This CVPR 2021 workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Single-Image HDR Reconstruction with Task-specific Network based on Channel Adaptive RDN

Guannan Chen Lijie Zhang Mengdi Sun Yan Gao Pablo Navarrete Michelini
YanHong Wu
AIOT CTO, BOE
No.9, Dize Road, BDA, Beijing, CHINA
{chenguannan, zhanglijie, sunmengdi, gaoyan, pnavarre, wuyanhong}@boe.com.cn

Abstract

We describe our solution for the NTIRE–2021 High Dynamic Range Challenge with Single Frame Track where we achieved the $\mathbf{3}^{rd}$ ***place in terms of muPSNR and the*** $\mathbf{1}^{st}$ ***place in terms of PSNR****. Aiming at this challenge we introduce the Task-specific Network based on Channel Adaptive RDN(**TCRDN**) that achieves good performance on the similarity with the ground truth. The network is divided into three subnets: Image Reconstruction(**IR**), Detail Restoration(**DR**) and Local Contrast Enhancement(**LCE**). Each subnet processes its own task, and results are fused to produce the HDR output. We carefully design these subnets so that they are properly trained for their intended purpose: detail restoration in the IR subnet and contrast enhancement in the LCE subnet. The Channel Adaptive RDN is a novel network working as the subnet backbone that combines the classic Residual Dense Network(RDN) and the Gate Channel Transformation layer. The L1 loss is used for training the network and the final model can balance the trade–off between PSNR and muPSNR for high performance in the competition's task.*

1. Introduction

HDR images are capable of capturing rich real-world scene appearances including lighting, contrast and details. However, consumer-grade digital cameras can only capture images within a limited dynamic range due to sensor constraints. The most common approach to generate HDR images is to merge multiple LDR images captured with different exposures [1]. Such a technique performs well on static scenes but often suffers from ghosting artifacts on dynamic scenes or hand-held cameras. Furthermore, capturing multiple images of the same scene may not always be feasible [8].Single-image HDR reconstruction aims to recover an HDR image from a single LDR input. The problem is challenging due to the missing information in under-/over- exposed regions. Recently, several methods [3, 4, 9, 15, 13] have been developed to reconstruct an HDR image from a given LDR input using deep convolution neural networks(CNNs). However, learning a direct LDR-to-HDR mapping is difficult as the variation of HDR pixels is significantly higher than that of LDR pixels.

The challenge on High Dynamic Range proposed within the 2021 CVPR workshop on New Trends in Image Restoration and Enhancement workshop and challenges (NTIRE–2021) was the first in its kind to tackle this problem. That is, the task of recovering an HDR image from one or multiple input Low Dynamic Range (LDR) images that are affected by noise, quantization errors, and might suffer from over- and under-exposed regions due to the sensor limitations. The provided datasets is composed of a varied number of scenes, comprising outdoors and indoors scenes, including both daylight and nightlight scenes, with an emphasis on complex moving lights and a very wide range of brightness levels within the same scene (e.g. high radiance light sources such as lamps combined with very dark shadowed areas). The datasets is a subset of images selected from the HdM HDR datasets (captured with a two Alexa Arri cameras with a mirror rig) where the respective LDR triplets are generated by applying a degradation model (e.g. exposure gain, noise addition and quantization, clipping) to three consecutive frames. Two tracks were proposed in the competition and we joined in the track 1:

- **Track 1: Single Frame HDR**, the aim is to obtain a solution capable to produce the HDR results with the best fidelity to the ground truth by one image input.
- **Track 2: Multiple Frames HDR**, the aim is to obtain a solution capable to produce HDR results with the best fidelity to the ground truth by three different exposure images input.

For the fidelity evaluation, the standard Peak Signal to

Noise Ratio (PSNR) is directly computed in the output images, and the muPSNR is also a kind of PSNR computed in the mu-law tone-mapped images. The PSNR and muPSNR are both expected above the average for the top rank HDR solutions in the competition.

In the Single Frame HDR track, we identified two major challenges: image content enhancement, due to the image denoising and image content similarity; and HDR enhancement, due to the dynamic range expanding and bit depth increasing. Our solution was built upon the Task-specific Network based on Channel Adaptive RDN (TCRDN) network [11], combining both the Task-specific Network and the RDN Network. First, **Task-specific network** is the architecture of JSI-GAN proposed by Kim *et al*, and gets favorable performance on single image HDR enhancement [6]. And second, the **Channel Adaptive RDN**, which combines the classic RDN [16] and the Gate Channel Transformation layer [14], can improve the fidelity on image enhancement and make training stable by hierarchical features learning and channels relationship analyzing.

