived by home viewers (Figure 1). In contrast, cinematic SDR content is produced using high-end cameras in the wider DCI-P3 color gamut and intended for projection in controlled dark-room environments. This setting demands precise recovery of fine details, where even minor artifacts can break

*Corresponding author.

Feature	Television	Cinema
SDR Gamut&EOTF	Rec.709, Gamma 2.2	DCI P3, Gamma 2.6
HDR Gamut&EOTF	Rec.2020, PQ/HLG	Rec.2020, CINITYLog/PQ
Capture Devices	Consumer-grade cameras	High-end cinema cameras
Encoded	YCbCr, RGB	DCI X'Y'Z'
Container	MP4, MKV, TS	DCDM, DCP
Viewing Environment	Various Environment	Cinema dark room

Table 1: Differences in SDR and HDR for TV and Cinema.

viewer immersion. Furthermore, existing methods are typically trained and optimized on datasets tailored to television content, rather than cinema-grade material, limiting their applicability in high-fidelity cinematic scenarios.

To address these challenges, we introduce HDRMovie7K, the first large-scale dataset of cinematic SDR-HDR frame pairs, sourced from professional DCDM workflows in lossless DPX format. This dataset captures the full fidelity and color richness of cinema production. Building on this foundation, we propose HDRMovieformer, a transformer-based framework specifically designed to recover fine-grained luminance and chroma information for cinematic HDR reconstruction. Unlike prior methods that apply coarse, global mappings, our approach leverages local context and luminance-aware guidance to preserve intricate scene characteristics. Additionally, with the popularity of online streaming media, we collect publicly available film clips to create an external dataset, HDRMovie1K, to further evaluate our model and enrich our data resources.

Our design begins with a Luminance Estimator (LE) module that produces a HDR luminance map, supplying guidance to the Luminance-Guided Transformer (LGT). The core of the LGT is the Luminance-Guided Multi-head Self-Attention (LG-MSA) mechanism, to focus on feature propagation on critical bright and dark regions, ensuring fine detail recovery. A subsequent Chroma Refiner (CR) then corrects subtle color shifts, and we optimize the network with a novel WCG-Loss that penalizes errors in the wide color-gamut space. Extensive experiments demonstrate that the proposed HDRMovieformer achieves HDR reconstructions whose contrast, highlight fidelity, and color accuracy meet the professional cinema standards, outperforming state-of-the-art (SOTA) SDR-to-HDR methods in both quantitative and perceptual evaluations. The contributions of this paper includes:

- We construct the first large-scale, high-quality cinematic dataset named HDRMovie7K, derived from professional DCDM workflows in DPX format. We also collect HDRMovie1K for HDR conventions in online streaming media movie content.
- We propose HDRMovieformer, a transformer-based framework with a Luminance Estimator, Luminance-Guided Multi-head Self-Attention, and Chroma Refiner, designed to recover fine-grained luminance and chroma details for cinematic HDR reconstruction.
- We develop a novel WCG-Loss function that penalizes errors in wide color-gamut space, optimizing the network to achieve high color accuracy in cinematic HDR content.
- Extensive experiments demonstrate that the proposed model achieves superior HDR restoration quality on both our curated dataset and external film clips.

Metrics on the extent of HDR	
FHLP	Fraction of HighLight Pixel (2023)
EHL	Extent of HighLight (2023)
Metrics on the extent of WCG	
FWGP	Fraction of Wide-Gamut Pixel (2023)
EWG	Extent of Wide-Gamut (2023)
Metrics on the overall-style	
ALL	Average Luminance Level (2023)
ASL	Average Saturation Level (2023)
HDRBQ	HDR Brightness Quantification (2023)
DR	Dynamic Range (2022) (log10)
Metrics on intra-frame diversity	
SI	Spatial Information (2019)
CF	Colorfulness (2003)
stdL	standard deviation of Luminance (2023)

Table 2: Metrics to assess the diversity of HDR video.

Related Work

SDR-to-HDR dataset. Kim et al. (Kim, Oh, and Kim 2019) collect the KAIST dataset, which comprises 10 4K-UHD HDR videos in BT.2020 (ITU-R 2015b) with PQ (SMPTE 2014) EOTF from YouTube, totaling 59,818 frames, using randomly cropped patches of size 160×160, and SDR pairs are obtained from YouTube's automatic conversion. Zeng et al. (Zeng et al. 2020) collect 23,229 training frames from 79 videos and 50 test frames, with SDR frames generated via Reinhard tone mapping from HDR frames to avoid mismatches. The HDRTV1K dataset is introduced by (Chen et al. 2021). It is constructed using HDR videos in the HDR10 standard and includes 1235 training images and 117 test images, with 22 video pairs sampled from YouTube. Guo et al. (Guo et al. 2023) introduce HDRTV4K, a dataset with 3878 high-quality, diversified HDRTV frames in BT.2020/PQ format, manually selected from various sources and re-graded to enhance diversity. However, these datasets are not well suited for cinematic HDR reconstruction tasks, as they often focus on short clips or frames extracted from web videos or demos, do not fit the high standards of the cinematic movies.

