Compressed-SDR to HDR Video Reconstruction

Hu Wang, Mao Ye*, Xiatian Zhu*, Shuai Li, Xue Li, and Ce Zhu

Abstract—The new generation of organic light emitting diode display is designed to enable the high dynamic range (HDR), going beyond the standard dynamic range (SDR) supported by the traditional display devices. However, a large quantity of videos are still of SDR format. Further, most pre-existing videos are compressed at varying degrees for minimizing the storage and traffic flow demands. To enable movie-going experience on new generation devices, converting the compressed SDR videos to the HDR format (i.e., *compressed-SDR to HDR* conversion) is in great demands. The key challenge with this new problem is how to solve the intrinsic many-to-many mapping issue. However, without constraining the solution space or simply imitating the inverse camera imaging pipeline in stages, existing SDR-to-HDR methods can not formulate the HDR video generation process explicitly. Besides, they ignore the fact that videos are often compressed. To address these challenges, in this work we propose a novel imaging knowledge-inspired parallel networks (termed as KPNet) for compressed-SDR to HDR (CSDR-to-HDR) video reconstruction. KPNet has two key designs: Knowledge-Inspired Block (KIB) and Information Fusion Module (IFM). Concretely, mathematically formulated using some priors with compressed videos, our conversion from a CSDR-to-HDR video reconstruction is conceptually divided into four synergistic parts: reducing compression artifacts, recovering missing details, adjusting imaging parameters, and reducing image noise. We approximate this process by a compact KIB. To capture richer details, we learn HDR representations with a set of KIBs connected in parallel and fused with the IFM. Extensive evaluations show that our KPNet achieves superior performance over the state-of-the-art methods. The dataset and code are available at https://wanghu178.github.io/KPNet/.

Index Terms—Standard Dynamic Range (SDR), High Dynamic Range (HDR), Compressed SDR video, Video Reconstruction.

1 Introduction

During the past few years, the popularity of high dynamic range (HDR) display devices in daily life is greatly increased. However, most of existing video contents are still in the standard dynamic range (SDR) format. In general, SDR refers to a dynamic range standard widely used in video and television industry, often the standard dynamic range defined under the ITU (International Telecommunication Union) and the ITU-R (International Television Union) standards. In comparison, LDR (Low Dynamic Range) refers to the images/videos with small dynamic range. Typically videos are highly compressed for saving the coding bit rate and storage space. SDR videos with compression artifacts cannot make the full use of the advantages of HDR display (wide color gamut, high peak brightness, high contrast, etc.), seriously reducing the quality of experience. Therefore, there is an urgent need to reconstruct the HDR version from compressed SDR videos. This task, denoted as ***Compressed-SDR to HDR (CSDR-to-HDR) video reconstruction***, is of great practical value, but receives little attention yet in the research community. There are two key causes. First, numerous standards are in place for HDR videos and they are not uniform. Second, no dataset exists for enabling model training and testing.

This work aims at promoting the development of this unexplored problem. CSDR-to-HDR video reconstruction

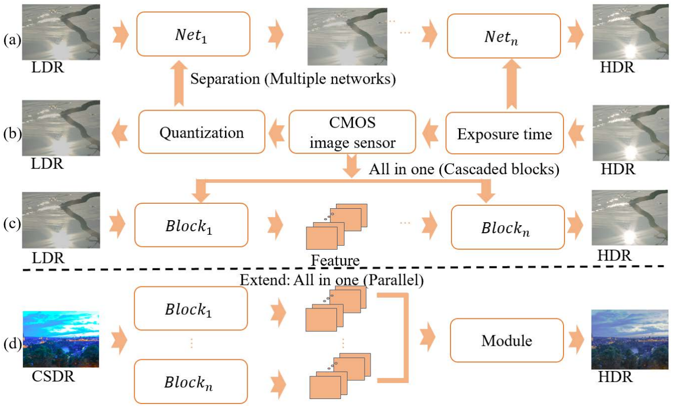


Fig. 1: Comparing our approach with the existing methods based on reversing the camera imaging pipeline. (a) Multiple neural networks are designed to approximate the stages of reversing camera imaging pipeline with (b) showing a representative camera imaging process. (c) Our preliminary method KUNet [1] uses a basic building block of UNet to simulate the LDR-to-HDR imaging formula. (d) Our KPNet introduced in this work further adopts a parallel structure to solve the shortcomings of the preliminary cascaded structure for solving the new more challenging CSDR-to-HDR video reconstruction task.

- *Hu Wang, Mao Ye are with School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China.*
E-mail: wanghu0833cv@gmail.com, maoye@uestc.edu.cn
- *Xiatian Zhu is with Surrey Institute for People-Centred Artificial Intelligence, CVSSP, University of Surrey, Guildford, UK.*
E-mail: xiatian.zhu@surrey.ac.uk
- *Shuai Li is with the School of Control Science and Engineering, Shandong University, Jinan 250000, PR China*
E-mail: shuaili@sdu.edu.cn
- *Xue Li is with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD 4072, Australia*
E-mail: xueli@itee.uq.edu.au
- *Ce Zhu is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, P.R. China*
E-mail: eczhu@uestc.edu.cn
- ** corresponding author.*

is a severely ill-posed problem since multiple mappings exist between the compressed SDR and HDR videos. While there are individual researches on SDR-to-HDR video reconstruction and compressed video quality enhancement, none is sufficiently effective for this more challenging CSDR-to-HDR video reconstruction problem. Moreover, going beyond simply combining the two constituent tasks, it presents a new many-to-many mapping problem.

For SDR-to-HDR video reconstruction, three different approaches have been proposed as follows. *Single-exposure HDR video reconstruction* uses single-exposure SDR frame to reconstruct HDR video [2]. The key of this method is how to restore the overexposed information and bit depth extension (see a common prior strategy in Fig. 1 (a)). *Multi-exposure stack HDR synthesis* constructs HDR video from sequences with alternating exposures. The problem with this approach [2] is how to align the video frames with different exposures and the acquisition of the data set. *The last methods* use different hardwares to construct HDR videos, for example, per-pixel exposure [3], modulo camera [4] and neuromorphic [5]. Obviously, these methods require special hardware for processing and are not universally applicable.

For the quality enhancement of compressed SDR videos, there exist two mainstream categories of methods: *single-frame* or *multi-frame* based. For the former (e.g., DNCNN [6], QE-CNN [7] and RBQE [8]), previous methods use the spatial information to enhance the quality, which can be adapted to videos by restoring each frame individually. For the latter, spatial-temporal information is used to enhance the frames [9], [10], typically achieving better performance than the single-frame based methods.

Although many works have emerged for the above two tasks respectively, their combination has not been investigated for appealing movie-going experience. In this work, we first investigate what is a good combination order of the two functions. When putting the quality enhancement before HDR video reconstruction, its causes the remaining compression artifacts are significantly amplified during video reconstruction. In case of the inverse order, the quality enhancement part becomes more challenging, since HDR video reconstruction is usually highly non-linear and tends to amplify the compression artifacts.

To address these challenges, in this work we propose a new ***Knowledge-inspired Parallel Network*** (KPNet) for CSDR-to-HDR video reconstruction, integrating both compressed SDR video prior knowledge and HDR imaging knowledge in a novel parallel structure (see Fig. 1 (d)). In particular, we model a compressed-SDR to HDR mathematical formula with a Knowledge-Inspired Block (KIB) module. For extracting strong HDR features, this KIB is designed compactly with the abilities for reducing compressing artifacts and recovering missing details in overexposed area, adjusting imaging parameters, reducing SDR imaging noise. With an elegant parallel design, we further fuse the output of multiple KIB modules for reconstructing richer details in HDR videos. Favorably, our model can solve the challenging ghosting artifacts problem in video reconstruction.

We summarize the *contributions* as follows. **(1)** We present a practical yet understudied problem, namely *CSDR-to-HDR video reconstruction*. This is a combination problem of quality enhancement and SDR-to-HDR video reconstruction, both of which have been studied in isolation, presenting more challenges than either. **(2)** By analyzing the camera imaging pipeline, we derive a HDR image restoration formula (see Fig. 1 (b) and (c)). We further extend this formula to the video domain. Combining with prior knowledge of compressed video enhancement, we reach a CSDR-to-HDR video reconstruction formula. This leads to the introduction of our KIB module design and finally a novel knowledge-inspired parallel network (KPNet). **(3)** For enabling quantitative evaluation of this new problem setting, we extend an existing dataset *HDRTV* [11] by applying additional standard compression (e.g., HM16.9). Extensive experiments show that our KPNet can outperform clearly the existing alternatives.

This work is an extended version of our IJCAI2022 paper [1]. We introduce several key differences: (1) Studying a more practical problem setting by additionally taking into account the video compression factor, typical to many pre-existing videos. This problem is largely ignored in the community but practically valuable and useful. (2) Extending our preliminary model KUNet [1] by architectural modification. In particular, our new model can address the problems of ghosting artifacts and color interruption suffered by KUNet. (3) Providing more comprehensive experimental analysis and discussion, including the rationale analysis of evolving KUNet to the model KPNet.

