KPN-MFI: A Kernel Prediction Network with Multi-frame Interaction for Video Inverse Tone Mapping

Gaofeng Cao[1,3] , **Fei Zhou**[2,3,5,6,7,*] , **Han Yan**[4] , **Anjie Wang**[1,3] and **Leidong Fan**[1,3]
[1]School of Electronic and Computer Engineering, Peking University
[2]College of Electronic and Information Engineering, Shenzhen University
[3]Peng Cheng Laboratory
[4]Department of Computer Science, Harbin Institute of Technology
[5]Guangdong Key Laboratory of Intelligent Information Processing
[6]Shenzhen Key Laboratory of Digital Creative Technology
[7] Shenzhen Institute for Artificial Intelligence and Robotics for Society (AIRS)
gaofengcao@pku.edu.cn, flying.zhou@163.com, 20b351014@stu.hit.edu.cn, {ajwang, fanleidong}@stu.pku.edu.cn

Abstract

Up to now, the image-based inverse tone mapping (iTM) models have been widely investigated, while there is little research on video-based iTM methods. It would be interesting to make use of these existing image-based models in the video iTM task. However, directly transferring the image-based iTM models to video data without modeling spatial-temporal information remains nontrivial and challenging. Considering both the intra-frame quality and the inter-frame consistency of a video, this article presents a new video iTM method based on a kernel prediction network (KPN), which takes advantage of multi-frame interaction (MFI) module to capture temporal-spatial information for video data. Specifically, a basic encoder-decoder KPN, essentially designed for image iTM, is trained to guarantee the mapping quality within each frame. More importantly, the MFI module is incorporated to capture temporal-spatial context information and preserve the inter-frame consistency by exploiting the correction between adjacent frames. Notably, we can readily extend any existing image iTM models to video iTM ones by involving the proposed MFI module. Furthermore, we propose an inter-frame brightness consistency loss function based on the Gaussian pyramid to reduce the video temporal inconsistency. Extensive experiments demonstrate that our model outperforms state-of-the-art image and video-based methods. The code is available at https://github.com/caogaofeng/KPN-MFI.

1 Introduction

High dynamic range is the most pervasive feature of the modern televisions (TVs). The effect of the HDR video rendered on the HDR display device is clearly approximating the human eye seeing the real world. Nevertheless, the majority of the transmitted visual contents are still in SDR format, e.g., the Digital TV and Internet Protocol TV (IPTV). Therefore, there is an urgent demand to convert the legacy SDR contents to their HDR version.

Transforming SDR contents into HDR ones is frequently referred to as inverse tone mapping (iTM). The problem of estimating HDR images from SDR photographs has been studied for many years [Liu *et al.*, 2020; Santos *et al.*, 2020; Cao *et al.*, 2021]. However, the video-based iTM task is challenging due to the video data always containing more complex scene contexts, motion, etc., in comparison with image data. It is difficult to model both spatial and temporal information simultaneously for video sequence data. As a result, some existing video-based iTM methods only focus on enhancing the quality of a single frame, ignoring the temporal context information between consecutive frames. In general, for video data, inter-frame temporal consistency is even more important than intra-frame visual quality, to a certain extent. However, these existing video-based iTM methods which are only based on single frame processing, cannot capture available information conveyed by the context information in the neighboring frames, resulting in compromised quality of results and cannot control the temporal consistency between frames.

To overcome this challenge, we propose a video iTM method taking advantage of both spatial and temporal information. Different from these previous image-based iTM methods which aim to predict the relative luminance of the scenes in the linear domain, our model directly predicts HDR videos in the HDR display format in the pixel domain, including the color gamut must be expanded from BT.709 [ITU-R, 2015a] to BT.2020 [ITU-R, 2015b], the bit-depth increased from 8 bit/pixel to 10 bit/pixel, and the optical-electro transfer function (OETF) also changes from gamma [ITU-R, 2011] to Perceptual Quantizer (PQ) [ITU-R, 2014]. To this end, we propose to model spatial-temporal information simulta-

[*] Corresponding author

neously in our video iTM method by integrating the multi-frame interaction (MFI) module into an encoder-decoder kernel prediction network (KPN). Firstly, a kernel prediction network framework is adopted in order to preserve the quality of the single frame video frames and improve the efficiency of the model by using a parameter sharing mechanism. Secondly, the multi-frame interaction module is designed to capture short-term spatial-temporal dependence between adjacent video frames. With the feature between adjacent frames exploited, not only the consistency between frames can be maintained, but also the quality of a single frame is improved. Additionally, any image-based method can capture the spatial and temporal features of a video sequence by introducing the MFI module, so that it can be readily extended to a video iTM method. Furthermore, a brightness consistency loss is developed to optimize our model to focus more on the brightness consistency between adjacent frame video frames.

