KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction

Hu Wang[1] , **Mao Ye**[1*] , **Xiatian Zhu**[2*] , **Shuai Li**[3] , **Ce Zhu**[4] and **Xue Li**[5]
[1]School of CSE, University of Electronic Science and Technology of China, Chengdu, China
[2]Surrey Institute for People-Centred Artificial Intelligence, CVSSP, University of Surrey, Guildford, UK
[3]School of Control Science and Engineering, Shandong University, Jinan, China
[4]School of ICE, University of Electronic Science and Technology of China, Chengdu, China
[5]School of ITEE, The University of Queensland, Brisbane, Australia
wanghu0833cv@gmail.com, maoye@uestc.edu.cn, xiatian.zhu@surrey.ac.uk

Abstract

Recently, with the rise of high dynamic range (HDR) display devices, there is a great demand to transfer traditional low dynamic range (LDR) images into HDR versions. The key to success is how to solve the many-to-many mapping problem. However, the existing approaches either do not consider constraining solution space or just simply imitate the inverse camera imaging pipeline in stages, without directly formulating the HDR image generation process. In this work, we address this problem by integrating LDR-to-HDR imaging knowledge into an UNet architecture, dubbed as **Knowledge-inspired UNet** (KUNet). The conversion from LDR-to-HDR image is mathematically formulated, and can be conceptually divided into recovering missing details, adjusting imaging parameters and reducing imaging noise. Accordingly, we develop a basic knowledge-inspired block (KIB) including three subnetworks corresponding to the three procedures in this HDR imaging process. The KIB blocks are cascaded in the similar way to the UNet to construct HDR image with rich global information. In addition, we also propose a knowledge inspired jump-connect structure to fit a dynamic range gap between HDR and LDR images. Experimental results demonstrate that the proposed KUNet achieves superior performance compared with the state-of-the-art methods. The code, dataset and appendix materials are available at https://github.com/wanghu178/KUNet.git.

1 Introduction

Due to the limitation of existing (imperfect) hardware devices, people can only get photos with a certain range of brightness, i.e., common LDR images, which leads to the capture of LDR pictures with overexposed and underexposed areas as well as indifferent colors [Wang and Yoon, 2021]. The HDR image itself has a rich dynamic range, so it can express rich scene brightness and vivid colors. Restoring HDR image from LDR image has a very important practical significance [Eilertsen *et al.*, 2017b], and this is an ill-posed problem since multiple mappings exist between the LDR and HDR images.

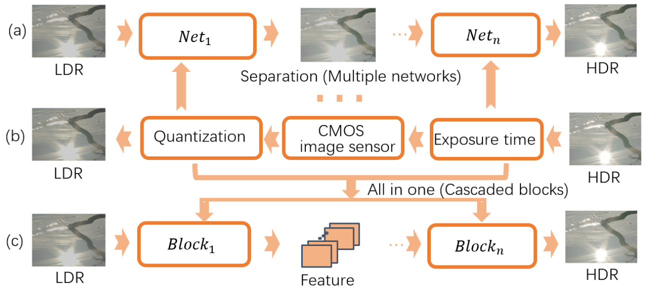

Figure 1: Comparison between the method of reversing the camera imaging pipeline and our method. (a) Multiple neural networks are designed to approximate the stages of reversing camera imaging pipeline where (b) describes a representative camera imaging process. (c) Our method uses a basic building block of UNet to simulate the LDR-to-HDR imaging formula.

There are three ways to construct HDR image [Kim *et al.*, 2020a]: *direct reconstruction*, *multi-exposure stack-based synthesis* and reconstruction by *reversing the camera imaging pipeline*. *Direct reconstruction* simply uses the modules in other research fields to generate HDR images in an end-to-end manner [Kim *et al.*, 2020b; Chen *et al.*, 2021a]. A pair of LDR-HDR image set is needed to train the model. These models are simple but do not in-depth consider HDR imaging mechanism; the noise reduction and recovery of the lost details are not enough.

The second line of methods use *multi-exposure LDR image stack* to reconstruct HDR image [Kim *et al.*, 2020a]. It uses a LDR image to generate multi-exposure LDR images to synthesize HDR image. In this way, different exposure information are used. The key to success for this line is how to generate accurate multi-exposure images and how to combine these multi-exposure images. For the last kind of methods, they *reverse the HDR-to-LDR image formation pipeline*, and then the HDR image generation task is decomposed into multiple subtasks [Liu *et al.*, 2020; Chen *et al.*, 2021b]. Multiple neural networks are used to ap-

[*] Mao Ye and Xiatian Zhu are corresponding authors. This work was supported by the National Key R&D Program of China (2018YFE0203900) and Sichuan Science and Technology Program (2020YFG0476).

proximate these subtasks to generate HDR images. However, the dependence between subtasks need to be carefully considered and how to train these cascaded networks becomes a new problem.

