MIEHDR CNN: Main Image Enhancement based Ghost-Free High Dynamic Range Imaging using Dual-Lens Systems

Xuan Dong[1], **Xiaoyan Hu**[1], **Weixin Li**[2*], **Xiaojie Wang**[1], **Yunhong Wang**[2]
[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876
[2]Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, 100191

Abstract

We study the High Dynamic Range (HDR) imaging problem using two Low Dynamic Range (LDR) images that are shot from dual-lens systems in a single shot time with different exposures. In most of the related HDR imaging methods, the problem is usually solved by Multiple Images Merging, i.e. the final HDR image is fused from pixels of all the input LDR images. However, ghost artifacts can be hardly avoided using this strategy. Instead of directly merging the multiple LDR inputs, we use an indirect way which enhances the main image, i.e. the short exposure image $\mathbf{I^S}$, using the long exposure image $\mathbf{I^L}$ serving as guidance. In detail, we propose a new model, named MIEHDR CNN model, which consists of three subnets, i.e. Soft Warp CNN, 3D Guided Denoising CNN and Fusion CNN. The Soft Warp CNN aligns $\mathbf{I^L}$ to get the aligned result $\mathbf{I^{LA}}$ using the soft exposed result of $\mathbf{I^S}$ as reference. The 3D Guided Denoising CNN denoises the soft exposed result of $\mathbf{I^S}$ using $\mathbf{I^{LA}}$ as guidance, whose result are fed into the Fusion CNN with $\mathbf{I^S}$ to get the HDR result. The MIEHDR CNN model is implemented by MindSpore and experimental results show that we can outperform related methods largely and avoid ghost artifacts.

Introduction

Generating High Dynamic Range (HDR) images by merging multiple Low Dynamic Range (LDR) images that are captured by consumer cameras is a popular approach for shooting high quality (signal-noise-ratio) images in HDR scenes. Currently commonly used systems, e.g. (Kalantari and Ramomoorthi 2017), usually use a single camera with multiple shots to shoot the series of LDR images with different exposure time. During the shooting, moving cameras or objects will lead to relative 2D movement between the corresponding pixels among the multiple LDR inputs. The chief challenge is to deal with the 2D movement so as to avoid ghost artifacts in the merged HDR result.

Nowadays, dual-lens systems are widely deployed in popular smart phones, e.g. iPhone, HuaWei, Sumsung, etc. So, in this paper, as shown in Fig. 1, we study how to generate the HDR image using two LDR images that are shot from dual-lens systems in a single shot time with different exposures. Since the imaging is finished within a single shot, the 2D movement caused by moving cameras and objects can be neglected. And the challenge is transferred to solve the relative 1D movement between the corresponding pixels among the two LDR inputs due to disparities, which is easier and less computationally costly.

(a) Input pair of images $\mathbf{I^S}$ and $\mathbf{I^L}$. (b) Our HDR result $\mathbf{I}^{HDR}$.

Figure 1: The input pair of short exposure image $\mathbf{I^S}$ and long exposure image $\mathbf{I^L}$ are shot by dual-lens respectively in a single shot time. We learn to enhance the main image $\mathbf{I^S}$ using $\mathbf{I^L}$ as guidance to generate the HDR result $\mathbf{I}^{HDR}$.

In the literature, starting from (Devevec and Malik 1997), most of existing methods, including recently proposed CNN based methods like (Kalantari and Ramomoorthi 2017), take the Multiple Images Merging strategy, as shown in Fig. 2. The input multiple LDR images are firstly aligned and the pixels of the final HDR image are fused from pixels of all the aligned LDR images. Under this strategy, the requirement for the accuracy of the alignment is very high. A small number of mis-aligned pixels between the aligned images may lead to obvious ghost artifacts into the merged HDR result.

Our insight is shown in Fig. 2. Instead of directly merging the multiple input images, we use an indirect way which enhances the main image, i.e. the short exposure image $\mathbf{I^S}$, using the long exposure image $\mathbf{I^L}$ serving as guidance. We perform soft exposure for $\mathbf{I^S}$ to simulate the camera exposure to get its long exposure version, which is then denoised using the aligned result of $\mathbf{I^L}$ as guidance. The denoising result and $\mathbf{I^S}$ are fused to obtain the final HDR result $\mathbf{I}^{HDR}$. With the Main Image Enhancement strategy, mis-alignment

*Corresponding Author. E-mail: weixinli@buaa.edu.cn

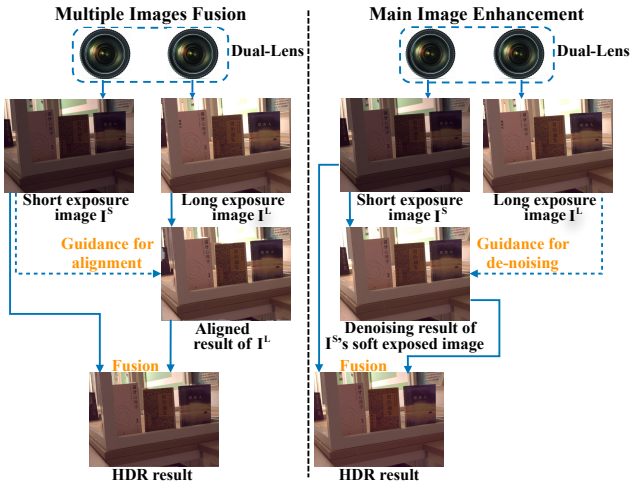


Figure 2: In traditional Multiple Images Fusion strategy, the HDR result is obtained by directly fusing all pixels of the aligned LDR input images. In this strategy, mis-aligned regions will generate ghost artifacts. In our Main Image Enhancement strategy, the main image, i.e. the short exposure image $\mathbf{I^S}$, and the denoising result of its soft exposed image are used for generating the HDR result. The long exposure image $\mathbf{I^L}$ is aligned and just used as the guidance for the denoising. In mis-aligned regions between the aligned images, the noises may not be removed completely but ghost artifacts within the final HDR result can be avoided.

pixels between the aligned images may reduce the denoising quality but the ghost artifacts in $\mathbf{I}^{HDR}$ can be avoided.