Our contributions are summarized as follows:

- A multi-frame interaction module is introduced for modeling spatial-temporal context information, which aims at the feature interaction between frames through leveraging feature re-calibration based on the spatial-temporal relation between consecutive frames.
- A kernel prediction network framework is proposed not only to ensure the intra-frame iTM ability of the encoder-decoder model for single frames, but also to keep the model efficient by using the parameters sharing mechanism.
- A brightness consistency loss function based on the Gaussian pyramid is proposed in order to preserve the brightness consistency between adjacent video frames.

2 Related Work

2.1 Image-based iTM

With the rapid development of Convolutional Neural Networks (CNNs) in many computer vision tasks, (e.g., super-resolution [Zhang *et al.*, 2021] and denoise [Lin *et al.*, 2021]), CNN-based iTM methods have been developed. For instance, Liu et al., [Liu *et al.*, 2020] use multiple sub-networks to model the reverse pipeline of the camera producing SDR images from HDR. [Santos *et al.*, 2020] reconstruct HDR images by using a CNN with masked features and perceptual loss. HDRUNet [Chen *et al.*, 2021b] predicts a non-linear 16-bit HDR image with three sub-networks. Cao et al., [Cao *et al.*, 2021] adopts pixel-wise kernel convolution of the input SDR image to generate the HDR results to avoid some artifacts in the under/over-exposed regions.

However, these image-based iTM methods usually aim to predict the relative luminance of a scene in the linear domain. Thereby, these image based methods can not directly output display format data and post-processing (e.g., gamut mapping, EOTF, quantization) must be performed.

2.2 Video-based iTM

In general, extant deep learning based iTM literature primarily focuses on HDR image reconstruction. Very few studies [Chen *et al.*, 2021c; Kim *et al.*, 2019; Kim *et al.*, 2020] have been proposed to address the video iTM tasks. One of the latest video iTM methods HDRTVNet [Chen *et al.*, 2021c] use three sub-networks to simulate the key steps in the HDR video generation process. [Kim *et al.*, 2020; Kim *et al.*, 2019] divided the input SDR frame into base and detail layers and then reconstruct the detail and enhancement local contrast, respectively. Nevertheless, all the above approaches can be seen as single-frame iTM approaches, as they do not use any advantageous information available in the neighboring frames. As a result, these single-frame methods may cause temporal inconsistency and has limited performance in single-frame.

To solve these problems of existing video iTM methods, in this paper, we design a lightweight model, in which the parameters are shared in the encoder and decoder model for adjacent frames. With the assistance of the multi-frame interaction module, our model exploits the spatial-temporal context information to improve single frame reconstruction quality and reduce temporal inconsistency between adjacent frames.

3 Proposed Approach

It is well known that, in video-based reconstruction tasks, we need to consider both the spatial structure within a single frame and the temporal consistency between adjacent frames. However, existing video-based iTM methods ignore the inter-frame information. Considering both intra-frame reconstruction ability and inter-frame brightness coherency, we design a kernel prediction network with multi-frame interaction, named KPN-MFI, which comprises an encoder–decoder backbone, multi-frame interaction module, and adaptive pixel-wise convolution module. Concretely, a basic encoder-decoder with an adaptive pixel-wise convolution module is adopted to guarantee the mapping quality within each frame. Moreover, the MFI module is designed to captures the inter-frame spatial-temporal context information and preserve consistency by exploiting the correction between adjacent frames. The detail of the proposed KPN-MFI is shown in Fig. 1. The KPN-MFI is organized under a parameter sharing mechanism. It means that adjacent frames share the same parameters of the backbone network.

3.1 Model Overview

As shown in Fig. 1, given three adjacent SDR frames (S_{t-1}, S_t, S_{t+1}), the KPN-MFI produces three HDR frames (H_{t-1}, H_t, H_{t+1}) simultaneously by passing them to the same encoder-decoder model. Specifically, for each frame of a sequential (S_{t-1}, S_t, S_{t+1}), multilevel features are extracted by the encoder. The feature extraction procedure can be written as:

$$\{\mathbf{X}_i^l, \mathbf{X}_i^h\} = \mathcal{E}(S_i), \quad i \in \{t-1, t, t+1\}, \tag{1}$$

where $\mathcal{E}(\cdot)$ denotes encoder model, $\mathbf{X}_i^l$ and $\mathbf{X}_i^h$ denote the extracted low-level and high-level feature maps, respectively.