In this work, we propose a new approach which integrates the HDR imaging knowledge to an UNet architecture, named as *Knowledge-inspired UNet*(KUNet). Our approach not only absorbs the advantages of the simple structure of UNet, but also considers the imaging principle, which greatly reduces the solution space. The LDR-to-HDR mathematical formula is modeled by *Knowledge-Inspired Block* (KIBs) composed of three parts in charge of recovering missing details in overexposed area, adjusting imaging parameters and reducing LDR imaging noise, respectively. Then, similarly as UNet, in order to gather the features needed in HDR reconstruction, the features from multiple KIBs are converged to reconstruct HDR image. Furthermore, the direct jump connection in the traditional UNet is adjusted to fit the dynamic range gap with a *Knowledge-Inspired jump Connection* (KIC), which transfers the LDR features to the HDR features. And the directly connection transported features are scored to assist final HDR reconstruction.

Our contributions are three-fold: (1) We proposed a new knowledge-inspired UNet for HDR reconstruction approach. By analyzing the camera imaging pipeline, the HDR image restoration formula is derived. Based on this knowledge, the KIB is constructed to obtain the HDR image needed features. (2) A new jump connection structure is further designed to address the dynamic range gap between LDR and HDR. More useful features can be combined in the final reconstruction layer. To the best of our knowledge, we are the first to propose a specific jump-connect structure for HDR image reconstruction. (3) Despite its simple design, extensive experiments on HDR image and video reconstruction datasets prove that our KUNet outperforms a wide variety of the state-of-the-art methods.

2 Related Work

Multi-exposure stack HDR synthesis. The most common way of HDR reconstruction is to capture a series of LDR images with different exposures and then fuse these images. Wang et al., [Wang and Yoon, 2021] divided these methods into five types, namely, flow-based alignment, direction feature concatenation, correlation-guided alignment, image translation-based alignment, and deep static image fusion. Although the technical route based on multi-exposure LDR image synthesis is significant, it is difficult to find similar multi-exposure images for most LDR images in the real world.

Single-exposure HDR image reconstruction. Previous works built models to fit the nonlinear HDR imaging process. Although high efficiency can be achieved, the performance is not satisfied because of the approximation ability of traditional models. With the development of deep learning in last decades, there emerged many methods to learn LDR-to-HDR mapping. They can be roughly divided into three categories as mentioned before.

For the first approach of *direct reconstruction* method, HDRCNN [Eilertsen *et al.*, 2017a] was first proposed to restore the overexposed areas in LDR image to construct HDR image which only considers the recovery of overexposed areas. After that, a lot of works proposed different neural networks to fit the LDR-to-HDR mapping [Kim *et al.*, 2020b; Chen *et al.*, 2021a]. Although these methods have achieved good results, these methods either directly use modules in other research fields, or use some of HDR imaging knowledge as conditions. There still exists a lot of room for improvement. For the line of *multi-exposure stack-based synthesis*, they pass an image through CNN to generate LDR images with different exposures, and then merge these images together to generate HDR images [Kim *et al.*, 2020a]. The key to success of this approach is the quality of the generated multi-exposure images which is affected by many factors and how to train these models is also not easy. The last approach [Liu *et al.*, 2020] *reverses camera imaging pipeline*, multiple networks are employed to learn the stages of inverse camera imaging pipeline to generate HDR images. After that, the method HDRTV [Chen *et al.*, 2021b] extends it for HDR TV task. This kind of methods have shown very good results. However, this type of method uses multiple network splicing which has a strong dependence between these networks. Except a large number of parameters, this architecture will bring troubles to model generalization and training.

Instead of this approach, we absorb the advantages of UNet which can easily integrate different scale features and use a basic building block to approximate the LDR-to-HDR formula. This knowledge-inspired UNet will merge different scale features which are needed for HDR image reconstruction. So the performance is further improved.

Jump-connect structure. Since the use of the jump-connect structure in the UNet network [Ronneberger *et al.*, 2015], the jump-connect structure has received widespread attention. MultiResUNet [Ibtehaz and Rahman, 2020] uses several convolutional blocks to construct Res path to improve the jump-connect structure. NBNet [Cheng *et al.*, 2021] also uses some convolutional blocks, but after convolution, the attention mechanism combined with the information on the decoding side is used. [Wang *et al.*, 2021] summarizes the information that needs to be jumped on the decoding end, and Transformer is employed.