Despite the good performance of JSI-GAN, it has obvious problems of high complexity and training difficulty. Our main contribution is to redesign an HDR network following the JSI-GAN, and make it easy and stable for training and deploying. We summarize our contributions as follows:

- We propose to combine the RDN and Gate Channel Transformation layer to generate the novel network of Channel Adaptive RDN.
- We propose to simplify all the subnet of JSI-GAN generator with Channel Adaptive RDN and reproduce a new Task-specific network.

2. Network Architecture

The network architecture is basically based on the classical theory of single image HDR enhancement. The theory can be described by the following equations [2]:

$$I_{blur} = BLUR(I_{ori}) \tag{1}$$

$$I_{detail} = I_{ori}/I_{blur} \tag{2}$$

$$I_{HDR} = I_{blur} * coef + I_{detail} \tag{3}$$

For single image HDR enhancement, the blur image I_{blur} is calculated from input image I_{ori} by classic blur filter such as Gaussian Filter or Bilateral Filter, and the detail image I_{detail} can be obtained by division between I_{ori} and I_{blur}. At last, the HDR image output I_{HDR} is calculated by equation (3).

The Task-specific network for single image HDR is designed with the equations above. The diagram of the architecture is shown in Fig 1. Similar with the JSI-GAN, the TCRDN architecture is composed of Image Reconstruction(IR) subnet that reconstructed the coarse HDR image, Detail Restoration(DR) subnet that restored the image details to be added on the coarse HDR image, and Local Contrast Enhancement(LCE) subnet that generated the local contrast mask to boost the contrast in this image. For each subnet, the LDR image is directly input to the IR subnet, the guided filter is used for getting the blur image by filtering the LDR image, and the blur image is input to LCE subnet, the detail image that input to DR subnet is generated by subtraction between LDR image and blur image, that is different with JSI-GAN which gets the detail image by division.

The Channel Adaptive RDN has been deployed for each subnet, that designed by classic RDN [16] network with the Gate Channel Transformation(GCT) layer added to each RDB block. The RDB block shown in Fig 2 is the classic residual dense block architecture and set by 6 dense-layers, channel growth rate of 32, the input and output channels of 64, the convolution filter of 3×3 and stride 1. Based on the classic RDN architecture, the image input is filtered by two 3x3 convolution layers for shallow feature extraction, and output 64 channels. After the three RDBs, the output channels of each RDB are concatenated. Therefore, the 1x1 convolution layer is used to reduce the channels to 64, and the 3x3 convolution layer is used to global feature fusion. At last, the output features by the first convolution and the global features are added to learn the global residual, and the global residual is filtered by the last 3x3 convolution layer to output the HDR result.

The GCT layer is a method that can create competition or cooperation relationship among the channels, and optimize the convolution weight towards more accurate for their task. It includes three parts called Global Context Embedding, Channel Normalization and Gating Adaptation, shown in Fig 3. The Global Context Embedding is used to aggregate global context information in each channel, as the information with a large receptive field is useful to avoid local ambiguities [5, 12] caused by the information with a small receptive field ($e.g.$, convolution layer). Given the embedding weight $\alpha = [\alpha_1, ..., \alpha_C]$, the module is defined as:

$$s_c = \alpha_c \parallel x_c \parallel_2 = \alpha_c \{[\sum_{i=1}^{H}\sum_{j=1}^{W}(x_c^{i,j})^2] + \epsilon\}^{0.5} \tag{4}$$

where x_c is the channel of features, ϵ is a small constant to avoid the problem of derivation at the zero point.

Channel Normalization is the method that can create competition relationship between neurons (or channels) [7] with lightweight computing resource and a stable training