In the encoder, low-level features $\mathbf{X}_i^l$ are with high resolution and sensitive to the local variations. It means they would be beneficial to recover fine-grained information. Thus, as shown in Fig. 1, we include a long skip connection in the

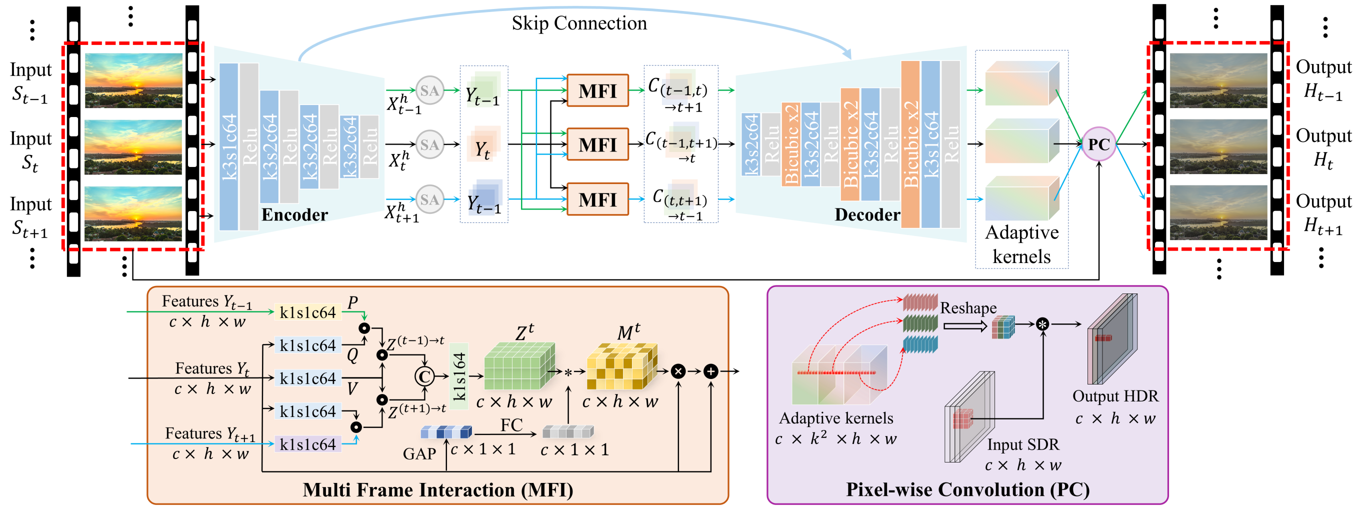


Figure 1: Overall architecture of the proposed KPN-MFI. k3s1c64 denotes a convolutional layer with kernel size 3, stride 1, and channels number 64. The same applies to k1s1c64. $\odot$ is matrix multiplication, $*$ is Hadamard product, $\otimes$ is element-wise multiplication, $\oplus$ is element-wise addition, $\circledast$ denotes convolution operation. Ⓒ denotes concatenate operation. c, h, w denote channel number, height, and width of frame or feature map respectively. GAP means global average poling operation and FC means fully connection operation.

backbone to pass $\mathbf{X}^l_i$ to the decoder for a fine intra-frame reconstruction. In contrast, high-level features are with larger receptive fields and thus more robust to feature displacements. Since feature displacements commonly appear between adjacent frames, we would like to apply the MFI modules to the high-level features. In this way, we can well capture the spatial-temporal context information without suffering from the alignment problem caused by moving objects in videos. Before the MFI modules, we first include self-attention (SA) modules [Zhang *et al.*, 2019b] to map the high-level features $\mathbf{X}^h_i$ to re-calibrated features denoted as $\mathbf{Y}_i$. Subsequently, the re-calibrated feature maps of sequential frames are further fed into the MFI module and output interacted feature maps. It can be expressed as:

$$\mathbf{C}_{(t-1,t+1)\to t} = MFI(\mathbf{Y}_{t-1}, \mathbf{Y}_t, \mathbf{Y}_{t+1}), \qquad (2)$$

where $\mathbf{C}_{(t-1,t+1)\to t}$ denotes the interacted feature of the frame S_t obtained from the intra-frame features $\mathbf{Y}_{t-1}$ and $\mathbf{Y}_{t+1}$. $MFI(\cdot)$ is the MFI module, which will be described in Section 3.2.

In the decoder, a set of adaptive kernels is produced from the interacted features and low-level features, which is formulated as $\mathcal{K}_t = \mathcal{D}(\mathbf{C}_{(t-1,t+1)\to t}, \mathbf{X}^l_t)$, where $\mathcal{D}(\cdot)$ is decoder model and $\mathcal{K}_t \in \mathbb{R}^{c\times k^2\times h\times w}$ is a tensor consisting of pixel-wise kernels. The output HDR frame is produced by a pixel-wise convolution (PC) operation on the input SDR frame, as shown in the purple block in Fig. 1. It has been demonstrated that potential artifacts can be effectively alleviated by predicting pixel-wise kernels rather than HDR pixel [Cao *et al.*, 2021]. The PC operation is in the form of

$$H^{c,i}_t = \sum_{j\in\Omega(i)} \mathcal{K}^{c,i}_t[\mathbf{p}_i - \mathbf{p}_j]\cdot S^{c,j}_t, \qquad (3)$$

where $H^{c,i}_t \in \mathbb{R}$ is the output HDR frame H at i^{th} pixel and c^{th} channel. $\Omega(i)$ denotes the $k\times k$ convolution window around i^{th} pixel. $\mathcal{K}^{c,i} \in \mathbb{R}^{k\times k}$ denoting the kernel at i^{th} pixel and c^{th} channel. $\mathbf{p}_i$ denotes 2D pixel coordinates and $[\mathbf{p}_i - \mathbf{p}_j] \in \{(-\frac{k-1}{2}, -\frac{k-1}{2}), (-\frac{k-1}{2}, -\frac{k-1}{2}+1), ..., (\frac{k-1}{2}, \frac{k-1}{2})\}$ denotes the position offset between i^{th} and j^{th} pixels.