2 Relate Works

2.1 Multi-exposure Stack HDR Imaging

Using a series of exposure images to compose HDR images is one of the most common approaches [12]. We can divide these methods into three types: *sequential exposure* [13], [14], [15], *alternate exposure* and *with specialized hardware*. The first category produces a single HDR image using a sequence of exposed SDR images. For example, Kalantari and Ramamoorthi [13] aligns several SDR images with different exposures and feeds them into neural networks, laying the foundation for multi-exposure synthesis of HDR images using deep learning. [15], [16] and [14] also adopt this pipeline, except that they use a more precise method to align or recover image detail. Despite giving good results, they are not suitable for HDR video recovery due to being limited to recovering the HDR version for only one input frame.

Alternate exposure is a promising direction for HDR video reconstruction. Kalantari and Ramamoorthi [17] proposed the first deep learning approach to produce HDR video from a sequence of alternating exposures. This reconstruction pipeline consists of two steps. First, optical flow is estimated using deep neural network, and then another network is used to obtain fusion weights for merging the aligned images. Although they exhibit good performance, ghosting artifacts will appear in the regions with large motions [2]. Inspired by the difficulty of alignment between SDR images with different exposure and deformable convolution, DeepHDRVideo [2] constructs a two-stage coarse-to-fine framework. Combining deformable convolution with HDR video reconstruction, this method achieves excellent results. Since most of the existing videos are single exposure SDR videos, it cannot be directly applied.

Another option relies on *specialized hardware* [18], e.g., per-pixel exposure [3], scanline exposure/ISO [19], [20], [21], internal [22], [23] or externalbeam splitter [24] that can split light to different sensors, modulo camera [4] and neuromorphic [5]. Despite their abilities to produce detailed HDR images/video efficiently on unique equipment. This specialized hardware also limits its widespread applications.

2.2 Single-exposure HDR Imaging

Single-exposure HDR imaging has a more flexible solutions, which can be roughly divided into three types as mentioned before. For the first approach of *direct reconstruction method*, HDRCNN [25] achieves excellent performance in the recovery of overexposed areas. However, this method ignores the dynamic range expansion in other regional. SR-ITM [26] is a representative framework for joint SR and single-exposure HDR. It is used to the video task. Then JSIGAN [27] develops this method and decomposes an LR SDR image into base layer and detail layer. Subsequently, these two layers are calculated separately by their respective modules and then joined together to produce a super-resolution HDR image. In general, they either directly use the modules from other research fields, or use some of HDR imaging knowledge as conditions, without theoretical innovation introduced. There still exits much room for advancement.

For the *multi-exposure stack-based synthesis* approach, DrTMO [28] is the first method of this kind which first passes an image through CNN to generate SDR images with different exposures, and then merge these images to generate HDR images. KIM [29] further extends this method. It is found that the quality of the generated multi-exposure images is a significant factor to reconstruct HDR image.

For the *reverses camera imaging pipeline* approach, three networks are used to learn the stages of inverse camera pipeline to generate HDR images in singleHDR [30]. Besides, HDRTV [11] also refers to the SDRTV imaging pipeline and a three-stage network for HDR reconstruction. Although these methods take advantages of the priori knowledge of HDR imaging, the HDR reconstruction component would introduce a large number of cumulative errors due to strong dependence between these networks.

The aforementioned methods can construct satisfactory HDR videos for conventional SDR. However, they do not take into account compressed SDR, thus less satisfactory for reconstructing HDR videos.

2.3 Single-frame Quality Enhancement

The existing methods can be divided into two categories *JPEG-compressed images quality enhancement* and *HEVC-compressed quality enhancement* [8]. For the first category, Dong et al., [31] proposed a pioneering Artifacts Reduction Convolution Neural Network (AR-CNN), which facilitates the development of CNN-based quality enhancement of JPEG-compressed images. Inspired by dual-domain sparse coding, deep Dual-Domain Convolution neural Network (DDCN) [6] is proposed for JPEG compressed artifacts removal. This model uses the quantization prior of JPEG compression and achieves good results. Recently, SwinIR [32] uses a swin transformer for JPEG compression artifact reduction and achieves art results. However, due to the different coding strategies of HEVC, these approaches cannot be directly used for HEVC-compressed image quality enhancement, especially those utilizing the prior of JPEG compression.

For HEVC-compressed image quality enhancement, the first deep learning-based method was proposed by Wang et al. [33]. The goal is for code rate savings. Then, Yang et al. [7] proposed a Quality Enhancement Convolutional Neural Network (QE-CNN), capable of handling not only I frames but also P/B frames. Recently, through a dynamic deep neural network which embeds an early-exit strategy to resource-efficient blind quality enhancement (RBQE), RBQE [8] achieves strong performance in terms of both blind quality enhancement and resource efficient.

Note, HDR image reconstruction is fundamentally different from compressed SDR image enhancement. Single-frame quality enhancement methods cannot be directly applied to the problem of CSDR-to-HDR video reconstruction. To overcome this challenge, in this work we propose to combine priori knowledge of compressed SDR image quality enhancement with a HDR image imaging pipeline.

3 Analysis of HDR Image Reconstruction

SDR image formation formula [34] is proposed as follows,

$$I_S = \begin{cases} \frac{t}{g}\phi + I_0 + n, & \text{Unsaturation;} \\ I_{\max}, & \text{Saturation} \end{cases} \tag{1}$$

where t is the exposure time, g is the sensor gain, and I_0 is the constant offset current. ϕ represents the scene brightness, as mentioned in [35], which can be assumed as HDR pixel value. I_S represents a SDR image pixel value and n is the sensor noise. Unsaturation represents the pixels that can be represented by the SDR image after camera imaging pipeline processing; while saturation represents sensor saturation occurs which is due to the limited capabilities of the current camera, so this pixel value will equal to a saturation point value $I_{\max}$ [35].

We consider both SDR and HDR in the linear (raw) domain in this work. From a definition perspective, a HDR file, a post-ISP (Image Signal Processor) image/video, should describe closely a natural scene. Using these files as input of the camera imaging pipeline (Eq. (1)) thus conforms to the actual scene approximately. Serving as an approximation of the output from the camera imaging pipeline, similarly considering the SDR image/video in the linear domain thus makes sense. Also, our treatment is consistent with the protocol of NTIRE 2021 Challenge [35]: Using the post-ISP HDR images to approximate the scene brightness ϕ, and Eq. (1) to generate SDR file (8 bits)

This formula, widely used in SDR imaging, inspires us to generate HDR images in a similar theoretical manner. Suppose we have a camera with infinite capture capabilities, the corresponding saturated pixel value in Eq. (1) of the SDR images can be represented as follows,

$$I_{\max} = I_{\text{overexposed}} - I_{\text{overflow}} \tag{2}$$

where $I_{\text{overexposed}}$ and I_{overflow} represent the pixel values captured by this infinitely capable camera, and the overflow values between the ideal and real cameras, respectively. Of course, if the pixel value is unsaturated, $I_{\text{overflow}} = 0$ since no difference exists between the ideal and real cameras. Combining Eq. (1) and Eq. (2), the SDR formation process can be unified as,

$$I_S = \frac{t}{g}\phi + I_0 + n - I_{\text{overflow}}. \tag{3}$$

By reversing Eq. (3), the true HDR pixel value can be obtained as follows,

$$\phi = \frac{g}{t}\left(I_S - I_0 + I_{\text{overflow}}\right) - \frac{g}{t}n. \tag{4}$$

Since the noise n also includes the impacts from g and t, without generality, we still can consider gn/t as the SDR image generation noise. From Eq. (4), we can conclude the restoration process from SDR-to-HDR image reconstruction by three parts: 1) inferring the pixel values in the overexposed area if $I_{\text{overflow}} \neq 0$; 2) adjusting sensor gain and exposure time; 3) reducing the noise caused by SDR image generation. In this way, we obtain an idealized HDR imaging pipeline.

It is demonstrated in KUNet [1] that the formula can be extended to HDR video reconstruction. Here, we will extend the formula to compressed video reconstruction task. As the first attempt at reconstructing compressed SDR videos to HDR counterparts, in this work we seek for a simpler solution based on single frame reconstruction, laying down a solid ground for more sophisticated approaches. Generally, quality enhancement for compressed frame can be denoted as follows:

$$I_S = E(\breve{I}_S) \tag{5}$$

where $\breve{I}_S$ represents the pixel values of the compressed SDR frame. E is the quality enhancement function. In video compression, the quantization process actively throws away some high frequency information that is far from the mean point, in order to save the bitrate [36]. In other words, the processing of high and low frequency information is not consistent. It is natural to think about the possibility of decomposing the video into low-frequency and high-frequency parts and then enhancing it with different functions. The details can be described as follows:

$$I_S = B[E_L(\breve{I}_L), E_H(\breve{I}_H)] \tag{6}$$

where $\breve{I}_L$ and $\breve{I}_H$ represent the low-frequency and high-frequency information of compress video. E_L and E_H are the corresponding quality enhancement functions. B is a blend function used to fuse the information of $\breve{I}_L$ and $\breve{I}_H$.

