Deep Joint Deinterlacing and Denoising for Single Shot Dual-ISO HDR Reconstruction

Uğur Çoğalan, Ahmet Oğuz Akyüz

***Abstract*—HDR images have traditionally been obtained by merging multiple exposures each captured with a different exposure time. However, this approach entails longer capture times and necessitates deghosting if the captured scene contains moving objects. With the advent of modern camera sensors that can perform per-pixel exposure modulation, it is now possible to capture all of the required exposures within a single shot. The new challenge then becomes how to best combine different pixels with different exposure values into a single full-resolution and low-noise HDR image. We propose a joint multi-exposure frame deinterlacing and denoising algorithm powered by deep convolutional neural networks (DCNN). In our algorithm, we first train two DCNNs, with one tuned for reconstructing low exposures and the other for high exposures. Each DCNN takes the same mosaicked dual-ISO input image and outputs either the low exposure or high exposure depending on the type of the network. The resulting exposures can be demosaicked and converted to the desired target color space prior to HDR assembly. Our evaluations indicate that the quality of our results significantly surpasses the state-of-the-art in single-image HDR reconstruction algorithms.**

***Index Terms*—Dual-ISO, HDR imaging, noise, deep learning**

I. Introduction

Creating HDR images by merging multiple low dynamic range (LDR) inputs is a popular approach. The chief limitation of this technique is the difficulty to deal with camera and scene motion during the capture process. Despite the development of various *deghosting* algorithms [1], [2], [3], [4], [5], [6], [7], a well-rounded solution that is efficient, robust, and reliable in all cases has been shown to be difficult [8], [9], [10], [11].

To avoid the ghosting problem altogether, an alternative approach is to capture all of the input exposures simultaneously through custom camera designs [12], [13], [14], [15], [16]. This also entails difficult problems such as the lack of commodity equipment, the difficulty of engineering, and the potential loss of energy as the light energy is distributed over multiple paths become limiting factors for these systems [17].

Several systems have also been proposed that allow per-pixel exposure control within a single shot [18], [19]. These algorithms typically achieve exposure patterns by using a spatially varying optical mask in front of the image sensor. The use of different reconstruction algorithms become the distinguishing factor for the quality of the results obtained by using these algorithms.

Recent camera hardware also support programming the sensor array such that each scanline in the sensor is set to a different sensitivity, i.e. ISO, value. The images obtained by using this technique become spatially interlaced in the exposure domain and must be resolved to obtain a full resolution HDR image.

U. Çoğalan and A. O. Akyüz are with Middle East Technical University.

The current work proposes an effective image reconstruction algorithm that not only deinterlaces an input dual-ISO frame but does so by considering the noise characteristics of each scanline. This is important as noise is the primary limiting factor of dynamic range especially in dark regions. The result becomes a joint deinterlacing and denoising algorithm accomplished within a single step. At the core of the proposed algorithm lies two DCNNs trained using a large number of image patches with simulated noise. Of these two DCNNs, one network is tuned for low exposures and the other for high exposures. The training and the reconstruction is done at the Bayer-pattern level prior to demosaicking. This reduces the information that needs to be learned by the DCNNs, while simultaneously decoupling demosaicking and HDR reconstruction from the former tasks.

An example result of our algorithm is illustrated in Figure 1. The RAW dual-ISO input in (a) is processed to obtain the low and high exposures in (b) and (c), which are merged to obtain the HDR image in (d). In this figure, network outputs (b) and (c) are shown as demosaicked for display purposes. In practice, our networks receives as input a Bayer image and produces Bayer domain outputs as well. To this end, we propose to leverage deep neural networks by taking noise characteristics into account. The primary contributions of our work are:

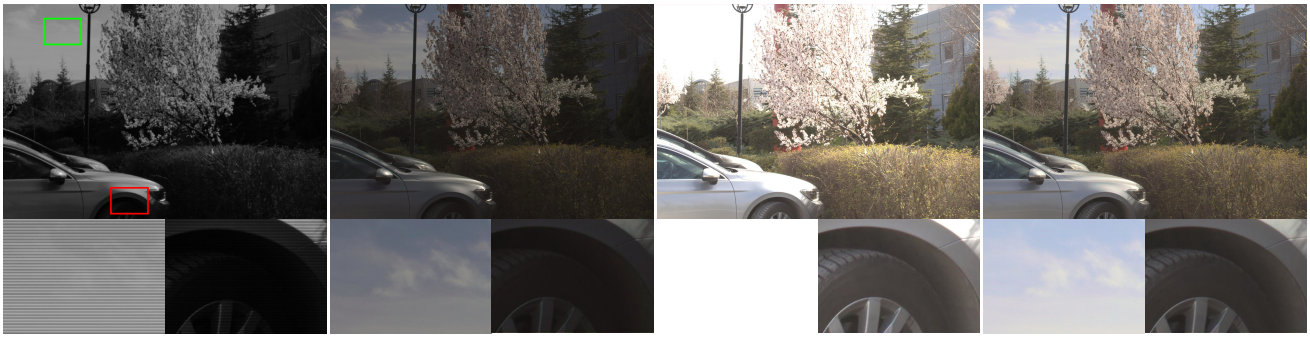
- Exploit a modern DCNN architecture for the first time to tackle the *dual-ISO* HDR image reconstruction problem,
- Provide an in-depth noise analysis of dual-ISO images and show how they contribute to dynamic range,
- Perform an analysis of alternative DCNN architectures and make extensive comparisons with the state-of-the art,
- Make available a dual-ISO dataset to stimulate further research in this direction.

II. Related Work

HDR imaging has received considerable attention in recent years. Of the various challenges posed by HDR imaging, capturing high quality HDR content has been among the most important ones [20]. Below we review various algorithms designed to overcome these challenges.

A. Multiple Exposures Techniques

HDR images and videos have traditionally been captured by using multiple exposures techniques. Mann and Picard



(a) Raw interlaced dual-ISO image (b) Reconstructed low exposure (c) Reconstructed high exposure (d) Tone-mapped HDR image

Fig. 1: Our algorithm takes as input a RAW dual-ISO input image (a) and reconstructs full resolution low (b) and high ISO (c) exposures in the Bayer domain. The resulting exposures can be demosaicked and merged to obtain an HDR image (shown as tone mapped in (d)). Note the visibility of details in both low and high luminance regions. The full resolution inputs and outputs of our algorithm as well as the final HDR image can be seen in electronic supplementary materials.

proposed to combine multiple exposures after linearizing them using a simple parametric camera response function (CRF) [21]. Debevec and Malik improved this technique by recovering the CRF directly from the input exposures allowing it to have arbitrary shapes [22]. They proposed to use a triangular weighting function during the HDR merging process to reduce the influence of under- and over-exposed pixels. Mitsunaga and Nayar, on the other hand, proposed to use a polynomial to model the CRF and use a noise aware weighting function [23]. Several related works followed that aim to recover the response function more accurately and obtain an HDR image with reduced noise [24], [25], [26], [27], [28].

A common assumption of these studies is that the captured scene as well as the capture device remain static during the capture process. Any motion breaks down the assumption that the same pixel measures the radiance of the same scene point. In order to circumvent this limitation, various methods known as *deghosting* algorithms have been proposed [5], [29], [2], [3], [30], [6], [7]. The performances of some of these algorithms have been evaluated through several studies [8], [9], [10], [11], with the overarching conclusion that although there are very high quality deghosting algorithms, they are computationally demanding and are still prone to artifacts especially when the reference exposures are not well-exposed.

Multiple exposures techniques are also used to capture HDR videos by alternating exposure times across frames. Dealing with motion artifacts is an integral part of these algorithms [31], [32], [4], [33].

B. Single-shot HDR Reconstruction

The difficulty of dealing with motion artifacts and the inherent complexities of the multiple exposures techniques motivated the development of single-shot HDR reconstruction methods. Nayar and Mitsunaga [18] and Schoberl et al. [34] proposed to use an optically varying mask in front of the sensor array to enable a spatially varying exposure. Alternatively, the read-out circuitry was modified to measure the *time-to-saturation* or perform *multiple non-destructive reads* during the capture process [35]. These schemes use standard CCD or CMOS sensors. One can also utilize photodetectors that are themselves HDR capable [36].

The aforementioned techniques require modifications to the camera hardware. An alternative approach is to use either single-aperture multi-sensor or multi-camera solutions. In the former, multiple exposures are captured simultaneously by impinging the incoming light onto multiple sensors within the camera [12]. Recent methods of this type aim to minimize the light loss due to separation of the incident light into multiple paths [37]. In the multi-camera solutions, the scene is captured from slightly different viewpoints, which must be brought into alignment during the HDR assembly [14], [38], [39].

Inverse tone mapping is another technique that can be used to create an HDR image from a single LDR image [40], [41]. Although this technique can be used to recover some of the lost details [42], it is primarily targeted for enhancing LDR content for HDR displays [43], [44] and image based lighting [45]. Several recent studies employ deep learning to perform inverse tone mapping [46], [47].

Of most related to our work are techniques that involve reconstructing an HDR image captured with a spatially varying exposure pattern. Among these Hajisharif et al. [48] use the dual-ISO module available in Magic Lantern firmware [49]. They create the HDR image by exploiting the adaptive kernel regression method [50]. Rodriguez et al. [51] use an inpainting based deinterlacing method to reconstruct two full resolution frames from a single dual-ISO image and combine them into an HDR image. Choi et al. also utilize the dual-ISO module in Magic Lantern and propose a joint dictionary learning method via sparse coding both for deinterlacing and denoising purposes [52]. However, their primary focus is in videos rather than individual images. Heide et al. [53] propose a digital camera imaging pipeline that performs demosaicking, denoising and deblurring jointly using a global optimization algorithm. They also show that their algorithm works for reconstructing an HDR image from an interlaced exposure image. Another group of algorithms utilize coded electronic shutter to obtain a row-wise varying exposure image [54], [55], [56]. Due to exposure time differences between the rows, they perform motion compensation to prevent deinterlacing artifacts on the

final HDR image. Hasinoff et al. [57] propose an alternative exposure bracketing method that uses multiple frames captured under a constant exposure. They merge multiple frames to reduce the noise and increase dynamic range. Finally, Serrano et al. propose a method that reconstructs an HDR image from an exposure-coded single LDR image employing convolutional sparse coding [19].