Based on our insight, we propose a new model, named MIEHDR CNN model. The overall structure is shown in Fig. 3, which consists of three sub-nets, i.e. Soft Warp CNN, 3D Guided Denoising CNN, and Fusion CNN. The Soft Warp CNN aligns $\mathbf{I^L}$ using the first-time soft exposure result of $\mathbf{I^S}$, i.e. $\mathbf{I^{SE'}}$, as reference to get the aligned result $\mathbf{I^{LA}}$. $\mathbf{I^{LA}}$ can provide better guidance than $\mathbf{I^L}$ for the denoising and is fed into the 3D Guided Denoising CNN with the second-time soft exposure result of $\mathbf{I^S}$, i.e. $\mathbf{I^{SE}}$. Using $\mathbf{I^{LA}}$ as guidance, this CNN uses a 3D U-Net to learn the filtering weights for $\mathbf{I^{SE}}$ with context and generates the spatially consistent de-noising result $\hat{\mathbf{I}}^{\mathbf{SE}}$. $\hat{\mathbf{I}}^{\mathbf{SE}}$ and $\mathbf{I^S}$ are finally fed into the Fusion CNN to generate the HDR result $\mathbf{I}^{HDR}$.

Experimental results show that we can outperform related methods largely for HDR imaging using dual-lens systems.

Contributions: 1) We propose the Main Image Enhancement strategy for HDR imaging. 2) We separate the process into three steps, i.e. alignment, guided de-noising, and fusion, and propose three CNN subnets respectively to finish the corresponding tasks. 3) We propose a 3D Guided Denoising CNN to learn the filtering weights with context for generating spatially consistent de-noising results. 4) We build a Dual-Lens HDR dataset.

Related Works

HDR imaging has been discussed for a long time and is still a hot topic, e.g. (Devevec and Malik 1997) (Kalantari and Ramomoorthi 2017). HDR imaging methods in static scenes, e.g. (Mertens, Kautz, and Reeth 2007) (Ma et al. 2020), perform well. In dynamic scenes and the dual-lens HDR systems, there exist relative movements between corresponding pixels in the multiple LDR images. And the most challenging problem is to avoid ghost artifacts when merging the multiple LDR images.

Recently, CNN based methods, e.g. (Kalantari and Ramomoorthi 2017) (Wu et al. 2018) (Yan et al. 2020), are proposed for solving this problem. To overcome the problem of relative movements between pixels, dense correspondence searching and alignment subnets are integrated into the CNN models, e.g. (Kalantari and Ramomoorthi 2017) (Chen et al. 2020) (Trinidad et al. 2019), to improve the reconstruction quality. Since the alignment results may not be perfect all the time, some recent works, e.g. (Yan et al. 2019) (Prabhakar et al. 2019) (Li et al. 2020), also propose to detect and/or correct the mis-alignment regions. But the detection/correction methods may still fail to repair all mis-alignment regions and thus cannot avoid ghost artifacts completely.

In short, most of the existing methods use the Multiple Images Merging strategy as discussed above and shown in Fig.2. Since human users are sensitive to ghost artifacts, this strategy has high demand for the alignment accuracy. Unfortunately, as mentioned in (Cogalan and Akyuz 2020), a robust and reliable ghost-free solution has been shown to be difficult.

The burst photography work in (Hasinoff et al. 2016) denoises the main frame using the other frames serving as reference and can avoid ghost artifacts. However, they assume that all input frames in the burst are under the same exposure time. The differences of the assumptions for the exposure time of input images make the method not proper for generating high quality HDR images in our problem.

Some recent marvelous works develope new hardware devices and propose corresponding software algorithms for HDR imaging, e.g. (Cogalan and Akyuz 2020) (Wang et al. 2019a). The algorithms are specifically designed for the new hardware devices while in our problem, the device is the widely deployed consumer dual-lens within smart phones. Due to different hardware devices, their algorithms are not proper for solving our problem.

Single image HDR imaging is also a hot problem and attracts many discussions recently, e.g. (Liu et al. 2020) (Santos, Ren, and Kalantari 2020). The main difficulty is to hallucinate the missing textures in under-/over-exposed regions. Instead of solving such a challenging problem, we choose the easier way to make full use of the two LDR input images from dual-lens systems for HDR imaging.

Besides HDR, there exist some other enhancement problems using multiple-camera systems, like video retargeting (Li et al. 2018), super resolution (Jeon et al. 2018; Wang et al. 2019b), deblur (Zhou et al. 2019), style transfer (Chen et al. 2018), colorization (Dong et al. 2019, 2020) and flow estimation (Pan et al. 2017). But, these methods cannot be directly used for our problem.

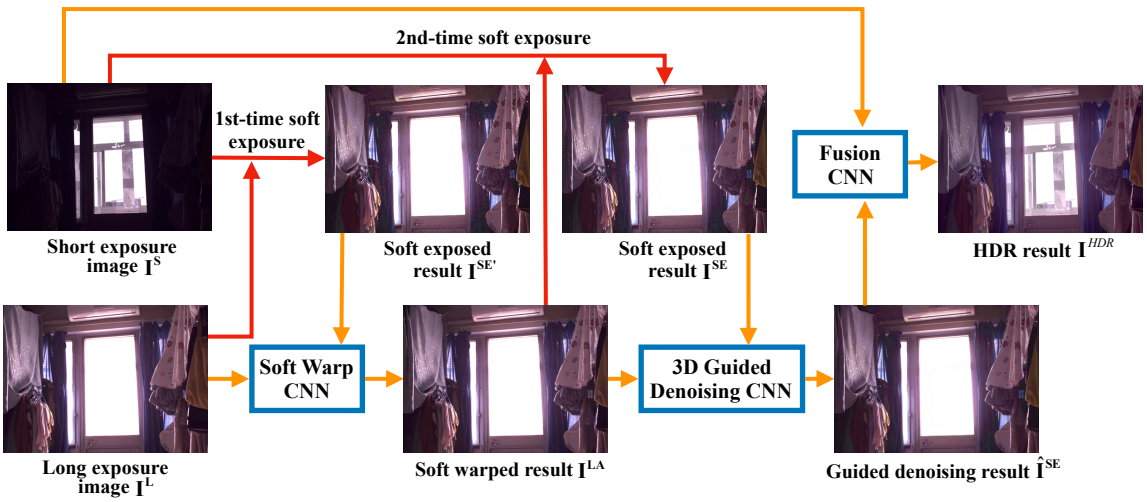


Figure 3: The overall structure of our MIEHDR CNN.

