KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction Technical Appendix

Hu Wang[1], **Mao Ye**[1]*, **Xiatian Zhu**[2], **Shuai Li**[3], **Ce Zhu**[4] and **Xue Li**[5]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, 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,eczhu}@uestc.edu.cn, xiatian.zhu@surrey.ac.uk, shuaili@sdu.edu.cn, xueli@itee.uq.edu.au

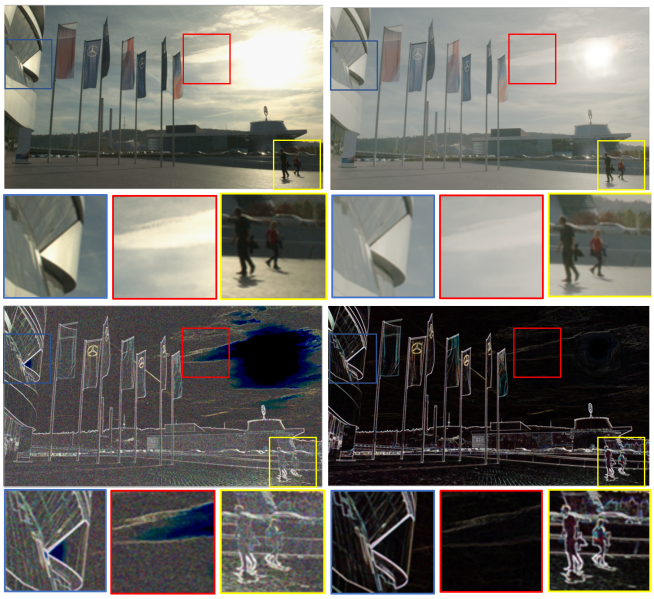

Figure 1: Gradient maps calculated by Scharr operator [Scharr, 2004]. The left column picture is the LDR image; the right column picture is the HDR image reconstructed by KUNet. *Zoom in for best view.*

1 Visual analysis

All data sets are stored in 16 bits "PNG" format. NTIRE image data set needs to be processed by tone mapping to display normally. Here we use a common method to process HDR image through gamma mapping, and map it to the LDR domain for visualization. The HDRTV data set belongs to the video data set. When it is displayed, the SDR screen has processed it. We follow the method in [Chen *et al.*, 2021b] using gamma *electro-optical transfer function* (EOTF). The actual effect on the SDR screen will be slightly darker than that on the HDR screen. For fairness comparison here, we do not have any additional processing on it.

*This work was supported by the National Key R&D Program of China (2018YFE0203900) and Sichuan Science and Technology Program (2020YFG0476). The corresponding author is Mao Ye.

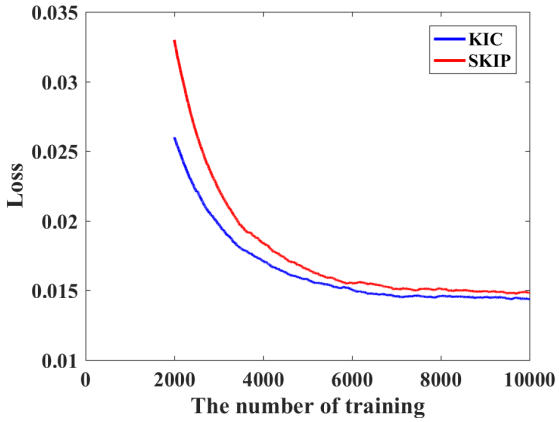

Figure 2: Illustration of training acceleration with KIC. For a clearer display, the loss values are smoothed and only displayed for 2000-10000 iterations. We can see that the convergence of the method with KIC is faster.

1.1 Denoising performance

An important task in the reconstruction of HDR image is denoising. In order to show the denoising performance of KUNet more intuitively. We use scharr operator [Scharr, 2004] to display the gradient maps (Fig.1) of the LDR and the corresponding HDR images. It can be found that our method can effectively remove LDR imaging noise. At the same time, due to the camera imaging pipeline, the areas where the gradient of the LDR image in the highlight area is not so obvious which can also be seen after restoration.

1.2 Training acceleration with KIC

In order to further show the effectiveness of KIC module. Here, we analyze the convergence speed of the model with (KIC) and without KIC module (SKIP), as shown in the Fig.2. It can be seen that the KIC module makes up the dynamic semantic gap between LDR-HDR features. The convergence speed of the model with KIC is accelerated and the performance is also improved.

