FTUnet: Feature Transferred U-Net For Single HDR Image Reconstruction

Shifeng Xie
Xidian University
Xi'an, China
shifeng.xie@telecom-paris.fr

Yi Liu*
Xidian University
Xi'an, China
yi_liu@xidian.edu.cn

Wenjing Shuai
Xidian University
Xi'an, China
wjshuai@xidian.edu.cn

ABSTRACT

The development of the display technology supports the application of High Dynamic Range (HDR) enabling devices. In order to meet the surging demand for the HDR media content, we propose a feature-transferred U-shaped network (FTUnet) to convert existing Standard Dynamic Range (SDR) images into their HDR counterparts. The proposed FTUnet is a feature transformation network that converts the encoded SDR features to the HDR features. This transformation network extracts features rich of spatial information by a self-attention mechanism, in order to improve the reconstruction of the over-exposed regions and avoid unreasonable patches. Besides, we propose an Excitation-Restoration (ER) sub-network to involve the inter-channel attention mechanism. The ER network is used to remove redundant information between channels and reserve the key features. Therefore, the proposed FTUnet can efficiently merge feature channels and contribute to the advantage in color accuracy for the generated HDR images. Experimental results show that our proposed FTUnet achieves state-of-the-art performance in both quantitative comparison and visual quality for the single HDR image reconstruction. The ablation study is also performed to demonstrate the effectiveness of each module of the proposed FTUnet.

CCS CONCEPTS

• **Human-centered computing** → **Visualization theory, concepts and paradigms**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

HDR Image Resconstruction, Feature Transferrred, Unet

ACM Reference Format:
Shifeng Xie, Yi Liu, and Wenjing Shuai. 2023. FTUnet: Feature Transferred U-Net For Single HDR Image Reconstruction. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3595916.3626431

*Corresponding author

MMAsia '23, December 6–8, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12…$15.00
https://doi.org/10.1145/3595916.3626431

1 INTRODUCTION

High Dynamic Range (HDR) imaging is able to capture more visual information than the conventional imaging from the real scene. Meanwhile, the captured HDR content can present a more vivid view than the conventional Standard Dynamic Range (SDR) content with the help of the HDR display devices. A huge demand for HDR content has emerged due to the well established support for HDR displays in consumer electronics. However, the SDR content is not able to be directly presented by the HDR device, due to the different display systems. Therefore, the research on the conversion from SDR to HDR is important to improve the current visual content to meet the emerging needs.

Recently, researchers have adopted convolutional neural networks (CNNs) to realize the conversion to HDR. HDRTVNet[4] was proposed with a multi-stage scheme simulating the imaging process by three networks, a global tone mapping network, a local image enhancement network and an image generation network. In the same year HDRUNet[3] was implemented for single image HDR reconstruction with denoising and dequantization. Later, KUNet[29] was reported as a knowledge-inspired HDR reconstruction.

In this work, we propose a Feature Transferred U-net (FTUnet) for the conversion from SDR to HDR. The proposed FTUnet consists of five parts: a header, an encoder, a feature transformation network, a decoder and a tail. The header and encoder extract features from the SDR images while avoid involving the imaging noise and quantization. The feature transformation network constructs the mapping from the SDR feature space to the HDR feature space. The decoder and the tail network decode the features and reconstruct the HDR images. The proposed FTUnet is an end-to-end network, which does not need to storage intermediate information that required by other cascaded multi-stages networks[22][4][1]. In the experimental section, we demonstrate the effectiveness of our solution by visual comparison and quantitative evaluation measured by two groups of objective quality metrics. First group: PSNR, SSIM, SR-SIM[33], $\triangle_{ITP}$ [15] and HDR-VDP[25]. Second group: PU21-PSNR[2], PU21-VSI[34], PU-FSIM[35], PU-SSIM and PU-MSSIM[31].

Our contribution can be summarized in three parts: firstly, we propose a novel end-to-end neural network that can efficiently convert SDR images to HDR ones. Secondly, in the proposed network, we design a scheme for SDR image encoding and HDR image decoding that emphasize on channel-wise operations. It differs from other approaches mainly extracting information on spatial-wise. Finally, we introduce a feature transformation network to map the SDR image features to HDR ones. This transformation is based on the self-attention mechanism and the intrinsic connection of features between SDR and HDR.