SDR-to-HDR Translation method. In recent years, there has been a surge in deep learning-based inverse tone mapping (ITM) methods for SDR-to-HDR video. Early approaches use end-to-end CNNs to decompose images into detail components, enhancing contrast and texture (Kim, Oh, and Kim 2020, 2019; Yao et al. 2023; He et al. 2023). JSI-GAN (Kim, Oh, and Kim 2020) integrates CNNs with pixel-level filters, while Deep SR-ITM (Kim, Oh, and Kim 2019) introduces modulation blocks for local contrast. Yao et al. (Yao et al. 2023) propose bidirectional translation for dynamic range management and He et al. (He et al. 2023)

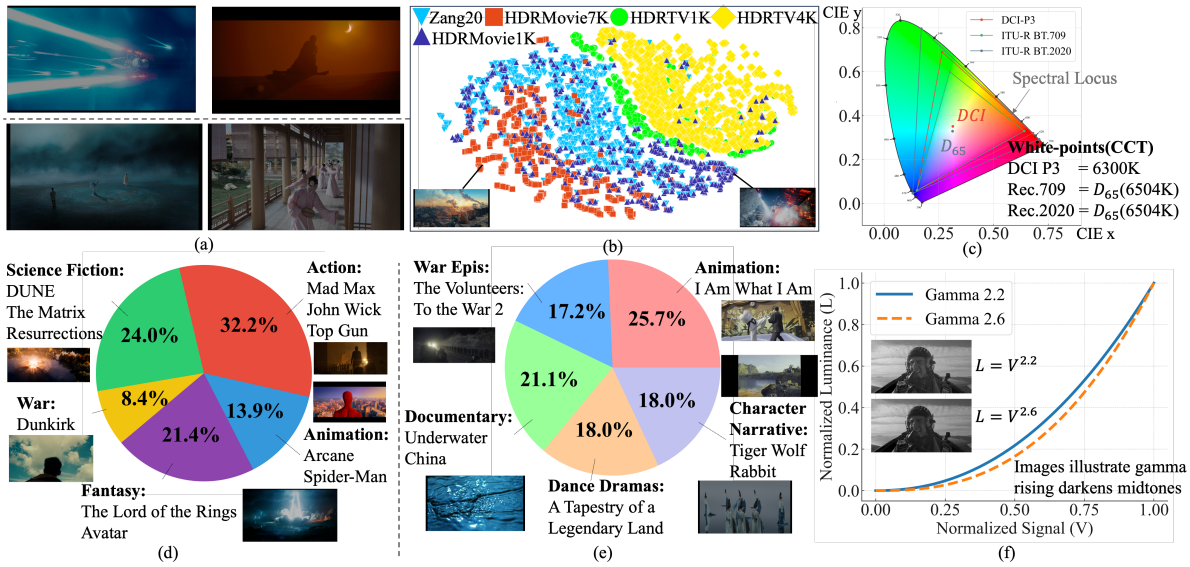

Figure 2: (a) Sampled movie hdr clips from HDRMovie1K (upper) HDRMovie7K (lower). (b) Diversity comparison: Our datasets vs. others. (c) Differences in color gamut chromaticity diagrams. (d) Movie genres percentage of HDRMovie1K. (e) Movie genres percentage of HDRMovie7K. (f) Different gamma curves cause midtone-darkening

Metric / Dataset	Extent of HDR		Extent of WCG		Intra-frame Diversity			Overall-style				Status	
	FHLP	EHL	FWGP	EWG	SI	CF	stdL	ASL	ALL	HDRBQ	DR	Pairs	HDR format
HDRTV1K	1.2843	0.1971	2.9089	0.1633	2.2378	11.0722	1.8006	10.9414	15.1626	2.7970	1.4838	1235	16 bit .png RGB
HDRTV4K	5.3083	0.9595	2.6369	0.5123	3.5508	10.5882	3.4837	9.8274	21.1996	5.1593	1.5799	3878	16 bit .tif RGB
Zeng20	0.0197	0.0012	0.4792	0.0034	0.1231	4.2048	0.3146	3.8061	6.0805	0.3781	4.7545	23229	10 bit .mp4 RGB
HDRMovie7K	0.2056	0.0819	0.6392	0.0643	0.1538	6.7955	0.7688	5.5327	5.9726	0.5856	3.0687	6775	16 bit .dpx X'Y'Z'
HDRMovie1K	0.2554	0.0281	2.4621	0.0744	0.4422	7.6540	0.7804	4.7012	9.4021	0.4854	4.0063	984	16 bit .png RGB

Table 3: Statistics of different datasets. Cinematic content tends to have lower values in these metrics due to more restrained tonal grading and artistic control, in contrast to HDRTV content which emphasizes visual extremes.

enhance global feature extraction using prior-guided modulation. Chen et al. (Chen et al. 2021) present a three-step pipeline with adaptive color mapping and local enhancement. KUNet (Wang et al. 2022) employs a U-Net with frequency-based separation, while HycondITM (Shao et al. 2022) fuses global and spatial modulation. Guo et al. (Guo et al. 2023) apply brightness-level segmentation and Transformer-style UNet for restoration. Xu et al. (Xu et al. 2022) leverage discrete cosine transform to reduce structural distortion. Huang et al. (Huang et al. 2023) operate in IC_TC_P space for luminance chrominance separation, and Liu et al. (Liu et al. 2024) embed metadata using multilayer perceptron (MLP) for HDR/WCG transmission. However, most existing methods are designed for general TV content and overlook the unique demands of cinematic or movie-grade HDR restoration, such as precise tone reproduction and compatibility with theatrical formats.

