FMSTFnet: Feature-Modulation Spatio-Temporal Fusion Network for HDR Video Reconstruction

Wei Zhang
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
350192287@qq.com

Yeyao Chen
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
chenyeyao@nbu.edu.cn

Gangyi Jiang
Faculty of Info. Sci. & Eng.
Ningbo University
Ningbo, China
jiangggangyi@126.com

Abstract—**This paper proposes a novel High Dynamic Range video (HDRv) reconstruction method from Standard Dynamic Range video (SDRv), with a Feature Modulation Spatio-Temporal Fusion network (FMSTFnet). FMSTFnet has low-frequency and high-frequency parts with a pyramid structure. The low-frequency part mainly includes a Combined Global and Local Feature Modulation module (CGLFM) and a Spatio-Temporal Fusion Module (STFM). CGLFM modulates global and local features of SDR frames to correct the detail deviation caused by brightness differences in different regions and obtain preliminary HDR frames. STFM is designed to enhance the preliminary HDR frames using inter-frame information, and eliminate possible inter-frame artifacts. Finally, an adaptive hybrid module is constructed to fuse the low-frequency HDR frames and gradually extend the processed high-frequency information from low resolution to the higher. The proposed network fully utilizes the inter-frame information of multiple SDR frames and the contextual information of previously predicted HDR frames to generate high-quality results that are consistent in the temporal domain. The experimental results show that compared with many representative methods, the proposed method can reconstruct higher quality HDR videos.**

Keywords-high dynamic range video reconstruction; feature modulation; spatio-temporal fusion; transformer block.

I. INTRODUCTION

New generation displays can display visual contents with High Dynamic Range (HDR) and wide color gamut, providing a higher visual experience quality. However, at present, most video resources are still stored as Standard Dynamic Range videos (SDRv), resulting in a shortage of HDR video (HDRv) resources. Thus, generating HDRv from SDRv (SDRv-to-HDRv) is a challenging task [1][2].

For learning-based SDRv-to-HDRv, Kim et al. [3] proposed a method with separating input SDR frame into base and detail layers for different processing, which has the advantage of being easier to restore fine details. Subsequently, they integrated video super-resolution with SDRv-to-HDRv task to enhance texture details [4]. Chen et al. [5] designed a deep learning network for a single SDRv-to-HDRv task, which includes global feature modulation, local enhancement, and over-exposure compensation, and achieved good results. Wang et al. [6] proposed an SDRv-to-HDRv method with three sub-networks corresponding to the three processes in HDR imaging pipeline, to generate HDR images with rich global information. Xu et al. [7] constructed a frequency-aware modulation network that enhances the contrast of SDR to HDR conversion in a frequency adaptive manner, for reducing structural distortion and artifacts in the low-frequency regions. Xue et al. [8] proposed an improved residual block for extracting and fusing multi-layer features for fine-grained HDR image reconstruction. Guo et al. [1] constructed an HDRTV4K dataset and an HDR to SDR degradation model, and proposed a brightness segmentation network consisting of a global mapping backbone and two Transformer branches on the brightness range. The above methods mainly perform SDRv-to-HDRv tasks spatially. Many SDRv-to-HDRv methods mainly utilize a single SDR frame to generate corresponding HDR frame, which may lead to temporal inconsistency of HDRvs and produce annoying artifacts. Cao et al. [9] presented a kernel prediction network based SDRv-to-HDRv method, which utilizes multi-frame interaction modules to capture spatial information of multi-frame data and uses correction between adjacent frames to maintain inter-frame consistency.

In this paper, a novel SDRv-to-HDRv method with the design of Feature Modulation Spatio-Temporal Fusion network (FMSTFnet) is proposed. Its main contributions are summarized as follows: (1) A Combined Global and Local Feature Modulation module (CGLFM) is designed to perform macroscopic global and detailed local modulation on the current frame to reduce the color deviation of HDR video frames; (2) A Spatio-Temporal Fusion Module (STFM) is constructed, which can process contextual information in spatio-temporal domain, enhancing spatial results while reducing temporal inconsistencies. (3) Low-frequency and high-frequency information of SDRv are processed separately using a pyramid structure and fused with each other to obtain high-resolution output. Experimental results demonstrate the effectiveness of the proposed method.

The rest of the paper is organized as follows. Section 2 describes the proposed method in detail, Section 3 gives experimental results and analyses, and finally Section 4 concludes the paper.