These structures are designed for their respective tasks, which cannot be directly used for HDR generation since LDR-to-HDR image reconstruction has the problems of dynamic range gap and overexposed areas. So we propose a new jump-connect structure which does mapping from LDR feature to HDR needed features, reduces the noise caused by HDR-to-LDR imaging process and decreases ghosting during LDR-to-HDR generation.

3 Analysis of HDR Image Reconstruction

LDR image formation formula is proposed in [Hasinoff *et al.*, 2010] as follows,

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

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

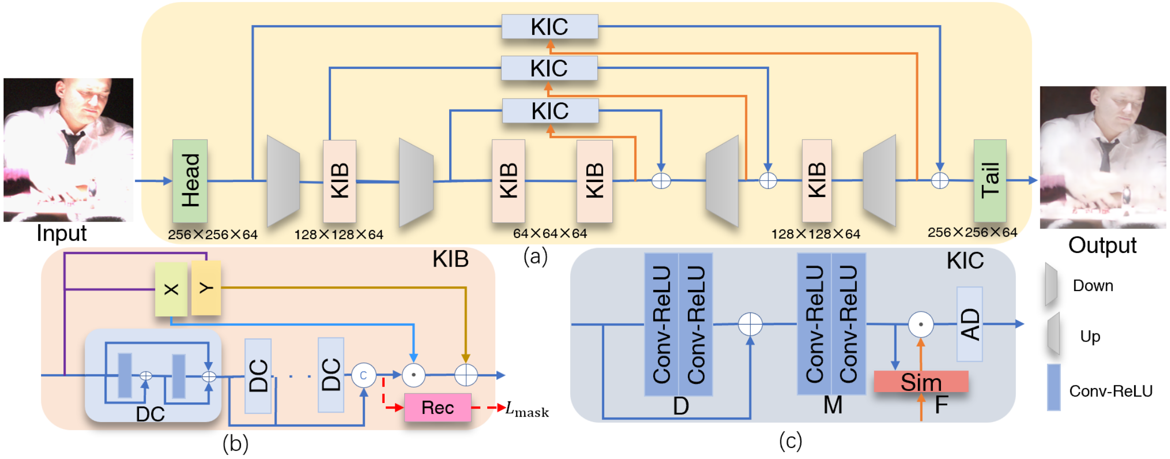

Figure 2: The overall structure of the proposed framework. (a) An Unet architecture is with a knowledge-inspired direct jump connection (KIC) where Head and Tail are feature extraction and reconstruction blocks respectively. (b) A knowledge-Inspired Block (KIB) is composed of three parts: imaging parameter adjusting (X), imaging noise reduction (Y) and missing overexposed features recovering (R). (c) KIC also consists of three main parts: high frequency feature enhancement (D), LDR-to-HDR feature mapping (M) and feature filtering (F). DC and AD mean dense connection and adaptive fine-tuning operation respectively.

mentioned in [Pérez-Pellitero *et al.*, 2021], which can be assumed as HDR pixel value. I_L represents a LDR image pixel value and n is the sensor noise. Unsaturation represents the pixels that can be represented by the LDR 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} [Pérez-Pellitero *et al.*, 2021].

This formula is widely used in LDR image formation which inspires us to generate the HDR images in a similar theoretically guided way. Suppose we have a camera with unlimited capture capabilities, the corresponding saturated pixel value in Eq. (1) of the LDR images can be represented as follows,

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

where $I_{overexposed}$ 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, $I_{overflow} = 0$ since no difference exists between the ideal and real cameras. By combining Eq.(2) and Eq.(1), the LDR formation process as can be unified as

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

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

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

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

The above Eq.(4) formulates the restoration of HDR images from the perspective of image pixel intensities. However, in the image domain, the direct estimation of $I_{overflow}$, g and t is very difficult. Considering the power of deep learning in feature representation and learning, in this paper, we formulate and estimate formula (4) of the HDR image restoration process in the feature domain using neural networks. Each function is formulated by a sub-network and together restores the HDR features, then generates HDR images. The detail of this knowledge-inspired model will be introduced in Section 4.

4 The Proposed Method

The framework of the proposed method is shown in Fig.2(a), which is based on an UNet-like architecture. To obtain more HDR needed features, the receptive field of a larger area by down sampling makes the convolution operation capture more information. So we choose the Unet-like architecture as the backbone of our base network.