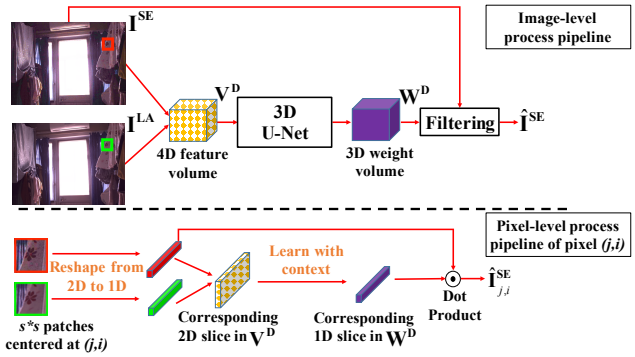


Figure 4: The structure of our 3D Guided Denoising CNN. Besides showing the image level process pipeline, we also show the detailed process pipeline for a single pixel (j, i) at the pixel level to better explain the process.

Method

The overall structure of our MIEHDR CNN model is shown in Fig. 3, which consists of three subnets, i.e. Soft Warp CNN , 3D Guided Denoising CNN, and Fusion CNN. The detailed layer information is shown in Table 1.

The goal of the Soft Warp CNN is to align the long exposure image $\mathbf{I}^{\mathbf{L}}$ so that the aligned result $\mathbf{I}^{\mathbf{LA}}$ can provide better guidance in the 3D guided denoising CNN. We use the soft exposed image $\mathbf{I}^{\mathbf{SE}'}$ of $\mathbf{I}^{\mathbf{S}}$ instead of $\mathbf{I}^{\mathbf{S}}$ itself as the reference for the alignment because $\mathbf{I}^{\mathbf{SE}'}$ has much more similar exposures with $\mathbf{I}^{\mathbf{L}}$, which makes the alignment easier. After getting $\mathbf{I}^{\mathbf{LA}}$, we use it as reference to perform the soft exposure in the second time to obtain $\mathbf{I}^{\mathbf{SE}}$, because the aligned image $\mathbf{I}^{\mathbf{LA}}$ can help generate more accurate soft exposure result. Within the Soft Warp CNN, due to the lack of ground-truth disparities for training for disparity estimation, we use the soft warp operation which utilizes multiple pixels in $\mathbf{I}^{\mathbf{L}}$ and perform weighted average of them for generating the aligned result of each pixel in $\mathbf{I}^{\mathbf{LA}}$.

The goal of the 3D guided denoising CNN is to use $\mathbf{I}^{\mathbf{LA}}$, which has low noises, as guidance for denoising $\mathbf{I}^{\mathbf{SE}}$, which has high noises. The guidance image $\mathbf{I}^{\mathbf{LA}}$ is used for helping estimate the filtering weight, and the denoising result $\hat{\mathbf{I}}^{\mathbf{SE}}$ is obtained by filtering pixels from $\mathbf{I}^{\mathbf{SE}}$ itself. Because the pixel values that contribute to the denoising results $\hat{\mathbf{I}}^{\mathbf{SE}}$ are from $\mathbf{I}^{\mathbf{SE}}$ itself, even if there is mis-alignment between $\mathbf{I}^{\mathbf{SE}}$ and $\mathbf{I}^{\mathbf{LA}}$, the pixel values of $\mathbf{I}^{\mathbf{LA}}$ will not pollute the result $\hat{\mathbf{I}}^{\mathbf{SE}}$. This helps our method avoid ghost artifacts. With the help of the 3D U-Net in this CNN, neighboring pixels will affect each other during learning the filtering weight so that we can obtain spatially consistent results.

Finally, we use $\mathbf{I}^{\mathbf{S}}$ and $\hat{\mathbf{I}}^{\mathbf{SE}}$ as the inputs of the Fusion CNN to obtain the HDR result $\mathbf{I}^{\mathbf{HDR}}$, because there is no mis-alignment between the inputs, this step is like exposure fusion in static scenes.

3D Guided Denoising CNN

The denoising result $\hat{\mathbf{I}}^{\mathbf{SE}}_{j,i}$ of each pixel (j, i) is obtained by weighted average of its neighboring pixels from $\mathbf{I}^{\mathbf{SE}}$ itself, i.e.

$$\hat{\mathbf{I}}^{\mathbf{SE}}_{j,i} = \sum_{(j',i')\in\Omega(j,i)} \mathbf{W}^{\mathbf{D}}_{j,i,(j'-j+r)\cdot s+(i'-i+r)} \cdot \mathbf{I}^{\mathbf{SE}}_{j',i'}, \quad (1)$$

where $\Omega(j, i)$ is the $s \cdot s$ (set as $5 \cdot 5$ in this paper) neighboring pixels centered at pixel (j, i), and r (set as 2 in this paper) is the radius of the neighboring window. The filtering weight $\mathbf{W}^{\mathbf{D}} \in \mathbb{R}^{h\times w\times s^2}$ is estimated by the 3D Guided Denoising CNN, as shown in Fig. 4. First, we use the input images $\mathbf{I}^{\mathbf{SE}}$ and $\mathbf{I}^{\mathbf{LA}}$ to build the 4D feature volume $\mathbf{V}^{\mathbf{D}} \in \mathbb{R}^{h\times w\times s^2\times m}$. For each pixel (j, i), its 2D $s \cdot s$ neighboring pixels $\Omega(j, i)$ are reshaped to 1D slice. So, for each pixel $(j', i') \in \Omega(j, i)$, its feature and learned weight values in $\mathbf{V}^{\mathbf{D}}$ and $\mathbf{W}^{\mathbf{D}}$ are denoted as $\mathbf{V}^{\mathbf{D}}_{j,i,(j'-j+r)\cdot s+(i'-i+r)}$ and $\mathbf{W}^{\mathbf{D}}_{j,i,(j'-j+r)\cdot s+(i'-i+r)}$ respectively. Between each pixel

$(j', i') \in \Omega(j, i)$ and its centered pixel (j, i), the feature is concatenated by the pixel values of (j', i') of $\mathbf{I^{SE}}$ and $\mathbf{I^{LA}}$, i.e. $\mathbf{I^{SE}_{j',i'}}$ and $\mathbf{I^{LA}_{j',i'}}$, the pixel value of the centered pixel (j, i) of $\mathbf{I^{SE}}$, i.e. $\mathbf{I^{SE}_{j,i}}$, and the geometry distance between (j', i') and (j, i) which is defined as $D_{gm}((j', i'), (j, i)) = (j - j')^2 + (i - i')^2$, i.e. $\mathbf{V^D_{j,i,(j'-j+r)\cdot s+(i'-i+r)}} = Concat(\mathbf{I^{SE}_{j',i'}}, \mathbf{I^{LA}_{j',i'}}, \mathbf{I^{SE}_{j,i}}, D_{gm}((j', i'), (j, i)))$.