Despite the significant progress that is achieved by these works, the limitation of the state-of-the-art methods lies in the fact that they do not exploit modern DCNN architectures to tackle the dual-ISO HDR image reconstruction problem. The only DCNN approach appears to have been proposed by An and Lee [56], but only for exposure interlacing – not for dual-ISO. As a result that study makes no attempt for noise reduction, which is a critical factor for reconstructing high quality HDR images from a dual-ISO setup.

III. Noise Model in Digital Imaging

Digital images are corrupted with noise caused by various sources. Assume that a pixel j records an irradiance of X_j for t_i seconds. Then the total collected charge (i.e. exposure) at that pixel can be represented as [26]:

$$E_{ij} = t_i(a_j X_j + D_j), \tag{1}$$

where a_j is the photo-response non-uniformity of the pixel and D_j represents the irradiance due to dark current. Assuming a linear camera (i.e. using the RAW output mode as in our case) with a combined gain of g_i, the output digital value will be equal to:

$$V_{ij} = \left[g_i E_{ij} + N_R\right], \tag{2}$$

where N_R represents the read-out noise [58]. The variance of V_{ij} is equal to:

$$\sigma^2_{V_{ij}} = g_i^2 \sigma^2_{E_{ij}} + \sigma^2_{N_R}. \tag{3}$$

Given a two exposures h and l with different gains but the same exposure time, the ratio of their noise variances can be written as:

$$\frac{\sigma^2_{V_{hj}}}{\sigma^2_{V_{lj}}} = \frac{g_h^2 \sigma^2_{E_{hj}} + \sigma^2_{N_{Rh}}}{g_l^2 \sigma^2_{E_{lj}} + \sigma^2_{N_{Rl}}}. \tag{4}$$

Note that $\sigma^2_{E_{hj}} = \sigma^2_{E_{lj}}$ as the exposure times for the both exposures are the same. This formula depends not only on the gains of the two exposures but also on the exposure values themselves. For low exposures, this ratio will approximate the ratios of the read-out noise variances. As the exposure values increase, they will start to dominate the result until when the exposure saturates and the result converges back towards unity.

The variance of the read noise can be modeled by two components, pre-amplifier and post-amplifier read noise:

$$\sigma^2_{N_R} = (g\sigma_{R_{\text{pre}}})^2 + \sigma^2_{R_{\text{post}}}. \tag{5}$$

Here, it is important to note that the post-amplifier read noise is not affected by the ISO setting of the camera. In other words, a high-ISO image normalized by its ISO value will actually have a lower noise variance compared to a low-ISO image. We discuss this issue in more detail in Section VII-A.

IV. Algorithm

The input to our algorithm is a single dual-ISO image that has scanlines with alternating ISO values. The alternation occurs not in each scanline but between pairs of scanlines to allow demosaicking using pixels with the same exposure. The low-ISO scanlines contain information for the light regions of the scene, whereas the high-ISO scanlines provide data for the dark regions. From this image, our algorithm reconstructs two full resolution exposures, one representing the low camera exposure and the other the high camera exposure. The resulting images can then be merged into an HDR image and tone mapped. The main steps of our algorithm are (see Figure 2):

1) Collect training data to obtain low and high exposure ground-truth images (§ IV-A),
2) Simulate dual-ISO by interleaving scanlines from these images and adding random noise commensurate with their noise characteristics (§ IV-B),
3) Train two DCNNs models with one aimed to recover a full resolution low-ISO exposure and the other a full resolution high-ISO exposure (§ IV-C).
4) Once the models are trained a dual-ISO input can be given to them to obtain two full resolution low- and high-ISO images. These images would be in the Bayer domain and any desired demosaicking and HDR assembly algorithm can be used to obtain the final image.

A. Training Set Collection

Although there are many existing HDR image datasets, none of them was suitable for our purpose as they contain images that have an unknown amount of noise. For this reason, we captured various images representing a diverse set of scenes using a Canon EOS 600d dSLR camera. The camera was programmed to capture in RAW format with an image resolution of 5202×3465. The dual-ISO feature of the camera was disabled for these images. Since the reconstruction of dark regions is important, we aimed to minimize the uncertainty in these regions. Training data was collected using two approaches. In the first approach, each scene was captured 40 times with half of them at ISO 100 and the other half at ISO 1600 setting of the camera. The resulting exposures were averaged separately using frame averaging [59] to obtain noise-reduced ground-truth exposures. 47 pairs of low- and high-ISO ground-truth images are created using this method.

However, for particularly dark scenes, averaging 20 images captured with ISO 1600 setting left some noticeable noise. As these training images were going to be used as ground-truth, we employed a second approach for darker scenes. Specifically, we captured 20 images at $-2EV$ and another 20 at $+2EV$ with all images captured at ISO 100 setting. Exposure value was altered by only changing the exposure time. Again, the exposures at each setting were averaged. 63 pairs of low- and high-exposure images that can serve as ground-truth are obtained using this approach. This gave rise to a total of 110 pairs of ground-truth images. As the goal of this stage was to obtain noise-free images that can serve as ground-truth, using two different approaches to obtain them did not have a negative impact for the training of our networks.

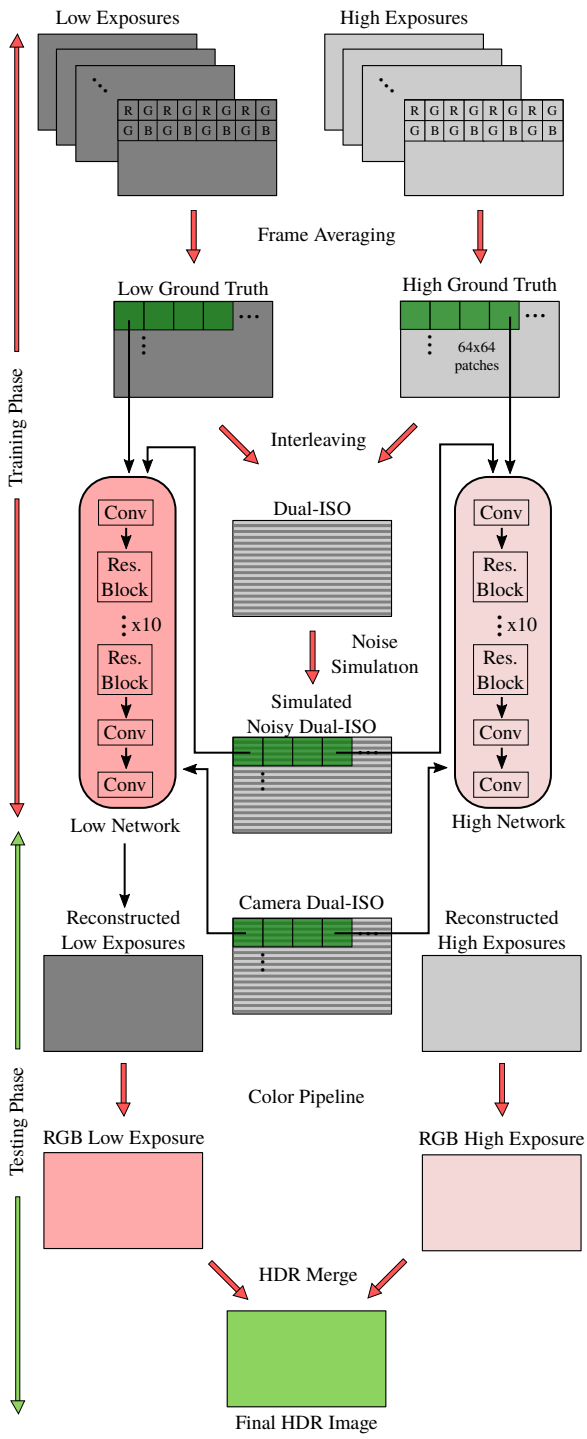


Fig. 2: The overall flow of our algorithm.

In all captures, the camera was controlled using a smart phone to allow rapid capture without user intervention. Static scenes were preferred and the camera was mounted on a tripod. Although images were captured using a tripod, an image registration algorithm [60] was still used to correct for minor misalignments.

B. Simulation of Dual-ISO and Noise

A dual-ISO image can be obtained from low- and high-ISO images by taking pairs of scanlines from each image

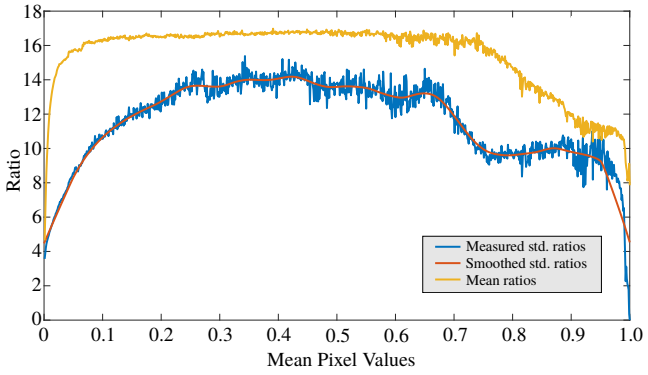


Fig. 3: Mean and noise ratios between high (ISO 1600) and low (ISO 100) ISO images computed over 64×64 patches of 10 test images and plotted as a function of mean pixel value of the high ISO image.

alternatingly. We perform this step in a manner that mimics the camera's dual-ISO mode. More precisely, the first two scanlines of the simulated image are retrieved from the high exposure, the third and fourth scanlines from the low exposure, and so on.