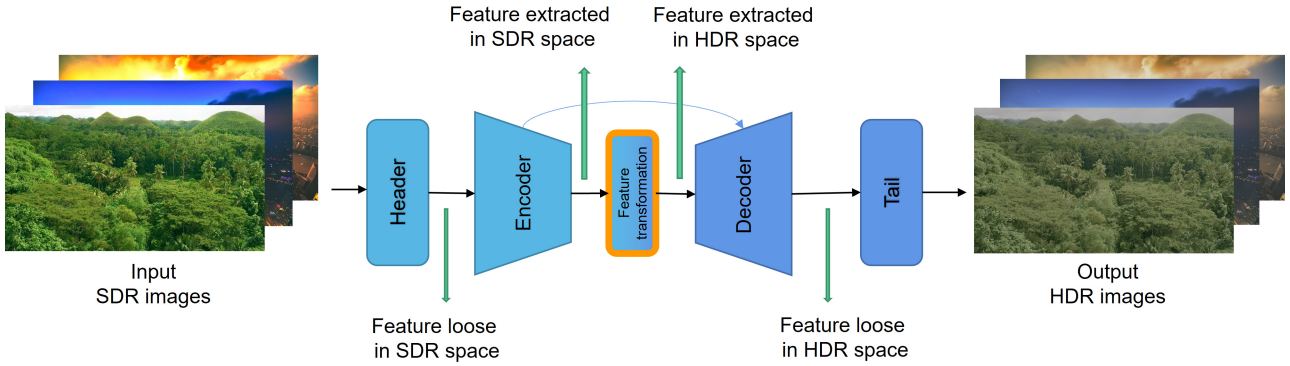

Figure 1: The overall structure of the proposed FTUnet.

2 RELATED WORK

Generating HDR images from the existing SDR ones is an economic and practical means to solve the HDR content shortage problem. There are two common ways to accomplish the generation: one is the HDR image reconstruction from multi-exposure SDR images; the other is the HDR reconstruction directly from a single SDR counterpart.

The HDR image reconstruction from multi-exposure SDR images requires a series of SDR images with different exposures and then fuse them together. *Wang et al.*[30] categorized these methods into five types: optical flow-based image alignment[16], direct feature concatenation[27], correlation-guided feature alignment[24], image translation-based alignment[23], and deep static exposure fusion[17]. Although this technical route is efficient. However, Taking multi-exposure images needs more time to capture and involves the risk of motion blur during the imaging.

Concerning HDR reconstruction from the single SDR image, conventional methods mainly focused on building models to fit the nonlinear HDR imaging process. *Huo et al.* [12] estimated the local adaptation luminosity based on the retinal response. In the same year, *Kovaleski and Oliveira* published their method[20] based on cross-bilateral filtering. Although the conventional methods have contributed efficient results, they are still limited in fitting capabilities. Recently, many SDR-to-HDR methods based on learning networks have emerged. Inspired by bilateral grid processing and local affine color transformation, *Gharbi et al.* proposed HDRNet[7], which is a convolutional neural network to predict the coefficients of a locally-affine model in bilateral space. In recent years, the method HDRTV [4] introduced a three-stage neural network based on the camera imaging process, which achieved good results. However, this method used multiple spliced networks. Each network needs self-training to generate a stage HDR image until the finally refined result. This architecture contains a large amount of trainable parameters thus prolongs the training time. HDRUnet[3] was proposed to introduce the UNet into the field of HDR image reconstruction, achieving decent reconstruction results. Later, KUNet[29] was reported to adopt the feature transformation, It was based on the HDRUnet and established a knowledge-inspired jump connection. However, in terms of measurement of the evaluation metrics, it scored lower than HDCFM[9].

3 METHODOLOGY

For the imaging of the same scene, the HDR images are able to record more visual details than SDR images. In order to explore their differences in both global and local view, we adopt a U-shaped neural network to investigate the features of HDR and SDR in a multi-scale way. The conversion from SDR to HDR is regarded as a feature transfer function completed through an encoding and decoding pipeline. In order to ensure sufficient conversion, unlike KUNet[29], which reconstructs the network in the skip connection, we specifically added a feature transformation network (TransNet) between the encoder and decoder, as highlighted in Figure 1. We know that the core task of HDR image reconstruction is to expand the luminance range. However, most related works have not paid enough attention to channel-wise operations and were based on convolution in the spatial dimension. In order to make use of the connection between channels, we introduced 1×1 convolution layers and channel-wise attention layers into the network. The overall structure of the proposed FTUnet is shown in Figure 2.