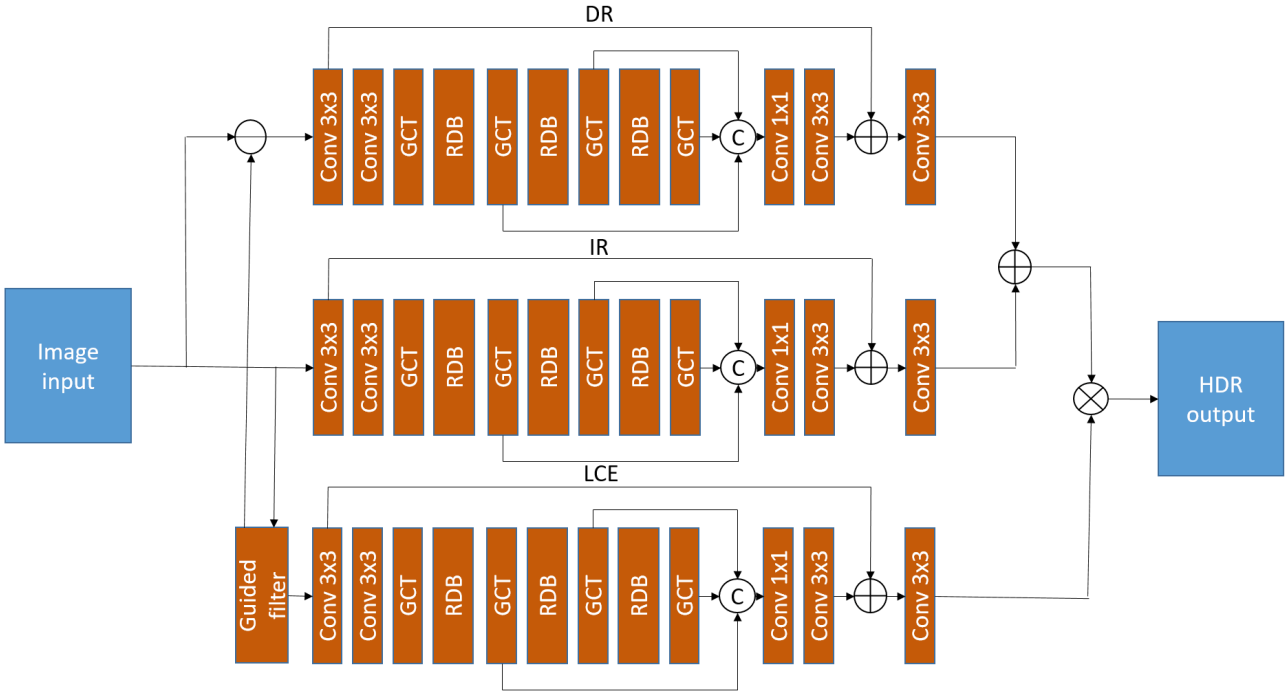

Figure 1. Task-specific Network based on Channel Adaptive RDN (TCRDN)

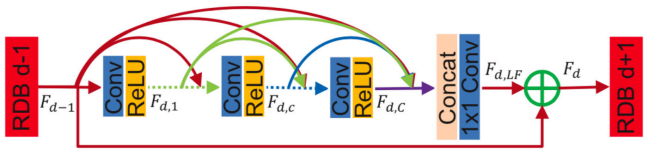

Figure 2. RDB Architecture [16]

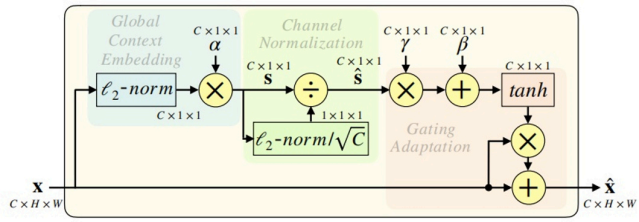

Figure 3. GCT Architecture [14]

performance. The l_2 normalization is created to operate across channels. Let $\boldsymbol{s} = [s_1, ..., s_C]$, the formula of channel normalization is:

$$\hat{s}_c = \frac{\sqrt{C} s_c}{\| \boldsymbol{s} \|_2} = \frac{\sqrt{C} s_c}{[(\sum_{c=1}^{C} s_c^2) + \epsilon]^{0.5}} \tag{5}$$

where ϵ is a small constant. The scalar $\sqrt{C}$ is used to normalize the scale of $\hat{s}$, avoiding a too small scale of $\hat{s}$ when C is large.

The gating mechanism is used to adapt the original feature. By introducing the Gating Adaptation, the GCT layer can facilitate both competition and cooperation during the training process. Let the trainable gating weight $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_C]$ and the trainable gating bias $\boldsymbol{\beta} = [\beta_1, ..., \beta_C]$, the gating function is:

$$\hat{x}_c = x_c[1 + \tanh(\gamma_c \hat{s}_c + \beta_c)] \tag{6}$$

With the combination of Task-specific architecture, RDN and GCT modules, the solution proposed can tackle the single image HDR task effectively and obtain favorable performance in the competition.