Understanding the noise characteristics of low and high-ISO exposures is important to simulate noise in a realistic manner. Assume that p_h represents a patch of certain size from a high-ISO exposure and q_h the corresponding patch in a noise-reduced image that may serve as ground-truth. The standard deviation (i.e. noise) of p_h can be computed using:

$$\sigma_{p_h} = \sqrt{\mathbb{E}((q_h - p_h)^2) - [\mathbb{E}(q_h - p_h)]^2} \tag{6}$$

The noise level of a patch from a low-ISO exposure can be found analogously. Using a patch size of 64×64, we conducted this analysis and generated noise levels of a large number of patches from exposures captured for a variety of scenes.

To compute noise ratios between high- and low-ISO patches as a function of mean intensity, we first subtracted the black level and divided by the saturation level to bring them to $[0, 1]$ range. We then found the mean value of a high patch, μ_{ph}, and computed its bin index in a table of N elements:

$$i = \text{bin}(\mu_{ph}) = \lfloor N \mu_{ph} \rfloor, \tag{7}$$

The noise ratio for this bin can then be computed by:

$$R(i) = \frac{1}{\mathbf{card}(\Omega_i)} \sum_{j \in \Omega_i} \frac{\sigma_{p_{hj}}}{\sigma_{p_{lj}}}, \tag{8}$$

$$\text{where} \quad \Omega_i = \left\{ j = 1 \dots P : \text{bin}(\mu_{phj}) = i \right\}. \tag{9}$$

Here P denotes the total number of patches used to obtain the ratio function. The resulting function R is plotted in Figure 3. It can be seen that it roughly follows an inverse U-shape due to the effect of the additive read-out noise term in Equation 4. Also the ratio of standard deviances stay below the mean ratios due to the influence of the additive terms as well.

We capitalize on this information in the following manner. To obtain a simulated noisy dual-ISO image, we select a uniformly distributed random number σ_{p_h} in $[0, 0.02]$. As shown by Burger et al. [61], simulating noise using a variety

of noise levels improves the robustness of networks when used to denoise images with unknown amount of noise. This value represents the standard deviation of the noise term that will be added to all pixels of patch p_h. The noise added high exposure is then computed as:

$$I'_{hj} = I_{hj} + \mathcal{N}(0, \sigma_{p_h}),\ \forall j \in p_h, \tag{10}$$

where I_h represents a ground-truth normalized high exposure image. The amount of noise that will be added to the corresponding pixels in the low exposure is computed by using the smoothed ratio function, R_s:

$$\sigma_{p_l} = \frac{\sigma_{p_h}}{R_s(\text{bin}(\mu_{ph}))}. \tag{11}$$

The noise added low exposure is then computed by:

$$I'_{lj} = I_{lj} + \mathcal{N}(0, \sigma_{p_l}),\ \forall j \in p_l. \tag{12}$$

Finally, I'_h and I'_l were interleaved to obtain the noise added dual-ISO frame, I'_{dl}. Patches from this image, $\{I'_{\text{dl}p}\}$, along with the corresponding patches from the ground-truth low and high exposures, $\{I_{\text{gt}_p}\}$, are used to train the network as explained below. It is important to note that the simulation of dual-ISO and noise was only done for the training of our networks. In the testing phase, we used real dual-ISO images that were directly captured by a camera.

C. Network Training

We treat joint deinterlacing and denoising as a supervised learning problem. Since our goal is to create two full-resolution exposures from a single dual-ISO image, we opted for training two networks one tuned for low-ISO exposures and the other for high-ISO ones. The input for each network is a single channel dual-ISO image, I'_{dl}, and the corresponding output, I_{gt}, is also a single channel image with the same resolution and reduced noise. This way both networks learn to fill the missing scanline information while simultaneously aiming to find a low noise solution.

1) Training: All training images were divided into 136×136 non-overlapping patches and noise was simulated as explained in Section IV-B. To obtain a whole number of patches both in horizontal and vertical directions, input images were cropped to 5168×3400 resolution. To further increase the number of patches, data augmentation was applied by rotating and shearing the image patches, creating 5 different versions for high and low patches from a single patch[1]. This resulted in a total training set size of 522500 patches for each network.

The training data of each network consisted of a set of dual-ISO and ground-truth patches, i.e. $D = \{\langle I'_{\text{dl}p}, I_{\text{gt}_p}\rangle\}$. The optimization function that minimizes the weights for both networks was:

$$\mathcal{L}(\Theta) = \frac{1}{s^2|D|} \sum_p || \widehat{I}_{\text{gt}_p} - I_{\text{gt}_p} ||^2, \tag{13}$$

where $\widehat{I}_{\text{gt}_p}$ is the estimated output and I_{gt_p} is the corresponding ground-truth patch. s stands for the patch size of 136×136 and Θ represents the network parameters. ADAM optimizer [62] was used for faster convergence and the parameters were set as $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\epsilon = 10^{-7}$ as recommended by the authors. The learning rate was set to 10^{-4} and batch size was 64 for each epoch. The higher initial learning rate did not provide good convergence. Training was conducted with Keras [63] using Tensorflow as the backend. A single NVIDIA Tesla P100 GPU was utilized for training. The training continued until there was no improvement on the validation loss. With this setup, the total training time for both networks took approximately 90 hours.

2) Network Architecture: We used a network architecture after He et al. [64]. In this architecture, rather than only stacking multiple convolutional layers, they introduced shortcut connections between layers to learn residuals. They proved that learning residuals is easier than direct mapping and learning residuals allows to go deeper by mitigating the degradation problem.

Our architecture starts with a convolutional layer followed by C residual blocks and ends with two more convolutional layers (Figure 4). Lim et al. [65] designed each residual block by removing batch normalization to speed up the training and improve performance. We followed the same idea and removed batch normalization. ReLU activation was applied only after the first layer inside the residual block. Residual blocks are followed by two convolutional layers which are activated by ReLU non-linear function. In addition to the skip connections inside the residual blocks, we also applied symmetric skip connections between residual blocks to better preserve details. As such, early layer features are propagated through the network which we found to improve the reconstruction quality.

Our architecture consists of $C = 10$ residual blocks and has a total depth of $D = 23$ convolutional layers. We followed the idea of Simonyan and Zisserman [66] and used a small kernel size of $K = 3$. This would result in a receptive field size of 47×47 for the whole network. As we wanted the receptive field to cover our entire patches without increasing the depth of our network, we opted to use dilated convolutions [67]. Dilated convolution with a dilation factor of $d = (d_w, d_h)$ can be formulated as:

$$(F *_d w)(p) = \sum_{i+jd=p} F(i) \cdot w(j), \tag{14}$$

where w is the convolution filter with the size of $K \times K$ and F is the input feature map. The dilation factors for each residual block are listed in Table I.

In this scheme, a dilated convolutional layer can be expressed as:

$$F_l = W_l *_d F_{l-1} + B_l, \tag{15}$$

where l stands for the layer index, W_l are the filters in this layer, and B_l are the corresponding biases. Within a residual block, the output after the application of a single convolution and ReLU can be written as:

$$F'_{2c} = \max(0, W_{2c} *_d F_{2c-1} + B_{2c}), \quad c \in \{1 \cdots C\}, \tag{16}$$

[1]We demosaicked the images first, rotated/sheared, and then remapped the results to the mosaic domain. This resembles what would have been obtained if the camera was rotated when the images were captured.

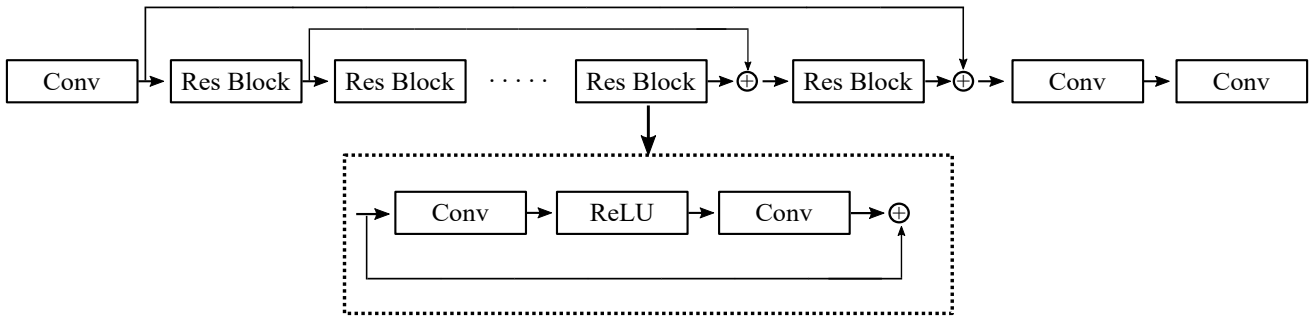


Fig. 4: The detail of the proposed architecture. ReLU activation layers are omitted for the sake of simplicity.

Residual Block	Dilation Factor	Output Receptive Field
c = 1	$(1, 1)$	7×7
c = 2	$(2, 2)$	15×15
c = 3	$(2, 2)$	23×23
c = 4	$(4, 4)$	39×39
c = 5	$(4, 4)$	55×55
c = 6	$(8, 8)$	87×87
c = 7	$(4, 4)$	103×103
c = 8	$(4, 4)$	119×119
c = 9	$(2, 2)$	127×127
c = 10	$(2, 2)$	135×135

TABLE I: Dilation factors $d = (d_w, d_h)$ for convolutional layers inside residual blocks and resulting receptive fields after each residual block. Convolutional layers outside residual blocks have dilation factor of $d = (1, 1)$ and the network has a total receptive field size of 139×139.

where c represents the residual block index. Finally, the output of a single residual block including skip connections becomes equal to:

$$F_{2c+1} = \begin{cases} (W_{2c+1} *_d F'_{2c} + B_{2c+1}) + F_{2c-1}, & c \leq C/2, \\ (W_{2c+1} *_d F'_{2c} + B_{2c+1}) + F_{2c-1} + F_{C-c+1}, & c > C/2. \end{cases} \quad (17)$$

We created two networks using the architecture in Figure 4 for reconstructing low and high exposures. Both networks share the same parameters stated above. All convolutional layers in both low and high networks have 32 filters except for the last layer which has only 1 filter that serves to accumulate the previous layer's outputs into a single result.