3.1 Header and tail

The purpose of the header stage is to map the input SDR image to a rich feature space. The input image is a three-channel matrix that contains densely rich features. The header stage releases these compact features to a relatively loose feature space. Our header network primarily consists of 1×1 convolutions and 3×3 convolutions, with residual links incorporated. This network can be expressed by the following formula:

$$F^S_{\text{loose}} = \sum_{i=0}^{i=n-1} (\text{ReLU} \circ \text{Conv}_{3\times3} \circ \text{ReLU} \circ \text{Conv}_{1\times1})^{n-i} (I_S) \quad (1)$$

Where $()^n$ denotes the cascading of n modules. We recommend $n = 2$. I_S represents the input SDR image, and F^S_{loose} represents the output feature. Similarly, the tail stage is designed to map the generated rich features to the three-channel HDR image with compact information. The tail stage can be represented by the following formula:

$$I_H = \sum_{i=1}^{i=n-1} (ReLU \circ Conv_{1\times1} \circ ReLU \circ Conv_{3\times3})^{n-i} \left(F^H_{loose}\right) \quad (2)$$

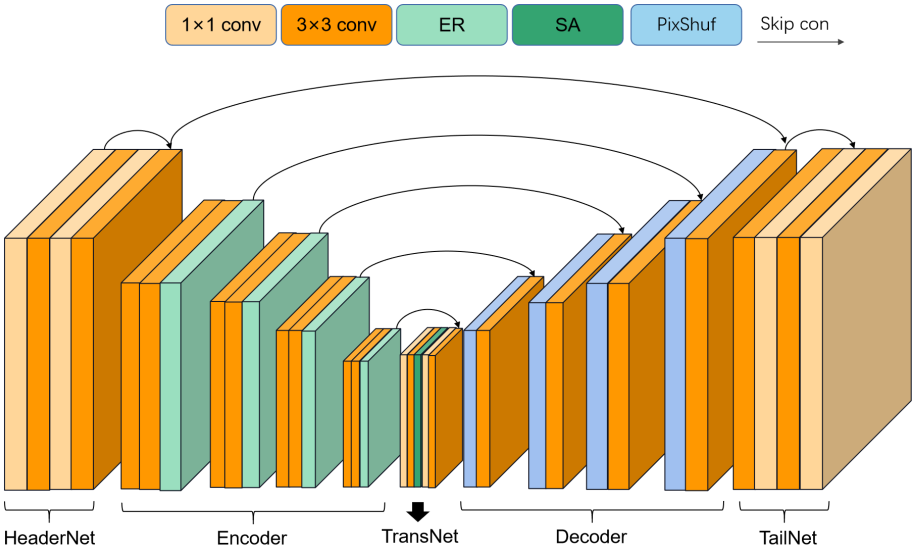

Figure 2: The overall structure of the proposed FTUnet. We introduced two channel-wise operations, 1×1 convolution and Excitation-Restoration network (ER) block. 1×1 convolution can be calculated directly for each channel, and ER module assigns values for each channel according to the connection between channels. We add a feature transformation network(TransNet) between the encoder and the decoder. The TransNet analyzes the connections between the extracted features and enhances the mapping ability. It makes use of the self-attention (SA) mechanism.

Where I_H represents the output HDR reconstructed image and F^H_{loose} represents the decoder output feature. We also recommend $n = 2$ for the tail stage.

3.2 The encoder network

The encoder network is also constituted by cascading modules, mainly using 3×3 convolutions with stride 1 and 2, followed by the Excitation-Restoration (ER block), which is a kind of channel attention mechanism. It can be expressed as follows:

$$F^s_{extracted} = (ER \circ ReLU \circ Conv_{3\times3,s=2} \circ ReLU \circ Conv_{3\times3,s=1})^N \left(F^s_{loose}\right) \tag{3}$$

where $F^S_{extracted}$ and F^S_{loose} respectively denote the output feature and input feature, the superscript s indicates the feature in the SDR space. We suggest setting $N = 4$. The 3×3 convolutional kernel with a stride of 1 can extract features while maintaining the original data size, while the 3×3 convolutional kernel with a stride of 2 can downsample the output features' width and height while extracting features.