In this work, we combine the above mentioned two issues for CSDR-to-HDR video reconstruction. However, in the image domain, the direct estimate of $I_{overflow}$, g and t is difficult. Thanks to the power of deep learning in feature representation and learning, we perform the formulate and estimate formula of the HDR video restoration process in the feature domain and separation of high and low frequency information in the feature domain to unify with the previous task for compressed SDR video using neural network. Each function is formulated by a sub-network and together restores the HDR features for generating HDR images. We present this knowledge-inspired module below.

4 Proposed Method

The proposed framework is shown in Fig. 2 (a). Compared with KUNet [1], a new parallel structure is adopted. The *Knowledge-inspired Parallel Net* (KPNet) consists of four stages. Head stage extracts features from the input compressed SDR video I_L and transforms the obtained SDR features into different scales; while tail stage reconstructs a HDR video from the refined HDR features. Between the head and tail stages, there is a knowledge-inspired block (KIB) cluster. It consists of three KIBs which are used to generate HDR features according to the Eq. (4) and Eq. (9). After transforming the compressed SDR video into different scales, a KIB cluster reconstructs three scales of HDR features. Then, these features are fed into an information fusion module (IFM) to generate detailed HDR features. We describe the key modules in detail below.

4.1 Head and Tail

The main function of head stage is transforming the compressed SDR video into feature space and scaling them into different levels. It does not require overly complex operations. First, a simple subnetwork of one layer convolution with a ReLU activation function is employed to complete this task, which is denoted as

$$F_{out} = Conv_{3\times3}(\breve{I}_S) \tag{7}$$

where F_{out} represents the output of head stage. To take full advantage of the intra-frame information and reduce the computational complexity, we use a three-branch scale network to transform the output features at three different scales. It can be denoted as the following,

$$\begin{aligned} F_S^1 &= Relu \circ Conv_{3\times3}(F_{out}), \\ F_S^2 &= Relu \circ Conv_{3\times3} \circ Down(F_{out}), \\ F_S^3 &= Relu \circ Conv_{3\times3} \circ Down \circ Down(F_{out}), \end{aligned} \tag{8}$$

where $F_S^1 \in \mathbb{R}^{C\times H\times W}$, $F_S^2 \in \mathbb{R}^{C\times H/2\times W/2}$, $F_S^3 \in \mathbb{R}^{C\times H/4\times W/4}$ represent the SDR features at different scales respectively. These features will be used to reconstruct HDR features in parallel.

The tail stage is used for upsampling HDR features and reconstructing HDR image. Besides, to refine the HDR features obtained by the IFM, we use two convolution operations before upsampling, which is described in detail by the following equation,

$$\hat{I}_H = Conv_{3\times3} \circ UP \circ Conv_{1\times1} \circ Conv_{3\times3} \circ ReLU(F_{fuse})$$

where F_{fuse} represents the feature generated by the IFM. $\hat{I}_H$ is the reconstructed HDR image.

4.2 Knowledge-inspired Block Cluster

As stated in Sec. 3, Eq. (4) can be used to reconstruct HDR videos at the image intensity perspective. However, it is hard to directly formulate and combine this function in the image intensity domain. Therefore we turn all these processes to feature space and take advantage of the representation ability of deep learning. We let $E(\cdot)$ represent

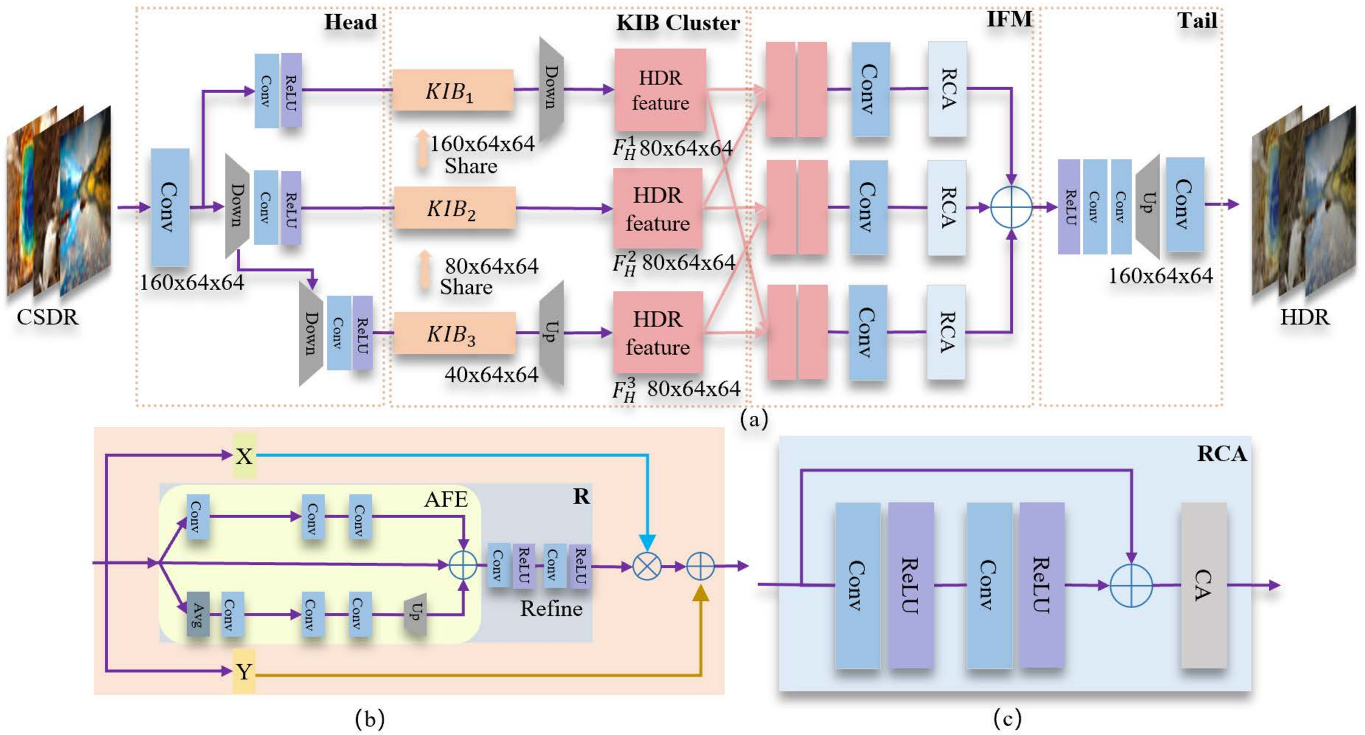


Fig. 2: Overview of the proposed KPNet. (a) The Knowledge-inspired Block (KIB) cluster is designed to generate HDR features with the Head for feature extraction, the Information fusion module (IFM) for feature refining, and the Tail for HDR image reconstruction blocks. (b) A KIB is composed of three parts: imaging parameter adjusting (X), imaging noise reduction (Y) and compressed image quality enhancement with missing overexposed features recovering (R). (c) The Residual with a Coordinate Attention (RCA) [37] module for filtering the fused HDR features.

an enhancement function of compressed SDR pixels. Consequently, the formula can be characterized in feature space as below,

$$\underbrace{\phi}_{H_F} = \underbrace{\frac{g}{t}}_{X(L_F)} \underbrace{(E(\breve{I}_S) - I_0 + I_{\text{overflow}})}_{\odot\ R(L_F)} \underbrace{-\frac{g}{t}n,}_{+Y(L_F)} \tag{9}$$

where $\odot$ denotes the element-wise multiplication. L_F and H_F represent the input compressed SDR features and the features to reconstruct the output HDR image, respectively. $R(\cdot)$ corresponds to the part of formula for the task of enhancing compressed SDR feature quality and inferring the missing overexposed feature to reconstruction HDR. $X(\cdot)$ is in charge of modulating the compensated feature $R(L_F)$ to HDR feature domain; $Y(\cdot)$ corresponds to the SDR imaging noise reduction part in Eq. (9). With the help of this expression in feature space, our knowledge-inspired block is developed as shown in Fig. 2 (b), consisting of three parts of networks which are fit to the functions $R(\cdot)$, $X(\cdot)$ and $Y(\cdot)$, respectively.

Note, $X(L_F)$ and $Y(L_F)$ can be used to approximate $\frac{g}{t}$ and $\frac{g}{t}n$ respectively without direct access to the camera configuration parameters. This is because SDR images are obtained from the HDR image conditioned on the camera configuration parameters (i.e., the camera parameters are implicitly embedded with SDR images). This implicit camera parameters have been also analyzed in previous works. For example, the inverse camera response function, related to the camera parameter configuration, can be obtained from SDR images [30].

For the network fitting the function $R(\cdot)$ (Fig. 2), two sub-modules are proposed to obtain the compensated features. For the compression enhancement task, guided by Eq. (6), we propose a new Asymmetric Frequency domain information Enhancement (AFE) module denoted as

$$(L_{F_h}, L_{F_l}) = S_F(L_F), \tag{10}$$

$$\hat{L}_F = E_H(L_{F_h}) + Upsample(E_L(L_{F_l})) + L_F, \tag{11}$$

where S_F represents a split function with two parts: a convolution layer and an average pool layer with a convolution layer. It can be used to separate high and low frequency information as shown in the OctConv [38]. Discrete Wavelet Transform (DWT) can accurately decompose the frequency domain information, it is not adopted here due to its computational cost and the confidence on fitting ability of neural networks. $E_H(\cdot)/E_L(\cdot)$ represent high/low frequency information enhancement function. They all consist of two layers of convolution and ReLU activation function ($Conv_{3\times3} \circ ReLU$). Then, the enhanced low frequency information is fused with the high frequency information through an upsampling operation. L_F is also incorporated into the enhanced features to ensure that the information is not corrupted. $\hat{L}_F$ represents the output feature of AFE module, which is fed to the refinement network for compensating the missing overexposed features to reconstruct

HDR image. It can be expressed as

$$R = Conv_{3\times3} \circ ReLU \circ Conv_{1\times1} \circ ReLU(\hat{L}_F), \tag{12}$$

where R represents the output of R module. Normalization is not applied in the process of learning high frequency, low frequency and original features, as it would hurt the reconstruction quality (see Tab. 4).