Since we want to preserve the resolution of the input image, we used a stride parameter of $S = 1$. Also the input images were zero padded in each dimension. The convolutional layers were initialized using the normalized uniform initialization of Glorot and Bengio [68] and the biases were set to zero. The total number of learnable parameters in each of our networks was 194817. One forward pass of an input patch resulted in 3,590,295,552 MACs and 7,193,630,784 FLOPs.

3) Discussion: We arrived at the aforementioned network architecture after experimenting with several alternatives. Initially, we used patch sizes of 32×32, and later 64×64, with network depths of 6, 11, and 16. These networks did not contain residual blocks or skip connections. Using a larger patch size with dilated convolutions improved this result. Initially, the dilation factors were less aggressive, which resulted in a smaller receptive field than the patch size. Increasing the dilation factors as in Table I improved the quality of the results. We also experimented with a pure residual network (termed as DualISONet hereafter), a residual network with dilated convolutions (DualISONet-D), and finally the residual network with both dilated convolutions and symmetric skip connections (DualISONet-DSC) as explained in this section. We compare the results of these three networks in the following section. We also tried using 64 filters per-layer instead of 32, but found it to not improve performance despite increasing computational complexity.

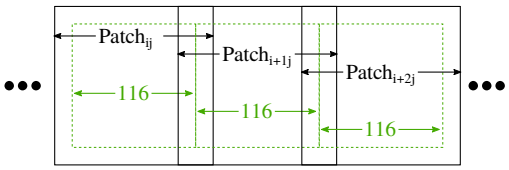


Fig. 5: Given a 136×136 input patch, only the center 116×116 region is considered to be a valid output. This overlap between the patches mitigates tiling artifacts that may be noticeable if patches were disjoint.

We also experimented with developing a single combined network that outputs both low- and high-ISO images instead of using two separate networks. This was motivated by the fact that there could be some joint features that are shared by both networks. Thus if these features are learned together by using data for both image types, better results could be obtained. To this end we tested two new network architectures. The first architecture was a fully joint architecture with its last layer designed to output 2 channels (called Joint-Full). The second architecture combined a joint and a forked design. It started with common layers followed by forking into separate paths. The paths join in the last layer which again outputs 2 channels for low- and high-ISO images. We call this architecture Joint-Forked. There are several sub-types of this architecture depending on how many layers are common before they fork into separate paths. We tried three alternatives namely Joint1-Forked, Joint3-Forked, and Joint5-Forked, where the number indicates the number of initial joint layers.

V. HDR Image Reconstruction

As mentioned above, the DCNNs learn to predict noise-reduced low-ISO and high-ISO frames given an input dual-ISO frame. To create an HDR image starting with a dual-

(a) Naive bicubic from a single dual-ISO image | (b) Full resolution two exposures | (c) Bicubic from denoised and interlaced two exposures | (d) Our result from a single dual-ISO image | (e) Ground-truth

Fig. 6: Comparison between several alternative approaches and our method.

ISO frame, we run both networks on the dual-ISO frame to reconstruct full resolution low- and high-ISO frames. Reconstruction is performed for each patch obtained by dividing the input image into 136×136 patches. To avoid blocking artifacts, we used a padding of 10 pixels in all directions (Figure 5). This number appeared to eliminate noticeable tiling in our results, but if necessary, can be increased at a small cost of increasing computational load.

Denoting these reconstructed full resolution frames respectively as E_l and E_h, the goal of HDR reconstruction is to combine these frames to obtain a relative irradiance estimate of the captured scene.

Remembering that these frames contain RAW data in Bayer domain, we pass them through a color image processing pipeline that resembles the built-in processing of a digital camera [69]. The images are first white-balanced using the RGGB multipliers stored in the RAW file of the original dual-ISO image. Demosaicking is applied to the white-balanced frames [70]. This is followed by a linear color transformation from the camera color space to the linear sRGB space. This process can be summarized as follows where $\mathbf{A}$ represents the white-balance matrix, $Z(.)$ demosaicking function, and $\mathbf{T}$ color transformation:

$$G = \mathbf{T}Z(\mathbf{A}E). \tag{18}$$

Both low and high reconstructed frames are processed to obtain G_l and G_h. The final HDR image H is created by:

$$H(x,y) = \frac{\sum_{i\in\{l,h\}} f(G_i(x,y))\dfrac{G_i(x,y)}{g_i}}{\sum_{i\in\{l,h\}} f(G_i(x,y))}, \tag{19}$$

where (x,y) denote the pixel coordinates, g_i is the gain, and f is a weighting function that is used to control the influence of under- and over-exposed pixels. This image can be tone mapped and gamma corrected for display purposes.

VI. Results

In this section, we present various results obtained with our algorithm and compare them to the state-of-the-art techniques. All of the presented results contain images that were outside of the training set of our algorithm, and obtained as dual-ISO images directly from the camera.

In Figure 6 we compare our results with simple alternative techniques noting that some of these alternative require multiple input images and therefore are not a replacement for a dual-ISO pipeline. In this figure, the image in (a) represents creating two exposures from a single dual-ISO input image by bicubic upsampling and merging them into an HDR image. This image contains both noise and interlacing artifacts as can be seen around the edges of the roof. To obtain the result in (b), two full-resolution exposures were captured with one having a low-ISO value and the other a high-ISO value. These exposures were then merged into an HDR image. The result contains noise but is devoid of interlacing artifacts due to using two separate exposures. Alternatively, in (c), a denoised dual-ISO input image is created by averaging a large number of low- and high-ISO frames and interlacing them as explained in Sections IV-A and IV-B. The resulting image is bicubic upsampled to two exposures and then merged into HDR. As can be seen in (c), the result has reduced noise but contains interlacing artifacts. (d) shows our result obtained from a single dual-ISO exposure. Finally, (e) is the ground-truth obtained by merging two full-resolution denoised low and high ISO images. It can be seen that our result most closely resembles the ground-truth despite having been obtained from a single dual-ISO exposure.

In Figure 7, we show the quality of detail recovery in under- and over-exposed regions of a scene captured using a single dual-ISO 100/1600 image. This scene represents a challenging capture condition due to having a lower illumination foreground combined with a background that receives direct sunlight. It can be seen that our algorithm successfully reconstructs both low and high luminance regions yielding an HDR image that contains visible details in both regions.

A. Comparison with the State-of-the-Art

First we compare our results with several state-of-the-art methods that are suitable for a dual-ISO HDR image reconstruction. We applied our method as well as the compared methods on high resolution input images and cropped several patches for a detailed visual illustration in Figure 8. We used the original source codes of the compared methods when available; otherwise we shared our inputs with the authors and were kindly provided the results. From left-to-right, the columns depict the results of bicubic interpolation, Hajisharif et al. [48], Serrano et al. [19][2], Choi et al. [52][3],

[2]This algorithm was designed for coded-exposure images. However, it was also shown to work on dual-ISO images by the original authors [19].

[3]This algorithm targets video reconstruction and therefore has a temporal denoising step. This step was skipped when processing individual images.

(a) Interlaced dual-ISO image (b) Reconstructed low exposure (c) Reconstructed high exposure (d) Tone-mapped HDR image

Fig. 7: The input image in (a) is processed by our method to produce the images in (b) and (c). The resulting images are merged to obtain an HDR image where details are visible in both low and high luminance regions. Images in (a), (b), and (c) are shown as demosaicked for display purposes – in practice, our algorithm operates in the Bayer domain.