Then, we use 3D U-Net network to learn the weight volume $\mathbf{W^D}$ from $\mathbf{V^D}$. The 3D U-Net is like the traditional U-Net style network (Ronneberger, Fischer, and Brox 2015) and the differences are that we use 3D convolution instead of 2D convolution. After getting the weight volume, we use Eq. 1 to get the denoising result.

In the training, we use SSIM (Wang et al. 2004) as the error metric and the loss is defined as

$$L_1 = 1 - SSIM(\alpha(\hat{\mathbf{I}}^{\mathbf{SE}}), \mathbf{G^{SE}}), \tag{2}$$

where $\mathbf{G^{SE}}$ is the ground-truth long exposure image of the main image $\mathbf{I^S}$, α is a global adjustment curve. We perform the global adjustment to minimize the exposure differences between $\hat{\mathbf{I}}^{\mathbf{SE}}$ and $\mathbf{G^{SE}}$ so as to avoid the difference affecting the evaluation of the denoising quality. α is estimated using $\hat{\mathbf{I}}^{\mathbf{SE}}$ and $\mathbf{G^{SE}}$. It contains 256 nodes in the dynamic range of $[0, 255]$. And the value at each node x is computed by $\alpha(x) = \frac{\sum_{j,i} D(\hat{\mathbf{I}}^{\mathbf{SE}}_{j,i}, x)\mathbf{G^{SE}_{j,i}}}{\sum_{j,i} D(\hat{\mathbf{I}}^{\mathbf{SE}}_{j,i}, x)}$, where $D(x', x) = e^{-\frac{(x'-x)^2}{2\sigma^2}}$, and σ is set to 5 in this paper.

Soft Warp CNN

The CNN structure is shown in Fig. 5. For the input images $\mathbf{I^L}$ and $\mathbf{I^{SE'}}$, first, we extract their deep features $\mathbf{F^L}$ and $\mathbf{F^{SE'}}$ respectively by a ResNet, named ResNet1. Then, between each pixel (j, i) in $\mathbf{I^{SE'}}$ and all of its candidate pixel $(j, i + k)$ in $\mathbf{I^L}$, we build the 4-D feature volume $\mathbf{V^A}$ by

$$\mathbf{V^A_{j,i,k}} = Concat(\mathbf{F^{SE'}_{j,i}}, \mathbf{F^L_{j,i+k}}). \tag{3}$$

Due to the 1D relative movement of corresponding pixels between the input pair of images, the range of candidate pixels for each pixel (j, i) is defined from (j, i) to $(j, i + d - 1)$, where the hyper-parameter d is the maximum disparity (set as 20% of the image width).

Next, the 3D regulation, which is proposed by (Alex et al. 2017), is performed to estimate the weight volume $\mathbf{W^A} \in \mathbb{R}^{h\times w\times d}$ from the feature volume $\mathbf{V^A}$. Once $\mathbf{W^A}$ is obtained, for each pixel (j, i), the aligned result $\mathbf{I^{LA}_{j,i}}$ is computed by the weighted average of its candidate pixels in image $\mathbf{I^L}$, i.e.

$$\mathbf{I^{LA}_{j,i}} = \sum_{k=0}^{d-1} \mathbf{W^A_{j,i,k}} \mathbf{I^L_{j,i+k}}. \tag{4}$$

The training loss is defined as

$$L_2 = 1 - SSIM(\mathbf{I^{LA}}, \mathbf{G^{SE}}), \tag{5}$$

where $\mathbf{G^{SE}}$ is the ground-truth long exposure image of the main image $\mathbf{I^S}$.

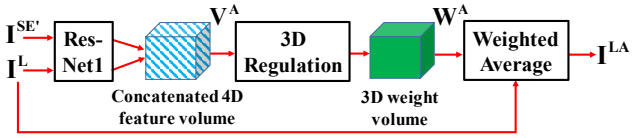


Figure 5: The structure of our Soft Warp CNN.

Fusion CNN

As shown in Fig. 3, the inputs of our Fusion CNN include $\mathbf{I^S}$ and $\hat{\mathbf{I}}^{\mathbf{SE}}$. And we build a straight ResNet, named ResNet2, to learn the blending weights $\mathbf{W^B}$ between $\mathbf{I^S}$ and $\hat{\mathbf{I}}^{\mathbf{SE}}$ so as to get the HDR result $\mathbf{I}^{HDR}$ by

$$\mathbf{I}^{HDR} = \mathbf{W^B} \cdot \mathbf{I^S} + (1 - \mathbf{W^B}) \cdot \hat{\mathbf{I}}^{\mathbf{SE}}. \tag{6}$$

The training loss is defined as

$$L_3 = 1 - SSIM(\mathbf{I}^{HDR}, \mathbf{G}^{HDR}), \tag{7}$$

where $\mathbf{G}^{HDR}$ is the ground-truth HDR image.

Soft Exposure

The soft exposure operation estimates a global adjustment curve to adjust the input short exposure image $\mathbf{I^S}$ so as to simulate the exposure process within the camera. We use histogram equalization (Bradski and Kaehler 2008) to perform the soft exposure because it does not need the input images to be well-aligned. As shown in Fig. 3, we perform the soft exposure two times. In the first time, we use the histogram of the input long exposure image $\mathbf{I^L}$ as the target histogram and adjust $\mathbf{I^S}$ to $\mathbf{I^{SE'}}$ so that the histogram of $\mathbf{I^{SE'}}$ approximately matches the target histogram of $\mathbf{I^L}$. After getting the aligned result $\mathbf{I^{LA}}$, because it can provide more accurate target histogram, we use the histogram of $\mathbf{I^{LA}}$ as the target histogram to soft expose $\mathbf{I^S}$ in the second-time to obtain more accurate soft exposure result $\mathbf{I^{SE}}$.