To realize X and Y, we use two 1×1 convolution layers to simulate these functions:

$$X = Conv_{1\times1} \circ Conv_{1\times1}(L_F), \tag{13}$$
$$Y = Conv_{1\times1} \circ Conv_{1\times1}(L_F). \tag{14}$$

All KIBs from different scales are used to generate richer HDR feature maps, i.e., forming a KIB cluster.

Remark. Compared with the previous methods, our KIB mimics the HDR imaging formula. The solution space is constrained and the many-to-many compressed-SDR to HDR mapping problem is well tackled. Under the adjustments of X and Y modules, the function of generating HDR features is adaptive to different SDR videos. Compared to the cascading design of KUNet [1] with the risk of accumulating errors, our KPNet exploits a parallel structure to mitigate ghosting artifacts with generated images (please see Sec. 5.3.4).

4.3 Information Fusion Module

Converging the learning of different scales of HDR features is critical. A straightforward method is feature weighing:

$$F_{fuse} = \alpha_1 F_R^1 + \alpha_2 F_R^2 + \alpha_3 F_R^3 \tag{15}$$

where F_R^1, F_R^2 and F_R^3 represent the features generated by the corresponding KIB clusters, with the corresponding weights as α_1, α_2, and α_3. Due to different natures (e.g., different scales), the weights should be not necessarily the same. Critically, this design does not make full use of different scales of features.

To address this issue. we propose a simple and efficient information fusion block. Specifically, taking the coarse HDR features as input, we first pair these features to allow the HDR elements with different proportions to interact with each other. Then, the fused HDR features are further weighted to obtain the final features with rich color and detail information. This process is described as

$$\begin{aligned} F_{fuse} \;&= T_1[Cat(F_R^1, F_R^2)] + T_2[Cat(F_R^1, F_R^3) \\ &+ T_3[Cat(F_R^2, F_R^3)], \end{aligned} \tag{16}$$

where T_i $(i = 1, 2, 3)$ represents the feature fusion function. It performs two tasks: one is feature fusion, and the other is adaptively weighting different fusion features. The first task is done by a residual network. The second task is realized by the attention mechanism with a strong focus on spatial information as well as positional information also exploited. Formally, this process is denoted as

$$T_i = \quad RCA(Conv_{1\times1}(D_i)) \tag{17}$$

where D_i $(i = 1, 2, 3)$ represents the feature after the operation $Cat(\cdot)$. RCA represents residual with an attention mechanism, for which we adopt Coordinate Attention (CA) [37] (see Fig. 2 (c)). Given an input X, each channel is encoded in horizontal and vertical coordinates using two pooling kernels $(H, 1)$ or $(1, W)$ as:

$$z_c^h(h) = \frac{1}{W} \sum_{0\le i<W} x_c(h, i), \tag{18}$$
$$z_c^w(w) = \frac{1}{H} \sum_{0\le j<H} x_c(j, w), \tag{19}$$

where z_c is the output associated with the c-th channel. h, w represent the vertical and horizontal directions respectively. This operation can effectively capture long-range correlation along one spatial direction, and keep accurate position information along another spatial direction [37]. Next, two attention maps $\mathbf{g}^h$ and $\mathbf{g}^w$ are generated as follows,

$$\mathbf{f} = \delta\left(Conv_{1\times1}\left(Cat(\mathbf{z}^h, \mathbf{z}^w)\right)\right), \tag{20}$$
$$\mathbf{f}^h, \mathbf{f}^w = split(f), \tag{21}$$
$$\mathbf{g}^h = \sigma\left(Conv_{1\times1}\left(\mathbf{f}^h\right)\right), \tag{22}$$
$$\mathbf{g}^w = \sigma\left(Conv_{1\times1}\left(\mathbf{f}^w\right)\right), \tag{23}$$

where δ is a non-linear activation function. Then, f is divided into two vectors $\mathbf{f}^h \in \mathbb{R}^{C/r\times H}$ and $\mathbf{f}^w \in \mathbb{R}^{C/r\times W}$ along the spatial dimension. r is used to reduce the channel number of f. After that, two 1×1 convolutions are used to convert $\mathbf{f}^h$ and $\mathbf{f}^w$ to the same dimension. σ is the sigmoid function. Finally, the output y from CA can be written as

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j). \tag{24}$$

This module takes advantage of the HDR features generated by the KIB clusters, leading to HDR image generated with significantly less ghosting artifacts. Specifically, IFM introduces pairwise feature interaction across scales, going beyond simple weighting. This design allows our knowledge inspired multi-scale features to encode added details in the target features, whilst reducing the inference difficulty.

Remark. Our information fusion module is simple and effective. With a skillful design, a simple interaction module is adopted to ensure the direct and close interactions between KIBs. The module is an amplification of the advantages of parallel construction, ensuring that our KPNet further reduces ghosting artifacts. This effect has been verified in our experiments (Tab. 2).

4.4 Loss Function

For simplicity and generality, we adopt the L_1 loss function for model training:

$$Loss(I_H, \hat{I}_H) = \|I_H - \hat{I}_H\|_1 \tag{25}$$

where I_H represents the real HDR image. Compared with our preliminary work KUNet [1], this loss function is rather simpler without any hyper-parameter.

5 Experiment

5.1 Experiment Setup

Dataset. As compressed-SDR to HDR (CSDR-to-HDR) video reconstruction is a new problem, there is no existing benchmark available. To enable model evaluation, we construct a test dataset from HDRTV [11]. This dataset is constructed by 22 HDR10 standard videos complying with the Rec. 2020 standard. It has 1235 paired training frames and 117 test frames. It contains information about moving light sources,

TABLE 1: Comparison with the state-of-the-art methods. QP: Quantization Parameter. *: Joint training applied. The red/blue/green indicate first/second/third best result.