II. THE PROPOSED METHOD WITH FMSFNET

A novel SDRv-to-HDRv method with the designed FMSTFNet is proposed, as shown in Figure 1. Aiming at the problem of color deviation, a CGLFM is designed by

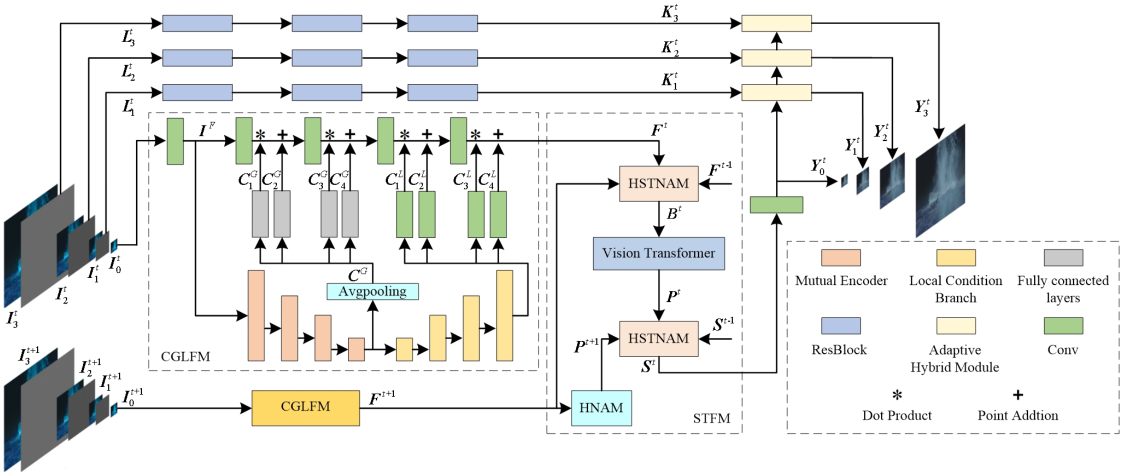


Figure 1. The proposed SDRv-to-HDRv method with the design of Feature Modulation Spatio-Temporal Fusion network (FMSTFnet).

combining adaptive feature modulation with Fourier convolution. For processing spatio-temporal information, a STFM is designed to fuse inter-frame features, and Transformer is employed to enhance the features, which can further reduce color deviation while eliminating temporal artifacts. The designed FMSFNet first establishes a pyramid structure and decomposes the input SDR frame into high-frequency component pyramids and low-frequency SDR frames. The low-frequency SDR frames are input to CGLFM and STFM to obtain low-frequency HDR frames. Residual blocks [10] are leveraged to reinforce the high-frequency components. Then, the enhanced high-frequency components are fused with low-frequency HDR frames using an Adaptive Hybrid Module (AHM), gradually expanding from low resolution to higher resolution results, and reconstructing the final high-resolution HDRv frame.

Specifically, for the t-th SDR frame $\boldsymbol{I}^t$, it is firstly decomposed into a Gaussian pyramid $\boldsymbol{M}_I^t$ =[$\boldsymbol{I}_0^t$, $\boldsymbol{I}_1^t$,… $\boldsymbol{I}_s^t$] and a high-frequency component pyramid $\boldsymbol{M}_L^t$ =[$\boldsymbol{L}_1^t$,… $\boldsymbol{L}_s^t$], where s is the number of downsampling. Similarly, $\boldsymbol{I}^{t+1}$ is also processed like $\boldsymbol{I}^t$. After that, the low-frequency components of $\boldsymbol{I}_0^t$ and $\boldsymbol{I}_0^{t+1}$ are respectively fed into CGLFM with weight sharing to obtain the preliminary HDR frames, denoted as $\boldsymbol{F}^t, \boldsymbol{F}^{t+1} = f_{CGLFM}(\boldsymbol{I}_0^t, \boldsymbol{I}_0^{t+1})$.

In Figure 1, $\boldsymbol{F}^t$ and $\boldsymbol{F}^{t+1}$ are then fed to STFM for spatio-temporal information enhancement; meanwhile, the (t-1)-th preliminary HDR frame $\boldsymbol{F}^{t-1}$ is also input to STFM to obtain the enhanced HDR frame $\boldsymbol{S}^t = f_{STFM}(\boldsymbol{F}^{t-1}, \boldsymbol{F}^t, \boldsymbol{F}^{t+1})$.