Proposed Dataset and Method

Dataset

HDR movie mastering follows a distinct pipeline tailored for cinematic environments (Digital Cinema Initiatives 2024), employing X'Y'Z' encoding and display characteristics optimized for dark theater settings. These technical and perceptual differences render HDRTV methods suboptimal for theatrical content. To address this, we introduce two novel datasets: HDRMovie7K and HDRMovie1K. For HDRMovie7K, it comprises 6,775 frames sourced from professional DCDM workflows. Developed in collaboration with the professional color grading team, this dataset is based on footage captured by high-end cinema cameras. All content was mastered using the Academy Color Encoding System (ACES) pipeline. SDR frames are encoded in 16-bit DCI-P3 X'Y'Z' with a Gamma 2.6 transfer function, while HDR frames are encoded in 16-bit BT.2020 X'Y'Z' with the PQ transfer function. The dataset spans resolutions 2K to 4K. All color grading was performed by professional colorists on a Sony BVM-X300 mastering monitor. HDRMovie7K encompasses a diverse range of cinematic genres: war epics, dance dramas, animations, documentaries, and character-driven narratives. Stored in DPX format to ensure lossless quality, all HDR content has been verified in Cinity LED cinema environments, guaranteeing perceptual fidelity under real-world theater conditions. To complement HDRMovie7K and address the growing prevalence

of online streaming, we curated HDRMovie1K, a dataset of 984 SDR-HDR image pairs sourced from publicly available HDR movie clips on YouTube. This dataset is designed to reflect the visual characteristics of web-distributed HDR content. SDR frames are encoded in 8-bit Rec.709 RGB with a Gamma 2.2 transfer function, while HDR frames are encoded in 16-bit BT.2020 RGB with the PQ transfer function. The resolutions are in 4K, covering genres such as action, animation, fantasy, science fiction, and war films.

Dataset Analysis. To quantitatively evaluate the superiority of our datasets, we analyze the diversity of the HDRTV1K, HDRTV4K, Zang20, and our datasets. Following (Guo et al. 2023; Tian et al. 2025), we utilize 10 metrics to assess the diversity of different HDR video datasets in terms of intra-frame diversity, extent, overall style. For each movie frame, 10 different metrics are calculated as outlined in Table 2. We then employ t-SNE (Van der Maaten and Hinton 2008) to project 10-D vector of each image into corresponding 2D coordinate to plot the dataset distribution of our dataset and the comparison datasets. As shown in Figure 2 (b) our dataset exhibits clear separation from other datasets, suggesting that HDR movie content offers distinct coverage, which is essential for training models. Furthermore, we present the different movie genre distributions of our datasets in Figure 2 (c) and (d). Detailed statistics of different datasets are shown in Table 3.

HDRMovieformer Architecture

Figure 3 illustrates the overall structure of our method, which consists of a Luminance Estimator, a Luminance-Guided Transformer (LGT) and a Chroma Refiner. LGT is a hierarchical encoder-decoder architecture. The core unit of LGT is the Luminance-Guided Attention Block (LGAB), which includes two Layer Normalization (LN) layers, a Luminance-Guided Multi-head Self-Attention (LG-MSA) module, and a multilayer perceptron (MLP).

Given an input SDR frame $\mathbf{I}_{\text{sdr}} \in \mathbb{R}^{H\times W\times 3}$, we first compute the channel-wise mean of the SDR frame as luminance prior $\mathbf{L}_p \in \mathbb{R}^{H\times W}$. Then, the luminance estimator $\mathcal{E}$ takes $\mathbf{I}_{\text{sdr}}$ and $\mathbf{L}_p$ as inputs. Leveraging the Retinex theory (Cai et al. 2023), $\mathcal{E}$ outputs the luminance boosted image $\mathbf{I}_{lu}$ and luminance feature $\mathbf{F}_{lu} \in \mathbb{R}^{H\times W\times C}$; where $H \times W$ denotes the spatial dimension and C is the number of channels. The boosted image is then refined by LGT $\mathcal{R}$, to produce the refined HDR output $\mathbf{I}_{\text{re}} \in \mathbb{R}^{H\times W\times 3}$. Finally, the refined output is enhanced by Chroma Refiner $\mathcal{H}$ to produce the final HDR frame $\mathbf{I}_{\text{hdr}} \in \mathbb{R}^{H\times W\times 3}$. The process is structured as:

$$(\mathbf{I}_{lu}, \mathbf{F}_{lu}) = \mathcal{E}(\mathbf{I}_{\text{sdr}}, \mathbf{L}_p), \mathbf{I}_{\text{re}} = \mathcal{R}(\mathbf{I}_{lu}, \mathbf{F}_{lu}), \mathbf{I}_{\text{hdr}} = \mathcal{H}(\mathbf{I}_{re}), \quad (1)$$

Architecture of Luminance Estimator. $\mathcal{E}$ begins by concatenating $\mathbf{I}_{\text{sdr}}$ with $\mathbf{L}_p$ and applying a 1×1 convolution for channel fusion. To better model local-global interactions across varying luminance regions, we apply a depth-wise separable 9×9 convolution to extract luminance sensitive features, forming $\mathbf{F}_{lu}$. Another 1×1 convolution generates a learned scaling map $\mathbf{L}_m \in \mathbb{R}^{H\times W\times 1}$, which is used to boost the SDR luminance by element-wise multiplication: $\mathbf{I}_{lu} = \mathbf{L}_p \odot \mathbf{L}_m$.

Luminance-Guided Transformer

Network Structure. LGT adopts a three-scale U-shaped Transformer architecture. The input $\mathbf{I}_{lu}$ is processed by a 3×3 convolution, followed by LGABs and strided 4×4 convolutions to downsample the features hierarchically. We denote the multiscale features as $\mathbf{F}_i \in \mathbb{R}^{\frac{H}{2^i}\times\frac{W}{2^i}\times 2^i C}$ for $i = 0, 1, 2$. After further processing by LGABs at the coarsest scale, we upsample the features using transposed convolutions and add skip connections for better detail preservation. The output of LGT is a residual image $\mathbf{I}_{res} \in \mathbb{R}^{H\times W\times 3}$, which is added to $\mathbf{I}_{lu}$ to generate the refined HDR luminance: $\mathbf{I}_{re} = \mathbf{I}_{lu} + \mathbf{I}_{res}$.