We observe that the change in each region from SDR to HDR is inconsistent, that is, the brightness and gamut transformation of each region are nonlinear, so we consider that the channels obtained by each convolution kernel should also be assigned different weight. The convolution operation for analysis here is inspired by the SE[11] Network(Squeeze-and-Excitation Networks), which is a channel attention mechanism realized by squeeze, expand and sigmoid activation operations. In contrast, we propose an Excitation-Restoration (ER) Network to implement the channel attention mechanism in our solution, as shown in Figure 3. The

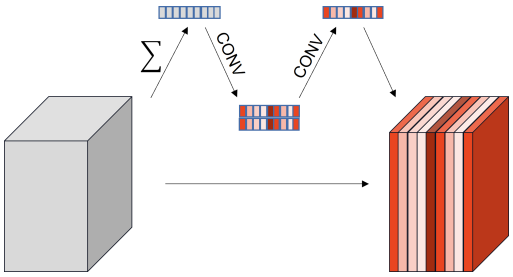

Figure 3: Overview of ER block.

reason is that we adopt the channel attention mechanism in encoder network, which contains the extracted feature data. The channels have contained rich information. If we squeeze in the channel attention, it will result in the information loss and difficulty to the analysis of the connections between the channels. So we propose Excitation-Restoration (ER) network to assign values to each channel. With the ER block, the encoder can combine each channel more reasonably and remove redundant information. This improvement aims to encode features exactly and it is helpful to recover the color information in the reconstructed HDR image. ER block is expressed as:

$$X = \sigma\left(\mathbf{W}_2\delta\left(\mathbf{W}_1\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_c(i,j)\right)\right)U, \tag{4}$$

where input tensor is $U = [u_1, u_2, ..., u_c, ..]$, output tensor is X. Eq. (4) firstly sums each channel, and then applies two convolution operations, W_1 and W_2. ER network excites the input U to a higher dimensional space via W_1 and then restore back to the original

size via W_2. The symbols δ and σ denote the ReLU and sigmoid activation functions, respectively.

3.3 The decoder network

The decoder is composed of a PixelShuffle operation and a cascade of 3×3 convolutions.

$$F^H_{loose} = \left(ReLU \circ Conv_{3\times3,s=1} \circ PixShuf\right)^N \left(F^H_{extracted}\right) \quad (5)$$

The output feature F^H_{loose} and the extracted feature $F^H_{extracted}$ are contained in the HDR feature space. The value of N is the same to that in the encoder network. The PixelShuffle operation reduces the number of channels and increases the output feature's size by merging. While deconvolution can achieve a similar effect, it may cause checkerboard effects, so we choose PixelShuffle. The convolutional network integrates the features, and we introduce the ReLU activation function to add non-linear transformations. Although the model can still perform well without an activation function and converge more quickly, we retain the activation function to increase the model's generalization ability.

3.4 The feature transformation network

The feature transformation network (TransNet) aims to transfers features from the SDR space to the HDR space. This network is composed of 3×3 convolutions, 1×1 convolution, and self-attention mechanism[28]. Although two fully connected layers can achieve similar effects, we chose to use a convolution-based network let the network handle any size of images and to increase the receptive field of the network.

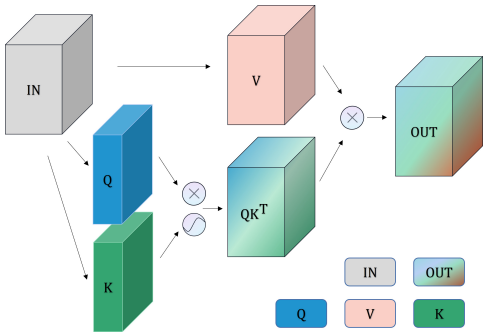

Figure 4: Definition of Self-Attention block

$$F^H_{extracted} = (ReLU \circ Conv_{3\times3} \circ ReLU \circ Conv_{1\times1} \circ SA \circ ReLU \circ Conv_{3\times3} \circ ReLU \circ Conv_{1\times1})^N \left(F^s_{extracted}\right) \quad (6)$$

The output feature $F^H_{extracted}$ is obtained by cascading N layers of 3×3 convolution, ReLU activation, 1×1 convolution, self-attention (SA) mechanism, 3×3 convolution, ReLU activation and 1×1 convolution. As shown in Figure 4. The self-attention mechanism consists of generating query (Q), key (K), value (V) matrices from the input data using 1×1 convolution. Q and K are multiplied, normalized using softmax activation, and then multiplied with V matrix.

$$Attention(Q, K, V) = softmax(QK^T)V \quad (7)$$

After processing by the self-attention mechanism, each value in the tensor has been reassigned. The values here represent features extracted by the encoder, and these features are queried to obtain the internal connection between each feature. We calculate the weight corresponding to each feature through these connections and assign the weight to each feature. That is to say, these extracted features not only contain local information obtained by convolution, but also obtain global information.

4 EXPERIMENT

4.1 Experiment Settings

Dataset. We used the HDRTV dataset, which consists of 1350 pairs of images for the training set and 117 pairs of SDR and HDR images for the test set. The images are in size of 3840×1260 pixels. The reference HDR content has been encoded in HDR10[5] format with BT.2020 color gamut[13].