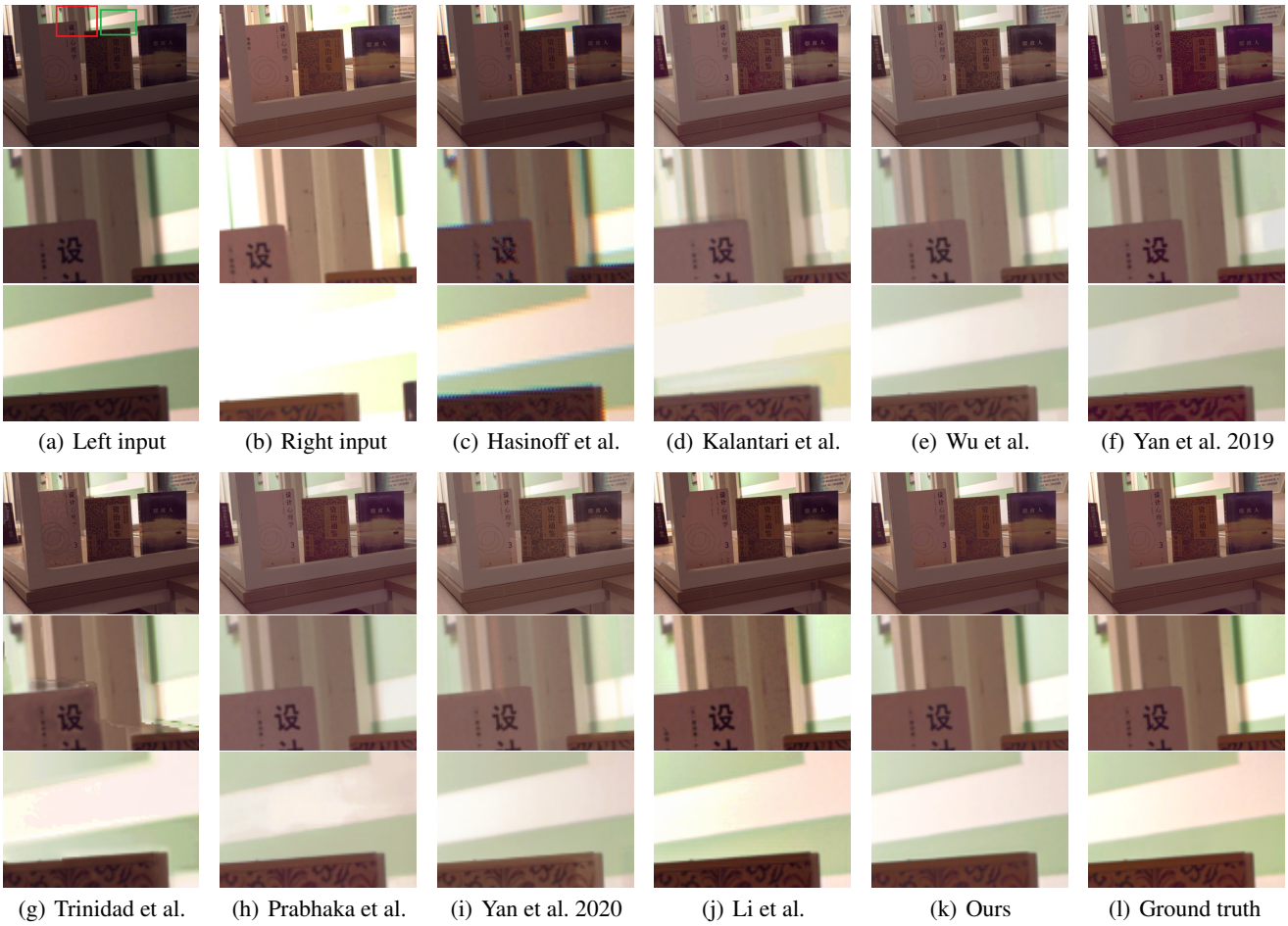
Experimental Results

Dataset

We use two the same color cameras, i.e. the JHSM500f, put them side by side on a tripod, rectify them like traditional stereo systems (Scharstein and Pal 2007), and shoot 1000 pairs of images in various scenes to build our dataset. The left and right cameras are programmed to shoot images in the same shot-time. We set short and long exposure time at each scene, and let the pair of cameras shoot images with the short and long exposure in two shots. Thus, we get 4 images in total at each scene. We follow the method in (Kalantari and Ramomoorthi 2017) to fuse the short and long exposure time images from the left camera to generate ground-truth HDR images. The fusion weight is a simple triangle weight. In addition, we use the short exposure image taken by the left camera and the long exposure image taken by the right camera as the input pair of images at this scene.