LG-MSA. LG-MSA incorporates $\mathbf{F}_{lu}$ into each attention block as an auxiliary modulation. For spatial alignment at smaller scales, $\mathbf{F}_{lu}$ is downsampled using 4×4 convolutions with stride 2. To reduce computational burden, we flatten single-channel maps into tokens and compute self-attention across them, allowing efficient long-range dependency modeling without full 2D global attention. Firstly, the input feature $\mathbf{F}_{in} \in \mathbb{R}^{H\times W\times C}$ is reshaped into tokens $\mathbf{T} \in \mathbb{R}^{HW\times C}$. Then, $\mathbf{T}$ is evenly split into k heads: $\mathbf{T} = [\mathbf{T}_1, \ \mathbf{T}_2, \cdots, \mathbf{T}_k]$, where $\mathbf{T}_i \in \mathbb{R}^{HW\times d_k}$, $d_k = \frac{C}{k}$, and $i = 1, 2, \cdots, k$. For each head i, three linear projections are applied via fully connected (fc) layers without bias to generate the $query$ ($\mathbf{Q}$), key ($\mathbf{K}$), and $value$ ($\mathbf{V}$) matrices:

$$\mathbf{Q}_i = \mathbf{T}_i\mathbf{W}_{\mathbf{Q}_i}^{\text{T}}, \ \mathbf{K}_i = \mathbf{T}_i\mathbf{W}_{\mathbf{K}_i}^{\text{T}}, \ \mathbf{V}_i = \mathbf{T}_i\mathbf{W}_{\mathbf{V}_i}^{\text{T}}, \quad (2)$$

where $\mathbf{W}_{\mathbf{Q}_i}$, $\mathbf{W}_{\mathbf{K}_i}$, and $\mathbf{W}_{\mathbf{V}_i} \in \mathbb{R}^{d_k\times d_k}$ are learnable projection matrices. To enhance attention modulation and guide HDR restoration, we leverage the intermediate luminance feature $\mathbf{F}_{lu} \in \mathbb{R}^{H\times W\times C}$, which encodes semantic and luminance priors derived from the SDR input. This feature is reshaped to $\mathbf{X} \in \mathbb{R}^{HW\times C}$ and similarly split into k heads: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_k]$, with each $\mathbf{X}_i \in \mathbb{R}^{HW\times d_k}$. These features modulate the value stream, enabling contextual enhancement from brighter or more informative regions to guide SDR-to-HDR mapping. The attention output of head i is computed as:

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i, \mathbf{X}_i) = (\mathbf{X}_i \odot \mathbf{V}_i)\text{softmax}\left(\frac{\mathbf{K}_i^{\text{T}}\mathbf{Q}_i}{\alpha_i}\right), \quad (3)$$

where $\odot$ denotes element-wise multiplication and $\alpha_i \in \mathbb{R}$ is a learnable scalar that controls the attention scaling for head i. After computing all heads, their outputs are concatenated and passed through an output fc layer. A learnable positional encoding $\mathbf{P} \in \mathbb{R}^{HW\times C}$ is added to form the final token sequence $\mathbf{X}_{\text{out}} \in \mathbb{R}^{HW\times C}$, which is reshaped back to the spatial format $\mathbf{F}_{\text{out}} \in \mathbb{R}^{H\times W\times C}$.

Chroma Refiner and Wide Color Gamut Loss

The Chroma Refiner module $\mathcal{H}$ is specifically designed to enhance the chrominance components. It begins with a 1×1 point-wise convolution to project the input features into a higher-dimensional space, enabling richer chroma representation. This is followed by a depth-wise separable 3×3 convolution that captures local spatial correlations within each

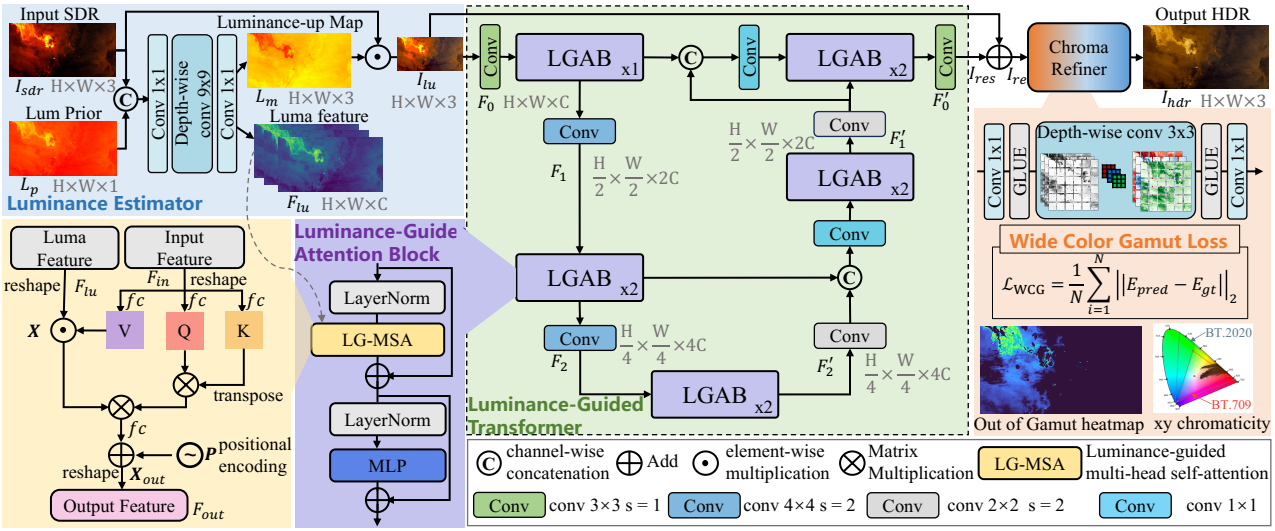

Figure 3: Architecture of the proposed HDRMovieformer framework.

channel, effectively refining chroma details. Finally, another 1×1 convolution restores the feature dimensionality, yielding the enhanced HDR chroma output $\mathbf{I}_{\text{hdr}}$ with improved color fidelity and structural coherence.