1.3 Qualitative comparison

The results of qualitative comparison are shown in Fig.5. We compare our method KUnet with 7 State-Of-The-Art (SOTA)

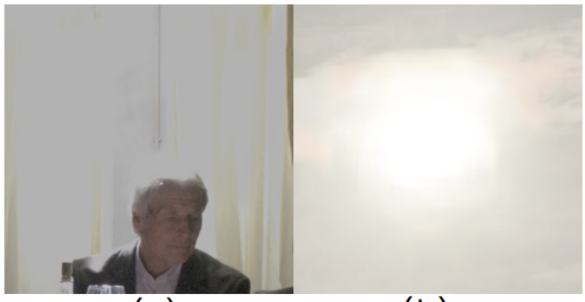

Figure 3: Failure case. (a) Trained by HDRTV dataset when applied on NTIRE 2021 dataset; (b) Over-large overexposed area.

methods. They are LandisEO [Landis, 2002], HuoEo [Huo *et al.*, 2014], HDRCNN [Eilertsenl *et al.*, 2017], SingleHDR [Liu *et al.*, 2020], DEEP SR-ITM [Kim *et al.*, 2019] and ResNet [He *et al.*, 2016], and HDRUnet [Chen *et al.*, 2021a]. The first two rows are the results on an image from NTIRE data set. We can see that most methods are powerless for the information lost in the overexposed area (blue box). Although HDRCNN can generate information about the overexposed area to a certain extent, it cannot solve the noise problem (purple box). The remaining rows are results on two images from HDRTV data set. It can be seen that our model can generate smoother colors in the sky area (green boxes in third and fourth rows). Our method has no faults. At the same time, we can ensure the smoothness of the entire image while ensuring its authenticity (red and yellow boxes in the last four rows).

1.4 Failure case discussion

A failure case is shown in Fig.3. (1) Limited training data might affect model generalization. For example, the model trained by HDRTV dataset does not work very well when applied on NITRE 2021 dataset (Fig.3 (a)). However, we note that, this limitation is mostly on the training data, typical for the majority of learning based methods, and should be better resolved by improving the training data, instead of model design. (2) In case of over-large overexposed area (e.g.,the sun image in Fig.3(b)), all existing methods do not work well. This is because large exposure area presents only much useless information whilst lacks useful contextua information, making the reconstruction extremely challenging. Although our model is superior over previous methods, this issue is still not completely solved. Potential solutions include constructing a larger and more diverse dataset, and further reducing the solution space.

2 Supplement experiment on HDRVDP3

As mentioned in the paper, KUnet focuses on the exact HDR image construction. We only choose the general L_1 loss function. However, as mentioned in [Santos *et al.*, 2020], just using pixel-wise loss function may produce blurry images. This leads to a drop in HDRVDP3 evaluation metric in some cases. Therefore, based on the inspiration of the work in [Santos *et al.*, 2020], we add a perceptual loss function (VGG loss function [Justin *et al.*, 2016]) to the total loss for improving HDRVDP3 index. The details of this loss function can be described as follows,

Table 1: The results of KUNet with perceptual loss on the HDRTV dataset. KUNet$_P$ represents the result with perceptual loss function

Method	KUNet	KUNet$_P$
PSNR↑	37.78	37.90
SSIM↑	0.9871	0.9868
SR-SIM↑	0.9973	0.9981
Δ_{ITP}↓	7.80	7.84
HDR-VDP3↑	8.393	8.533

$$L_P(I_H, \hat{I}_H) = f_{vgg19}(I_H, \hat{I}_H), \tag{1}$$

where f_{vgg19} assess how similar the features extracted from the reconstructed HDR image are to the real HDR image. With the help of this loss function, KUNet can produce texture details similar to real images at a deep level. After using the perceptual loss function, our total loss function is as follows:

$$L_{improve} = \text{Loss}(I_H, \hat{I}_H) + L_P(I_H, \hat{I}_H). \tag{2}$$

With the introduction of the perceptual loss function into our model, we perform experiment on HDRTV dataset. From Table 1, it can be found that HDRVDP3 and many indicators have been significantly improved under the premise of the small performance decrease of SSIM and Δ_{ITP} loss indicators. This phenomenon is reasonable, because emphasizing the perception effect will weaken the performance of the accuracy indicator. At the same time, this also inspires us future research direction. According to the characteristics of HDR image/video reconstruction, a new loss function is required to balance the reconstruction quality and visual effects.