It is worth noting that if we exclude the MFI modules in Fig. 1, it is essentially a deep architecture designed for image iTM. The MFI module is proposed to encourage inter-frame interactions and capture spatial-temporal context information. Consequently, the developed MFI module can be embedded in any other image iTM architecture to achieve a successful video iTM, as demonstrated in Section 3.2.

3.2 Multi-Frame Interaction Module

Inspired by the non-local block in [Wang *et al.*, 2018] capturing long-range dependencies, a multi-frame interaction (MFI) module is proposed to fully utilize the spatial-temporal context information. Different from the original non-local block, the MFI pays attention to the relationship between adjacent frames (S_{t-1}, S_{t+1}) and the current frame S_t. Intuitively, adjacent frame information can compensate for the current frame information reconstruction in both spatial and channel domains, which motivates us to highlight the relationship between multiple frames in spatial and channel domains. Therefore, the spatial and channel interaction operation (i.e., cosine distance and channel attention) between adjacent frames is introduced to make the model able to capture the spatial-temporal context information. Specifically, as shown in the orange block in Fig. 1, we first use 1×1 convolution layer to extract the features $\mathbf{P} \in \mathbb{R}^{c\times h\times w}$ from the re-calibrated high-level feature $\mathbf{Y}_{t-1}$, and $(\mathbf{Q}, \mathbf{V}) \in \mathbb{R}^{c\times h\times w}$ from $\mathbf{Y}_t$. Then, we reshape $(\mathbf{P}, \mathbf{Q})$ to $\mathbb{R}^{c\times m}$ and $\mathbf{V}$ to $\mathbb{R}^{m\times c}$, i.e., $\mathbf{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_m]$, $\mathbf{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, ..., \boldsymbol{q}_m]$, and $\mathbf{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_c]$, where $m = h\times w$ is the number of spatial positions on each feature map. $(\boldsymbol{p}_i, \boldsymbol{q}_i) \in \mathbb{R}^c$ and $\mathbf{v}_i \in \mathbb{R}^m$ are the feature vectors at the i^{th} spatial position in $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{V}$ respectively.

The relevance between $\mathbf{P}$ and $\mathbf{Q}$ is computed with cosine distance and get the correlation map $\mathbf{R} \in \mathbb{R}^{m\times m}$:

$$\mathbf{R}_{i,j}^{(t-1)\to t} = \left(\frac{\boldsymbol{p}_i}{||\boldsymbol{p}_i||_2}\right)^T \left(\frac{\boldsymbol{q}_i}{||\boldsymbol{q}_i||_2}\right), \quad i,j = 1,...,m. \quad (4)$$

The interaction relationship between frame S_{t-1} and frame S_t is measured as:

$$\mathbf{Z}_{i,j}^{(t-1)\to t} = \sum \boldsymbol{r}_i \cdot \boldsymbol{v}_j, \quad i = 1,...,m, j = 1,...,c, \quad (5)$$

where $\mathbf{Z}^{(t-1)\to t} \in \mathbb{R}^{m\times c}$, $\boldsymbol{r}_i$ is i^{th} row of $\mathbf{R}$, and $\boldsymbol{v}_j$ is j^{th} column of $\mathbf{V}$. The feature $\mathbf{Z}^{(t-1)\to t}$ is reshaped to $\mathbb{R}^{c\times h\times w}$. Likewise, $\mathbf{Z}^{(t+1)\to t} \in \mathbb{R}^{c\times h\times w}$ can be calculated in the same manner.

Then, the reshaped interaction relationship features $\mathbf{Z}^{(t-1)\to t}$ and $\mathbf{Z}^{(t+1)\to t}$ are fusion using 1×1 convolution layer and obtain multi-frame interaction relationship map $\mathbf{Z}^t$, which can be regarded as spatial attention in multi-frame context information. In order to further preserve the channel attention between $\mathbf{Z}^t$ and Y_t, a channel attention module is adopted as function:

$$\mathbf{M}^t = \omega * \mathbf{Z}^t, \quad \omega = FC(GAP(\mathbf{Y}_t)), \quad (6)$$

where $*$ presents hadamard product, GAP and FC denote global average pooling operation and fully connection operation, respectively. $\mathbf{M}^t \in \mathbb{R}^{c\times h\times w}$, $\omega \in \mathbb{R}^{c\times 1\times 1}$ is employed to learn the channel-wise weights for $\mathbf{Y}_t$ in terms of preserving global properties.