Wide Color Gamut Loss. To encourage the model to reconstruct chromaticities that respect wide color gamut (WCG) fidelity, we introduce a differentiable loss computed directly in the CIE 1931 XYZ color space. Both the predicted output and ground truth are assumed to be linear RGB values defined in the BT.2020 color space. These are converted to XYZ using the standard 3×3 matrix transformation:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{M} \begin{bmatrix} R \\ G \\ B \end{bmatrix}, \mathbf{M} = \begin{bmatrix} 0.6370 & 0.1446 & 0.1689 \\ 0.2627 & 0.6780 & 0.0593 \\ 0.0000 & 0.0281 & 1.0610 \end{bmatrix}, \quad (4)$$

Let $\mathbf{E}_{\text{pred}}$ and $\mathbf{E}_{\text{gt}}$ denote the predicted and ground truth XYZ images, respectively. The WCG-Degree Loss is defined as the average Euclidean distance between the predicted and ground truth XYZ values over all pixels:

$$\mathcal{L}_{\text{WCG}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{E}_{\text{pred}}^{(i)} - \mathbf{E}_{\text{gt}}^{(i)} \right\|_2, \quad (5)$$

where N is the total number of pixels. This loss provides a perceptually grounded supervision signal that reflects chromatic differences beyond traditional RGB space and encourages the network to produce color values more consistent with the BT.2020 reference.

Experiments and Analysis

Datasets. We conduct our experiments on our HDRMovie7K and HDRMovie1K datasets. For HDRMovie7K, we convert all original DPX files into lossless TIFF and split the resulting 6,775 frame pairs into 5,420 for training and 1,355 for testing. As for HDRMovie1K, it is partitioned into 884 training pairs and 100 testing pairs.

Implementation details. We implement HDRMovieformer with PyTorch (Paszke 2019). The model is trained on paired 1024×1024 SDR-HDR patches cropped from our datasets using a single RTX 4090 GPU. Optimization is performed using the Adam (Kingma and Ba 2015) optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for 1.5×10^5 iterations. The learning rate is initially set to 2×10^{-4} and then steadily decreased to 1×10^{-6} using the cosine annealing scheme (Loshchilov and Hutter 2017) during the training process. We use L1 loss and WCG loss with a weighting of 0.1. The batch size is set to 2.

Evaluation metrics. Since HDR values are nonlinearly related to perceived differences, we follow prior work (Liu et al. 2024; Azimi et al. 2021) and transform HDR images into PU space using the perceptual uniform (PU) encoder. We evaluate in PU space using PSNR (Y), MS-SSIM (Wang, Simoncelli, and Bovik 2003), VSI (Zhang, Shen, and Li 2014), and FSIM (Zhang et al. 2011). The perceptual HDR-specific metric HDR-VDP3 (Mantiuk, Hammou, and Hanji 2023) is also used. For chromatic and perceptual quality, we report chromaticity error (ITU-R 2019) and SR-SIM (Zhang and Li 2012). Conventional PSNR and SSIM (Wang et al. 2004) are included for reference.

Comparison With Other Methods. We compare our method with four categories: the joint SR and SDRTV-to-HDRTV (SR-ITM-GAN (Zeng et al. 2020)), image translation (Pixel2Pixel (Isola et al. 2017)), image restoration (OKNet (Cui, Ren, and Knoll 2024), UHDformer (Wang et al. 2024)), and SDRTV-to-HDRTV (HDRTVNet (Chen et al. 2021), FMNet (Xu et al. 2022), ICTCPNet (Huang et al. 2023)). To ensure fair comparison, we adapt their inputs/outputs as needed, normalize 16-bit SDR inputs by $2^{16} - 1$, and retrain all models using their default settings.

Quantitative comparison. As shown in Table 4, we conduct quantitative evaluations on HDRMovie1K and HDRMovie7K datasets. On the HDRMovie1K dataset, our method achieves the best performance across most metrics, notably attaining the highest PSNR(Y) of 43.79, MS-SSIM

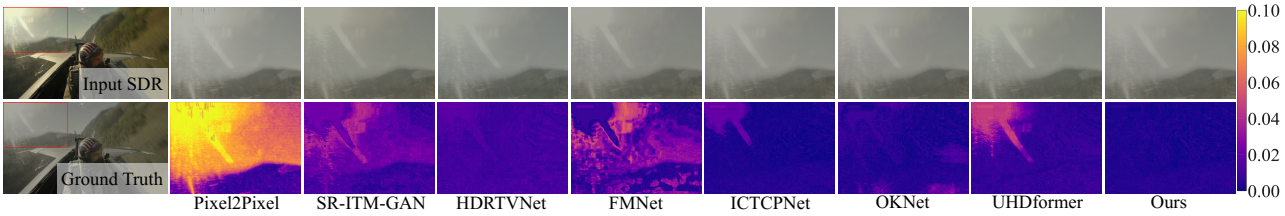

Figure 4: Qualitative comparisons. We visualize the error maps below the results.