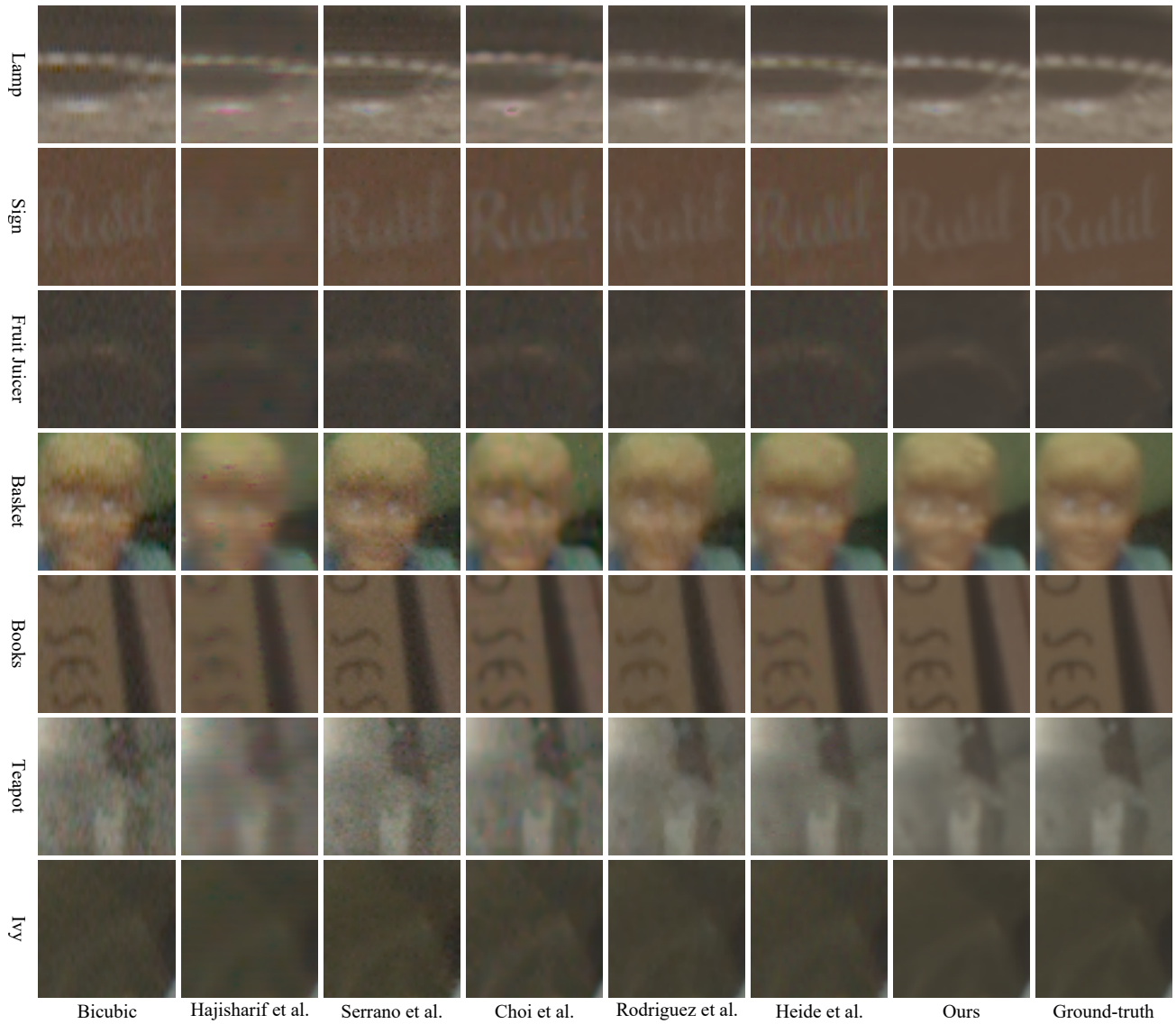


Fig. 8: As shown in these examples, the noise reduction and deinterlacing quality of our approach closely resembles the ground-truth and outperforms the state-of-the-art methods.

PSNR/HDR-VDP-2.2	Lamp	Sign	Fruit Juicer	Basket	Books	Teapot	Ivy	Mean
Bicubic	31.65 / 55.40	37.39 / 55.76	37.82 / 54.47	31.42 / 54.60	36.41 / 56.19	32.60 / 53.36	40.11 / 56.80	35.34 / 55.23
Hajisharif et al.	33.45 / 57.76	38.43 / 54.68	39.95 / 59.04	32.57 / 55.03	36.52 / 55.23	34.36 / 55.52	41.61 / 59.53	36.70 / 56.68
Serrano et al.	36.47 / 57.43	39.09 / 56.98	39.28 / 55.98	32.82 / 57.39	38.60 / 61.36	31.50 / 55.96	38.13 / 56.54	36.55/ 57.38
Choi et al.	31.85 / 56.10	38.34 / 56.13	39.59 / 56.37	33.24 / 56.64	38.47 / 59.64	34.30 / 56.27	39.96 / 57.80	36.53 / 57.00
Rodriguez et al.	37.27 / 59.23	42.42 / 58.58	42.24 / 57.96	34.96 / 57.70	40.53 / 59.42	37.80 /57.92	42.79 / 59.88	39.72 / 58.67
Heide et al.	36.74 / 60.28	39.40 / 59.29	39.98 / 57.94	34.57 / 59.42	40.94 / 61.77	38.17 / 58.21	41.54 / 59.18	38.76 / 59.44
DualISONet-DSC	**39.96 / 63.21**	**44.94 / 61.20**	**46.75 / 63.08**	**37.42 / 59.85**	**44.55 / 63.89**	**41.26 / 60.97**	**48.61 / 63.68**	**43.35 / 62.27**

TABLE II: PSNR and HDR-VDP-2.2 [71] values of the compared methods with respect to the ground-truth. Computations are made using the tone mapped [72] (key value was set to 0.15) reconstructed and ground-truth images. The PSNR and HDR-VDP-2.2 values of our method are consistently larger than the comparisons.

Rodriguez et al. [51], Heide et al. [53], our approach, and the ground-truth. The PSNR and HDR-VDP-2.2 [71] values with respect to ground-truths are reported in Table II. As can be seen in the figure and the corresponding table, the proposed algorithm's reconstruction quality outperforms all of the compared methods with the resulting images exhibiting less noise and being sharper than those of the comparison. Its run-time performance also surpasses the compared algorithms (Table IV).

In order to get a more detailed understanding about the image quality afforded by all of these algorithms, we also performed a statistical analysis. To this end, we partitioned each of the high resolution input images to 400×400 patches. We visually inspected these them to remove highly correlated patches. Also some patches exhibited motion artifacts (as the ground-truths are computed by frame averaging), and these patches were removed as well. This gave rise to a total number of 466 patches. Using these patches and the corresponding ground-truths, we computed the PSNR and HDR-VDP-2.2 results for all algorithms. As this yields a distribution of scores, we applied ANOVA to understand whether statistically significant differences exist between these distributions. The corresponding mean scores using 95% confidence intervals are shown in Figure 9. The ANOVA results revealed that there were significant differences between the algorithms: $F(8, 4185) = 508.1, p \ll 0.001$. Post-hoc tests using Bonferonni correction indicated the significance groups shown in Figure 10. According to these results, our residual network with dilated convolutions and symmetric skip connections (DualISONet-DSC) produced the best result, achieving a mean PSNR value of 43.55.

The comparison between the joint architectures using the same evaluation framework is summarized in Table III. Here, it can be seen that the proposed dual architecture has a slightly better performance than those that involve common layers. For fair comparison, we tuned the number of filters in each layer so that all architectures contained approximately equal number of parameters.

B. Further Comparisons

In addition to PSNR measurement on tone mapped images, we applied perceptually linear transform to HDR images [73]. This transform allows standard image quality metrics to be applied on HDR images without tone mapping them first. We also used the HDR-VDP-2.2 metric directly on the HDR images. As shown in Table V, in both evaluations our mean

Network Type	Filters	Filters (last)	Parameters	PSNR
Joint1-Forked	33	2	413,954	43.38
Joint3-Forked	33	2	394,286	43.19
Joint5-Forked	34	2	397,598	43.08
Joint-Full	46	2	402,180	43.41
DualISONet-DSC	32	1	389,634	43.55

TABLE III: The comparison between our dual network architecture and alternatives that use joint layers. The third column indicates the filter count in the last layer.

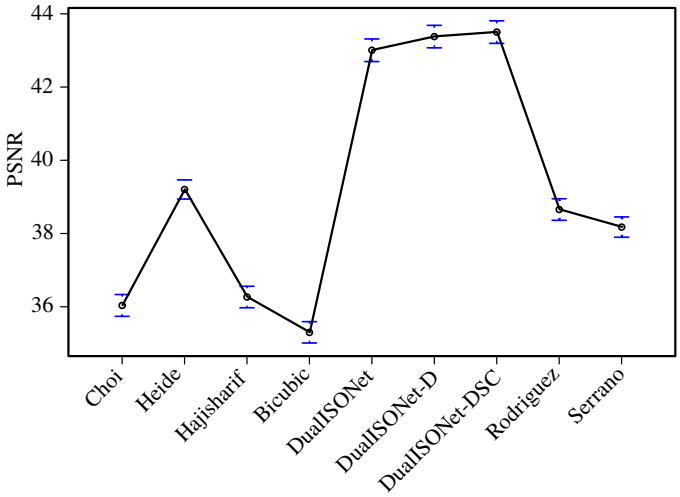


Fig. 9: Mean plot with 95% confidence intervals.

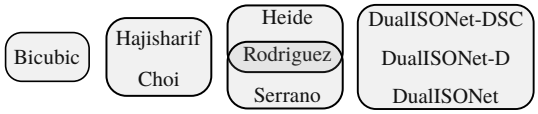


Fig. 10: Algorithms enclosed in the same box are found to be in the same statistical significance group.

Time (sec.)	512×512	1024×1024
Serrano et al.	1783.62	19286.23
Hajisharif et al.	22.61	94.68
Choi et al.	≈ 18.75	≈ 75
Rodriguez et al.	≈ 25	≈ 200
Heide et al.	53.38	258.53
DualISONet-DSC (CPU)	6.87	32.51
DualISONet-DSC (GPU)	0.76	2.26

TABLE IV: Running times of the compared methods. Tests were executed on an Intel i7-7700HQ CPU and NVIDIA GTX 1050 Ti GPU system. Choi et al. and Rodriguez et al.'s results are extrapolated from the timings shared with us by the authors on full-resolution images using a comparable test platform.

Scores	Perceptual Transform PSNR [73]	HDR-VDP-2.2
Bicubic	48.16	81.71
Serrano et al.	63.51	81.00
Hajisharif et al.	62.33	81.98
Choi et al.	60.34	80.96
Rodriguez et al.	63.41	83.83
Heide et al.	67.17	84.51
DualISONet-DSC	69.49	86.10

TABLE V: The middle column shows the PSNR scores computed after perceptually linear transform [73]. The right column shows the HDR-VDP-2.2 scores computed directly between the HDR images.

PSNR	**0.045**	**0.09**	**0.18**	**0.36**	**0.72**
Heide et al.	43.00	40.79	38.48	36.21	34.35
DualISONet-DSC	46.90	44.80	42.88	41.06	39.67

TABLE VI: PSNR results for different key value parameter of the photographic tone mapping operator [72].

scores surpass the comparisons. Finally, to show that our results on tone mapped images are not specific to a single parameter setting, we changed the value of the key value parameter in photographic TMO [72], which controls the overall brightness of the image. We show the comparison of our results with our closest competitor [53] in Table VI.