QP	Category	Model	Venue	Params	PSNR↑	SSIM↑	MSE↓	ΔE_{ITP} ↓	HDR-VDP3↑
32	Image Translation	CSRNet [39]	ECCV20	36K	34.3815	0.9682	0.00050	12.7086	7.1291
		Pixel2Pixel [40]	CVPR17	11.38M	30.9305	0.9595	0.00197	19.4141	7.5197
		DNCNN [31]	TIP17	559K	33.8340	0.9670	0.00062	14.4560	7.8556
		DPIR [41]	TPAMI22	32.64M	34.9800	0.9720	0.00043	10.6693	7.9842
	SDR-to-HDR	JSI-GAN [27]↓	AAAI20	1.06M	34.7672	0.9635	0.00056	13.6780	-
		HDRTV [11]	ICCV21	1.41M	34.3060	0.9713	0.00517	12.3924	8.0818
		KUNet [1]	IJCAI22	1.12M	34.8270	0.9713	0.00046	11.0500	7.8574
	Cascading	DNCNN [31]+HDRTV [11]	-	2.00M	34.7056	0.9676	0.00047	12.1986	8.0526
		DNCNN [31]+KUNet [1]	-	1.68M	34.5778	0.9687	0.00048	11.8109	7.9765
		DNCNN [31]+KUNet [1]*	-	1.68M	34.9668	0.9713	0.00045	11.3471	7.8999
	CSDR-to-HDR	**KPNet**	Ours	1.55M	35.5253	0.9724	0.00038	10.7900	8.0960
37	Image Translation	CSRNet [39]	ECCV20	36K	33.3495	0.9567	0.00060	14.1212	7.5784
		Pixel2Pixel [40]	CVPR17	11.38M	30.5000	0.9724	0.00121	19.8403	7.3802
		DNCNN [31]	TIP17	559K	32.5919	0.9554	0.00739	15.5816	7.4464
		DPIR [41]	TPAMI22	32.64M	34.0380	0.9610	0.00051	11.9210	7.5593
	SDR-to-HDR	JSI-GAN [27]↓	AAAI20	1.06M	33.6724	0.9635	0.00056	13.6780	-
		HDRTV [11]	ICCV21	1.41M	33.3980	0.9603	0.00058	13.6280	7.6575
		KUNet [1]	IJCAI22	1.12M	33.5860	0.9601	0.00057	12.6073	7.4081
	Cascading	DNCNN [31]+HDRTV [11]	-	2.00M	33.6102	0.9563	0.00058	13.6091	7.6208
		DNCNN [31]+KUNet [1]	-	1.68M	33.3744	0.9571	0.00060	13.3998	7.5447
		DNCNN [31]+KUNet [1]*	-	1.68M	33.6512	0.9599	0.00057	12.8425	7.5108
	CSDR-to-HDR	**KPNet**	Ours	1.55M	34.1907	0.9608	0.00050	12.4292	7.6379
42	Image Translation	CSRNet [39]	ECCV20	36K	31.9878	0.9381	0.00081	16.2704	6.9567
		Pixel2Pixel [40]	CVPR17	11.38M	27.9251	0.9102	0.00384	25.7382	6.5744
		DNCNN [31]	TIP17	559K	31.4434	0.9372	0.00091	17.3354	6.7819
		DPIR [41]	TPAMI22	32.64M	32.3987	0.9427	0.00071	14.1976	6.9014
	SDR-to-HDR	JSI-GAN [27]↓	AAAI20	1.06M	32.2527	0.9445	0.00074	15.6759	-
		HDRTV [11]	ICCV21	1.41M	32.1151	0.9424	0.00080	15.5980	6.9329
		KUNet [1]	IJCAI22	1.12M	32.1384	0.9418	0.00076	14.8065	6.8324
	Cascading	DNCNN [31]+HDRTV [11]	-	2.00M	32.1693	0.9383	0.00077	15.5662	6.9717
		DNCNN [31]+KUNet [1]	-	1.68M	31.7923	0.9383	0.00084	15.6402	6.9023
		DNCNN [31]+KUNet [1]*	-	1.68M	32.0765	0.9414	0.00081	15.2456	6.8924
	CSDR-to-HDR	**KPNet**	Ours	1.55M	32.6943	0.9421	0.00071	14.5567	6.9209
47	Image Translation	CSRNet [39]	ECCV20	36K	30.2894	0.9127	0.00117	19.3696	6.1670
		Pixel2Pixel [40]	CVPR17	11.38M	24.6152	0.8548	0.01081	37.4932	5.7147
		DNCNN [31]	TIP17	559K	29.8220	0.9123	0.00129	20.7242	6.0758
		DPIR [41]	TPAMI22	32.64M	30.7398	0.9185	0.00104	17.2544	6.1113
	SDR-to-HDR	JSI-GAN [27]↓	AAAI20	1.06M	30.7269	0.9173	0.00102	18.5330	-
		HDRTV [11]	ICCV21	1.41M	29.4684	0.9159	0.00161	21.6288	6.0346
		KUNet [1]	IJCAI22	1.12M	30.5031	0.9174	0.00112	17.8971	6.0753
	Cascading	DNCNN [31]+HDRTV [11]	-	2.00M	30.4798	0.9125	0.00112	18.7283	6.1689
		DNCNN [31]+KUNet [1]	-	1.68M	30.1021	0.9125	0.00121	19.0349	6.1259
		DNCNN [31]+KUNet [1]*	-	1.68M	30.5722	0.9171	0.00108	18.1048	6.0338
	CSDR-to-HDR	**KPNet**	Ours	1.55M	30.9391	0.9182	0.00102	17.6343	6.0975

rich colors, highlights and bright. We compress all the images by HM16.9[1] under the intra-coding configuration, setting the Quantization Parameters (QPs) to 32, 37, 42, 47, respectively. These QPs can reflect the dramatically varying quality of compressed videos. As the very first exploration, we leave the investigation of inter-frame information based video decoders for future work.

Evaluation metrics. We employ five evaluation metrics for comprehensive comparisons, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [42], Mean squared error (MSE), ΔE_{ITP} [43], HDR-VDP3 [44]. PSNR, SSIM and MSE are commonly used to measure image similarity. An important task with CSDR-to-HDR video reconstruction is the recovery of the color gamut (Rec. 709->Rec. 2020), we thus adopt color difference metric ΔE_{ITP}. Following [11], we choose HDR-VDP3, a visual metric for predicting visibility (discrimination) and quality (mean-opinion-score) [44].

Implementation details. For the Head, Tail, R block of KIB and IFM, we use 3×3 convolution with a step size of 1 unless specifically stated. For the X and Y branches in the KIB, we use 1×1 convolution. The activation function of all networks is ReLU function. Except for input and output layers, the channels of feature maps are all set to 64. The down-sampling operation uses a 3×3 convolution operation with a step size of 2. Up-sampling operation uses PixelShuffle [45]. We use ADAM optimizer [46] with the learning rate of $2e^{-4}$ decayed by a factor of 2 after $20K$ iterations and then decayed by a factor of 2 after every $10K$. The total number of iterations is 550000. The batch size is 12. All models are built on the PyTorch framework

5.2 Comparison with State-of-the-art Methods

Due to no existing works, we conduct extensive evaluation by comparing three different approaches: (1) existing image translation methods (e.g., Pixel2Pixel [40], DNCNN [31], DPIR [41], CSRNet [39]); (2) SDR-to-HDR image/video reconstruction methods (e.g., JSI-GAN [27], HDRTV [11], KUNet [1]); and (3) cascading compressed SDR video enhancement (e.g., DNCNN [31] and DPIR [41]) with HDR

[1] https://hevc.hhi.fraunhofer.de/

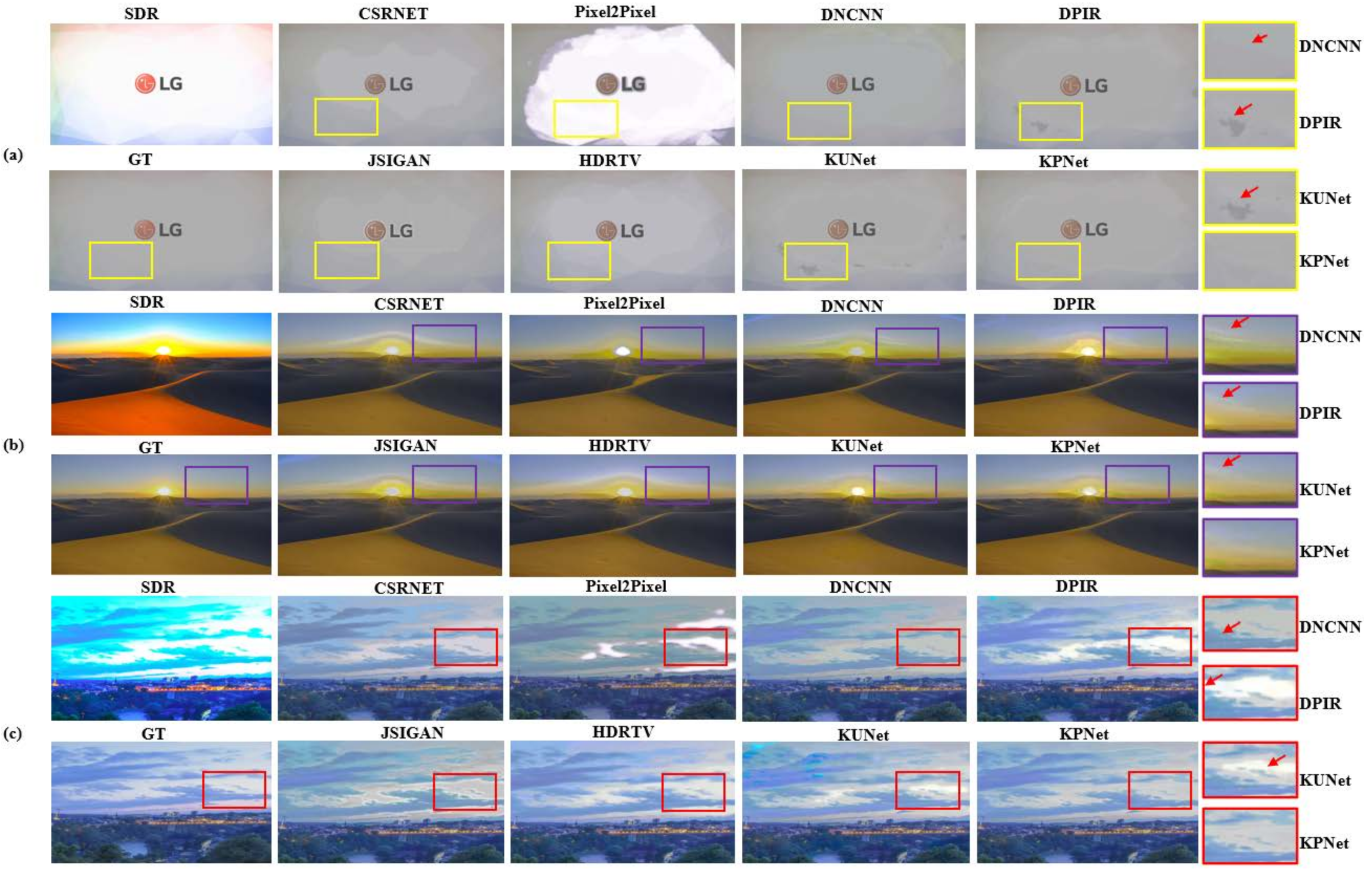


Fig. 3: Qualitative comparisons. We highlight the key differences with bounding boxes and arrows. GT: The real HDR sample.

reconstruction (e.g., HDRTV [11], KUNet [1]) with and without joint training. For fair comparison, we follow the original optimal configuration of prior methods during their training. Note, as JSI-GAN is originally designed for super-resolution, we additionally apply a down-sampling operation to ensure the integrity, while keeping the input and output sizes consistent.