Metric / Method	Perceptual Uniform (2021)				Luma	HDR	Chroma	Conventional	
	PSNR(Y) ↑	MS-SSIM ↑	VSI ↑	FSIM ↑	SR-SIM ↑	HDR-VDP3 ↑	ΔE_{ITP} ↓	PSNR ↑	SSIM ↑
HDRMovie1K dataset									
Pixel2Pixel (2017)	34.8433	0.9884	0.9978	0.9916	0.9945	8.2412	11.821	36.14	0.9784
SR-ITM-GAN (2020)	36.7389	0.9904	0.9969	0.9920	0.9913	7.7649	13.341	36.18	0.9803
HDRTVNet (2021)	40.1686	0.9917	0.9985	0.9945	0.9965	8.6287	7.316	41.14	0.9847
FMNet (2022)	36.8597	0.9872	0.9966	0.9868	0.9944	8.5070	6.816	38.69	0.9851
ICTCPNet (2023)	41.4888	0.9932	0.9987	0.9950	**0.9973**	8.7363	6.517	42.34	0.9854
OKNet (2024)	40.8165	0.9936	0.9986	0.9952	0.9970	**8.7956**	5.484	41.89	0.9887
UHDformer (2024)	38.2332	0.9916	0.9980	0.9934	0.9953	8.5866	10.30	39.18	0.9827
Ours	**43.7938**	**0.9939**	**0.9989**	**0.9960**	0.9971	8.7830	**5.037**	**44.41**	**0.9896**
HDRMovie7K dataset									
HDRTVNet (2021)	53.6309	0.9996	0.9969	0.9988	0.9979	9.8550	3.167	51.91	0.9854
FMNet (2022)	51.4693	0.9984	0.9975	0.9952	0.9970	9.2977	3.995	50.06	0.9823
ICTCPNet (2023)	51.2232	0.9993	0.9977	0.9980	0.9957	9.7408	2.954	50.23	**0.9957**
OKNet (2024)	50.0044	0.9996	0.9978	0.9996	0.9958	9.8517	6.135	47.62	0.9781
UHDformer (2024)	55.8492	0.9994	0.9978	0.9994	0.9973	9.8438	2.065	54.73	0.9775
Ours	**57.3160**	**0.9997**	**0.9979**	**0.9998**	**0.9981**	**9.8935**	**1.461**	**55.15**	0.9941

Table 4: Quantitative comparisons. The best and second best results are marked in **bold** and underlined.

of 0.9939, VSI of 0.9989, FSIM of 0.9960, and the lowest chroma error ΔE_{ITP} of 5.037. It also reaches the best conventional scores (PSNR 44.41, SSIM 0.9896) and ranks second in SR-SIM and HDR-VDP3, demonstrating superior perceptual and fidelity performance. Similarly, on the larger HDRMovie7K dataset, our method consistently outperforms all competitors, yielding the top scores in PSNR(Y), MS-SSIM, VSI, FSIM, SR-SIM, HDR-VDP3, and ΔE_{ITP}. Notably, it surpasses the recent UHDformer (2024) by 1.47 dB in PSNR and reduces color distortion with a ΔE_{ITP} of only 1.461. These results confirm the effectiveness of our method in preserving both luminance and chrominance information on both cinematic and streaming data.

Qualitative comparison. Figures 4 and 6 provide qualitative comparisons highlighting the effectiveness of our method. Figure 4 presents error maps on a fighter jet cockpit scene with strong glare. Our method shows the lowest deviation from the ground truth, especially in the glare region, outperforming Pixel2Pixel, SR-ITM-GAN, HDRTVNet, FMNet, and UHDformer in preserving scene fidelity under challenging lighting. Figure 6 shows pixel distributions in CIE xyY space for a dramatic sky scene. Our results best match the ground truth in both chromaticity and luminance, indicating accurate color reproduction and dynamic range. Overall, our method demonstrates superior spatial accuracy, glare handling, and color fidelity.

Subjective experiment. Inspired by BT.500 (ITU-R 2023)

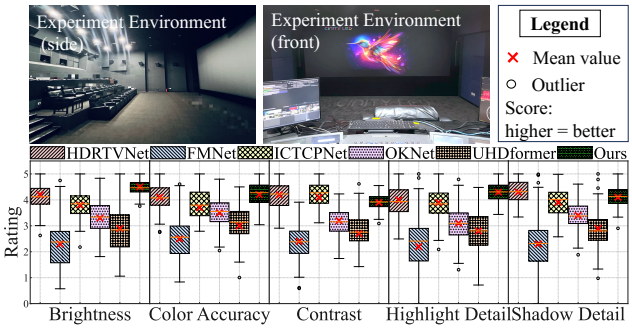

Figure 5: Results and environment of subjective experiment. The Cinity Film Lab calibrated for correct SDR/HDR visualization. Results are shown in quartile charts.

guidelines, we recruit 20 non-expert participants. The experiment is conducted in a dark room provided by Cinity Film Lab (Figure 5). We select a diverse set of 10-20 second movie clips from each method. Participants rate each clip via the Absolute Category Rating (ACR) (ITU-R 2008) method, evaluating brightness, contrast, color accuracy, highlight detail, and shadow detail on 5-point scales. Clips and conversion conditions are randomized and counterbalanced to avoid order effects, with hidden repeated trials included to check rating consistency. Results indicate that our method generally performs well across various metrics, achieving