At last, we use a residual attention mechanism, where the initial feature maps $\mathbf{Y}_t$ are element-wisely weighted by (1 + $\mathbf{M}^t$) for preserving feature consistency and compensating information. Thus, the multi-frame interaction feature map with spatial-temporal information can be formulated as:

$$\mathbf{C}_{(t-1,t+1)\to t} = (\mathbf{M}^t + 1) \otimes \mathbf{Y}_t. \quad (7)$$

In this way, we capture inter-frame correlation according to the spatial-temporal pixel-wise relation between multiple sequential frames. Similarity, we can get the multi-frame interaction feature map $\mathbf{C}_{(t,t+1)\to t-1}$ or $\mathbf{C}_{(t-1,t)\to t+1}$.

3.3 Loss Function

Our loss function contains an intra-frame content loss $\mathcal{L}_\mathcal{C}$ and an inter-frame brightness consistency loss $\mathcal{L}_\mathcal{T}$. Specifically, the intra-frame content loss function is formed as:

$$\mathcal{L}_\mathcal{C} = \sum_i \left|\left|Tanh(H_i) - Tanh(H_i^{'})\right|\right|_1, i \in \{t-1, t, t+1\}, \quad (8)$$

where $H^{'}$ denotes the ground truth HDR frame, $Tanh(\cdot)$ function is adopted to balance the low and high luminance values impact.

In order to ensure the temporal consistency further, we propose an inter-frame brightness consistency loss $\mathcal{L}_\mathcal{T}$. Some existing inter-frame consistency loss uses the optical flow to align the frame, e.g., temporal stability loss in [Lai *et al.*, 2018] and Long-term temporal loss in [Zhang *et al.*, 2019a]. However, the current best inter-frame alignment methods (e.g., [Hui and Loy, 2020] and [Chen *et al.*, 2021a]) have a limited effect on complex scenes. If we calculate the consistency loss between unaligned frames, it will reduce the model performance. To alleviate the impact of misalignment, we propose decomposing the predicted frame with a Gaussian pyramid. With the decomposition process of the Gaussian pyramid, the misaligned high-frequency information between the misaligned frames is gradually eliminated. Concretely, we first align the predicted frame H_{t-1} and H_{t+1} to H_t (defined as $H_{(t-1)\to t}$ and $H_{(t+1)\to t}$) by an optical flow model Liteflownet3 [Hui and Loy, 2020]. Then, we obtain the illumination component ($B_{(t-1)\to t}$, $B_{(t+1)\to t}$) of aligned frames ($H_{(t-1)\to t}$, $H_{(t+1)\to t}$) according to the classic Retinex theory [Land and McCann, 1971]. Finally, the illumination component is decomposed in a Gaussian pyramid, and the loss is calculated for each layer of the pyramid composition. The inter-frame brightness consistency loss $\mathcal{L}_\mathcal{T}$ is defined as:

$$\mathcal{L}_\mathcal{T} = \sum_{j=1}^{l} \gamma_j \left[\left|\left|B_{(t-1)\to t}^j - B_t^j\right|\right|_1 + \left|\left|B_{(t+1)\to t}^j - B_t^j\right|\right|_1 \right], \quad (9)$$

where j means the number of pyramid decomposition layers and l is the total number of the pyramid decomposition layers (we set $l = 5$ in our experiments), B^j is j^{th} illumination component, we use bilateral filter (with parameter $d = 9, sigmaColor = 10, sigmaSpace = 5$) to filter the frame to approximate the illuminance image in logarithm domain. γ_j is coefficients to measure the importance of the low-frequency pyramid component. During the pyramid decomposition process, the component with the smaller resolution carries the less high-frequency detail information and the more accurate representation of low-frequency information (such as global brightness), and the less misaligned information between frames. Therefore, we set $[\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5] = [1/16, 1/8, 1/4, 1/2, 1]$. The Gaussian function with ($ksize = 9$, $\sigma = 1.7$).

The overall loss function to optimize the model is given by:

$$\mathcal{L} = \mathcal{L}_\mathcal{C} + \alpha \mathcal{L}_\mathcal{T}, \quad (10)$$

where α is a constant to balance $\mathcal{L}_\mathcal{C}$ and $\mathcal{L}_\mathcal{T}$ ($\alpha = 0.05$ in our implementation).

4 Experiments

4.1 Experimental Settings

Dataset. We collect 21 pairs of 4K-UHD HDR videos under HDR10 standard and their SDR counterpart from YouTube, as in [Chen *et al.*, 2021c; Kim *et al.*, 2020]. 17 video pairs are used for training and the left 4 videos for testing. We sample three consecutive frames every two seconds of each video, and then we delete the frames with scene transitions. Thus, 3,241 pairs are generated for training while 482 pairs for testing are generated. The frames are cropped to 1024×1024 with overlapping, and 42, 608 pairs are generated totally for training.