Each layer of the high-frequency component pyramid $\boldsymbol{M}_L^t$ is fed to multiple residual blocks $f_{Res}(\cdot)$, to enhance the high-frequency information, denoted as $\boldsymbol{K}_L^t = f_{Res}(\boldsymbol{M}_L^t)$. By relying on the high-frequency information $\boldsymbol{K}_L^t$ and the enhanced pyramid low-frequency HDR frame $\boldsymbol{S}^t$, high-resolution results can be reconstructed. Adaptive Hybrid Module (AHM) is used to fuse high-frequency component pyramids with low-frequency HDR frames, the final output pyramid $\boldsymbol{E}_L^t$ =[$\boldsymbol{Y}_0^t$, $\boldsymbol{Y}_1^t$,… $\boldsymbol{Y}_s^t$] is obtained, where $\boldsymbol{Y}_s^t$ denotes the reconstructed HDR frames, $\boldsymbol{Y}_s^t = f_{AHM}(\boldsymbol{K}_L^t, \boldsymbol{S}^t)$.

A. CGLFM

In the SDRv-to-HDRv task, there may be a phenomenon of uneven repair of over-exposed and under-exposed regions, as well as uneven color mapping from standard color gamut to wide color gamut. To address this issues, CGLFM, as shown in Figure 1, is designed, in which the global rough modulation is for roughness adjustment on images, while the local detail fine-tuning is for local detail enhancement. Specifically, the input SDR frame $\boldsymbol{I}^t$ is processed through two-layer convolution to obtain low dynamic range features $\boldsymbol{I}^F$, which will be modulated into high dynamic range features $\boldsymbol{F}^t$. CGLFM has two parts, namely, conditional generation module and feature modulation module. The conditional generation module can extract global and local information from features for modulation. Global conditional generation module uses Fourier convolution to perform global operations on input features, and then uses average pooling to downsample while reducing information loss, so as to obtain global information of the image. After five downsampling and global pooling, the feature $\boldsymbol{C}^G$ is get, denoted by $\boldsymbol{C}^G = f_{AVG}(f_{CGFM}(\boldsymbol{I}^F))$, $f_{CGFM}(\cdot)$ and $f_{AVG}(\cdot)$ are the global operation and global pooling, respectively.

By processing $\boldsymbol{C}^G$, global conditional features $\boldsymbol{C}_V^G$ (V=A,B) are obtained, which are used as the global modulation vectors. Local modulation requires local features that represent the corresponding pixel positions in the image. Here, through upsampling the global features five times and decoding from the encoded global information, the local conditional features $\boldsymbol{C}_V^L$ is obtained and expressed by $\boldsymbol{C}_V^L = f_{CLFM}(f_{CGFM}(\boldsymbol{I}^F))$, and $f_{CLFM}(\cdot)$ is the local operation.

Then, perform global rough modulation and local detail fine-tuning on the features. The former uses global features $\boldsymbol{C}_A^G$ to point-multiply the SDR feature $\boldsymbol{H}_G$ to achieve global

Copyright (c) IARIA, 2025. ISBN: 978-1-68558-245-6

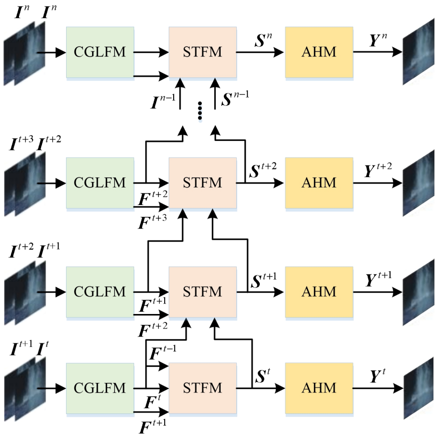
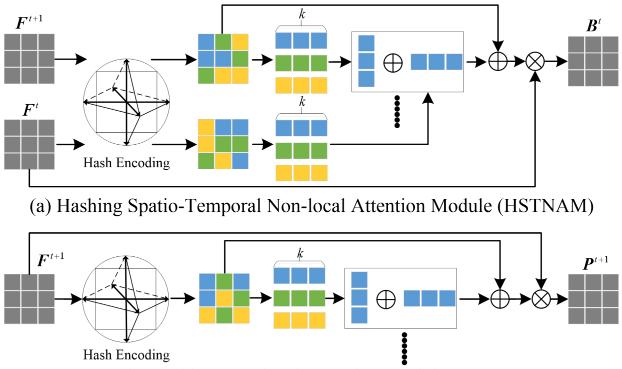


Figure 2. Information transmission approach of FMSTFNet.