	Layer Description	Output Tensor Dim.
	Input and output images	$h \times w \times 3$
3D U-Net in 3D Guided Denoising CNN (Fig. 4)		
1	3D conv, $3 \times 3 \times 3$, n feat.	$h \times w \times s^2 \times n$
2	3D conv, $3 \times 3 \times 3$, n feat.	$h \times w \times s^2 \times n$
3	3D conv, $3 \times 3 \times 3$, n feat.	$h \times w \times s^2 \times n$
4	3D conv, $3 \times 3 \times 3$, n feat.	$h \times w \times s^2 \times n$
5	3D conv, $3 \times 3 \times 3$, n feat.	$h \times w \times s^2 \times n$
6-14	(repeat layer 3, 4, 5)$\times$3	$h \times w \times s^2 \times n$
15	$3 \times 3 \times 3$, 3D trans conv, n feat.	$h \times w \times s^2 \times n$
	add layer 15 and 11 (residual connection)	$h \times w \times s^2 \times n$
16	$3 \times 3 \times 3$, 3D trans conv, n feat.	$h \times w \times s^2 \times n$
	add layer 16 and 8 (residual connection)	$h \times w \times s^2 \times n$
17	$3 \times 3 \times 3$, 3D trans conv, n feat.	$h \times w \times s^2 \times n$
	add layer 17 and 5 (residual connection)	$h \times w \times s^2 \times n$
18	$3 \times 3 \times 3$, 3D trans conv, n feat.	$h \times w \times s^2 \times n$
	add layer 18 and 2 (residual connection)	$h \times w \times s^2 \times n$
19	$3 \times 3 \times 3$, 3D trans conv, 1 feat.	$h \times w \times s^2$
20	softmax	$h \times w \times s^2$
ResNet1 in Soft Warp CNN (Fig. 5)		
1	5×5 conv, n feat., stride 2	$\frac{h}{2} \times \frac{w}{2} \times n$
2	3×3 conv, n feat.	$\frac{h}{2} \times \frac{w}{2} \times n$
3	3×3 conv, n feat.	$\frac{h}{2} \times \frac{w}{2} \times n$
	add layer 1 and 3 feat. (residue connection)	$\frac{h}{2} \times \frac{w}{2} \times n$
4-17	(repeat layers 2,3 and residual connection)$\times$7	$\frac{h}{2} \times \frac{w}{2} \times n$
18	3×3 conv, n feat.	$\frac{h}{2} \times \frac{w}{2} \times n$
3D regulation in Soft Warp CNN (Fig. 5)		
21	3D conv, $3 \times 3 \times 3$, n feat.	$\frac{h}{2} \times \frac{w}{2} \times \frac{d}{2} \times n$
22	3D conv, $3 \times 3 \times 3$, n feat.	$\frac{h}{2} \times \frac{w}{2} \times \frac{d}{2} \times n$
23	3D conv, $3 \times 3 \times 3$, $2n$ feat., stride 2	$\frac{h}{4} \times \frac{w}{4} \times \frac{d}{4} \times 2n$
24	3D conv, $3 \times 3 \times 3$, $2n$ feat.	$\frac{h}{4} \times \frac{w}{4} \times \frac{d}{4} \times 2n$
25	3D conv, $3 \times 3 \times 3$, $2n$ feat.	$\frac{h}{4} \times \frac{w}{4} \times \frac{d}{4} \times 2n$
26-34	(repeat layer 23, 24, 25)$\times$3	$\frac{h}{32} \times \frac{w}{32} \times \frac{d}{32} \times 2n$
35	$3 \times 3 \times 3$, 3D trans conv, $2n$ feat., stride 2	$\frac{h}{16} \times \frac{w}{16} \times \frac{d}{16} \times 2n$
	add layer 35 and 31 (residual connection)	$\frac{h}{16} \times \frac{w}{16} \times \frac{d}{16} \times 2n$
36	$3 \times 3 \times 3$, 3D trans conv, $2n$ feat., stride 2	$\frac{h}{8} \times \frac{w}{8} \times \frac{d}{8} \times 2n$
	add layer 36 and 28 (residual connection)	$\frac{h}{8} \times \frac{w}{8} \times \frac{d}{8} \times 2n$
37	$3 \times 3 \times 3$, 3D trans conv, $2n$ feat., stride 2	$\frac{h}{4} \times \frac{w}{4} \times \frac{d}{4} \times 2n$
	add layer 37 and 25 (residual connection)	$\frac{h}{4} \times \frac{w}{4} \times \frac{d}{4} \times 2n$
38	$3 \times 3 \times 3$, 3D trans conv, n feat., stride 2	$\frac{h}{2} \times \frac{w}{2} \times \frac{d}{2} \times n$
	add layer 38 and 22 (residual connection)	$\frac{h}{2} \times \frac{w}{2} \times \frac{d}{2} \times n$
39	$3 \times 3 \times 3$, 3D trans conv, 1 feat.	$h \times w \times d$
40	softmax	$h \times w \times d$
ResNet2 in Fusion CNN		
1	5×5 conv, n feat.	$h \times w \times n$
2-17	repeat layers 2-17 in **ResNet1**	$h \times w \times n$
18	3×3 conv, 1 feat., Sigmoid	$h \times w$

Table 1: Summary of the layer information of MIEHDR CNN. Each 2D or 3D convolution layer represents a block of convolution, batch normalization and ReLu (unless otherwise specified).



(a) Left input (b) Right input (c) Hasinoff et al. (d) Kalantari et al. (e) Wu et al. (f) Yan et al. 2019

(g) Trinidad et al. (h) Prabhaka et al. (i) Yan et al. 2020 (j) Li et al. (k) Ours (l) Ground truth

Figure 6: Examples to compare the state-of-the-art HDR imaging methods with ours. The regions marked with the red and green boxes are shown in the second and third rows respectively.

Implementation Details

Our network is implemented by MindSpore (MindSpore 2020) with a constant learning rate of 0.001. The images of the dataset are randomly divided into the training set with 700 pairs of images and the testing set with 300 pairs of images. All the models are run on a server with an Intel I7 CPU and 4 NVIDIA 1080Ti GPUs. In the training step, we use the images with the resolution level of 416x576 from the dataset. In the testing step, we test three resolution levels, i.e. level1 (832x1184), level2 (416x576) and level3 (192x288).

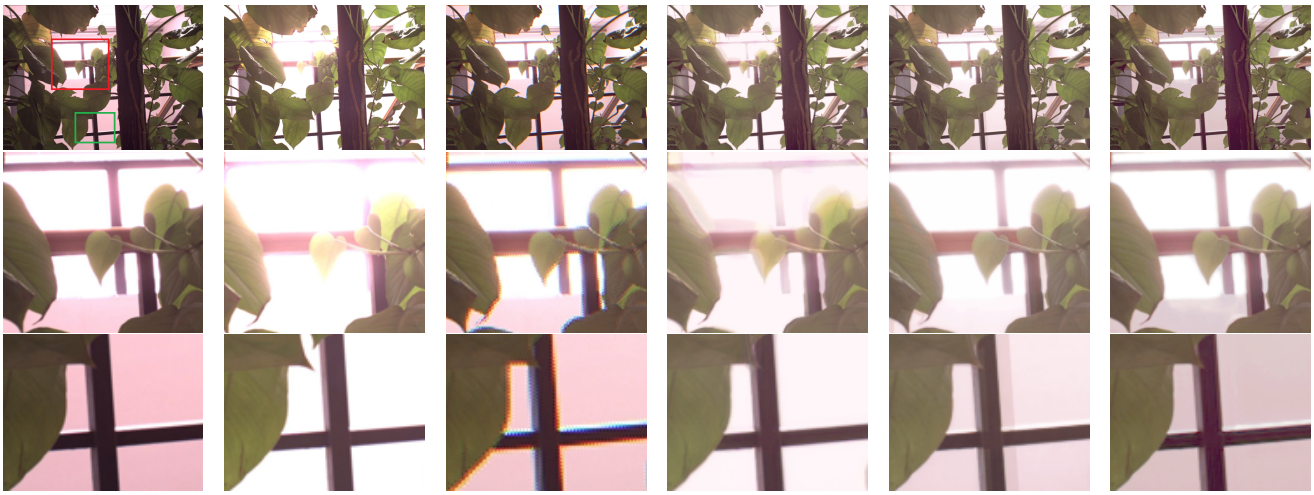
Comparison Algorithms

We compare with the state-of-the-art high dynamic range imaging methods of (Yan et al. 2020) (Wu et al. 2018) (Kalantari and Ramomoorthi 2017) (Trinidad et al. 2019) (Yan et al. 2019) (Li et al. 2020) (Prabhakar et al. 2019) and (Hasinoff et al. 2016). For the learning-based methods, we fine-tune them on our dataset for fair comparison.