5.2.1 Quantitative Evaluation

We present the comparison with prior art methods in Tab. 1. We draw several key observations: (1) Our KPNet achieves the best performance for CSDR-to-HDR video reconstruction across different QP settings, validating the superiority of our design and formulation over all the alternative approaches. (2) The best image translation method, DNCNN [31], is less satisfactory, because it only addresses compressed SDR enhancement and does not take HDR image reconstruction into account. (3) Despite giving competitive results, DPIR [41] is significantly more heavy and more expensive. (4) Compared with SDR-to-HDR image/video reconstruction methods, our preliminary model KUNet [1] already performs well. After incorporating the priori knowledge of compressed video, our KPNet can further improve the performance. (5) HDRTV [11] is less competitive than ours due to ignoring information loss of compressed SDR videos. (6) The cascading approach is effective in improving the constituent models but the gain is limited. In particular, the combination of DNCNN+KUNet even performs worse than KUNet alone. A plausible cause is error accumulation – error information from quality enhancement may interfere with HDR video reconstruction. Further, cascading leads to more complex pipelines in both training and design.

5.2.2 Qualitative Evaluation

We provide qualitative comparisons in Fig. 3. For diverse evaluation, three different scenes are selected: (a) low information density, (b) salient object with overexposed observation, (c) a wealth of color information with complex patterns. The real HDR image is denoted as GT (ground-truth). We summarize the following observations: (1) For the first scene, within the white area adjacent to the logo “LG”, CSRNet [39], Pixel2Pixel [40], DNCNN [31], and JSIGAN [27] all cannot recover the compression ghosting artifacts accurately. As indicated by red arrows, DNCNN yields two colors (red and white) alien to the original scene. HDRTV [11] instead suffers color breaks. Whilst no color breakage, KUNet [1] and DPIR [41] both present compression ghosting artifacts in the white area. (2) For the second scene, KPNet performs well in recovering the sky colors (see purple boxes) and the color breaks near sunlight. It is evident that DNCNN is inferior in coloring. (3) For the last scene, we similarly observe the inferior ability of prior models in dealing with the fine-grained details of the nature, e.g., bright regions. Overall, our method performs the best across all scenarios, matching the real scene closely.

TABLE 2: Ablation study. '✓−': Only one KIB is used for HDR image reconstruction with about 512K parameters, a lightweight variant. The red/blue indicate the first/second best result.

Index	KIB			Information fusion design			Loss			PSNR↑	SSIM↑	$\Delta E_{ITP}\downarrow$	MSE↓
	$R(\cdot)$	$X(\cdot)$	$Y(\cdot)$	IFM$_D$	IFM$_N$	IFM$_{ours}$	L_1	L_2	L_{KIB}				
1	✓	✓	✓	-	-	-	✓	-	-	35.0584	0.97117	10.9900	0.00042
2	✓	✓	✓	-	-	✓	✓	-	-	35.5253	0.97242	10.7900	0.00038
3	✓	✓	✓	-	-	✓	-	✓	-	35.2212	0.97110	11.3522	0.00041
4	✓	✓	✓	-	-	✓	-	-	✓	35.5293	0.97224	10.8489	0.00040
5	✓−	✓−	✓−	-	-	-	✓	-	-	34.8204	0.97088	11.0500	0.00046
6	✓	-	-	-	-	-	✓	-	-	34.8852	0.97059	11.1434	0.00044
7	✓	✓	-	-	-	-	✓	-	-	34.9780	0.97076	11.1495	0.00044
8	✓	✓	-	-	-	✓	✓	-	-	35.4696	0.97187	10.9964	0.00041
9	✓	-	✓	-	-	✓	✓	-	-	35.3151	0.97201	10.9685	0.00039
10	✓	✓	✓	✓	-	-	✓	-	-	35.0091	0.97178	11.0780	0.00042
11	✓	✓	✓	-	✓	-	✓	-	-	33.2560	0.96742	13.2365	0.00069

5.3 Ablation Study

5.3.1 Effect of key modules

We evaluate the effect of each key module (e.g., KIB, information fusion module (IFM), and loss) with our KPNet. We use the QP = 32 setting. We make the following observations from Tab. 2. **(1)** Row 1 *vs.* 2: With L_1 loss, our IFM can clearly improve the performance.

We further compare with two alternative loss functions, L_2 and L_{KIB}, with the latter meaning the total loss function has a supervision training term on KIB: $Loss(I_H, \hat{I}_H) = \|I_H - \hat{I}_H\|_1 + \gamma \|I_H - K_i\|_1$, where $\{K_i | i = 1, 2, 3\}$ represent the HDR images reconstructed from the features generated by the corresponding KIBs respectively, and γ is a balance term. **(2)** From Row 2 *vs.* 3 *vs.* 4, it can be seen that using L_1 loss can obtain satisfactory results. L_{KIB} gives a slight gain on PSNR but hurts the others; Also, reconstructed images are often less accurate and the γ needs to set small to avoid over-penalty. We thus simply uses L_1 loss.

Next we analyze the effect on the number of KIB. To that end, we construct a lightweight variant – only one KIB is used for HDR image reconstruction with about 512K parameters. **(3)** From Row 1 *vs.* 5, we see clear performance degradation, suggesting the importance of having a stronger model. Note that our lightweight model achieves similar result as HDRTV but using less than a half of parameters. We find using three KIBs strikes a good trade-off between efficiency and accuracy.

We examine the design of KIB, including the core module R, and the adaptive X and Y branches for further HDR feature refinement. **(4)** From Row 6, we show that only using the R module can already achieve an acceptable result. Adding the X and Y branches leads to stronger expressive ability. As suggested in Eq. (9) that X and Y branches should be used simultaneously, which has been verified in the contrast of Row 6/7 *vs.* 1 and 8/9 *vs.* 2. Our X only variant achieves a good performance gain, regardless of using IFM or not. Note, without IFM, KPNet degrades to KUNet.

We examine the design of information fusion. We further compare our IFM with two more variants: (i) IFM$_D$ - directly weighing the features with Eq. (15), with the best setting we empirically find as $\alpha_1 = 0.3$, $\alpha_2 = 0.4$ and $\alpha_3 = 0.3$; (ii) IFM$_N$ - fusion without pairwise interaction as $F_{\text{fuse}} = T_1 \left[F_R^1\right] + T_2 \left[F_R^2\right] + T_3 \left[F_R^3\right]$. **(5)** From Row 10/11 *vs.* 2, it is evident that our design is the best. This is because both variants fail to explore the mutual relationships between different features. As a result, inferior results could be generated, e.g., IFM$_D$ gives a plenty of ghosting artifacts (see blue boxes in Fig. 4), whilst IFM$_N$ is poor in color recovery (see orange boxes in Fig. 4).

TABLE 3: Evaluation of information loss during modeling.

Iterations	$\breve{K}_1$	$\breve{K}_2$	$\breve{K}_3$
500	0.000905	0.003068	0.001962
5000	0.000091	0.000252	0.000419
10000	0.000020	0.000050	0.000133

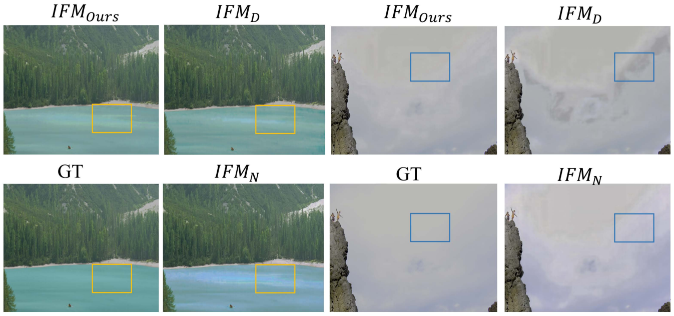


Fig. 4: Visual evaluation of different information fusion strategies. GT: The real HDR sample.

We finally evaluate the information loss (e.g., physical meaning loss of input) during modeling. Specifically, we reconstruct the CSDR input from the three scaled CSDR features (after nonlinear mapping): $\breve{K}_1 = Conv_{3\times3}(F_S^1)$, $\breve{K}_2 = Conv_{3\times3} \circ Up(F_S^2)$, $\breve{K}_3 = Conv_{3\times3} \circ Up \circ Up(F_S^3)$, where $\{\breve{K}_i | i = 1, 2, 3\}$ represent the CSDR frames reconstructed from the CSDR features F_S^1, F_S^2, F_S^3, respectively. We use the L_1 reconstruction loss for training: $L(\breve{I}_S) = \|\breve{K}_i - \breve{I}_S\|_1$. For reconstruction quality, we measure the mean square errors (MSE) between the reconstructed CSDR frames and the original ones. We observe from Tab. 3 that very small reconstruction errors can be achieved, decreasing rapidly along with the training iterations. This validates little information loss from our modeling.

5.3.2 Feature normalization

As mentioned in Sec. 4.2, feature normalization is not applied. Tab. 4 shows that the performance of KPNet drops to

TABLE 4: Effect of feature normalization with our KPNet. QP: Quantization Parameter.