3. Training and Inference

For the HDR competition, the PSNR is the key parameter for model evaluation. Therefore, we use L1 loss function and Adam optimizer to train the model. The L1 loss function is defined as follow:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} | I^i_{pred} - I^i_{gt} | \tag{7}$$

For fidelity of the competition, PSNR and muPSNR are available for the model performance evaluation. The PSNR

Ground Truth **Input - medium** **Ours (PSNR/mu-PSNR 42.01/36.72dB)**

Figure 4. Example outputs and performance for image 00647.png in the HDR training set that were used only for validation

Ground Truth **Input - medium** **Ours (PSNR/mu-PSNR 41.02/32.15dB)**

Figure 5. Example outputs and performance for image 00969.png in the HDR training set that were used only for validation

of the normalized images is defined as follow:

$$PSNR = -10\log_{10}[\frac{1}{N}\sum_{i=1}^{N}(I^i_{pred} - I^i_{gt})^2] \quad (8)$$

where the images of I_{pred} and I_{gt} are the prediction HDR image and ground truth HDR image normalized to the max value of the ground-truth HDR image.

The muPSNR is the PSNR of normalized images by mu-law tone mapping, that is defined as follow:

$$MU(I) = \frac{\log(1 + mu * \tanh(I))}{\log(1 + mu)} \quad (9)$$

$$muPSNR = -10\log_{10}[\frac{1}{N}\sum_{i=1}^{N}(MU(I^i_{pred}) - MU(I^i_{gt}))^2] \quad (10)$$

where $MU(I)$ is the tone mapping function of image I, $mu = 5000$ is the parameter controlling the compression performed during tone mapping. The images of I_{pred} and I_{gt} are the prediction HDR image and ground truth HDR image normalized to the 99 percentile of the ground-truth HDR image.

During the training stage, we used the validation images to calculate the PSNR and muPSNR for every epoch, and saved the model with corresponding PSNR and muPSNR. It is convenient to observe the of training convergence and finding the best model for fidelity performance.

For inference stage, we first produce four different images as input, that are original input, flip horizontal input, flip vertical input and 180 degree rotating input. We input the four images to the model, align the output images to the original image, and calculate the average image as the final output.

4. Experiments

4.1. Settings

We run training and testing the task on Tesla V100 GPUs. We randomly chose 60 images in different scenes from the training set for validation during training. We trained using patch size 160×160. We trained our model using Adam optimizer with constant learning rate 10^{-4}. We set the batch size to 1 patch, and training epoch is 300. After

Rank	muPSNR	PSNR
$\mathbf{1}^{st}$	**34.80**	32.87
$\mathbf{2}^{nd}$	34.74	32.29
$\mathbf{3}^{rd}(ours)$	34.41	**33.49**

Table 1. NTIRE 2021 HDR competition result with top 3 on Single Frame track [11].

every epoch we run the validation of our 60 images.

4.2. Performance

We report a running time about 1.0 [s] to process a LDR image (1900×1060) on a Tesla V100 GPU, without using the multi-input inference approach. For our submissions we use the system in the slowest mode, which uses multi-input inference approach. By using this approach it takes $4.7[s]$ to process a 1900×1060 image. The multi-input approach gives a slight increase in PSNR(about 0.15dB). The settings used for the competition are purposely not practical for applications as they focus exclusively on image quality.

4.3. Challenge Results

The Single Frame HDR track of the NTIRE 2021 HDR challenge had 120 participants, with 7 finalists submitting results for the test stage. As shown in Table 1, our results obtained the $\mathbf{3}^{rd}$ **place**, with an average muPSNR of 34.41 dB in the full output images of the test set. This is, 0.39 dB below the top score winner. For PSNR score, we obtained the $\mathbf{1}^{st}$ **place**, with an average PSNR of 33.49 dB in the full output images of the test set, that is 0.62 dB ahead to the $\mathbf{2}^{nd}$ place. Compared with the top 3 scores, the difference between muPSNR and PSNR of our solution is about 1dB, and it is more than 2dB for the other two solutions. Therefore, our solution seems more balance on fidelity performance than other solutions for Single Frame HDR track.

Figures 4 and 5 show examples of our best results for the fidelity of Single Frame HDR track on images used for validation during our training process. These images show the values of PSNR/muPSNR for measuring fidelity, and the significant PSNR proves that the TCRDN has good ability for image de-noising. For the images `00647.png` and `00969.png`, we observe that our results restore more details at over-exposure area and reduce noise obviously. For image `00647.png`, the brightness is very high especially at the area of lamp bulb, but our output shows that the over-exposure area is restored and more details are displayed which similar to the ground truth. For image `00969.png`, the flame area of the input image is saturated and lost the detail information. Therefore, our model can not restored all the details well. The saturation problem can be solved by GAN [6]. According to the fidelity evaluation for competition, we didn't use GAN to generate more details as the GAN may randomly generate details at other area that lead to the PSNR reduction. However, the GAN is still available for practical application of single image HDR.