Results

The quantitative results are shown in Table 3. We use PSNR, SSIM (Wang et al. 2004) and HDR-VDP-2 (Mantiuk et al. 2011) as the quality metrics. And **the qualitative results** are shown in Figs. 6 and 7. The processing time of different methods is shown in Table 2. We are faster than all the comparison methods because, as explained in Introduction, they are designed to solve the 2D movement of pixels between images while our method is designed to solve the 1D movement of pixels caused by disparity which is less computationally costly.

As shown in Table 3 and Figs. 6 and 7, we have better results than the comparison methods. Ghosting artifacts can be avoided by our method while the comparison methods usually generate ghosting artifacts, e.g. the marked red and greed box regions in Figs. 6 and 7. The works of (Yan et al. 2020) (Wu et al. 2018) let a single CNN finish all the alignment and fusion work which makes it difficult to learn correct results, especially in our problem where most of pixels between the input images have pixel movement due to disparities. As a result, their results have many ghost artifacts.



(a) Left input (b) Right input (c) Hasinoff et al. (d) Kalantari et al. (e) Wu et al. (f) Yan et al. 2019

(g) Trinidad et al. (h) Prabhaka et al. (i) Yan et al. 2020 (j) Li et al. (k) Ours (l) Ground truth

Figure 7: Examples to compare the state-of-the-art HDR imaging methods with ours. The regions marked with the red and green boxes are shown in the second and third rows respectively.

(a) Input images. (b) No denoising. (c) No alignment. (d) BM3D. (e) Guided filter. (f) Ours. (g) Ground truth.

Figure 8: Example results in the ablation study. The regions marked with the red boxes are shown in the second row.

The works of (Kalantari and Ramomoorthi 2017) (Trinidad et al. 2019) firstly perform alignment, and then fuse the aligned images into the HDR results. However, the alignment can be hardly perfect in every case, especially in occlusion regions and over-exposed regions, and ghost artifacts are introduced into the final results. The works of (Yan et al. 2019) (Li et al. 2020) (Prabhakar et al. 2019) propose to detect and/or correct the mis-alignment regions so as to reduce the mis-alignment and ghost artifacts. They can reduce the ghost artifacts in some cases, but the detection method itself may not be accurate all the time. Failing to detect the mis-alignment correctly will lead to ghost artifacts. The work of (Hasinoff et al. 2016) assumes the input images are with the same exposure level. When using it in our case, the re-

Time(s)	Hasinoff	Kalantari	Wu	Yan 2019	Trinidad	Prabhakar	Yan 2020	Li	Ours
level1	2.91	14.29	7.87	1.60	0.78	1.96	1.89	1.92	**0.71**
level2	2.48	2.76	1.95	0.47	0.48	0.98	0.89	0.47	**0.31**
level3	2.31	0.71	0.46	0.17	0.43	0.43	0.47	0.12	**0.11**

Table 2: Average processing time of different methods at three resolution levels, i.e. level1 (832x1184), level2 (416x576), and level3 (192x288).

	PSNR(dB)			SSIM			HDR-VDP-2		
	level1	level2	level3	level1	level2	level3	level1	level2	level3
Hasinoff	21.72	21.23	20.69	0.870	0.838	0.801	59.44	60.89	58.96
Kalantari	27.34	27.32	27.30	0.914	0.907	0.902	61.05	59.44	58.64
Wu	28.61	28.79	28.88	0.928	0.942	0.941	62.99	61.37	60.41
Yan 2019	25.26	25.26	25.28	0.880	0.873	0.871	62.18	61.05	60.25
Trinidad	24.38	24.30	24.18	0.902	0.896	0.888	58.64	57.99	57.67
Prabhakar	28.86	28.81	28.80	0.944	0.941	0.939	63.63	62.02	60.57
Yan 2020	28.11	28.24	28.30	0.919	0.932	0.934	62.50	60.25	58.80
Li	26.88	27.19	27.03	0.907	0.915	0.922	61.05	62.66	63.63
Ours	**33.01**	**33.08**	**33.19**	**0.963**	**0.964**	**0.966**	**69.91**	**69.27**	**68.78**

Table 3: Average PSNR, SSIM and HDR-VDP-2 values of different methods on three levels of resolutions, i.e. level1 (832x1184), level2 (416x576), and level3 (192x288).

	PSNR (dB)	SSIM
No denoising	25.37	0.890
No alignment	27.75	0.912
BM3D for the denoising	27.16	0.910
Guided filter for the denoising	28.02	0.917
Ours	**33.08**	**0.964**

Table 4: Ablation study. We show average PSNR and SSIM values of the HDR results of different variants of our model at the resolution level2.

sults usually have blur artifacts. In addition, the hand-crafted tone-mapping method cannot recover the lighting as good as the learning based methods.

Ablation Study

We compare a number of different model variants at the key parts of our model. The key ideas of our method includes the Main Image Enhancement strategy, which consists of alignment, guided denoising and fusion, and the 3D guided noising CNN for the denoising. So we 1) remove the Soft Warp CNN and use $\mathbf{I^L}$ directly as the guidance for the 3D guided noising CNN, 2) remove the 3D guided noising CNN and feed $\mathbf{I^S}$ and $\mathbf{I^{SE}}$ into the Fusion CNN, 3) replace the 3D guided noising CNN by the baseline guided denoising method, i.e. the guided filter (He, Sun, and Tang 2010), and 4) replace the 3D guided noising CNN by the baseline single image denoising method, i.e. BM3D (Dabov et al. 2007). Table 4 shows the summary performance of different model variants. Fig. 8 shows some subjective examples. The results show that any of these variants will degrade the HDR imaging quality. This verifies the contributions of our model.