	Normalization	PSNR↑	SSIM↑	MSE↓	ΔE_{ITP}↓
QP32	✓	35.2725	0.9701	0.00040	11.0262
	✗	**35.5253**	**0.9724**	**0.00038**	**10.7900**
QP37	✓	33.9337	0.9584	0.00053	12.7788
	✗	**34.1907**	**0.9608**	**0.00050**	**12.4293**
QP42	✓	32.5473	0.9409	0.00070	14.6997
	✗	**32.6943**	**0.9421**	**0.00071**	**14.5567**
QP47	✓	30.7764	0.9180	0.00103	17.9768
	✗	**30.9391**	**0.9182**	**0.00102**	**17.6343**

TABLE 5: Ablation on the feature scales..

Methods	PSNR ↑	SSIM ↑	MSE ↓	ΔE_{ITP} ↓
B1	35.3915	0.9718	0.00040	10.9694
B2	35.5108	0.9718	0.00040	10.8633
B3	**35.5253**	**0.9724**	**0.00038**	**10.7900**

some extent from normalization. We consider feature normalization might alter the brightness distribution of video frames, leading to brightness distortion and finally hurting the accuracy of HDR video reconstruction.

5.3.3 *Downsampling*

As discussed in Sec. 4.1, the output feature F_{out} is transformed to three different scales / branches. We evaluate the choice of this design selection. We compare three designs: `B1`, `B2` and `B3` representing one branch without downsampling, two branches with single downsampling and three branches with two downsampling, respectively. Tab. 5 shows that the best results are obtained with three branches. The same is visually validated in Fig. 5. Specifically, we observe significant color difference between `B1` and `B2` (see Group A); The block effect is not eliminated on the "HDR". Group B shows that `B1` and `B2` present noticeable boundaries in color transition regions, especially in the contrast between `B1` and `B3`. The trend suggests further adding more branches would be marginally beneficial with lower cost-effectiveness.

5.3.4 *KPNet vs. KUNet*

Our preliminary model KUNet [1] tends to present color breakdowns and ghosting artifacts in CSDR-to-HDR video reconstruction. This is the motivation for design the KPNet. With careful investigation and experimental analysis, we realize that the ghosting artifacts are caused by two factors: image compression and the cascaded design of KIBs (propagating and amplifying the error information). For empirical validation, we design a test with hand-designed color cards for easy visual examination. We compare KPNet with DPIR [41], KUNet, and two more variants: (i) Single - an improved KIB with a new R module combining the prior knowledge of compressed video reconstruction as KPNet; (ii) KUNet-New - the KUNet architecture with the improved KIB.

We observe from Fig. 6 that, (1) KUNet-New can reduce the occurrence of ghost compared with KUNet. Due to the use of cascaded KIBs, it still accumulates the errors and gives consequently color breakdown. In contrast, the model Single does not. This suggests the limitation of the cascade structure. (2) With a a parallel structure, KPNet solves the above issue. (3) DPIR cannot excel in visual analysis though it yields a competitive quantitative result (Tab. 1). Also, it suffers from color discontinuity due ot using the UNet.

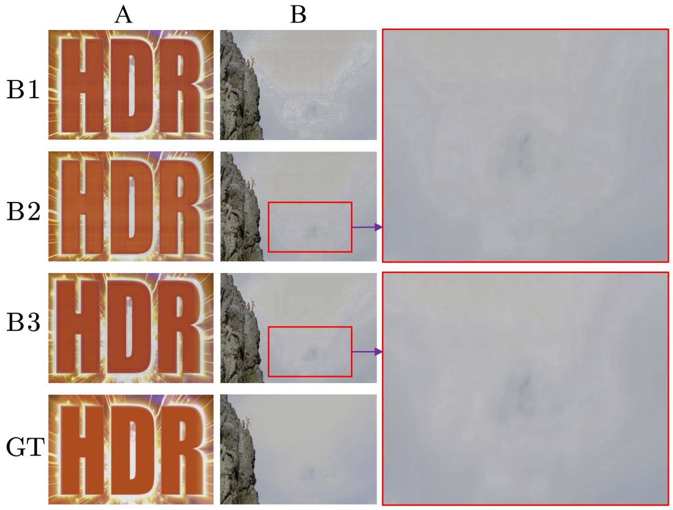


Fig. 5: Visual examples using different feature scales. GT: The real HDR sample.

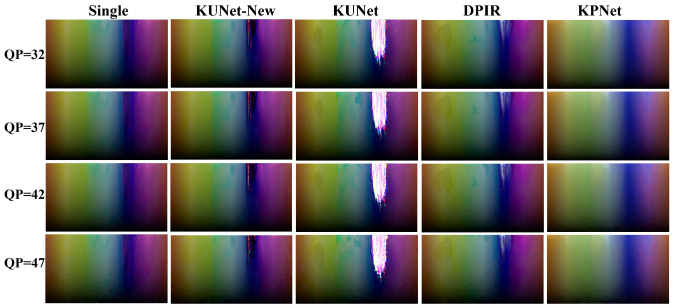


Fig. 6: Analysis on color breakdowns and ghosting artifacts. QP: Quantization Parameter.

5.4 Preliminary video evaluation

We further conduct video based visualization evaluation. This test focuses on both the reconstruction quality of individual video frames and the coherence through the frames over time. Considering that the dataset available for HDR video reconstruction is relatively limited and HDRTV is a single frame based dataset, we sample a random Internet video[2] with typical natural illumination conditions like the sun and water surface reflections. This video comprises 600 frames at a resolution of 3840*2160. The first five frames are used for presentation. For comparative evaluation, we select a compressed image enhancement method (DNCNN [31]) and a HDR reconstruction method (KUNet [1]).

We make several observations from Fig. 7: (1) It is evident that DNCNN can not cause any ghosting in the vicinity of the sun, but the sun has no any edges and the colour turns red. (2) Although KUNet [1] can preserve the sun's shape, ghosting appears across the first, fourth and fifth frames, resulting in discontinuous observation over time. This is because KUNet only solves the problem of HDR image

[2] https://youtu.be/mxdyKT8WEZM?si=HExWQaMxRGnzk9yp

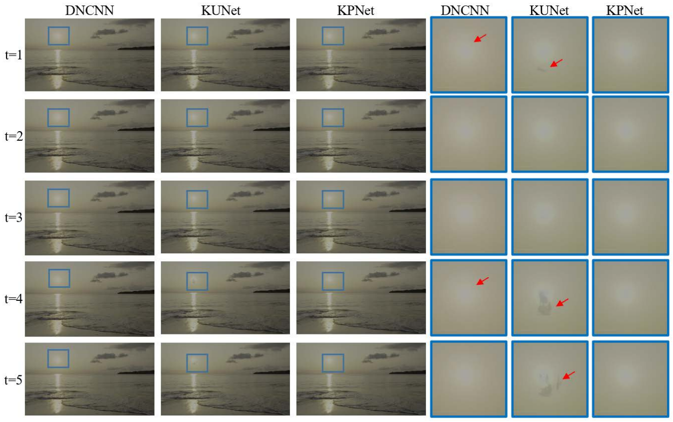


Fig. 7: Visualization on HDR video reconstruction.

reconstruction but not the compressed video reconstruction. (3) Our KUNet not only preserves the silhouette of the sun but also eliminates ghosting artifacts, yielding the best continuity of video playback and enhancing the viewing experience significantly. This is achieved by leveraging the knowledge inspired block to incorporate a priori knowledge of compressed video in a parallel structure without error accumulation across operations.

6 Conclusion

In this paper, we have studied the largely ignored problem of CSDR-to-HDR video reconstruction. By analyzing the HDR-to-SDR imaging process and obtaining the HDR image formulation formula, we formulate a novel model KPNet, characterized by incorporating the knowledge of compressed video reconstruction into the knowledge-inspired block in an interactive parallel structure. For enabling model evaluation, the very first benchmark has been constructed. Experiments demonstrate that our KPNet achieves the state-of-the-art performance for CSDR-to-HDR video reconstruction, along with detailed ablation study and analysis.

The research on compressed-SDR to HDR (CSDR-to-HDR) video reconstruction task is still in its nascent stage, leaving ample scope for advancements in both theoretical frameworks and methodological approaches. For example, leveraging inter-frame information to reduce ghosting artifacts caused by video compression is a promising area. Except the traditional frame or feature alignment based approaches, highly exploratory routes include to use multiple compressed SDR frames as the input conditions and to explore the potential of recent diffusion models.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2018YFE0203900) and National Natural Science Foundation of China (62276048, U19A2052, 62020106011).

References

[1] H. Wang, M. Ye, X. Zhu, S. Li, C. Zhu, and X. Li, “KUNet: Imaging knowledge-inspired single hdr image reconstruction,” in *International Joint Conference on Artificial Intelligence and European Conference on Artificial Intelligence*, 2022.

[2] G. Chen, C. Chen, S. Guo, Z. Liang, K.-Y. K. Wong, and L. Zhang, “HDR video reconstruction: A coarse-to-fine network and a real-world benchmark dataset,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2502–2511.

[3] S. K. Nayar and T. Mitsunaga, “High dynamic range imaging: Spatially varying pixel exposures,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2000, pp. 472–479.

[4] H. Zhao, B. Shi, C. Fernandez-Cull, S.-K. Yeung, and R. Raskar, “Unbounded high dynamic range photography using a modulo camera,” in *IEEE International Conference on Computational Photography*. IEEE, 2015, pp. 1–10.