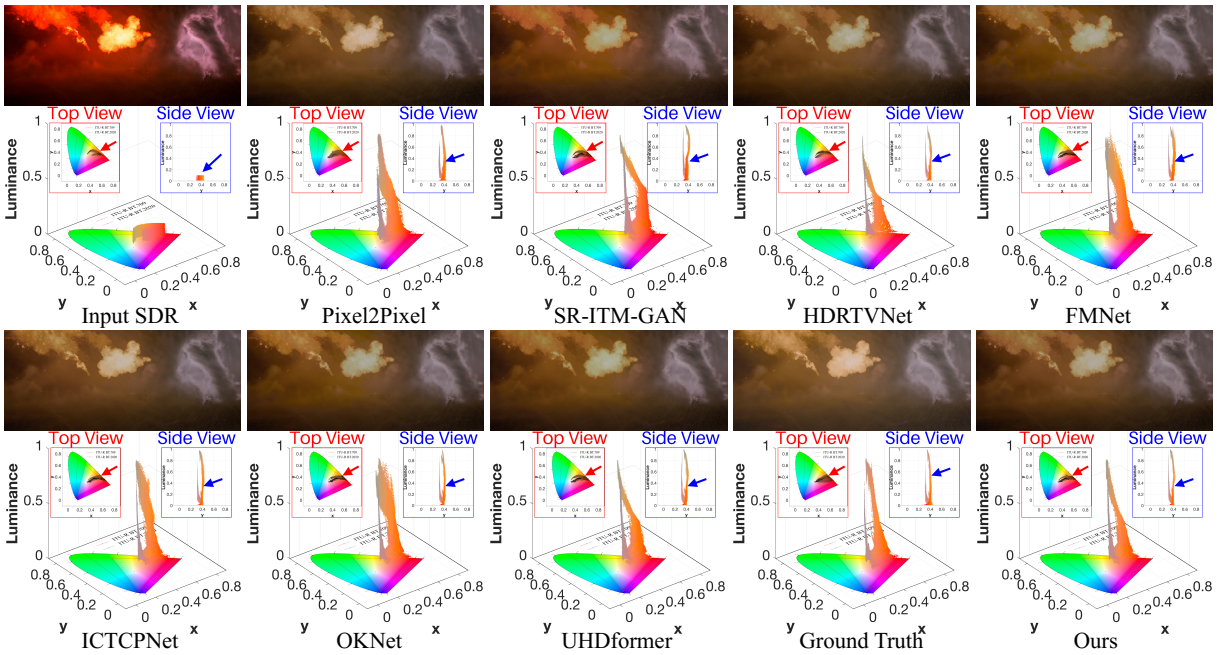

Figure 6: Qualitative comparison. Pixel distributions are visualized in CIE xyY space (Schanda 2007): red illustrations show chromaticity (xy) distribution, blue illustrations show luminance (Y).

Experiments	PSNR ↑	SSIM ↑	HDR-VDP-3 ↑	ΔE_{ITP} ↓
w/o LE	53.70	0.9915	9.8905	1.654
w/o LG-MSA	54.45	0.9952	9.8933	1.489
w/o CR	54.74	0.9936	9.8893	1.494
w/o WCG Loss	54.72	0.9962	9.8904	1.511
Full Model	55.15	0.9941	9.8935	1.461

Table 5: Ablation study on HDRMovieformer.

Methods	Params (M)	FLOPs (G)
Pixel2Pixel (2017)	10.85	185.20
SR-ITM-GAN (2020)	0.49	180.92
HDRTVNet (2021)	37.20	206.85
FMNet (2022)	1.24	86.41
ICTCPNet (2023)	0.76	70.69
OKNet (2024)	2.40	17.86
UHDformer (2024)	0.34	12.10
Ours	1.53	62.31

Table 6: Model complexity of different methods.

the highest mean ratings in brightness, highlight detail, and color accuracy. However, in contrast and shadow detail, our method slightly underperforms compared to HDRTVNet. This suggests that while our approach excels in overall image enhancement, there may be room for improvement in preserving contrast and shadow details.

Ablation Studies. We conduct ablation studies on the HDRMovie7K dataset to evaluate the contribution of each key component in HDRMovieformer. The results are summarized in Table 5. Removing the Luminance Estimator (LE) leads to the most significant drop in performance, with a PSNR decrease of 1.45 dB and an increase in ΔE_{ITP} of 0.193, confirming its critical role in luminance-aware HDR reconstruction. Replacing the LG-MSA with a standard global MSA results in a PSNR reduction of 0.70 dB and a ΔE_{ITP} increase of 0.028, demonstrating that LG-MSA better captures long-range spatial dependencies under varying luminance. To assess the importance of the Chroma Refiner (CR) module, we replace it with a standard convolutional head. This change reduces PSNR by 0.41 dB and increases ΔE_{ITP} by 0.033, showing its contribution to effective chroma detail reconstruction. Finally, removing the WCG Loss $\mathcal{L}_{\text{WCG}}$ and using only an RGB ℓ_1 loss causes a PSNR drop of 0.43 dB and a ΔE_{ITP} rise of 0.050, indicating that $\mathcal{L}_{\text{WCG}}$ is essential for maintaining chromatic fidelity under the wide gamut BT.2020 color space. These results validate that each component plays a crucial role in achieving SOTA HDR reconstruction quality.

Conclusion

In this study, we present HDRMovie7K, the first large-scale dataset of cinematic SDR-HDR frame pairs and the HDRMovie1K for streaming scenarios, addressing the gap in cinema-grade resources for HDR model training. Additionally, we propose HDRMovieformer, a transformer-based approach with a Luminance Estimator for coarse HDR guidance, a LG-MSA mechanism for fine luminance recovery, a Chroma Refiner for enhanced color accuracy, and a novel WCG Loss further improves color fidelity. It achieves SOTA performance, validated by experiments and subjective studies, laying a foundation for future cinematic HDR research.

Acknowledgements

This work is supported in part by the National Key R&D Program of China under No.2023YFF0904800, the NSFC under No.62272059 and No.U24B20176, the Beijing Natural Science Foundation under No.JQ24020, and the Beijing Nova Program under No.20230484406.

References

Azimi, M.; et al. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*, 1–5. IEEE.

BT.1788, I.-R. R. 2019. Methodology for the subjective assessment of video quality in multimedia applications.

Cai, Y.; Bian, H.; Lin, J.; Wang, H.; Timofte, R.; and Zhang, Y. 2023. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12504–12513.