Evaluation Metrics. In our experiments, we assess the performance in terms of the Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) [Wang *et al.*, 2004], ΔE_{ITP} [ITU-R, 2019] and HDR-VDP3 [Mantiuk *et al.*,

Method		PSNR ↑	SSIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑	MABD ↓	Params ↓	GFLOPs ↓
image-based	HuoPhysEO	21.0278	0.9133	69.6043	7.4001	0.0147	–	–
	DeepHDR	28.0664	0.9518	29.6662	7.0987	<u>0.0122</u>	51.5419	**18.9459**
	KPNiTM	33.6523	0.9591	12.3999	8.3062	0.0150	35.8732	73.4167
	HDRUnet	36.6332	0.9823	<u>9.4729</u>	8.3054	0.0125	<u>1.6515</u>	<u>23.4223</u>
video-based	SRiTM	36.3970	0.9838	10.6946	8.0513	0.0131	2.6338	182.5195
	JSINet	36.2092	0.9787	10.4287	7.9656	0.0127	**1.2493**	81.8757
	HDRTVNet	<u>37.2615</u>	**0.9868**	9.5365	<u>8.4408</u>	0.0126	38.1977	55.5621
	Ours	**37.7260**	<u>0.9856</u>	**9.1975**	**8.5516**	**0.0119**	3.3735	29.0289

Table 1: Quantitative results of our method comparison with state-of-the-art methods. Bold text indicates the best result, and underlined text indicates the best performing state-of-the-art method.

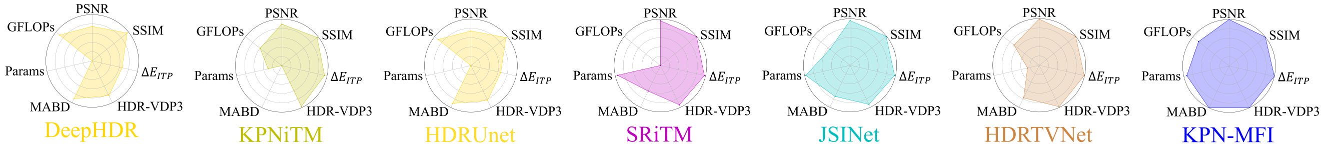


Figure 2: Radar charts visualizing Table 1. Values are normalized to the unit range, and axes is inverted so that a higher value is always better.

2011]. ΔE_{ITP} and HDR-VDP3 are designed for HDR videos. Furthermore, we choose the MADB in [Jiang and Zheng, 2019] to validate the temporal consistency of models.

4.2 Experimental Results

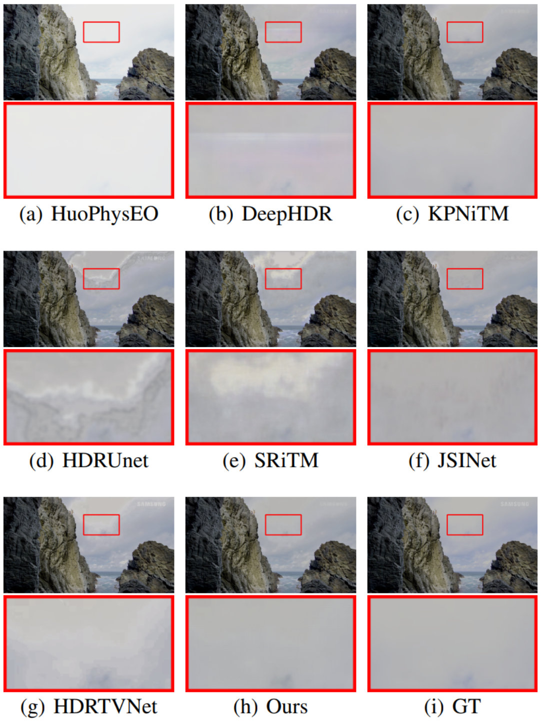
Qualitative Evaluations

To evaluate the performance of the proposed approach, we compare our approach against several existing state-of-the-art video-based iTM methods, including SRiTM (ICCV) [Kim *et al.*, 2019], JSINet (AAAI) [Kim *et al.*, 2020] and HDRTVNet (ICCV) [Chen *et al.*, 2021c]. Since there are few video-based iTM methods, we also compare our approach with several existing image-based methods, including HuoPhysEO [Huo *et al.*, 2014], DeepHDR (TOG) [Santos *et al.*, 2020], KPNiTM [Cao *et al.*, 2021], and HDRUnet (CVPR) [Chen *et al.*, 2021b]. All these methods are retrained on our dataset.