Training Setup. We use the L1 loss function. The Adam optimizer is used, the initial learning rate is set to 0.0005, and the learning rate is set to $\frac{1}{2}$ of the initial rate every 200000 iterations.

Evaluation Setup. In order to verify the effectiveness of our proposed method in a fair way, we use two group of evaluation metrics. In the first group, we use the same evaluation metrics as HDRTVnet and KUNet, where they are measured by PSNR, SSIM, SR-SIM[33], $\triangle_{ITP}$[15], HDR-VDP[25], as shown in Table 1. Among these metrics, a higher score measured by PSNR, SSIM, SR-SIM or HDR-VDP indicated a better quality, whereas a lower score evaluated by $\triangle_{ITP}$ demotes the better quality in color accuracy. In the second group, following the work of *Hanji et al.* on quality assessment of single image HDR reconstruction methods[8], we select their five recommended full-reference metrics (PU21-PSNR[2], PU21-VSI[34], PU-FSIM[35], PU-SSIM and PU-MSSIM[31]), as shown in Table 2.

4.2 Experiment Results

Quantitative Results. According to the Table 1 data, our proposed FTUnet ranks first in PSNR, SR-SIM, $\triangle_{ITP}$ and HDR-VDP, and ranks second in SSIM. Moreover, the value of $\triangle_{ITP}$ indicator of our method is much better than that of other models, indicating a superiority of the proposed FTUnet in color accuracy. Compared to KUNet and HDCFM introduced in 2022, we have improved in HDR-VDP, indicating that our proposed FTUnet has contributed better visual quality. According to the data in Table 2, we find that our proposed method maintains its advantage after PU encoding.

Visual Comparison. Visual comparisons of the generated images are shown in Figure 5 and Figure 6, where our results are compared with recent HDRTVNet[4] and KUNet[29]. Figure 5 presents reconstructed HDR images directly and makes the comparison in texture. According to Figure 5, the results of HDRTVnet and KUNet show enhanced contours of flames in the second row, light edges in the third row. These contours have been stronger than the original HDR images. Meanwhile, KUNet produces color shift to the water in the first and fourth row. In contrast, our results have not generated the contour and color shift. In order to check the real performance, we present their results on Sony HDR television and take photos shown in Figure 6, where HDRTVnet and KUNet tend to produce color shift and blob-like distortion, respectively. Compared to that, the result images of the proposed FTUnet remain less artifacts and

Table 1: Quantitative comparisons on the HDRTV dataset. Red text indicates the best. Blue text indicates the second.

Method	Venue	PSNR↑	SSIM↑	SR-SIM↑	$\triangle_{ITP}\downarrow$	HDR-VDP↑
HuoPhyEo[12]	TVC14	25.90	0.9296	0.9981	38.06	7.893
KovaleskiEO[21]	SIBGRAPI4	27.89	0.9273	0.9809	28.00	7.431
Pixel2Pixel[14]	CVPR17	25.80	0.8777	0.9871	44.25	7.136
HDRCNN[6]	TOG17	19.33	0.8704	0.7774	75.24	5.807
CyleGAN[36]	ICCV17	21.33	0.8496	0.9595	77.74	6.941
HDRNet[7]	ACMTOG17	35.73	0.9664	0.9957	11.52	8.462
EXPANDNet[26]	CGF18	20.76	0.8580	0.8600	73.79	7.163
CSRNet[10]	ECCV20	35.04	0.9625	0.9955	14.28	8.400
Ada-3DLUT[32]	TPAMI20	36.22	0.9658	0.9967	10.89	8.423
Deep SR-ITM[18]	ICCV19	37.10	0.9686	0.9950	9.24	8.233
JSI-GAN[19]	AAAI20	37.01	0.9694	0.9928	9.36	8.169
HDRTVnet(AGCM+LE)[4]	ICCV21	37.61	0.9726	0.9967	8.89	8.613
HDRTVnet(AGCM+LE+HG)[4]	ICCV21	37.21	0.9699	0.9968	9.11	8.569
KUNet[29]	IJCAI22	37.78	**0.9868**	0.9971	7.80	8.393
HDCFM[9]	ACMMM22	38.42	0.9732	0.9974	7.83	8.571
OURS	OURS	**39.13**	0.9859	**0.9993**	**7.31**	**8.637**

Table 2: Quantitative comparisons with Perceptually Uniform (PU) encoding. Red text indicates the best. Blue text indicates the second.