Variant	muPSNR	PSNR
IR	34.18	35.87
IR+DR	33.72	36.22
IR+LCE	33.97	**36.59**
IR+DR+LCE	**34.83**	36.41

Table 2. Ablation study on the subnets.

Method	muPSNR	PSNR
ExpandNet [10]	30.44	31.49
TCRDN	**34.83**	**36.41**

Table 3. Quantitative Comparison.

We also performed the an ablation study on three subnets in TCRDN, by retraining different variants of the network. Table 2 shows the muPSNR/PSNR performance of four combinations(variants) of the three subnets, where the IR subnet is essential for all cases. The muPSNR/PSNR performance is evaluated by the same validation images of the competition datasets.

As shown in Table 2, employing the DR subnet to the IR subnet brings 0.35 dB gain in PSNR, and additionally using the LCE subnet, further increases the PSNR by 0.19 dB, resulting in a total 0.54 dB gain over only using the IR subnet. It is also noted that employing the LCE subnet to the IR subnet achieve the better PSNR than three subnets combination. That means, the LCE subnet is significantly beneficial with the IR subnet in PSNR. However, the three subnets combination achieves significantly better muPSNR than other variants of subnets. Employing the LCE subnet or the DR subnet to the IR subnet suffers the muPSNR reduction compared to the IR subnet only. Therefore, the architecture of the TCRDN with the three subnets combination is more suitable for muPSNR than other variants of the network, and that's why the TCRDN architecture is decided to deploy in the competition.

The Table 3 shows the quantitative comparison between TCRDN and ExpandNet [10]. The ExpandNet is also a Task-specific architecture for single image HDR enhancement with Local Branch, Dilation Branch and Global Branch. The standard ExpandNet is trained with the HDR competition datasets, and test the model with the best performance epoch as the TCRDN. As shown in Table 3, TCRDN is obviously better than ExpandNet in both PSNR/muPSNR. That means, TCRDN has better performance of fidelity than ExpandNet.

5. Conclusion

In this paper, we proposed a novel network called TCRDN that achieved the good performance of fidelity in NTIRE 2021 HDR competition. The TCRDN is combined the Channel Adaptive RDN to Task-specific architecture designed according to the classic HDR enhancement theory. The result of the competition shows that, for the single image HDR enhancement, TCRDN has its own advantage for fidelity performance and over-exposure area adjustment. However, for saturation area, the TCRDN can't restore the losing details very well. Therefore, we will focus on the GAN training method that can make the model generate more details in saturation area, and optimize the network to make it suitable for practical applications.

References

[1] Paul Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. *ACM SIGGRAPH'97*, 97, 09 1997. 1

[2] Frdo Durand and Julie Dorsey. Fast bilateral filtering for the display of high-dynamic-range images. volume 21, pages 257–266, 07 2002. 2

[3] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafa K. Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *international conference on computer graphics and interactive techniques*, 36(6):178, 2017. 1

[4] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics*, 36(6):177, 2017. 1

[5] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 9401–9411, 2018. 2

[6] Soo 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:11287–11295, 04 2020. 2, 5

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of The ACM*, 60(6):84–90, 2017. 2

[8] Yu-Lun Liu, Wei-Sheng Lai, Yu Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. pages 1648–1657, 06 2020. 1

[9] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet : a deep convolutional neural network for high dynamic range expansion from low dynamic range content. *Computer Graphics Forum*, 37(2):37–49, 2018. 1

[10] Demetris Marnerides, Thomas Bashford-Rogers, Jon Hatchett, and Kurt Debattista. ExpandNet: A Deep Convolutional Neural Network for High Dynamic Range Expansion from Low Dynamic Range Content. *Computer Graphics Forum*, 2018. 5

[11] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, Radu Timofte, et al. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. 2, 5

[12] Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003. 2

[13] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 120–135, 2018. 1

[14] Zongxin Yang, Linchao Zhu, Yu Wu, and Yi Yang. Gated channel transformation for visual recognition. pages 11791–11800, 06 2020. 2, 3

[15] Jinsong Zhang and Jean-Francois Lalonde. Learning high dynamic range from outdoor panoramas. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4529–4538, 2017. 1

[16] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. pages 2472–2481, 06 2018. 2, 3