As shown in Table 1 and Fig. 2, our method KPN-MFI outperforms other methods by a large margin on PSNR, ΔE_{ITP}, and HDR-VDP3, but a slight decrease on SSIM. Quantitative results show the effectiveness of the proposed method. Besides, the calculation cost and parameters are compared. Our model can achieve a good balance between efficiency and performance. We can also see in Fig. 2, our method is with balanced performances in the context of all the 7 metrics.

Quantitative Evaluations

Furthermore, visual results are listed in Fig. 3 and Fig. 4. In comparison with other models, the visual results show that our model reduces the noticeable artifacts, i.e., blurring, oversharping, false contour, and incorrect color. We believe that this is attributed to the fact that our model introduces the MFI module to fully utilize spatial-temporal context information of adjacent frames to reduce these artifacts and synthesize visually pleasing textures. Besides, our approach adopts an adaptive pixel-wise kernel prediction strategy, which predicts an adaptive kernel for each pixel of the input SDR frame rather than directly estimate the HDR pixel values. More visual results can be found in our online website[1].



(a) HuoPhysEO (b) DeepHDR (c) KPNiTM

(d) HDRUnet (e) SRiTM (f) JSINet

(g) HDRTVNet (h) Ours (i) GT

Figure 3: Qualitative comparisons on over brightness scenes.

4.3 Ablation Study

Effectiveness of MFI module for image-based model. In order to assess the effectiveness of the MFI module, we introduce the MFI module into the existing image-based deep

[1] http://www.vista.ac.cn/kpn-mfi/

Method	PSNR ↑	SSIM ↑	ΔE_{ITP} ↓	HDR-VDP3 ↑	MABD ↓
MFI-DeepHDR	30.0085 (+1.9421)	0.9472 (-0.0046)	24.4903 (+5.1759)	6.9060 (-0.1927)	0.0121 (+0.0001)
MFI-KPNiTM	35.0250 (+1.3727)	0.9681 (+0.0090)	10.9689 (+1.4301)	8.2842 (-0.0220)	0.0136 (+0.0014)
MFI-HDRUnet	36.9134 (+0.2802)	0.9846 (+0.0023)	9.4474 (-0.0255)	8.3632 (+0.0578)	0.0125 (+0.0000)

Table 2: Quantitative results of the image-based method with MFI module. The value in parentheses represents the amount of change in the metrics compared to the original methods.

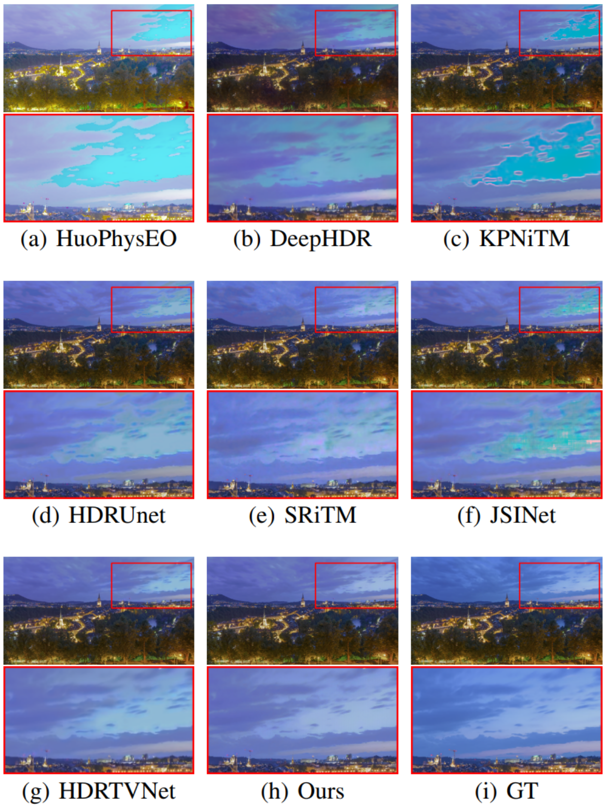


Figure 4: Qualitative comparisons on dark light scenes.

learning methods (DeepHDR, KPNiTM, and HDRUnet) to capture the spatial-temporal information between adjacent SDR frames, named MFI-DeepHDR, MFI-KPNiTM, and MFI-HDRUnet, respectively. We simply introduce the MFI module into the last layer of the down-sampling for these three models. As shown in Table 2, the performance of the above three methods with the MFI module has improved. It clearly demonstrates the effectiveness of the MFI module.

Effectiveness of multi-frame interaction module. As shown in Table 3, in comparison with baseline model #1, the model #2 with the MFI module achieves and approximately yields 6.82%, 1.18%, 19.13%, and, 10.22% performance gain in terms of PSNR, SSIM, HDR-VDP3, and MABD on the testing dataset. It is clearly demonstrated that the MFI module plays a crucial role in improving performance. As the features of adjacent frames interact in the MFI module, the spatial-temporal information is captured simultaneously and the multi-frame information is leveraged to generate a frame HDR output.

	model #1	model #2	model #3
MFI module	×	✓	✓
Consistency loss	×	×	✓
PSNR	34.9014	37.2804	**37.7260**
SSIM	0.9697	0.9811	**0.9856**
ΔE_{ITP}	12.3194	9.9622	**9.1975**
HDR-VDP3	8.0115	8.3858	**8.5516**
MABD	0.0137	0.0123	**0.0119**

Table 3: Ablation study on the proposed KPN-MFI.