C. Comparison with iTMO

Inverse tone mapping operators (iTMOs) provide an alternative approach to create an HDR image from a single LDR image. In Figure 11 we compare the results of two deep learning based iTMOs, namely Eilertsen et al. [46] and Endo et al. [47], with our algorithm. For this comparison, in addition to using low- and high-ISO inputs we also captured a medium-ISO image to provide a better exposed input for these algorithms. The results for two regions with different light conditions are shown in top and bottom rows. As can be seen from the figure, our results are closer to the ground-truth (computed by frame averaging), than both of the iTMOs.

D. Comparison with direct HDR reconstruction

In addition to creating separate networks for low- and high-ISO images, we also experimented with reconstructing an HDR image directly with a single network using the same network architecture. We hypothesized that this would pose a more difficult learning challenge as it necessitates learning of a higher bit-depth HDR image created by using a weighting function. As can be seen in Figure 12, the direct reconstruction does not fully eliminate deinterlacing artifacts. Furthermore, the overall contrast of the image is reduced. The same observation was made by Endo et al. for the purpose of inverse tone mapping [47].

E. Dynamic Scenes and Video

Thanks to the simultaneous capture of low- and high-ISO scanlines, our algorithm enables capture of highly dynamic HDR scenes without resorting to deghosting algorithms. Results for various scenes that are typically considered as challenging for HDR imaging are shown in Figure 13.

Finally, applying our algorithm individually to each dual-ISO frame of a video sequence produces plausible results as shown in Figure 14. A temporally coherent tone mapping operator can be used to prevent potential flickering in the tone mapped videos [74].

VII. Discussion

In this paper, we proposed a simple yet effective dual-ISO HDR reconstruction algorithm using DCNNs that can noticeably outperform the state-of-the-art techniques. We attribute the quality of our results to three primary factors:

- Teach the DCNN the minimum amount of information that it needs to learn leaving out the tasks such as demosaicking and white balancing,
- Train two network models for reconstructing low- and high-ISO images separately, instead of using a single network to reconstruct the HDR image,
- Simulate noise in a realistic manner by measuring noise not only across different types of scanlines but also across different intensity levels.

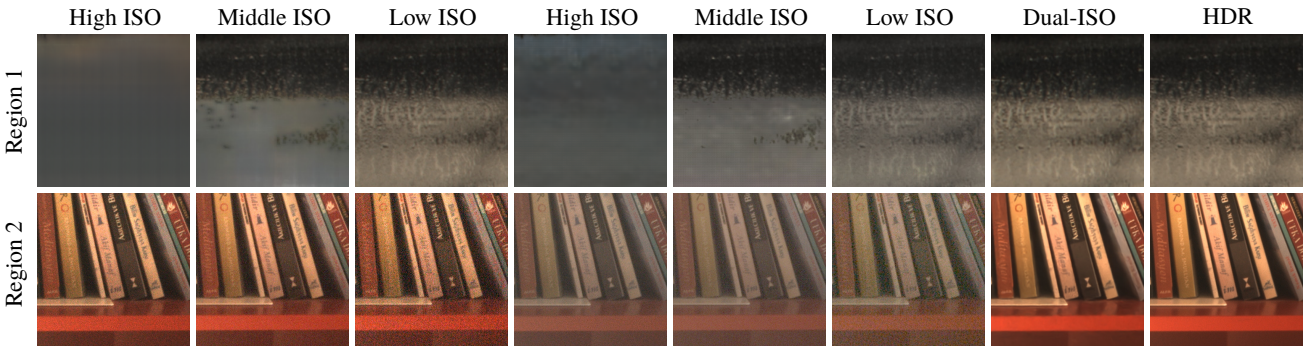
We observed that all three factors allow for a reduction in training data that is necessary to achieve high quality results. Below, we analyze two more issues related to dual-ISO systems in general and the generalizability of our networks for other dual-ISO combinations.

A. Why Dual-ISO Matters?

It may be argued that a dual-ISO exposure does not have benefits over a single low-ISO exposure due to the fact that high-ISO scanlines are simply obtained by amplification of the collected charges at the CCD/CMOS elements. It is true that unlike an exposure time bracketed sequence, in which a long exposure has *less* photon-shot noise with respect to a short exposure [59], a high-ISO exposure exhibits the same amount of photon-shot noise as the corresponding low-ISO exposure. This is because both the low- and high-ISO images are obtained by using the same exposure time.

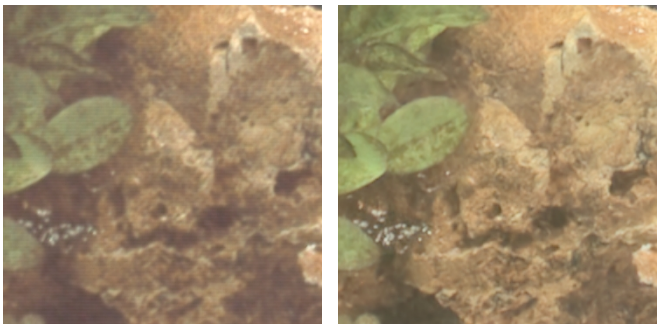
However, when *normalized* by their gains, a high-ISO image actually has less noise variance compared to its low-ISO counterpart. This can be understood from Equation 4 and observed in Figure 3. In this figure, the blue line represents the ratio of standard deviations (i.e., noise) of ISO 1600 and ISO 100 exposures. Although the two exposures are related by a gain ratio of 16, the ratio of the standard deviations remain below this number. This can be explained by the additive read noise term to be not directly proportional to the gains of the exposures due to the post-amplifier read noise shown in Equation 5.

As a consequence, an HDR image obtained from a dual-ISO exposure actually has a lower noise floor compared to a single low-ISO exposure. As dynamic range of an imager is typically defined as the ratio between the value that saturates its sensor and the value at the minimum acceptable noise level [75], it can be argued that a dual-ISO exposure improves dynamic range over a single low-ISO exposure. Furthermore, when training our networks we provide nearly noise-free ground-truth images as the target to be learned by these networks.



(a) Eilertsen et al. (b) Endo et al. (c) Ours (d) Ground-truth

Fig. 11: Comparison between iTMO algorithms and our method. We used ISO 100/400/1600 images as input to the iTMO algorithms of Eilertsen et al. [46] and Endo et al. [47]. Our algorithm takes a single dual-ISO image as input.



(a) Direct HDR reconstruction (b) Proposed approach

Fig. 12: Using a single network to reconstruct an HDR image produces lower quality results than the proposed approach. Note the deinterlacing artifacts on the left image as well as an overall loss of contrast.

Fig. 13: The results of our algorithm for various dynamic scenes. Original dual-ISO inputs are available in accompanying electronic materials.

(a) Frame 35 (b) Frame 71

Fig. 14: Our algorithm can also be used to create HDR videos from dual-ISO input frames. A video tone mapping operator can be used to eliminate flickering artifacts [74].

The networks learn to minimize the loss between a noise-free ground-truth and the corresponding noisy interlaced input. As a result, as input patches are transformed by the layers of the network, they tend to converge to the ground-truth which is nearly devoid of multiple noise sources, not merely the post amplifier read-noise. This includes Poisson-distributed photon-shot noise as well as it can also be reduced by averaging multiple frames [59].

In Figure 15, we illustrate this by comparing our results to a state-of-the-art demosaicking and denoising algorithm [76]. In this figure, (a) shows Gharbi et al.'s results applied on an ISO 100 exposure. In (b), we show the tone mapped HDR image created from a single dual-ISO image using our approach. The ground-truth (c) was obtained by frame averaging as explained in Section IV-A. All images were scaled to bring them to the same mean intensity in order to clearly illustrate the noise differences.

Given that a dual-ISO system can actually increase the dynamic range of an imager, the significance of this setup should be emphasized. A dual-ISO system does not contain motion artifacts between its different exposures, except if rolling shutter artifacts are significant enough to cause neighboring scanlines to capture different objects. Motion artifacts pose major challenges for exposure-time bracketed images. Furthermore, a dual-ISO system is also devoid of vignetting differences between images, which can be a problem for aperture-size bracketed sequences. However, a dual-ISO system requires addressing deinterlacing and denoising problems, for which an effective solution is proposed in this paper.

B. Are the Networks Generalizable?

For certain scenes, it may be desirable to capture dual-ISO images using combinations other than ISO 100 and ISO 1600. For example, if the dynamic range of the scene is not very high or a higher overlap is preferred between the two images, one may choose to use ISO 800 (or lower) for the high exposure. On the other hand, for higher dynamic range scenes it may be necessary to use ISO 3200 or higher. In order to understand whether our networks are generalizable to such conditions, we first performed an experiment using a synthetically generated scene followed by experiments using real photographs.

a) Gharbi et al. b) Ours c) Ground-truth

Fig. 15: The comparison of Gharbi et al.'s [76] result (a) with ours (b). The ground-truth is shown in (c). The PSNR values of (a) and (b) are 29.59 and 45.68 respectively.

To this end, we generated a dual-ISO image (in the same format as would be generated by a real camera) using constant values in its low and high scanlines. For instance, if the low scanlines with a gain of unity were set to a value of L, high scanlines were set to gL, with g representing the high gain. We varied the value of g from 8 to 32 to simulate high-ISO exposures from ISO 800 to ISO 3200. Noise was added to each type of scanline as explained in Section IV-B. The simulation results showed that, in all cases, our low and high networks were able to reconstruct the low and high exposures around the expected values of L and gL. This led us to conclude that the trained network models actually learned to *scale* the input scanlines rather than hardcoding the ratio of 16.

We put this to test using real dual-ISO images with ISO 100/800 and ISO 100/3200 combinations. In Figure 16, we show a sample result together with the ISO 100/1600 combination as a reference. It can be seen that although the ISO 100/1600 result has the highest PSNR, the other combinations also produce very high PSNR values that are visually indistinguishable from each other.