fileName	PU-PSNR	PU-VSI	PU-FSIM	PU-SSIM	PU-MSSSIM
KovaleskiEO[21]	13.25	0.9832	0.9568	0.7385	0.7319
HDRCNN[6]	11.75	0.9491	0.8498	0.6901	0.6616
EXPANDNET[26]	13.54	0.9704	0.9202	0.7433	0.7362
HDRTVnet[4]	30.40	0.9950	0.9866	0.9433	0.9680
KUNet[29]	30.65	0.9945	0.9832	0.9402	0.9682
OURS	31.89	0.9961	0.9901	0.9429	0.9684

Table 3: Ablation analysis

Method	PSNR↑	SSIM↑	SR-SIM↑	$\triangle_{ITP}\downarrow$	HDR-VDP↑
Net_o.ER	36.91	0.9707	0.9985	9.71	8.39
Net_o.FT	38.07	0.9716	0.9991	8.42	8.51
Net_w.SE	38.84	0.9853	0.9992	7.40	8.57
OURS	**39.13**	**0.9859**	**0.9993**	**7.31**	**8.63**

keep better similarity to the ground truth.

Ablation Study. At the same time, in order to prove the effectiveness of the channel-wise operation and feature transformation network we introduced, we conducted ablation experiments. We removed the channel wise operation, that is, the 1×1 convolution in the network is replaced by 3×3 convolution, and the ER block is removed. We name it Net_o.ER. We took our network and deleted the feature transformation network and named it Net_o.FT. To prove the effectiveness of ER block and compare it with SE block,we replace the ER block with SE block (Net_w.SE). The results are shown in Table 3. With the addition of feature transformation network (Net_o.FT), the objective quality is improved due to the new feature transformation method, which can add globle information to each feature, PSNR, SSIM, SR-SIM and HDR-VDP are improved by 1.06, 0.0143, 0.0002 and 0.12 respectively, and $\triangle_{ITP}$ is reduced by 1.11. With the addition of channel-wise options (Net_o.ER), the objective quality is improved due to the new channel-wise attention mechanism, which removes redundancy and increases the effectiveness of the encoder. PSNR, SSIM, SR-SIM and HDR-VDP are improved by 2.22, 0.0152, 0.0008 and 0.24 respectively, and $\triangle_{ITP}$ is reduced by 2.40. By using ER block instead of SE block[11], the encoder can have more data to analyze the connections between channels. This approach improves the scores of the PSNR, SSIM, SR-SIM and HDR-VDP metrics by 0.29, 0.0006, 0.0001 and 0.06, respectively, and reduces $\triangle_{ITP}$ by 0.09.

5 CONCLUSION

For the task of expanding standard dynamic range (SDR) images to high dynamic range (HDR) ones, previous methods performed global conversion of HDR images without taking into account features transmations and the connections in the channels. In this paper, we provide a novel Feature Transferred U-Net (FTUnet) to perform the single image HDR reconstruction task. The architecture of the proposed network incorporates intra-channel as well as cross-channel attention mechanism to enable the encoder to capture the global information and channel-wise dependencies. Moreover, the feature transformation network, which adds global information to each feature, is designed between the encoder and the decoder to map the SDR features to HDR feature space. Experiments show that our method generates HDR contents with higher color accuracy and less unfavorable artifacts. Overall, our methods outperforms state-of-the-art methods in quantitative and visual comparisons.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No.61901337 and 61906148), the Fundamental Research Funds for the Central Universities (XJSJ23029).