(a) Hashing Spatio-Temporal Non-local Attention Module (HSTNAM)

(b) Hashing Non-local Attention Module (HNAM)

Figure 3. The used two non-local attention modules.

scaling, and directly adds $\boldsymbol{C}_B^G$ to achieve global displacement. The latter uses $\boldsymbol{C}_A^L$ to point-multiply the feature $\boldsymbol{H}_L$ to achieve local scaling, followed by adding $\boldsymbol{C}_B^L$ to achieve local displacement. After implementing local and global modulation, the features are converted to the HDR domain to obtain the preliminary HDR frame, which is expressed as

$$\boldsymbol{H}_L = \boldsymbol{C}_A^G * (\boldsymbol{H}_G) + \boldsymbol{C}_B^G \tag{1}$$

$$\boldsymbol{F}^t = \boldsymbol{C}_A^L * (\boldsymbol{H}_L) + \boldsymbol{C}_B^L \tag{2}$$

B. Spatio-Temporal Fusion Module (STFM)

STFM includes spatial and temporal reinforcement, mainly relying on the non-local attention mechanism. As shown in Figure 1, STFM mainly includes Hashing Spatio-Temporal Non-local Attention Module (HSTNAM), Hashing Non-local Attention Module (HNAM) [11], and Vision Transformer (ViT). To reduce resource consumption, when fusing inter-frame information in the temporal domain, only the information transmitted from the previous frame is used. Only the *t*-th and (*t*+1)-th frames are processed, and the (*t*-1)-th frame is obtained from the previous processing, as shown in Figure 2. Note that the (*t*-1)-th frame transmitted in the network is the intermediate feature rather than image. This processing can reduce the used memory while allowing the network to learn the entire sequence information. The input (*t*-1)-th frame contains the content of the previous video frames. As the number of input video frames increases, the network can learn all the early video frames.

STFM has four input features, i.e., $\boldsymbol{F}^t$, $\boldsymbol{F}^{t+1}$, $\boldsymbol{F}^{t-1}$ and $\boldsymbol{S}^{t-1}$. It has conducted two inter-frame information fusions, and with the deepening of the network, more deep level information is carried in the features. HSTNAM in Figure 3(a) is constructed to fuse the features of the *t*-th, (*t*-1)-th and (*t*+1)-th frames to obtain inter-frame information. Figure 3(b) represents the hashing non-local attention module, which differs from HSTNAM in that it only calculates spatial domain non-local attention. The purpose of STFM is to enhance features from both spatial and temporal perspectives, learn global inter-frame information to improve the temporal correlation of videos.

C. HDR Reconstruction and Loss Function

The FMSTFNet employs a pyramid structure, and the proposed method mainly focuses on handling the low-frequency components of the pyramid, which are processed using the above modules. For the high-frequency components, the stacked residual blocks are directly used for processing. AHM is constructed to facilitate rapid scaling of low resolution results. A lightweight module is designed as

$$\boldsymbol{Y}_{s+1}^t = h(\phi_2(cat(up(\phi_1(\boldsymbol{Y}_s^t)), \boldsymbol{K}_s^t))) \tag{3}$$

where $up(\cdot)$ is the bilinear interpolation, $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are two 3×3 convolutional layers, $cat(\cdot)$ is the channel concatenation, $\boldsymbol{K}_s^t$ is the high-frequency component of $\boldsymbol{I}^t$, and $h(\cdot)$ is the ReLU activation function.

The proposed loss function includes a multi-scale HDR reconstruction loss L_r and a perceptual loss L_p, expressed as

$$loss: L = \lambda_1 L_r + \lambda_2 L_p \tag{4}$$

where L_r represents the L_1 loss between the ground truth HDR image pyramid $\boldsymbol{H}_L$ and the predicted HDR image pyramid $\boldsymbol{Y}_L$. L_p is the L_1-norm difference between the intermediate feature maps when $\boldsymbol{Y}_L$ and $\boldsymbol{H}_L$ are separately fed into the pre-trained VGG19.