A similar situation exists when using cameras other than the one used to capture our training images. As an illustration, we show the reconstruction results for a dual-ISO image captured by a Canon 5D Mark III camera in Figure 17. In this figure, we compare our results with that of Hajisharif et al. [48] and Rodriguez et al. [51]. It can be argued that our results offer a better visual quality for this camera as well, although our network was trained using images from a Canon EOS 600d camera. It may, however, be possible that if a camera that has a vastly different noise characteristic is used, retraining the network models may be necessary.

VIII. Conclusions

In this paper, we proposed a deep-learning based dual-ISO reconstruction approach that can achieve high quality HDR reconstructions with PSNR values $40dB$ or more. We have shown that our results outperform the state-of-the-art in this field, both in terms of quality as well as efficiency. This improvement was made possible by a judicious training phase of a DCNN which can accurately learn to model the relationship between input and output patches for the tasks of denoising and deinterlacing. We also proposed an in-depth noise model analysis for dual-ISO exposures and showed why they can be used to improve dynamic range.

Currently, dual-ISO modulation is made possible by custom firmware such as Magic Lantern. It is hoped that the potential of this approach will render it available in a much broader range of consumer grade capture devices.

Acknowledgements

We are greatly thankful to Ana Serrano for explaining how to execute their code for dual-ISO images, Saghi Hajisharif for sharing their source code, Raquel Gil Rodriguez for sharing the output of their algorithm, Ronan Boitard for sharing their zonal brightness coherency video TMO, Wolfgang Heidrich and Felix Heide for help with FlexISP, and Min H. Kim and Inchang Choi for producing their algorithm's results for us.

References

[1] G. Ward, "Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures," *Journal of Graphics Tools*, vol. 8, no. 2, pp. 17–30, 2003.

[2] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes." *ACM Trans. Graph.*, vol. 31, no. 6, p. 203, 2012.

[3] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation," in *CVPR, 2013 IEEE Conf. on*, 2013.

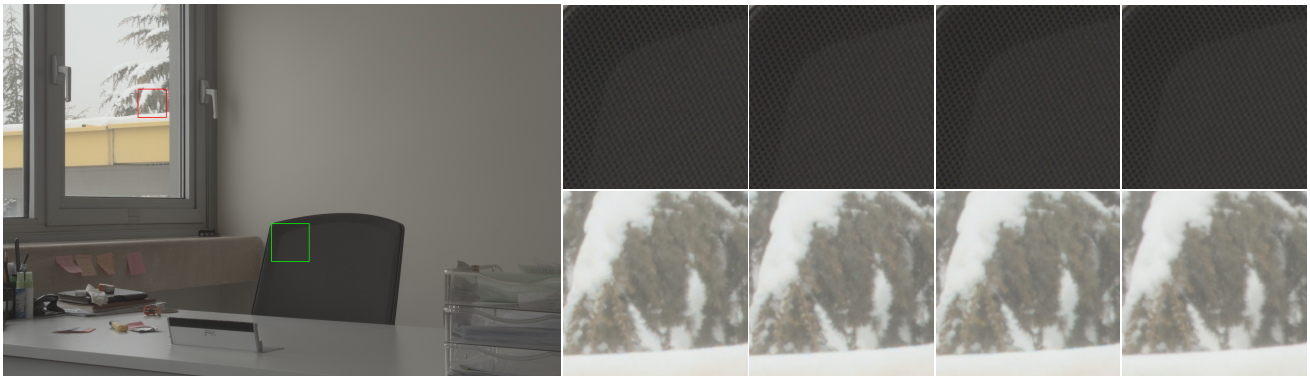
[4] N. K. Kalantari, E. Shechtman, C. Barnes, S. Darabi, D. B. Goldman, and P. Sen, "Patch-based high dynamic range video," *ACM Trans. Graph.*, vol. 32, no. 6, 2013.

[5] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph*, vol. 36, no. 4, p. 144, 2017.

[6] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. of ECCV*, 2018, pp. 117–132.

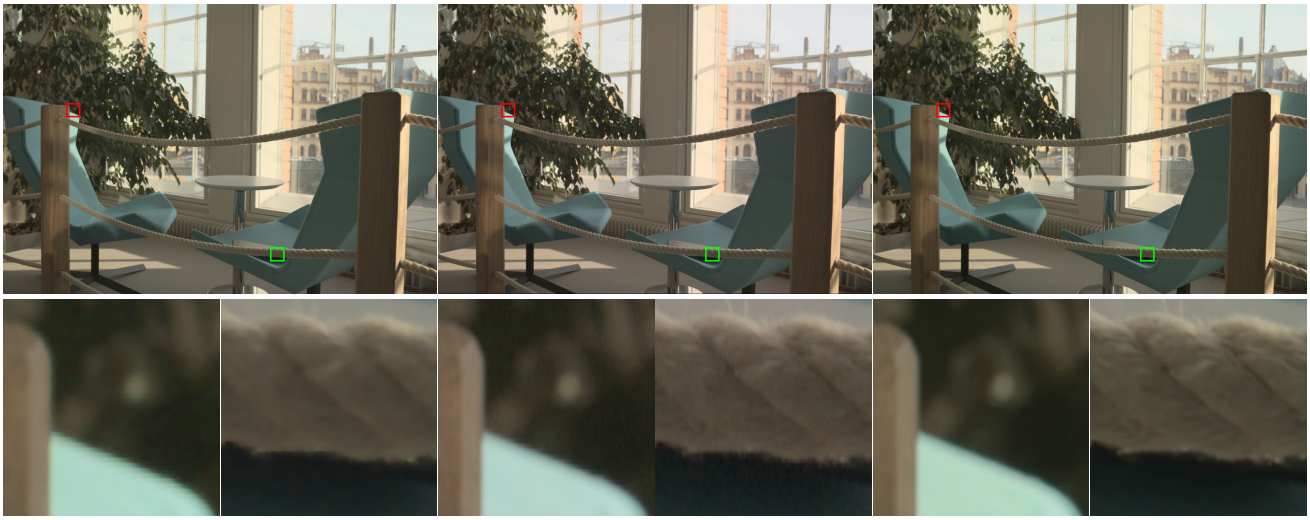
[7] Q. Yan, D. Gong, Q. Shi, A. v. d. Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. of the IEEE CVPR*, 2019, pp. 1751–1760.

[8] A. Srikantha and D. Sidibé, "Ghost detection and removal for high dynamic range images: Recent advances," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 650–662, 2012.



Ground-truth full resolution | ISO100 - ISO 800 | ISO100 - ISO 1600 | ISO100 - ISO 3200 | Ground-truth

Fig. 16: Comparison of different dual-ISO combinations. The PSNR values are measured using the ground-truths obtained from averaging a large number of full resolution images (Section IV-A) captured using the corresponding ISO values. The PSNRs are 43.93, 44.60 and 42.66 for ISO 100/ISO 800, ISO 100/ISO 1600 and ISO 100/ISO 3200 settings, respectively.



Hajisharif et al. | Rodriguez et al. | Ours

Fig. 17: A dual-ISO image captured with a Canon 5D Mark III camera was reconstructed using Hajisharif et al.'s [48], Rodriguez et al.'s [51] algorithms and our method. The insets show that the deinterlacing quality as well as the noise reduction achieved by our method are better than the comparisons.

[9] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: A survey and evaluation," *Computer Graphics Forum*, 2015.

[10] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "An objective deghosting quality metric for HDR images," in *Computer Graphics Forum*, vol. 35, no. 2. Wiley Online Library, 2016, pp. 139–152.

[11] K. Karadjuzović-Hadžiabdić, J. H. Telalović, and R. K. Mantiuk, "Assessment of multi-exposure HDR image deghosting methods," *Computers & Graphics*, vol. 63, pp. 1–17, 2017.

[12] M. Aggarwal and N. Ahuja, "Split aperture imaging for high dynamic range," *Int. Journal of Computer Vision*, vol. 58, no. 1, pp. 7–17, 2004.

[13] H. Wang, R. Raskar, and N. Ahuja, "High dynamic range video using split aperture camera," in *IEEE 6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras, Washington, DC*, 2005.

[14] M. Bätz, T. Richter, J.-U. Garbas, A. Papst, J. Seiler, and A. Kaup, "High dynamic range video reconstruction from a stereo camera setup," *Signal Processing: Image Comm.*, vol. 29, no. 2, pp. 191–202, 2014.

[15] A. Manakov, J. Restrepo, O. Klehm, R. Hegedus, E. Eisemann, H.-P. Seidel, and I. Ihrke, "A reconfigurable camera add-on for high dynamic range, multispectral, polarization, and light-field imaging," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 47–1, 2013.

[16] J. Froehlich, S. Grandinetti, B. Eberhardt, S. Walter, A. Schilling, and H. Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Digital Photography X*, vol. 9023. Intl. Soc. for Optics and Photonics, 2014, p. 90230X.

[17] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar, "Optical splitting trees for high-precision monocular imaging," *Comp. Graph. and Applications, IEEE*, vol. 27, no. 2, pp. 32–42, 2007.

[18] S. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *Computer Vision and Pattern Recognition, 2000. IEEE Conf. on*, vol. 1, 2000, pp. 472–479.

[19] A. Serrano, F. Heide, D. Gutierrez, G. Wetzstein, and B. Masia, "Convolutional sparse coding for high dynamic range imaging," in *Computer Graphics Forum*, vol. 35, no. 2, 2016, pp. 153–163.

[20] A. Chalmers, P. Campisi, P. Shirley, and I. G. Olaizola, *High dynamic range video: concepts, technologies and applications*. Academic Press, 2016.

[21] S. Mann and R. Picard, "Being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures," 1995.

[22] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *SIGGRAPH 97 Conference Proceedings*, August 1997, pp. 369–378.