3 Implementation and network details

3.1 Implementation details

For the Head, Tail, D block of KIB module and KIC module, we use 3×3 convolution with a step size of 1, and for the X and Y branches in the KIB module, 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 64. The down-sampling operation uses a 3×3 convolution operation with a step size of 2. Up-sampling operation uses PixelShuffle [Shi *et al.*, 2016]. We use ADAM optimizer [Kingma and Ba, 2014] with the learning rate of $2e^{-4}$ decayed by a factor of 2 after every $20K$ iterations. The batch size is 12. All models are built on the PyTorch framework and trained with NVIDIA GeForce RTX 2080 SUPER. The total training time is about 6 days.

3.2 Network details

As we mentioned in Section 4.1 of paper, we adapt an UNet-like architecture. In this section, we will present the details of our architecture. During the construction of the model, KUNet does not use any complex mechanisms. All key modules are obtained by using 3×3 convolution and 1×1 convolution. The "strides" and "padding" of these convolution

Table 2: The detail of KUnet. PS represents the Pixel Shuffle [Shi *et al.*, 2016] layer.

KUNet					
Block	Layer(filter size)	filters	Block	Layer(filter size)	filters
Head	Conv2D(3,3)*3	64	Tail	Conv2D(3,3)*3	64
DC	Conv2D(3,3) Conv2D(3,3)	64 64	KIC	Conv(3,3)*4 Conv2D(1,1)	64 64
KIB 1 KIB 4	DC*4 Conv2D(1,1)*2	64 64	Rec 1 Rec 4	Conv2D(3,3)*2 PS(2)	256 64
KIB 2 KIB 3	DC*6 Conv2D(1,1)*2	64 64	Rec 2 Rec 3	Conv2D(3,3)*3 PS(2)*2	256 64
Down	Conv2D(3,3)	64	Up	Conv2D(3,3) PS(2)	256 64

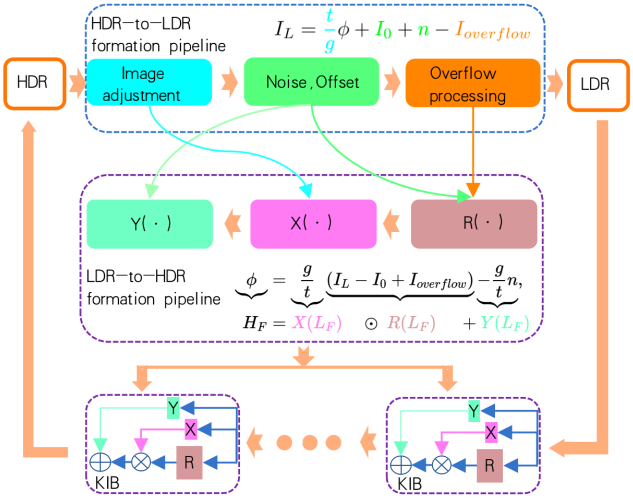

Figure 4: Illustration of the connection among the HDR-to-LDR formation pipeline (top), our derived LDR-to-HDR formation (middle), and our entire KUNet (bottom).

blocks are both 1. The detail of each module can be seen in Table 2.

4 The connection between KUNet and the LDR-to-HDR formation pipeline

As shown in Fig.4, the HDR-to-LDR formation pipeline (top) is related to our derived formation (middle) at the component level. As a whole, our KUNet conducts a series of LDR-to-HDR processes as each KIB unit is designed specially to simulate an individual such process (bottom). This design takes a divide-and-conquer strategy.

References

[Chen *et al.*, 2021a] X. Chen, Y. Liu, Z. Zhang, Y. Qiao, and C. Dong. HDRUnet: Single image HDR reconstruction with denoising and dequantization. In *CVPR*, pages 354–363, 2021.

[Chen *et al.*, 2021b] X. Chen, Z. Zhang, J. Ren, L. Tian, Y. Qiao, and C. Dong. A new journey from SDRTV to HDRTV. In *ICCV*, pages 4500–4509, 2021.

[Eilertsenl *et al.*, 2017] G. Eilertsenl, J. Kronanderl, G. Denes, R. Mantiuk, and J. Unger. HDR image reconstruction from a single exposure using deep CNNs. *ACM TOG*, 36(6):1–15, 2017.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer, 2016.

[Huo *et al.*, 2014] Y. Huo, F. Yang, L. Dong, and V. Brost. Physiological inverse tone mapping based on retina response. *TVC*, 30(5):507–517, 2014.

[Justin *et al.*, 2016] J. Justin, A. Alexandre, and F. Li. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[Kim *et al.*, 2019] S. Kim, J. Oh, and M. Kiml. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4k UHD HDR applications. In *ICCV*, pages 3116–3125, 2019.