III. Experimental Results

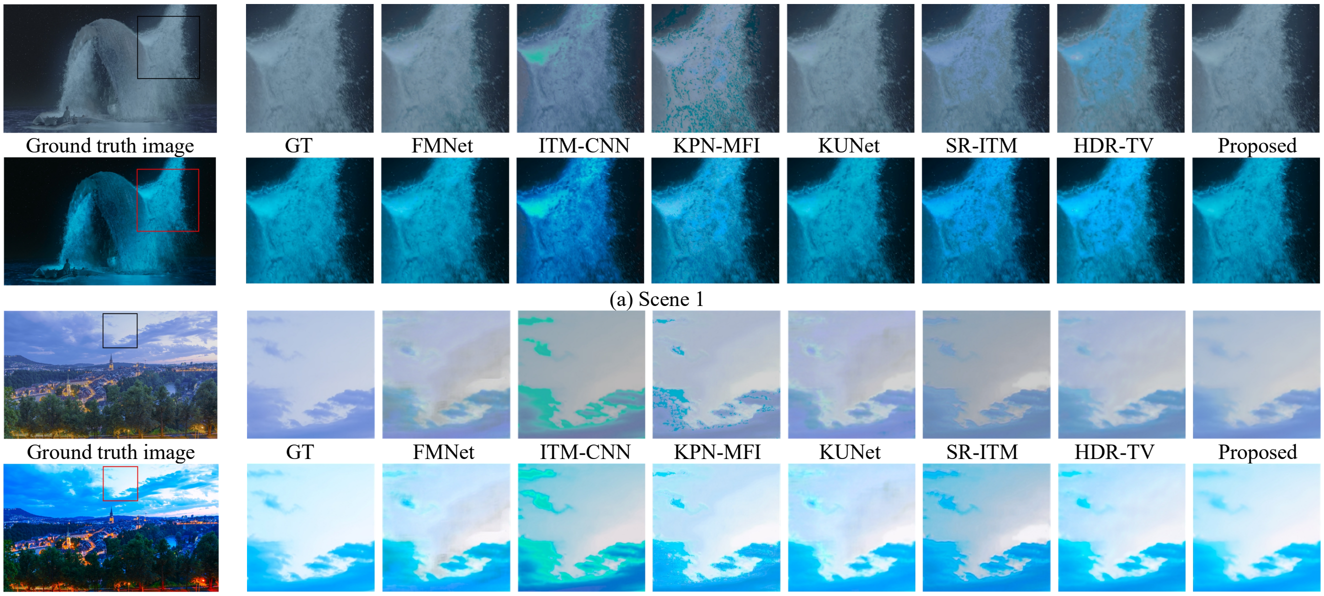
This section verifies and compares the proposed method with some representative methods including ITM-CNN [3], FMNet [7], KPN-MFI [9], KUNet [6], SR-ITM [4] and HDR-TV [5], and so on. Moreover, ablation experiment is constructed to investigate the role of the core modules of the proposed method. The proposed method is implemented with Pytorch, and the environment is configured with an Intel(R) Xeon(R) Silver4210 CPU, NVDIA RTX 3090Ti GPU. The proposed FMSTFnet is trained by the Adam optimizer, with β_1=0.9 and β_2=0.999. The batch size is 7, the initial learning rate is set to 0.0002, and it decays to 0.00001 after 100 epochs. The network parameters are initialized by the MSRA tool. A multi-frame SDRv-to-HDRv dataset is constructed for training and evaluation. 20 HDR10 standard HDR videos with 2160×3840 are collected from YouTube,

TABLE I. THE RESULTS OF THE PROPOSED METHOD COMPARED TO THE EXISTING REPRESENTATIVE METHODS

Methods	PSNR↑	SSIM↑	SR-SIM↑	LPIPS↓	ΔEITP↓	HDR-VDP↑
ITM-CNN [3]	29.96	0.9622	0.9358	12.73	22.354	8.0753
FMNet [7]	35.70	0.9811	0.9367	8.78	9.621	8.1787
KPN-MFI [9]	34.73	0.9645	0.9592	14.85	9.733	8.4039
KUNet [6]	35.72	0.9743	0.9419	9.58	10.458	8.2122
SR-ITM [4]	33.89	0.9782	0.9494	10.15	15.522	8.1667
HDR-TV [5]	37.45	0.9858	0.9650	6.53	8.947	8.6111
Proposed	**38.53**	**0.9880**	**0.9710**	**5.34**	**7.517**	**8.6806**

TABLE II. THE RESULTS OF AVERAGE PSNR, SSIM AND ΔE_{ITP} FOR DIFFERENT MODULES

CGFM	AHM	CLFM	HSTNAM1	ViT	HSTNAM2	PSNR↑	SSIM↑	ΔEITP↓
√						36.51	0.9824	9.730
√	√					37.60	0.9862	8.574
√	√	√				37.60	0.9866	8.647
√	√	√	√			37.70	0.9863	8.434
√	√	√	√	√		37.70	0.9869	8.415
√	√	√	√	√	√	**38.53**	**0.9880**	**7.517**