The *Knowledge-inspired UNet* (KUNet) consists of five stages. Head stage extracts features from the input LDR image I_L; while Tail stage reconstructs a HDR image based on the merged features. Between the Head and Tail stages, there is a jump-connection (KIC). After transforming the LDR image into the feature space, four KIBs are used to transfer the LDR features to HDR features according to formula (4). To fully explore the global information with a large receptive field, two KIBs are used in a smaller scale in a UNet structure. Finally, the features from the fourth KIB are integrated with the features from LDR image, transferred through KIC to formulate the HDR features, and a final Tail module is used to generate the HDR image from the features, to reconstruct HDR image $\hat{I}_H$. In the following, we will introduce the main parts separately.

4.1 Head and Tail

The main function of Head stage is to transfer the LDR image to a rich feature space for the following LDR-to-HDR feature conversion. It does not require overly complex operations. A simple subnetwork of three layer convolutions and ReLU activation functions are employed to complete this task, which is denoted by

$$F_{out} = (ReLU \circ Conv_{3\times3})^3(I_L) \tag{5}$$

where $(\cdot)^n$ represents the serial cascade of n modules. F_{out} represents the output of Head. Correspondingly, the Tail stage reconstructs the image from the features. So a symmetrical structure is adopted at the Tail stage as follows,

$$\hat{I}_H = (ReLU \circ Conv_{3\times3})^3(F^p_{out}) \tag{6}$$

where $\hat{I}_H$ is the reconstructed HDR image as mentioned before; $F^p_{out} = f(F_{out})$ represents the output features of the intermediate process of KUNet.

4.2 Knowledge-Inspired Block

As stated in Section 3, formula (4) can be used to reconstruct the HDR image from the image intensity perspective. However, it is hard to direct formulate the functions in the image intensity domain. Therefore, we turn to the feature space and take advantage of the representation ability of deep learning. Consequently, the formula can be characterized in feature space as below,

$$\underbrace{\phi}_{H_F} \;=\; \underbrace{\frac{g}{t}}_{X(L_F)} \underbrace{(I_L - I_0 + I_{\text{overflow}})}_{\odot\; R(L_F)} - \underbrace{\frac{g}{t}}_{+Y(L_F)} n, \tag{7}$$

where $\odot$ denotes the element-wise multiplication. L_F and H_F represent the input LDR feature from previous module and the feature to reconstruct the output HDR image, respectively. $R(\cdot)$ corresponds to the part of formula (4) for compensating the missing overexposed features to reconstruct HDR; $X(\cdot)$ is in charge of modulating the compensated features $R(L_F)$ to HDR feature domain; $Y(\cdot)$ corresponds to the LDR imaging noise reduction part of formula (4). With the help of this form of expression in feature space, our knowledge-inspired block is developed as shown in Fig.2(b), consisting of three parts of networks which fit the functions $R(\cdot)$, $X(\cdot)$ and $Y(\cdot)$, respectively.

For the network fitting the $R(\cdot)$ function, as shown in Fig.2(b), k Densely Connected (DC) blocks are cascaded and concatenated to obtain the compensated features. It is denoted by:

$$R = Cat(DC(L_F), \cdots, DC^k(L_F)), \tag{8}$$

where $Cat(\cdot)$ means the concatenation operation. The DC block consists of two layers of convolution and ReLU activation function ($ReLu \circ Conv_{3\times3}$), which are also densely connected to guarantee the information from upper layer will not be lost. In this way, the information from L_F is retained as much as possible. For recovering the overexposed area, a mask loss is proposed for the R neural network of the ith KIB block as the following,

$$\begin{aligned} L^i_{\text{mask}} =& \|I_H \odot M - K_i \odot M\|_1 \\ & + \gamma \|I_H - K_i\|_1 \end{aligned} \tag{9}$$

where I_H represents the real HDR image. γ controls the recovery of overexposed area under a stage-wise consistency regularization between a reconstructed image and the true HDR image. Often, reconstructed images are not highly accurate and we set small γ to avoid over-penalty. So the parameter γ = 0.1 is fixed in our implementation. K_i is a reconstructed image through a reconstruction branch. When $i = 1$ or 4, it is composed of an upsampling and a convolution layers ($Up \circ Conv_{3\times3}$); while for other cases, it is composed of two upsampling and a convolution layers ($Up^2 \circ Conv_{3\times3}$). M represents an overexposed area mask. As mentioned in [Yu *et al.*, 2021], the overexposure mask is calculated as the following,

$$M = \begin{cases} 1, & if \;\; \frac{1}{3}\sum_c I(x,y,c) < \tau, \\ 0, & \textbf{otherwise} \end{cases} \tag{10}$$

where c is the color channel index and τ is 0.83 [Yu *et al.*, 2021]. Through this loss function, our R networks pays more attention to recover the information of overexposed area, which much better fits formula (4). It should be noted that the mask branch is only used in training phase, which does not bring any additional overhead at inference phase.