[23] T. Mitsunaga and S. K. Nayar, "Radiometric self calibration," in *Proceedings of CVPR*, vol. 2, June 1999, pp. 374–380.

[24] M. A. Robertson, S. Borman, and R. L. Stevenson, "Estimation-theoretic approach to dynamic range enhancement using multiple exposures," *Journal of Electronic Imaging*, vol. 12, p. 2003, 1999.

[25] A. O. Akyüz and E. Reinhard, "Noise reduction in high dynamic range imaging," *Journal of Visual Communication and Image Representation*, vol. 18, no. 5, pp. 366–376, 2007.

[26] M. Granados, B. Ajdin, M. Wand, C. Theobalt, H.-P. Seidel, and H. Lensch, "Optimal HDR reconstruction with linear digital cameras," in *CVPR, 2010 IEEE Conference on*, 2010, pp. 215–222.

[27] O. Gallo, M. Tico, R. Manduchi, N. Gelfand, and K. Pulli, "Metering for exposure stacks," in *Computer Graphics Forum*, vol. 31, no. 2pt2. Wiley Online Library, 2012, pp. 479–488.

[28] S. W. Hasinoff, F. Durand, and W. T. Freeman, "Noise-optimal capture for high dynamic range photography," in *CVPR, IEEE Conf. on*, 2010, pp. 553–560.

[29] H. Zimmer, A. Bruhn, and J. Weickert, "Freehand HDR imaging of moving scenes with simultaneous resolution enhancement," *Computer Graphics Forum*, vol. 30, no. 2, pp. 405–414, 2011.

[30] M. Granados, K. I. Kim, J. Tompkin, and C. Theobalt, "Automatic noise modeling for ghost-free HDR reconstruction," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 201:1–201:10, Nov. 2013.

[31] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High dynamic range video," in *ACM Trans. Graph.*, vol. 22, no. 3, 2003, pp. 319–325.

[32] S. Mangiat and J. Gibson, "High dynamic range video with ghost removal," in *SPIE Optical Engineering Applications*. Intl. Soc. for Optics and Photonics, 2010.

[33] Y. Li, C. Lee, and V. Monga, "A MAP estimation framework for HDR video synthesis," in *IEEE Intl. Conf. on Image Proc.*, 2015.

[34] M. Schoberl, A. Belz, J. Seiler, S. Foessel, and A. Kaup, "High dynamic range video by spatially non-regular optical filtering," in *2012 19th IEEE International Conference on Image Processing*, 2012.

[35] S. Kavusi and A. El Gamal, "Quantitative study of high-dynamic-range image sensor architectures," in *Electronic Imaging 2004*. International Society for Optics and Photonics, 2004, pp. 264–275.

[36] Q. Zhou, S. Guo, G. Du, Y. Wang, and Y. Chang, "High-dynamic-range photodetecting scheme based on pept with a large output swing," *Electron Devices, IEEE Trans. on*, vol. 59, no. 5, pp. 1423–1429, 2012.

[37] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," *ACM Trans. Graph.*, vol. 30, no. 4, p. 41, 2011.

[38] K. Venkataraman, D. Lelescu, J. Duparré, A. McMahon, G. Molina, P. Chatterjee, R. Mullis, and S. Nayar, "Picam: An ultra-thin high performance monolithic camera array," *ACM Trans. Graph.*, vol. 32, no. 6, pp. 166:1–166:13, Nov. 2013.

[39] W.-J. Park, S.-W. Ji, S.-J. Kang, S.-W. Jung, and S.-J. Ko, "Stereo vision-based high dynamic range imaging using differently-exposed image pair," *Sensors*, vol. 17, no. 7, p. 1473, 2017.

[40] F. Banterle, P. Ledda, K. Debattista, and A. Chalmers, "Inverse tone mapping," in *Proc. of GRAPHITE '06*, 2006, pp. 349–356.

[41] B. Masia, S. Agustin, R. W. Fleming, O. Sorkine, and D. Gutierrez, "Evaluation of reverse tone mapping through varying exposure conditions," *ACM Trans. on Graph.*, vol. 28, no. 5, pp. 160:1–160:8, 2009.

[42] L. Wang, L.-Y. Wei, K. Zhou, B. Guo, and H.-Y. Shum, "High dynamic range image hallucination." in *Rendering Techniques*, 2007, pp. 321–326.

[43] A. O. Akyüz, R. Fleming, B. E. Riecke, E. Reinhard, and H. H. Bülthoff, "Do HDR displays support LDR content?: A psychophysical evaluation," *ACM Trans. on Graph.*, vol. 26, no. 3, p. 38, 2007.

[44] A. G. Rempel, M. Trentacoste, H. Seetzen, H. D. Young, W. Heidrich, L. Whitehead, and G. Ward, "Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs," in *ACM Trans. Graph.*, vol. 26, no. 3, 2007, p. 39.

[45] F. Banterle, P. Ledda, K. Debattista, M. Bloj, A. Artusi, and A. Chalmers, "A psychophysical evaluation of inverse tone mapping techniques," in *Computer Graphics Forum*, vol. 28, no. 1, 2009, pp. 13–25.

[46] G. Eilertsen, J. Kronander, G. Denes, R. K. Mantiuk, and J. Unger, "HDR image reconstruction from a single exposure using deep CNNs," *ACM Trans. on Graph.*, vol. 36, no. 6, p. 178, 2017.

[47] Y. Endo, Y. Kanamori, and J. Mitani, "Deep reverse tone mapping," *ACM Trans. on Graph.*, vol. 36, no. 6, Nov. 2017.

[48] S. Hajisharif, J. Kronander, and J. Unger, "Adaptive dualiso HDR reconstruction," *EURASIP Journal on Image and Video Processing*, vol. 2015, no. 1, p. 41, Dec 2015.

[49] a1ex, "Dynamic range improvement for some canon dslrs by alternating ISO during sensor readout," 2013, http://acoutts.com/a1ex/dual_iso.pdf.

[50] H. Takeda, S. Farsiu, and P. Milanfar, "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 349–366, Feb 2007.

[51] R. G. Rodríguez and M. Bertalmío, "High quality video in high dynamic range scenes from interlaced dual-iso footage," in *Digital Photography and Mobile Imaging*, 2016.

[52] I. Choi, S. H. Baek, and M. H. Kim, "Reconstructing interlaced high-dynamic-range video using joint learning," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5353–5366, Nov 2017.

[53] F. Heide, M. Steinberger, Y.-T. Tsai, M. Rouf, D. Pająk, D. Reddy, O. Gallo, J. Liu, W. Heidrich, K. Egiazarian, J. Kautz, and K. Pulli, "FlexISP: A flexible camera image processing framework," *ACM Trans. Graph.*, vol. 33, no. 6, 2014.

[54] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar, "Coded rolling shutter photography: Flexible space-time sampling," in *IEEE ICCP*, March 2010, pp. 1–8.

[55] H. Cho, S. J. Kim, and S. Lee, "Single-shot high dynamic range imaging using coded electronic shutter," *Computer Graphics Forum*, vol. 33, no. 7, pp. 329–338, Oct. 2014.

[56] V. G. An and C. Lee, "Single-shot high dynamic range imaging via deep convolutional neural network," in *2017 APSIPA ASC*, Dec 2017, pp. 1768–1772.

[57] S. W. Hasinoff, D. Sharlet, R. Geiss, A. Adams, J. T. Barron, F. Kainz, J. Chen, and M. Levoy, "Burst photography for high dynamic range and low-light imaging on mobile cameras," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 192:1–192:12, Nov. 2016.

[58] J. R. Janesick *et al.*, *Scientific charge-coupled devices*. SPIE press Bellingham, 2001, vol. 117.

[59] A. C. Bovik, *Handbook of image and video processing*. Academic Press, 2010.

[60] B. S. Reddy and B. N. Chatterji, "An FFT-based technique for translation, rotation, and scale-invariant image registration," *IEEE Transactions on Image Processing*, vol. 5, no. 8, pp. 1266–1271, Aug 1996.

[61] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *IEEE CVPR*, June 2012, pp. 2392–2399.

[62] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[63] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[65] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," *CoRR*, vol. abs/1707.02921, 2017.

[66] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[67] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *CoRR*, vol. abs/1511.07122, 2016.

[68] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. of the 13th Intl. Conf. on Artificial Intelligence and Statistics*, vol. 9. PMLR, May 2010, pp. 249–256.

[69] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew, "Color image processing pipeline," *IEEE Signal Processing Magazine*, vol. 22, no. 1, pp. 34–43, Jan 2005.

[70] H. S. Malvar, L. wei He, and R. Cutler, "High-quality linear interpolation for demosaicing of bayer-patterned color images," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, May 2004, pp. iii–485–8 vol.3.

[71] M. Narwaria, R. Mantiuk, M. P. Da Silva, and P. Le Callet, "Hdr-vdp-2.2: a calibrated method for objective quality prediction of high-dynamic range and standard images," *Journal of Electronic Imaging*, vol. 24, no. 1, p. 010501, 2015.

[72] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 267–276, Jul. 2002.

[73] T. O. Aydın, R. Mantiuk, and H.-P. Seidel, "Extending quality metrics to full luminance range images," in *Human Vision and Electronic Imaging XIII*, vol. 6806. SPIE, 2008, pp. 109 – 118.

[74] R. Boitard, R. Cozot, D. Thoreau, and K. Bouatouch, "Zonal brightness coherency for video tone mapping," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 229–246, 2014.

[75] B. Karr, K. Debattista, and A. Chalmers, "Calibrated measurement of imager dynamic range," in *High Dynamic Range Video*, A. Chalmers, P. Campisi, P. Shirley, and I. G. Olaizola, Eds. Academic Press, 2017, pp. 87 – 108.

[76] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 191:1–191:12, Nov. 2016.