Conclusions

We have presented a novel CNN model for HDR imaging using dual-lens systems. We use the Main Image Enhancement strategy and propose a new model, named MIEHDR CNN model, which consists of three subnets, i.e. Soft Warp CNN, 3D Guided Denoising CNN and Fusion CNN. Experimental results show that our method achieves superior performance than the state-of-the-art methods.

Acknowledgments

This work is sponsored by the National Nature Science Foundation of China (No. 61802026, 61806016 and 62076032) and CAAI-Huawei MindSpore Open Fund.

References

Alex, K.; Martirosyan, H.; Dasgupta, S.; Henry, P.; Kennedy, R.; Bachrach, A.; and Bry, A. 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. *International Conference on Computer Vision* .

Bradski, G.; and Kaehler, A. 2008. *Learning OpenCV : Computer Vision with the OpenCV Library*. O'Reilly Media, 1st edition.

Chen, D.; Yuan, L.; Liao, J.; Yu, N.; and Hua, G. 2018. Stereoscopic Neural Style Transfer. *The IEEE Conference on Computer Vision and Pattern Recognition* .

Chen, Y.; Jiang, G.; Yu, M.; Yang, Y.; and Ho, Y. 2020. Learning Stereo High Dynamic Range Imaging from A Pair of Cameras with Different Exposure Parameters. *IEEE Transactions on Computational Imaging* .

Cogalan, U.; and Akyuz, A. 2020. Deep Joint Deinterlacing and Denoising for Single Shot Dual-ISO HDR Reconstruction. *TIP* .

Dabov, K.; Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Image Denoising by Sparse 3D Transform-domain Collaborative Filtering. *TIP* .

Devevec, P. E.; and Malik, J. 1997. Recovering High Dynamic Range Radiance Maps from Photographs. *TOG* .

Dong, X.; Li, W.; Wang, X.; and Wang, Y. 2019. Learning a Deep Convolutional Network for Colorization in Monochrome-Color Dual-Lens System. *AAAI Conference on Artificial Intelligence* .

Dong, X.; Li, W.; Wang, X.; and Wang, Y. 2020. Cycle-CNN for Colorization towards Real Monochrome-Color Camera Systems. *AAAI Conference on Artificial Intelligence* .

Hasinoff, S.; Sharlet, D.; Geiss, R.; Adams, A.; Barron, J.; Kainz, F.; Chen, J.; and Levoy, M. 2016. Burst Photography for High Dynamic Range and Low-light Imaging on Mobile Cameras. *TOG* .

He, K.; Sun, J.; and Tang, X. 2010. Guided Image Filtering. *ECCV* .

Jeon, D.; Baek, S.; Choi, I.; and Kim, M. 2018. Enhancing the Spatial Resolution of Stereo Images using a Parallax Prior. *CVPR* .

Kalantari, N. K.; and Ramomoorthi, R. 2017. Deep High Dynamic Range Imaging of Dynamic Scenes. *TOG* .

Li, B.; Lin, C.; Shi, B.; Huang, T.; Gao, W.; and Kuo, C. 2018. Depth-Aware Stereo Video Retargeting. *CVPR* .

Li, H.; Ma, K.; Yong, H.; and Zhang, L. 2020. Fast Multi-scale Structural Patch Decomposition for Multi-Exposure Image Fusion. *TIP* .

Liu, Y.; Lai, W.; Chen, Y.; Kao, Y.; Yang, M.; Chuang, Y.; and Huang, J. 2020. Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline. *CVPR* .

Ma, K.; Duanmu, Z.; Zhu, H.; Fang, Y.; and Wang, Z. 2020. Deep Guided Learning for Fast Multi-exposure Image Fusion. *TIP* .

Mantiuk, R.; Kim, K.; Rempel, A.; and Heidrich, W. 2011. HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics* .

Mertens, T.; Kautz, J.; and Reeth, F. 2007. Exposure Fusion. *Pacific Graphics* .

MindSpore. 2020. http://www.mindspore.cn/.

Pan, L.; Dai, Y.; Liu, M.; and Porikli, F. 2017. Simultaneous Stereo Video Deblurring and Scene Flow Estimation. *CVPR* .

Prabhakar, K.; Arora, R.; Swaminathan, A.; Singh, K.; and Babu, R. 2019. A Fast, Scalable and Reliable Deghosting Method for Extreme Exposure Fusion. *ICCP* .

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention* .

Santos, M.; Ren, T.; and Kalantari, N. 2020. Single Image HDR Reconstruction Using a CNN with Masked Features and Perceptual Loss. *TOG* .

Scharstein, D.; and Pal, C. 2007. Learning conditional random fields for stereo. *IEEE Conference on Computer Vision and Pattern Recognition* 1–8.

Trinidad, M.; Brualla, R.; Kainz, F.; and Kontkanen, J. 2019. Multi-view Image Fusion. *ICCV* .

Wang, J.; Xue, T.; Barron, J.; and Chen, J. 2019a. Stereoscopic Dark Flash for Low-light Photography. *ICCP* .

Wang, L.; Wang, Y.; Liang, Z.; Lin, Z.; Yang, J.; An, W.; and Guo, Y. 2019b. Learning Parallax Attention for Stereo Image Super-Resolution. *CVPR* .

Wang, Z.; Bovik, A. C.; Sheikh, H.; and Simoncelli, E. P. 2004. Image quality assessment from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4): 600–612.

Wu, S.; Xu, J.; Tai, Y.; and Tang, C. 2018. Deep High Dynamic Range Imaging with Large Foreground Motions. *ECCV* .

Yan, Q.; Gong, D.; Shi, Q.; Hengel, A.; Shen, C.; Reid, I.; and Zhang, Y. 2019. Attention-guided Network for Ghost-Free High Dynamic Range Imaging. *CVPR* .

Yan, Q.; Zhang, L.; Liu, Y.; Zhu, Y.; Sun, J.; Shi, Q.; and Zhang, Y. 2020. Deep HDR Imaging via A Non-local Network. *TIP* .

Zhou, S.; Zhang, J.; Zuo, W.; Xie, H.; Pan, J.; and Ren, J. 2019. DAVANet: Stereo Deblurring with View Aggregation. *CVPR* .