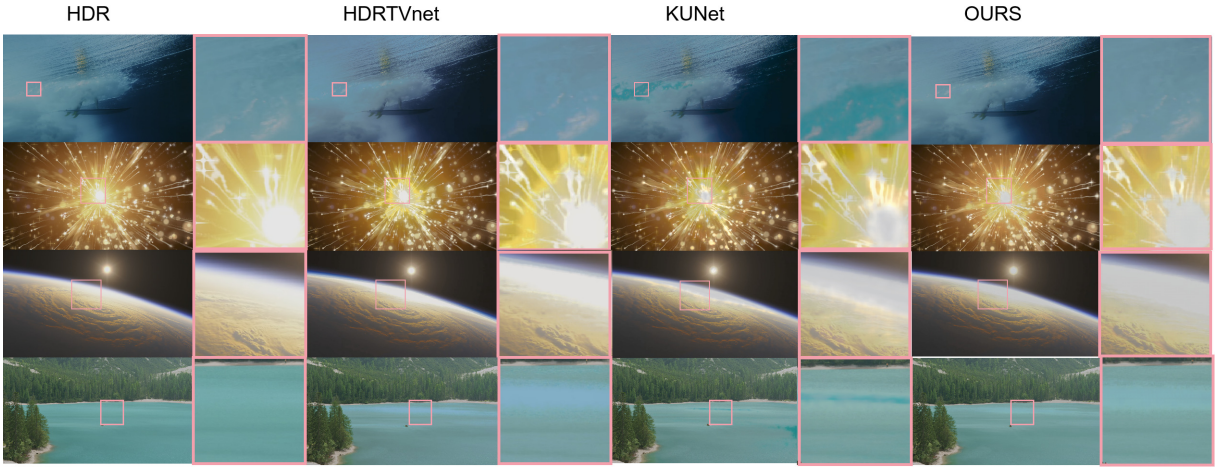

Figure 5: HDR reconstruction of frame results. The proposed FTUnet has avoided much unreal artifacts in the reconstructed HDR images. All images have not been additionally processed to preserve all detail of the HDR images, an HDR display is required to fully display the visual quality of the HDR images, and playback on an SDR display will be dark.

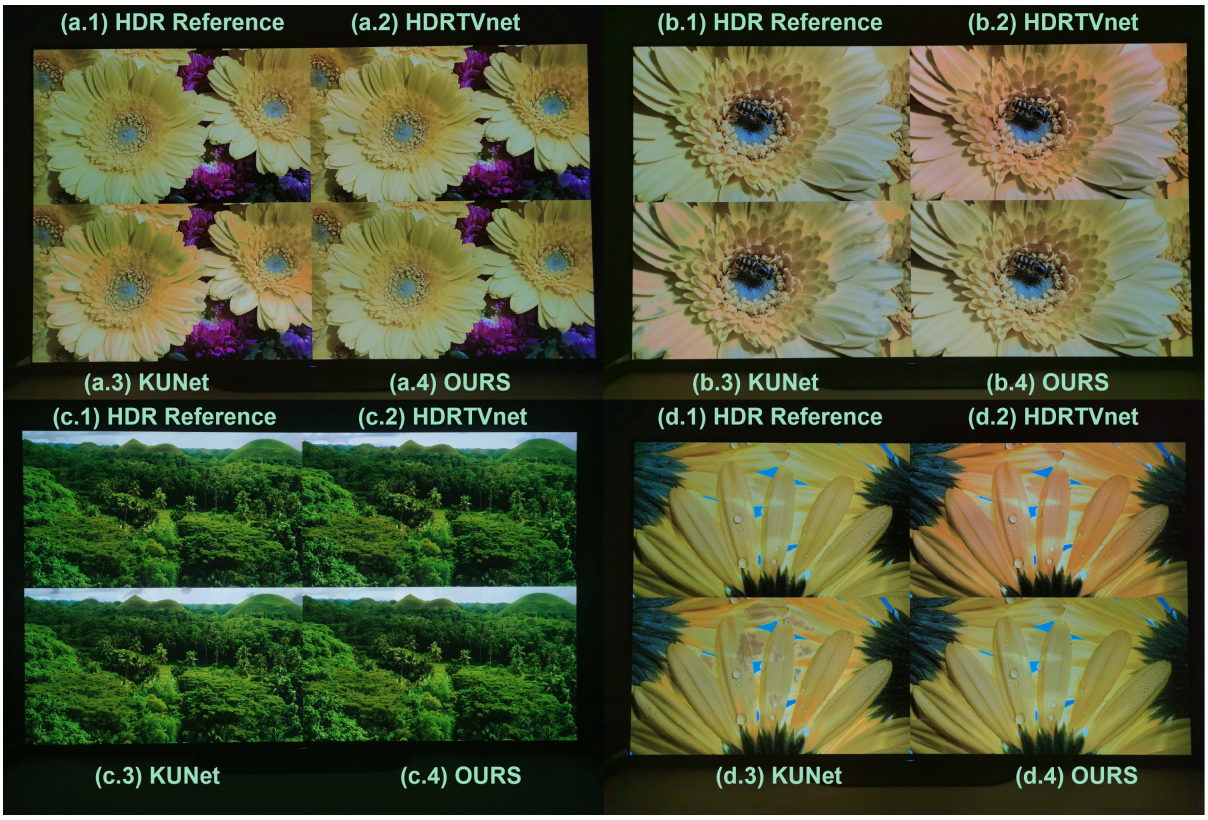

Figure 6: Visual comparison presented by the Sony HDR television. Our advantage is more obvious on HDR display devices. We present our proposed method and some state-of-the-art methods on an HDR display device, and then take photos. In group a, the flowers in a.3 have color distortion. In group b and d, b.2 and d.2 show color distortion, b.3 and d.3 have obvious blob-like distortion. In group c, our method has the best reconstruction effect on cloud.