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

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

[Liu *et al.*, 2020] Y. 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 *CVPR*, pages 1651–1660, 2020.

[Santos *et al.*, 2020] M. Santos, T. Ren, and N. Kalantari. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *SIGGRAPH*, 2020.

[Scharr, 2004] H. Scharr. Optimal filters for extended optical flow. In *IWCM*, pages 14–29. Springer, 2004.

[Shi *et al.*, 2016] 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 *CVPR*, pages 1874–1883, 2016.

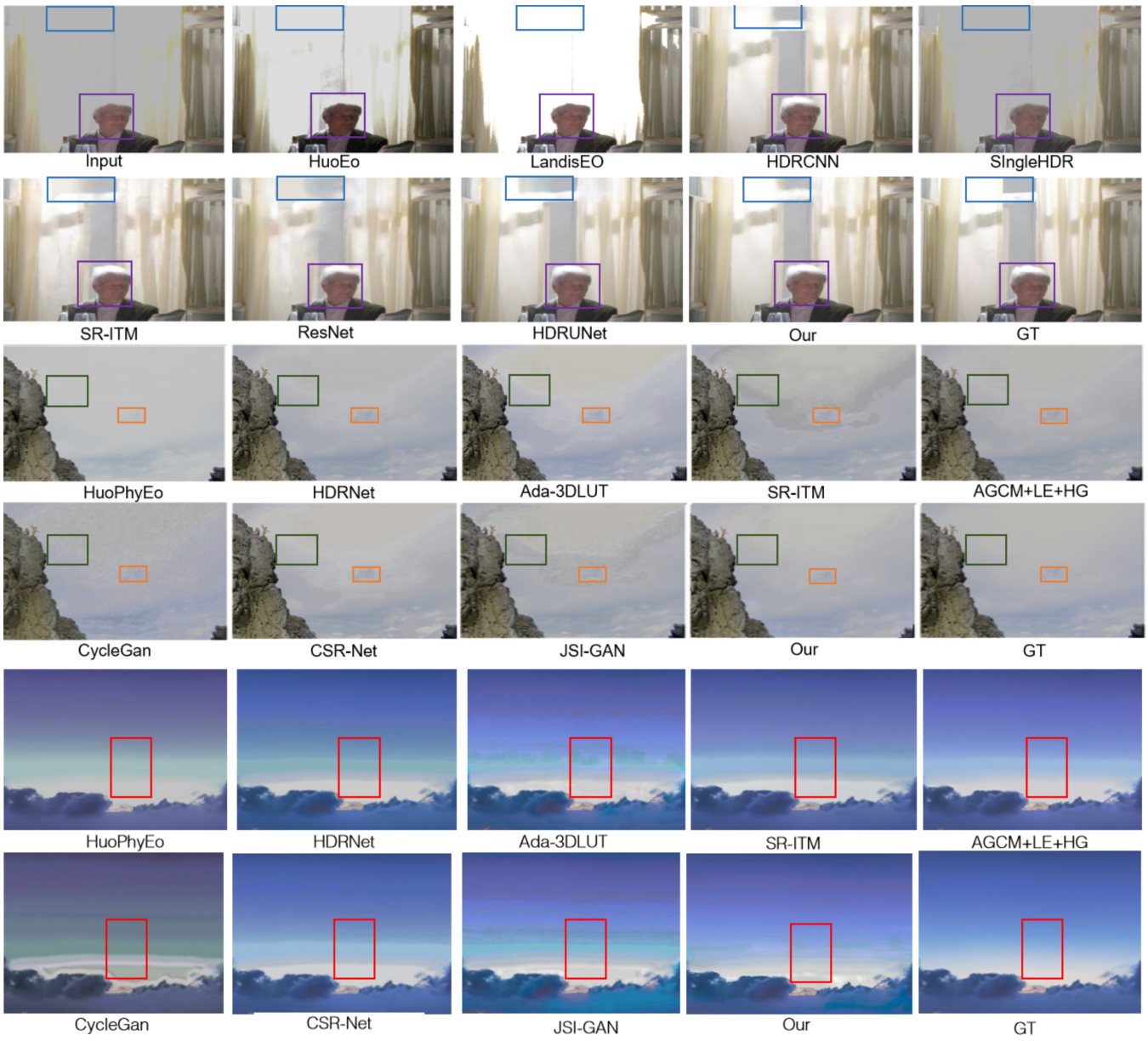

Figure 5: Qualitative comparison. The first two rows are results on an image from NTIRE data set, while the remaining rows are the results on two images from HDRTV data set. *Zoom in for best view.*