For the X and Y networks, it is denoted in [Liu *et al.*, 2020] that the L_F feature contains information to approximate the camera imaging parameter and noise. Inspired by [He *et al.*, 2020], we use two 1×1 convolution layers to simulate these functions. The process can be described as follows:

$$X = Conv_{1\times1} \circ Conv_{1\times1}(L_F), \tag{11}$$

$$Y = Conv_{1\times1} \circ Conv_{1\times1}(L_F). \tag{12}$$

Remark. Compared with the previous methods, our KIB fits the HDR imaging formula. The solution space is constrained and the many-to-many LDR-to-HDR mapping problem is reduced. In addition, under the adjustments of X and Y modules, the function of generating HDR features is adaptive to the different LDR images.

4.3 Knowledge-Inspired Jump Connection

The information at the front end of the network are also useful for HDR image reconstruction. As UNet does, we also use direct jump connection to transport this information. Assume the front end feature is F and the feature at the transport destination is P, tradition direct jump connection formula is

$$\hat{F} = F + P, \tag{13}$$

where $\hat{F}$ represents the fused feature. This type of connection does not fit our LDR-to-HDR problem. There exist two defects. (1) As denoted in Eq.(1), for LDR image formation pipeline, a lot of noise will be produced which should be eliminated when transported to the destination. (2) The dynamic range gap exists between the LDR and HDR images which will cause the feature space not consistent between the front and back ends. Based on this knowledge, we design a new KIC jump-connect structure which pays attention to the noise reduction and feature mapping.

For reducing LDR imaging noise, we use a residual structure as shown in Fig.2(c). The correct pixels in the unsaturated regions are directly connected to the output; while the

Method	Venue	PSNR ↑	PSNR-μ ↑
LandisEO[Landis, 2002]	SIGGRAPH02	17.88	23.30
HuoPhyEO [Huo *et al.*, 2014]	TVC14	32.40	17.35
SingleHDR[Liu *et al.*, 2020]	CVPR20	32.32	19.54
HDRCNN [Eilertsen *et al.*, 2017a]	ACM TOG17	39.47	26.05
Deep SR-ITM [Kim *et al.*, 2019]	ICCV19	43.29	26.25
ResNet [He *et al.*, 2016]	ECCV16	41.92	33.24
HDRUNet [Chen *et al.*, 2021a]	CVPRW21	44.10	33.59
KUNet	Ours	44.83	33.67

Table 1: Quantitative comparisons on the NTIRE2021 dataset. Red text indicates the best.

high frequency or the information in the over-exposure area are processed by two convolutional layers. Here, since most of the HDR processing has already been conducted in the KIB branch, for simplicity, all the features are bypassed through the direction connection and the convolution processing for the over-exposure area learns the difference between the original features and the over-exposed features. It is the D network in Fig.2(c) which is denoted by

$$D = Conv_{3\times3} \circ ReLU \circ Conv_{3\times3} \circ ReLU(F) + F. \qquad (14)$$

For the feature mapping from LDR-close feature space to HDR-close feature space which is shown as the M part in Fig.2(c), two simple two convolutional layers are employed as

$$\tilde{F} = Conv_{3\times3} \circ ReLU \circ Conv_{3\times3} \circ ReLU(D). \qquad (15)$$

It has also been noted that a small number of convolution blocks are effective for HDR image reconstruction [Chen *et al.*, 2021b].

Furthermore, in order to suppress useless information and reduce the visual ghosting, we further filter the features processed by the above networks. Scoring mechanism is employed. The final filtered feature is

$$\hat{F} = AD(Sim(\tilde{F}, P) \odot \tilde{F}) + P \qquad (16)$$

where AD represents a 1×1 convolution which is used to adjust the scored feature. We adopt Cosine similarity in [Zhang *et al.*, 2020] as follows,

$$Sim(\tilde{F}, P) = \frac{\tilde{F} \cdot P}{\max(||\tilde{F}||_2 \cdot ||P||_2, \varepsilon)}$$

where ε represent a very small parameter which is used to prevent division by zero.