[5] J. Han, C. Zhou, P. Duan, Y. Tang, C. Xu, C. Xu, T. Huang, and B. Shi, “Neuromorphic camera guided high dynamic range imaging,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1730–1739.

[6] J. Guo and H. Chao, “Building dual-domain representations for compression artifacts reduction,” in *European Conference on Computer Vision*. Springer, 2016, pp. 628–644.

[7] R. Yang, M. Xu, T. Liu, Z. Wang, and Z. Guan, “Enhancing quality for hevc compressed videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 7, pp. 2039–2054, 2018.

[8] Q. Xing, M. Xu, T. Li, and Z. Guan, “Early exit or not: Resource-efficient blind quality enhancement for compressed images,” in *European Conference on Computer Vision*. Springer, 2020, pp. 275–292.

[9] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang, “Mfqe 2.0: A new approach for multi-frame quality enhancement on compressed video,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 43, no. 3, pp. 949–963, 2019.

[10] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, “Deformable convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.

[11] X. Chen, Z. Zhang, J. S. Ren, L. Tian, Y. Qiao, and C. Dong, “A new journey from SDRTV to HDRTV,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4500–4509.

[12] L. Wang and K.-J. Yoon, “Deep learning for HDR imaging: State-of-the-art and future trends,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[13] N. K. Kalantari, R. Ramamoorthi *et al.*, “Deep high dynamic range imaging of dynamic scenes.” *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 144–1, 2017.

[14] Z. Liu, W. Lin, X. Li, Q. Rao, T. Jiang, M. Han, H. Fan, J. Sun, and S. Liu, “ADNet: Attention-guided deformable convolutional network for high dynamic range imaging,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 463–470.

[15] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. Zhang, “Attention-guided network for ghost-free high dynamic range imaging,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1751–1760.

[16] F. Peng, M. Zhang, S. Lai, H. Tan, and S. Yan, “Deep HDR reconstruction of dynamic scenes,” in *International Conference on Image, Vision and Computing*. IEEE, 2018, pp. 347–351.

[17] N. K. Kalantari and R. Ramamoorthi, “Deep HDR video from sequences with alternating exposures,” in *Computer Graphics Forum*, vol. 38, no. 2. Wiley Online Library, 2019, pp. 193–205.

[18] Z. Pu, P. Guo, M. S. Asif, and Z. Ma, “Deep exposure fusion with deghosting via homography estimation and attention learning,” in *International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 1464–1468.

[19] C. Wang, J. Zhou, and S. Liu, “Adaptive non-local means filter for image deblocking,” *Signal Processing: Image Communication*, vol. 28, no. 5, pp. 522–530, 2013.

[20] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pająk, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian *et al.*, “Flexisp: A flexible camera image processing framework,” *ACM Transactions on Graphics*, vol. 33, no. 6, pp. 1–13, 2014.

[21] I. Choi, S.-H. Baek, and M. H. Kim, “Reconstructing interlaced high-dynamic-range video using joint learning,” *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5353–5366, 2017.

[22] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, “A versatile hdr video production system,” *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 1–10, 2011.

[23] J. Kronander, S. Gustavson, G. Bonnet, A. Ynnerman, and J. Unger, “A unified framework for multi-sensor HDR video reconstruc-

tion," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 203–215, 2014.

[24] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar, "Optical splitting trees for high-precision monocular imaging," *IEEE Computer Graphics and Applications*, vol. 27, no. 2, pp. 32–42, 2007.

[25] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Transactions on Graphics*, vol. 36, no. 6, pp. 1–15, 2017.

[26] S. Y. Kim, J. Oh, and M. Kim, "Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3116–3125.

[27] Soo Ye Kim and Jihyong Oh and Munchurl Kim, "JSI-GAN: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for UHD HDR video," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 287–11 295.

[28] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Transactions on Graphics*, vol. 36, no. 6, Nov. 2017.

[29] J. H. Kim, S. Lee, and S. Kang, "End-to-end differentiable learning to HDR image synthesis for multi-exposure images," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[30] Y.-L. Liu, W. Lai, Y. Chen, Y. Kao, M. Yang, Y. Chuang, and J. Huang, "Single-image HDR reconstruction by learning to reverse the camera pipeline," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1651–1660.

[31] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[32] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

[33] T. Wang, M. Chen, and H. Chao, "A novel deep learning-based method of improving coding efficiency from the decoder-end for hevc," in *Data Compression Conference*. IEEE, 2017, pp. 410–419.

[34] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 553–560.

[35] E. Pérez-Pellitero, S. Catley-Chandar, A. Leonardis, and R. Timofte, "NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 691–700.

[36] C. Huang, J. Li, B. Li, D. Liu, and Y. Lu, "Neural compression-based feature learning for video restoration," 06 2022, pp. 5862–5871.

[37] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 713–13 722.

[38] Y. Chen, H. Fan, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3435–3444.

[39] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1091–1100.

[40] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[41] K. Zhang, Y. Li, W. Zuo, L. Zhang, L. Van Gool, and R. Timofte, "Plug-and-play image restoration with deep denoiser prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[43] ITU-R, "Tobjective metric for the assessment of the potential visibility of colour differences in television," ITU-R Rec, BT.2124-0, Tech. Rep., 2019.

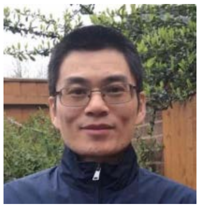
[44] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on Graphics*, vol. 30, no. 4, pp. 1–14, 2011.

[45] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1874–1883.

[46] D. P. Kingma and J. Ba, "ADAM: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

Hu Wang received the B.S. degree from the Zhengzhou University, Zhengzhou, China, in 2020. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include video enhancement, image processing and HDR imaging.

Mao Ye received the B.S. degree from Sichuan Normal University, Chengdu, China, in 1995, and the M.S degree from University of Electronic Science and Technology of China, Chengdu, China, in 1998 and Ph.D. degree from Chinese University of Hong Kong, China, in 2002, all in mathematics. He has been a short-time visiting scholar at University of Queensland, and University of Pennsylvania. He is currently a professor and director of CVLab with University of Electronic Science and Technology of China, Chengdu, China. His research interests include machine learning and computer vision. In these areas, he has published over 90 papers in leading international journals or conference proceedings. He has served on the editorial board of ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE. He was a co-recipient of the Best Student Paper Award at the IEEE ICME 2017.



Xiatian Zhu is a Senior Lecturer with Surrey Institute for People-Centred Artificial Intelligence, and Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, UK. He received his Ph.D. degree from the Queen Mary University of London. He won the Sullivan Doctoral Thesis Prize 2016. He was a research scientist at Samsung AI Centre, Cambridge, UK. His research interests include computer vision, and machine learning.

Shuai Li is currently with the School of Control Science and Engineering, Shandong University (SDU), China, as a Professor and QiLu Young Scholar. He was with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, China, as an Associate Professor from 2018-2020. He received his Ph.D. degree from the University of Wollongong, Australia, in 2018. His research interests include image video coding, 3D video processing and computer vision. He was a co-recipient of two best paper awards at the IEEE BMSB 2014 and IIH-MSP 2013, respectively.

Xue Li (Member, IEEE) received his Bachelor of Science (Computer Software) in 1981 from Chongqing University, China, Master of Computer Science in 1989 from the University of Queensland, and PhD of Information Systems in 1997 from Queensland University of Technology, Australia. He is currently a Professor with the School of Information Technology and Electrical Engineering, at The University of Queensland, Australia. His major areas of research interests and expertise include health data analytics, data mining, social computing, and intelligent web information systems.

Ce Zhu (M'03–SM'04–F'17) received the B.S. degree from Sichuan University, Chengdu, China, in 1989, and the M.Eng and Ph.D. degrees from Southeast University, Nanjing, China, in 1992 and 1994, respectively, all in electronic and information engineering. He held a post-doctoral research position with the Chinese University of Hong Kong in 1995, the City University of Hong Kong, and the University of Melbourne, Australia, from 1996 to 1998. He was with Nanyang Technological University, Singapore, for 14 years from 1998 to 2012, where he was a Research Fellow, a Program Manager, an Assistant Professor, and then promoted to an Associate Professor in 2005. He has been with University of Electronic Science and Technology of China, Chengdu, China, as a Professor since 2012. His research interests include video coding and communications, video analysis and processing, 3D video, visual perception and applications. He has served on the editorial boards of a few journals, including as an Associate Editor of IEEE Transactions on image processing, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Broadcasting, IEEE Signal Processing Letters, an Editor of IEEE Communications Surveys and Tutorials, and an Area Editor of Signal Processing: Image Communication. He has also served as a Guest Editor of a few special issues in international journals, including as a Guest Editor in the IEEE Journal of Selected Topics in Signal Processing. He is an APSIPA Distinguished Lecturer (2021-2022), and was also an IEEE Distinguished Lecturer of Circuits and Systems Society (2019-2020). He is a co-recipient of multiple paper awards at international conferences, including the most recent Best Demo Award in IEEE MMSP 2022, and the Best Paper Runner Up Award in IEEE ICME 2020.