Cai, Y.; Xiao, Z.; Liang, Y.; Qin, M.; Zhang, Y.; Yang, X.; Liu, Y.; and Yuille, A. 2024. HDR-GS: Efficient High Dynamic Range Novel View Synthesis at 1000x Speed via Gaussian Splatting. In *NeurIPS*.

Chen, X.; Zhang, Z.; Ren, J. S.; Tian, L.; Qiao, Y.; and Dong, C. 2021. A new journey from sdrtv to hdrtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4500–4509.

Cui, M.; Wang, Z.; Wang, D.; Zhao, B.; and Li, X. 2024. Color event enhanced single-exposure HDR imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1399–1407.

Cui, Y.; Ren, W.; and Knoll, A. 2024. Omni-Kernel Network for Image Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1426–1434.

Digital Cinema Initiatives. 2024. High Dynamic Range Digital Cinema Addendum Version 1.2.1. https://documents.dcimovies.com/HDR-Addendum/release/1.2.1/. DCI Specification Addendum, published October 10, 2018.

Guo, C.; Fan, L.; Xue, Z.; and Jiang, X. 2023. 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*, 22231–22241.

Hasler, D.; and Suesstrunk, S. E. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, 87–95. SPIE.

He, G.; Long, S.; Xu, L.; Wu, C.; Yu, W.; and Zhou, J. 2023. Global priors guided modulation network for joint super-resolution and SDRTV-to-HDRTV. *Neurocomputing*, 554: 126590.

Hu, X.; Shen, L.; Jiang, M.; Ma, R.; and An, P. 2022. LA-HDR: Light adaptive HDR reconstruction framework for single LDR image considering varied light conditions. 25: 4814–4829.

Huang, P.; Cao, G.; Zhou, F.; and Qiu, G. 2023. Video Inverse Tone Mapping Network with Luma and Chroma Mapping. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1383–1391.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

ITU-R. 2008. Subjective video quality assessment methods for multimedia applications. *ITU-R Rec, P.910, Tech. Rep.*

ITU-R. 2015a. Parameter values for the hdtv standards for production and international programme exchange. *ITU-R Rec, BT.709-6, Tech. Rep.*

ITU-R. 2015b. Parameter values for ultra-high definition television systems for production and international programme exchange,. *ITU-R Rec, BT.2020-2, Tech. Rep.*

ITU-R. 2019. Objective metric for the assessment of the potential visibility of colour differences in television. *ITU-R Rec, BT.2124-0, Tech. Rep.*

ITU-R. 2023. *Recommendation ITU-R BT.500-15: Methodologies for the subjective assessment of the quality of television images*. ITU, Geneva, Switzerland, 15 edition.

Kim, S. Y.; Oh, J.; and Kim, M. 2019. 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*, 3116–3125.

Kim, S. Y.; Oh, J.; and Kim, M. 2020. 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, 11287–11295.

Kingma, D. P.; and Ba, J. L. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Liu, P.; Li, J.; Wang, L.; Zha, Z.-J.; and Xiong, Z. 2024. MLP Embedded Inverse Tone Mapping. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, 9 pages. New York, NY, USA: ACM.

Loshchilov, I.; and Hutter, F. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*.

Mantiuk, R. K.; Hammou, D.; and Hanji, P. 2023. 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*.

Paszke, A. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Schanda, J. 2007. *Colorimetry: Understanding the CIE system*. John Wiley & Sons Press.

Shao, T.; Zhai, D.; Jiang, J.; and Liu, X. 2022. Hybrid conditional deep inverse tone mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1016–1024.

SMPTE. 2014. *ST 2084:2014 - SMPTE Standard - High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*. SMPTE, NY, USA.

Tian, Z.; Wang, F.; Wang, S.; Zhou, Z.; Zhu, Y.; and Shen, L. 2025. High Dynamic Range Video Compression: A Large-Scale Benchmark Dataset and A Learned Bit-depth Scalable Compression Algorithm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 7320–7330.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, C.; Pan, J.; Wang, W.; Fu, G.; Liang, S.; Wang, M.; Wu, X.-M.; and Liu, J. 2024. Correlation Matching Transformation Transformers for UHD Image Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5336–5344.

Wang, H.; Ye, M.; Zhu, X.; Li, S.; Zhu, C.; and Li, X. 2022. Kunet: Imaging knowledge-inspired single hdr image reconstruction. In *The 31st International Joint Conference On Artificial Intelligence (IJCAI/ECAI 22)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Xu, G.; Hou, Q.; Zhang, L.; and Cheng, M.-M. 2022. Fmnet: Frequency-aware modulation network for sdr-to-hdr translation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6425–6435.

Yang, X.; and Khoo, O. 2024. Projecting China to the world: Cinity, high frame rate cinema and the future of Chinese screening technology. *Asian Cinema*, 35(1-2): 63–79.

Yao, M.; He, D.; Li, X.; Pan, Z.; and Xiong, Z. 2023. Bidirectional translation between uhd-hdr and hd-sdr videos. *IEEE Transactions on Multimedia*.

Zeng, H.; Zhang, X.; Yu, Z.; and Wang, Y. 2020. SR-ITM-GAN: Learning 4K UHD HDR with a generative adversarial network. *IEEE Access*, 8: 182815–182827.

Zhang, L.; and Li, H. 2012. SR-SIM: A fast and high performance IQA index based on spectral residual. In *2012 19th IEEE international conference on image processing*, 1473–1476. IEEE.

Zhang, L.; Shen, Y.; and Li, H. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10): 4270–4281.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8): 2378–2386.

Zhang, X.; Zhu, Q.; Hu, T.; and Yan, Q. 2024. Eiffhdr: An efficient network for multi-exposure high dynamic range imaging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6560–6564. IEEE.