Remark. Ideally, the jump-connect structure should solve the problems of noise reduction, feature mapping, and over-exposure pixel value discovering. The over-exposure area recovery is not implemented in here due to two reasons: (1) The cascaded KIBs already try to find features to recover the over-exposure area; (2) A simple jump-connect structure cannot handle this situation very well.

4.4 Loss Function

Our total loss function is divided into a main loss function and a mask loss function, which is denoted by

$$\text{Loss}(I_H, \hat{I}_H) = L_{\text{main}}(I_H, \hat{I}_H) + \beta L_{\text{mask}}(I_H, \hat{I}_H) \qquad (17)$$

where $L_{\text{main}} = \|I_H - \hat{I}_H\|_1$ and $L_{\text{mask}} = \sum_{i=1}^{4} L^i_{\text{mask}}$. Similar as KIB's R network, the mask loss with the weight controlled by β is also designed to recover the overexposed area, playing an auxiliary role. Given that, β is also set small. So the parameter $\beta = 0.01$ is fixed in our implementation. Compared with the approaches based on reversing the camera imaging pipeline, our loss function is rather simple and no hyper-parameter exists. The training of our method is also simple and end-to-end.

5 Experiments

5.1 Experiment Setup

Datasets. We use two data sets to evaluate our method, i.e., for image and video tasks. Both of these two data sets contain information about moving light sources, rich colors, highlights and bright. For the image task, following the works in [Chen *et al.*, 2021a; Liu *et al.*, 2021], the NITRE 2021 dataset is used which was proposed in **NITRE 2021 HDR Challenge** [Pérez-Pellitero *et al.*, 2021] selected from HDM HDR dataset [Froehlich *et al.*, 2014]. This data set only contains HDR images. Since the ground truth of test and validation images are not available, by similar operation in [Chen *et al.*, 2021a], the original training set is decomposed into two parts for training and test. They are 1416 paired training images and 78 test images. For the video task, we conduct experiment on **HDRTV** [Chen *et al.*, 2021b]. This dataset is obtained from 22 HDR10 standard videos and these videos comply with the Rec.2020 standard. It contains 1235 paired training pictures and 117 test pictures. It is mentioned in [Chen *et al.*, 2021b] that the LDR-to-HDR image task is different from the SDRTV-to-HDRTV task. However, for our method, the restoration task is limited to a certain brightness interval, which is consistent with the restoration goal of the SDRTV-to-HDRTV task. So the HDRTV is also selected as our comparison data set.

Evaluation metrics. Since the video and image tasks have different recovery goals, they have different measurement methods. For the image data set, we follow the evaluation methods in [Chen *et al.*, 2021a; Liu *et al.*, 2021; Pérez-Pellitero *et al.*, 2021] to use PSNR and PSNR-μ [Demetris *et al.*, 2018]. For the video data set, we follow the comparison method in [Chen *et al.*, 2021b], using PSNR, SSIM, SR-SSIM [Zhang and Li, 2012], HDRVDP3 [Mantiuk *et al.*, 2011] and δE_{ITP}.

In order to be consistent with other video measurement methods, we use the same HDRVDP3 parameters in [Chen *et al.*, 2021b].

Implementation details. All models are built on the PyTorch framework. Due to space limitations, more details can be obtained from the Appendix.

5.2 Comparison with State-of-the-art

Compared methods. For the HDR image data set, we compare our method KUnet with 7 State-Of-The-Art (SOTA) methods. They are LandisEO [Landis, 2002], HuoEo [Huo *et al.*, 2014], HDRCNN [Eilertsen *et al.*, 2017a], SingleHDR [Liu *et al.*, 2020], DEEP SR-ITM [Kim *et al.*, 2019]