REFERENCES

[1] S M A Sharif, Rizwan Ali Naqvi, Mithun Biswas, and Sungjun Kim. 2021. A Two-Stage Deep Network for High Dynamic Range Image Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 550–559.

[2] Maryam Azimi et al. 2021. PU21: A novel perceptually uniform encoding for adapting existing quality metrics for HDR. In *2021 Picture Coding Symposium (PCS)*. IEEE, 1–5.

[3] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. 2021. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 354–363.

[4] Xiangyu Chen, Zhengwen Zhang, Jimmy S. Ren, Lynhoo Tian, Yu Qiao, and Chao Dong. 2021. A New Journey From SDRTV to HDRTV. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4500–4509.

[5] Consumer Technology Association. 2015. HDR10: High Dynamic Range Format. Consumer Technology Association Standard. https://www.cta.tech/Standards/Standard-Detail.aspx?Id=5434 CTA-861-G.

[6] Eilertsen Gabriel, Kronander Joel, Denes Gyorgy, Mantiuk Rafał, and Unger Jonas. 2017. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions on Graphics (TOG)* 36, 6, Article 178 (2017).

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

[8] Param Hanji, Rafał K. Mantiuk, Gabriel Eilertsen, Saghi Hajisharif, and Jonas Unger. 2022. Comparison of single image HDR reconstruction methods — the caveats of quality assessment. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings)*. https://doi.org/10.1145/3528233.3530729

[9] Gang He, Kepeng Xu, Li Xu, Chang Wu, Ming Sun, Xing Wen, and Yu-Wing Tai. 2022. SDRTV-to-HDRTV via hierarchical dynamic context feature mapping. In *Proceedings of the 30th ACM International Conference on Multimedia*. 2890–2898.

[10] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. 2020. Conditional sequential modulation for efficient global image retouching. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 679–695.

[11] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

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

[13] International Telecommunication Union. 2020. BT.2020: Parameter values for ultra-high-definition television systems for production and international programme exchange. ITU-R Recommendation. https://www.itu.int/rec/R-REC-BT.2020 Rec. ITU-R BT.2020.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

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

[16] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. 2017. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.* 36, 4 (2017), 144–1.

[17] Harpreet Kaur, Deepika Koundal, and Virender Kadyan. 2021. Image fusion techniques: a survey. *Archives of computational methods in Engineering* 28 (2021), 4425–4447.

[18] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2019. 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*. 3116–3125.

[19] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. 2020. 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. 11287–11295.

[20] Rafael P. Kovaleski and Manuel M. Oliveira. 2014. High-Quality Reverse Tone Mapping for a Wide Range of Exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*. 49–56. https://doi.org/10.1109/SIBGRAPI.2014.29

[21] Rafael P Kovaleski and Manuel M Oliveira. 2014. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 49–56.

[22] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. 2023. Single-Image HDR Reconstruction by Multi-Exposure Generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 4063–4072.

[23] Sang-Hoon Lee, Haesoo Chung, and Nam Ik Cho. 2020. Exposure-structure blending network for high dynamic range imaging of dynamic scenes. *IEEE Access* 8 (2020), 117428–117438.

[24] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. 2021. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 463–470.

[25] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on graphics (TOG)* 30, 4 (2011), 1–14.

[26] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. 2018. ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content. *CoRR* abs/1803.02266 (2018). arXiv:1803.02266 http://arxiv.org/abs/1803.02266

[27] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. 2021. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Transactions on Image Processing* 30 (2021), 3885–3896.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[29] Hu Wang, Mao Ye, Xiatian Zhu, Shuai Li, Ce Zhu, and Xue Li. 2022. KUNet: Imaging Knowledge-Inspired Single HDR Image Reconstruction. In *IJCAI-ECAI 2022*.

[30] Lin Wang and Kuk-Jin Yoon. 2021. Deep learning for hdr imaging: State-of-the-art and future trends. *IEEE transactions on pattern analysis and machine intelligence* 44, 12 (2021), 8874–8895.

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

[32] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. 2020. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2058–2073.

[33] Lin Zhang and Hongyu Li. 2012. SR-SIM: A fast and high performance IQA index based on spectral residual. In *2012 19th IEEE international conference on image processing*. IEEE, 1473–1476.

[34] Lin Zhang, Ying Shen, and Hongyu Li. 2014. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing* 23, 10 (2014), 4270–4281.

[35] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20, 8 (2011), 2378–2386.

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.