Effectiveness of brightness consistency loss. Compared with model #2, in Table 3, the model #3 not only has a performance gain of 3.25% in MABD score, but also has increases of 1.13%, 0.45%, 8.37%, and 1.98% in PSNR, SSIM, ΔE_{ITP} and HDR-VDP3, respectively. It is observed that performance can be further augmented by introducing a consistency loss. It may be due to the brightness consistency loss tends to encourage the MFI module to primarily focus on the related information between two frames and ignore the irrelevant information. Therefore, the brightness consistency loss based on Gaussian pyramid can not only improve the consistency of the model, but also help to improve the performance of the model to a certain extent.

5 Conclusion

In this paper, we propose a video iTM model which aims at reconstructing the brightness consistency of HDR video from SDR video. To this end, an MFI module is designed to capture the temporal-spatial feature from adjacent frames. Notably, we can readily extend the existing successful image-based iTM method to capture the temporal and spatial features of the video by using the MFI module. Meanwhile, a brightness consistency loss function is adopted to further improve the inter-frame brightness consistency of the model. Extensive experimental results demonstrate that our method consistently achieves superior performance in terms of various metrics and renders visually pleasing results. We believe the proposed KPN-MFI model can be generalized to other low-level vision video tasks, which will be explored in future work.

Acknowledgements

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011584).

References

[Cao *et al.*, 2021] Gaofeng Cao, Fei Zhou, Kanglin Liu, and Liu Bozhi. A brightness-adaptive kernel prediction network for inverse tone mapping. *Neurocomputing*, 464(13):1–14, 2021.

[Chen *et al.*, 2021a] Weitao Chen, Zhibin Wang, and Hao Li. Get better 1 pixel pck: Ladder scales correspondence flow networks for remote sensing image matching in higher resolution. In *IEEE/CVF CVPR*, pages 742–751, 2021.

[Chen *et al.*, 2021b] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *IEEE/CVF CVPRW*, pages 354–363, 2021.

[Chen *et al.*, 2021c] Xiangyu Chen, Zhengwen Zhang, Jimmy S Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from sdrtv to hdrtv. In *IEEE/CVF ICCV*, pages 4500–4509, 2021.

[Hui and Loy, 2020] Tak-Wai Hui and Chen Change Loy. Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In *ECCV*, pages 169–184. Springer, 2020.

[Huo *et al.*, 2014] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5):507–517, 2014.

[ITU-R, 2011] ITU-R. Reference electro-optical transfer function for flat panel displays used in hdtv studio production. *ITU-R Rec, BT.1886*, 2011.

[ITU-R, 2014] ITU-R. High dynamic range electro-optical transfer function of mastering reference displays. *SMPTE ST2084:2014*, 2014.

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

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

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

[Jiang and Zheng, 2019] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *IEEE/CVF CVPR*, pages 7324–7333, 2019.

[Kim *et al.*, 2019] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep sr-itm: Joint learning of super-resolution and inverse tone-mapping for 4k uhd hdr applications. In *IEEE/CVF ICCV*, pages 3116–3125, 2019.

[Kim *et al.*, 2020] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Jsi-gan: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for uhd hdr video. In *AAAI*, volume 34, pages 11287–11295, 2020.

[Lai *et al.*, 2018] Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *ECCV*, pages 170–185, 2018.

[Land and McCann, 1971] Edwin H. Land and John J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, 1971.

[Lin *et al.*, 2021] Huangxing Lin, Yihong Zhuang, Yue Huang, Xinghao Ding, Xiaoqing Liu, and Yizhou Yu. Noise2grad: Extract image noise to denoise. In *IJCAI*, pages 830–836, 2021.

[Liu *et al.*, 2020] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *IEEE/CVF CVPR*, pages 1651–1660, 2020.

[Mantiuk *et al.*, 2011] 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 TOG*, 30(4):1–14, 2011.

[Santos *et al.*, 2020] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM TOG*, 39(4):1–10, 2020.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[Wang *et al.*, 2018] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE/CVF CVPR*, pages 7794–7803, 2018.

[Zhang *et al.*, 2019a] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep exemplar-based video colorization. In *IEEE/CVF CVPR*, pages 8052–8061, 2019.

[Zhang *et al.*, 2019b] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, volume 97, pages 7354–7363, 2019.

[Zhang *et al.*, 2021] Guoqing Zhang, Yuhao Chen, Weisi Lin, Arun Chandran, and Xuan Jing. Low resolution information also matters: Learning multi-resolution representations for person re-identification. In *IJCAI*, pages 1295–1301, 2021.