	Method	Venue	Params	PSNR ↑	SSIM ↑	SR-SIM ↑	Δ_{ITP} ↓	HDR-VDP3 ↑
LDR-HDR	HuoPhyEO[Huo *et al.*, 2014]	TVC14	-	25.90	0.9296	0.9881	38.06	7.893
	KovaleskiEO[Kovaleski and Oliveira, 2014]	SIBGRAPI14	-	27.89	0.9273	0.9809	28.00	7.431
Image-to-image translation	ResNet[He *et al.*, 2016]	ECCV16	1.37M	37.32	0.9720	0.9950	9.02	8.391
	Pixel2Pixel [Isola *et al.*, 2017]	CVPR17	11.38M	25.80	0.8777	0.9871	44.25	7.136
	CycleGAN [Zhu *et al.*, 2017]	ICCV17	11.38M	21.33	0.8496	0.9595	77.74	6.941
Photo retouching	HDRNet [Gharbi *et al.*, 2017]	ACM TOG17	482K	35.73	0.9664	0.9957	11.52	8.462
	CSRNet [He *et al.*, 2020]	ECCV20	36K	35.04	0.9625	0.9955	14.28	8.400
	Ada-3DLUT [Zeng *et al.*, 2020]	TPAMI20	594K	36.22	0.9658	0.9967	10.89	8.423
SDRTV-to-HDRTV	Deep SR-ITM [Kim *et al.*, 2019]	ICCV19	2.89M	37.10	0.9686	0.9950	9.24	8.233
	JSI-GAN [Kim *et al.*, 2020b]	AAAI20	1.06M	37.01	0.9694	0.9928	9.36	8.169
	AGCM+LE [Chen *et al.*, 2021b]	ICCV21	1.41M	37.61	0.9726	0.9967	8.89	8.613
	AGCM+LE+HG [Chen *et al.*, 2021b]	ICCV21	37.20M	37.21	0.9699	0.9968	9.11	8.569
Ours	KUNet	Ours	1.12M	37.78	0.9868	0.9971	7.80	8.393

Table 2: Quantitative comparisons on the HDRTV dataset. Red text indicates the best.

and ResNet [He *et al.*, 2016], and HDRUnet [Chen *et al.*, 2021a].

For the HDR video data set. we combine our methods with four types of methods including SDRTV-to-HDRTV, image-to-image translation, photo retouching and LDR-to-HDR [Chen *et al.*, 2021b].

Quantitative comparison. It can be seen from Table 1 and Table 2 that our model shows excellent performance in the data sets for both of image and video tasks. For the NTIRE dataset, SingleHDR and traditional algorithms [Huo *et al.*, 2014; Landis, 2002] are not very suitable for this problem. This is because their goals are to restore the relative brightness instead of restoring the absolute brightness. So the performance is not satisfied. While DEEP SR-ITM and HDRCNN do not consider the denoising issue, the performance is slightly worse. Although HDRUNET does considering the problem of reducing noise, it does not consider the difference between the traditional image and HDR generation problems. In general, our performance is SOTA.

For HDRTV dataset, the Δ_{ITP} of KUnet far exceeds other SOTA algorithms. It shows that our model is satisfactory on color gamut recovery in this video data set. The reason is that the original intention of our model considers that the generated images contain different color gamuts, but the color gamut in the HDRTV set is consistent, i.e., it has a smaller solution space compared to different color gamuts, so we can achieve excellent performance. From Table 2, it is worth noting that KUNet achieves SOTA performance except the HDRVDP3 index. The reason is that KUNet is trying to reconstruct a HDR image that is close to the original HDR image in terms of color, bit depth, etc., however, HDRVDP3 is used to evaluate human visual perception in a specific range. Therefore, KUNet does not achieve the SOTA performance on this index, while it is still competitive compared with other methods. As shown in Appendix, with an additional perception loss, the HDRVDP3 index will be improved.

Due to space limitations, the visual analysis are shown in Appendix. There are three basic observations. (1) Compared with the existing methods, KUNet model can show satisfactory visual effects for both image and video datasets. (2) KUNet can effectively remove imaging noise. (3) The KIC module can accelerate model training compared with only direct jump connection.

R	X	Y	Skip	DM	F	L_{mask}	PSNR↑	PSNR-μ ↑
✓	-	-	✓	-	-	-	44.17	33.35
✓	✓		✓	-	-	-	44.37	33.63
✓	-	✓	✓	-	-	-	44.16	33.48
✓	✓	✓	✓	-		-	44.50	33.66
✓	✓	✓	-	✓	-	-	44.79	33.60
✓	✓	✓	-	✓	✓	-	44.80	33.67
✓	✓	✓	-	✓	✓	✓	44.83	33.67

Table 3: Ablation analysis.

5.3 Ablation Studies

In this part, we will do ablation analysis on KIB and KIC modules. The experiments are performed on the NTIRE2021 dataset. As mentioned before, **KIB** has three parts. One part is the core module D, which is also our base block, and other parts are the adaptive X and Y branches for fine-tuning the generated HDR features from R module. From the first row in Table 3, it can be seen that only using the D module can also achieve an acceptable result where 'Skip' means using directly jump connection. Then, with the X and Y branches, the module has stronger expressive ability and the performance is increased. As shown in formula (4) that the X and Y branches must be used at the same time. Through experiments, we have also proved that our analysis is correct. **KIC** consists of two parts: the noise reduction (D) and feature mapping and filter (F). From the fifth and sixth rows in Table 3, we can find that these two parts work very well. The last row shows that the model with the loss L_{mask} can achieve the best results.