(a) Scene 1

(b) Scene 2

Figure 4. Visual effects of videos obtained by different SDRv-to-HDRv methods (Two partially enlarged regions are water splashes and the sky).

each with a corresponding SDR video. All videos are encoded using PQ curves and BT.2020 color gamut. 16 pairs of videos are used for training, and the remaining 4 pairs are used for testing. To evaluate the quality of generated HDR videos, six quality metrics are used, namely PSNR, SSIM, spectral residual based similarity (SR-SIM), learned perceptual image patch similarity (LPIPS), color difference indicator (ΔE_{ITP}), and HDR visual difference predictor (HDR-VDP).

Table I presents the objective comparison between the proposed method and representative methods, and the best results are presented in bold. The proposed method achieves better HDR video reconstruction performance, resulting in higher fidelity in spatial details and dynamic range of the reconstructed HDR video. The proposed method combines local and global features in the spatial domain and fuses inter-frame features in the temporal domain, this can better fit the nonlinear mapping process required for SDR frame to HDR frame reconstruction. The proposed method also achieves the best performance in ΔE_{ITP}, demonstrating the superiority of the proposed method in color restoration.

Figure 4 shows the visual effects of videos obtained by different methods. For each scene, the upper row shows the original HDR frames without tone mapping, while the lower is the tone mapped frames, similar to [4]. It can be found that the proposed method reconstructs the HDR images with higher visual quality and effectively restores the color information. For example, in the cloud region of the sky, the comparison methods produce significant visual artifacts. In contrast, the proposed method utilizes both local and global information to enhance the reconstruction results, thus more realistically reproducing the information of cloud region.

For the ablation experiments, Table II shows the results of average PSNR, SSIM and ΔE_{ITP} for different modules and their combination. Clearly, the proposed full network

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

achieves the best performance, which verifies the effectiveness of each module.

IV. Conclusions

We have proposed a new HDR video reconstruction method from SDR video method based on the design of Feature-Modulation Spatio-Temporal Fusion network (called FMSTFnet). The proposed method can fully utilize temporal and spatial information to reconstruct HDR video, improve the visual effect of the HDR video, and reduce its color deviation. The designed FMSTFnet has low-frequency and high-frequency parts with a pyramid structure, and combined global and local feature modulation module and spatio-temporal fusion module are constructed for eliminating possible inter-frame artifacts and color deviation. In future work, it will be extended to HDR light field reconstruction and angular consistency constraint will be explored to ensure better quality of reconstructed HDR light field images.

Acknowledgment

The work was supported by the National Natural Science Foundation of China under Grant 62271276.

References

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

[2] D. Vo, C. Liu and M. Nelson, “Extremely light-weight learning based LDR to PQ HDR conversion using bernstein curves,” IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2024, pp. 3910–3914

[3] S. Kim, D. Kim, and M. Kim, “ITM-CNN: Learning the inverse tone mapping from low dynamic range video to high dynamic range displays using convolutional neural networks,” Asian Conf. Comput. Vis., 2019, pp. 395–409

[4] S. Kim, J. Oh, and M. Kim, “Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications,” IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3116–3125.

[5] X. Chen, Z. Zhang, J. Ren, L. Tian, Y. Qiao, and C. Dong, “A new journey from SDRTV to HDRTV,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 4500–4509.

[6] H. Wang, M. Ye, X. Zhu, S. Li, C. Zhu, and X. Li, “KUNet: Imaging knowledge-inspired single HDR image reconstruction,” Int. Joint Conf. Artif. Intell. Eur. Conf. Artif. Intell., 2022, pp. 1408–1414.

[7] G. Xu, Q. Hou, L. Zhang, and M. Cheng, “FMNet: Frequency-aware modulation network for SDR-to-HDR translation,” ACM Int. Conf. Multimedia, 2022, pp. 6425–6435.

[8] L. Xue, T. Xu, Y. Song, Y. Liu, L. Zhang, X. Zhen, and J. Xu, “Lightweight improved residual network for efficient inverse tone mapping,” arXiv preprint arXiv:2307.03998, 2023.

[9] G. Cao, F. Zhou, H. Yan, A. Wang, and L. Fan, "KPN-MFI: A kernel prediction network with multi-frame interaction for video inverse tone mapping", Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, 2022, pp. 806–812.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

[11] N. Kitaev, L. Kaiser, and A. Levskaya, “Reformer: The efficient transformer,” arXiv preprint arXiv: 2001.04451, 2020.