6 Conclusion

In this paper, we analyzed the HDR-to-LDR imaging process and obtained the HDR image formulation formula, which inspired us to propose a new model KUnet. The KIB block is combined with the UNet network to reconstruct the HDR image. In addition, we made a preliminary improvement on the incompatibility of the UNet jump-connect structure to the problem of HDR image restoration, and proposed the KIC branch. It is used to assist HDR image restoration, and achieved very good results. Experiments demonstrated that our method KUnet has achieved the SOTA results on HDR image and video datasets.

References

[Chen *et al.*, 2021a] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. HDRUnet: Single image HDR reconstruction with denoising and dequantization. In *CVPR*, pages 354–363, 2021.

[Chen *et al.*, 2021b] Xiangyu Chen, Zhengwen Zhang, Jimmy S Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. A new journey from SDRTV to HDRTV. In *ICCV*, pages 4500–4509, 2021.

[Cheng *et al.*, 2021] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. NBNet: Noise basis learning for image denoising with subspace projection. In *CVPR*, pages 4896–4906, 2021.

[Demetris *et al.*, 2018] Marnerides Demetris, Bashford-Rogers Thomas, Hatchett Jonathan, and Debattista Kurt. EXPNet:A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *CGF*, volume 37, pages 37–49. Wiley Online Library, 2018.

[Eilertsen *et al.*, 2017a] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM TOG*, 36(6):1–15, 2017.

[Eilertsen *et al.*, 2017b] Gabriel Eilertsen, Rafal Konrad Mantiuk, and Jonas Unger. A comparative review of tone-mapping algorithms for high dynamic range video. In *CGF*, volume 36, pages 565–592. Wiley Online Library, 2017.

[Froehlich *et al.*, 2014] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, volume 9023, page 90230X. SPIE, 2014.

[Gharbi *et al.*, 2017] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM TOG*, 36(4):1–12, 2017.

[Hasinoff *et al.*, 2010] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *CVPR*, pages 553–560. IEEE, 2010.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016.

[He *et al.*, 2020] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *ECCV*, pages 679–695. Springer, 2020.

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

[Ibtehaz and Rahman, 2020] Nabil Ibtehaz and M Sohel Rahman. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks*, 121:74–87, 2020.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017.

[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 *ICCV*, pages 3116–3125, 2019.

[Kim *et al.*, 2020a] Jung Hee Kim, Siyeong Lee, and SukJu Kang. End-to-end differentiable learning to HDR image synthesis for multi-exposure images. *AAAI*, 2020.

[Kim *et al.*, 2020b] 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.

[Kovaleski and Oliveira, 2014] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *SIBGRAPI*, pages 49–56. IEEE, 2014.

[Landis, 2002] Hayden Landis. Production-ready global illumination. In *Siggraph*, volume 5, pages 93–95, 2002.

[Liu *et al.*, 2020] Yu-Lun Liu, WeiSheng Lai, YuSheng Chen, YiLung Kao, MingHsuan Yang, YungYu Chuang, and JiaBin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *CVPR*, pages 1651–1660, 2020.

[Liu *et al.*, 2021] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *CVPR*, pages 463–470, 2021.

[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.

[Pérez-Pellitero *et al.*, 2021] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *CVPR*, pages 691–700, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Wang and Yoon, 2021] Lin Wang and Kuk-Jin Yoon. Deep learning for HDR imaging: State-of-the-art and future trends. *arXiv preprint arXiv:2110.10394*, 2021.

[Wang *et al.*, 2021] Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. UCTransNet: Rethinking the skip connections in U-Net from a channel-wise perspective with transformer. *arXiv preprint arXiv:2109.04335*, 2021.

[Yu *et al.*, 2021] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Zou Qin, and Chunxia Xiao. Luminance attentive networks for HDR image and panorama reconstruction. *arXiv preprint arXiv:2109.06688*, 2021.

[Zeng *et al.*, 2020] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *TPAMI*, 2020.

[Zhang and Li, 2012] Lin Zhang and Hongyu Li. Sr-sim: A fast and high performance iqa index based on spectral residual. In *ICIP*, pages 1473–1476. IEEE, 2012.

[Zhang *et al.*, 2020] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. SG-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Trans Cybern*, 50(9):3855–3865, 2020.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.