HDR Video Reconstruction: A Coarse-to-fine Network and A Real-world Benchmark Dataset

Guanying Chen Chaofeng Chen Shi Guo Zhetong Liang Kwan-Yee K. Wong Lei Zhang
The University of Hong Kong DAMO Academy, Alibaba Group
The Hong Kong Polytechnic University

Abstract

High dynamic range (HDR) video reconstruction from sequences captured with alternating exposures is a very challenging problem. Existing methods often align low dynamic range (LDR) input sequence in the image space using optical flow, and then merge the aligned images to produce HDR output. However, accurate alignment and fusion in the image space are difficult due to the missing details in the over-exposed regions and noise in the under-exposed regions, resulting in unpleasing ghosting artifacts. To enable more accurate alignment and HDR fusion, we introduce a coarse-to-fine deep learning framework for HDR video reconstruction. Firstly, we perform coarse alignment and pixel blending in the image space to estimate the coarse HDR video. Secondly, we conduct more sophisticated alignment and temporal fusion in the feature space of the coarse HDR video to produce better reconstruction. Considering the fact that there is no publicly available dataset for quantitative and comprehensive evaluation of HDR video reconstruction methods, we collect such a benchmark dataset, which contains 97 *sequences of static scenes and* 184 *testing pairs of dynamic scenes. Extensive experiments show that our method outperforms previous state-of-the-art methods. Our dataset, code and model will be made publicly available.*

1. Introduction

Compared with low dynamic range (LDR) images, high dynamic range (HDR) images can better reflect the visual details of a scene in both bright and dark regions. Although significant progress has been made in HDR image reconstruction using multi-exposure images [23, 58, 60], the more challenging problem of HDR video reconstruction is still less explored. Different from HDR image reconstruction, HDR video reconstruction has to recover the HDR for every input frame (see Fig. 1), but not just for a single reference frame (*e.g.*, the middle exposure image). Existing successful HDR video reconstruction techniques often rely on costly and specialized hardware (*e.g.*, scanline exposure/ISO, or internal/external beam splitter) [56, 31, 63], which hinders their wider applications among ordinary consumers. A promising direction for low-cost HDR video reconstruction is to utilize video sequences captured with alternating exposures (*e.g.*, videos with a periodic exposure of {EV-3, EV+3, EV-3, . . . }). This is practical as many off-the-shelf cameras can alternate exposures during recording.

Figure 1. HDR video reconstruction from sequences captured with three alternating exposures. Row 1 shows four input LDR frames. Rows 2–3 are the reconstructed (tonemapped) HDR frames.

Conventional reconstruction pipeline along this direction often consists of two steps [26]. In the first step, neighboring frames with different exposures are aligned to the current frame using optical flow. In the second step, the aligned images are fused to produce the HDR image. However, accurate alignment and fusion are difficult to achieve for LDR images with different exposures as there are saturated pixel values in the over-exposed regions, and noise in the under-exposed regions. Recently, Kalantari and Ramamoorthi [24] proposed to estimate the optical flow with a deep neural network, and used another network to predict the fusion weights for merging the aligned images. Although improved results over traditional methods [25, 39, 26, 33] have been achieved, their method still

relies on the accuracy of optical flow alignment and pixel blending, and suffers from ghosting artifacts in regions with large motion (see the second row of Fig. 1). It remains a challenging problem to reconstruct ghost-free HDR videos from sequences with alternating exposures.

Recently, deformable convolution [8] has been successfully applied to feature alignment in video super-resolution [57, 55]. However, they are not tailored for LDR images with different exposures. Motivated by the observation that accurate image alignment between LDR images with different exposures is difficult, and the success of deformable feature alignment for videos with constant exposure, we introduce a two-stage coarse-to-fine framework for this problem. The first stage, denoted as *CoarseNet*, aligns images using optical flow in the image space and blends the aligned images to reconstruct the coarse HDR video. This stage can recover/remove a large part of missing details/noise from the input LDR images, but there exist some artifacts in regions with large motion. The second stage, denoted as *RefineNet*, performs more sophisticated alignment and fusion in the feature space of the coarse HDR video using deformable convolution [8] and temporal attention. Such a two-stage approach avoids the need of estimating highly accurate optical flow from images with different exposures, and therefore reduces the learning difficulty and removes ghosting artifacts in the final results.

As there is no publicly available real-world video dataset with ground-truth HDR for evaluation, comprehensive comparisons among different methods are difficult to achieve. To alleviate this problem, we create a real-world dataset containing both static and dynamic scenes as a benchmark for quantitative and qualitative evaluation.

In summary, the key contributions of this paper are as follows:

- We propose a two-stage framework, which first performs image alignment and HDR fusion in the image space and then in feature space, for HDR video reconstruction from sequences with alternating exposures.
- We create a real-world video dataset captured with alternating exposures as a benchmark to enable quantitative evaluation for this problem.
- Our method achieves state-of-the-art results on both synthetic and real-world datasets.

2. Related Work

HDR image reconstruction Merging multi-exposure LDR images is the most common way to reconstruct HDR images [9, 40]. To handle dynamic scenes, image alignment is employed to reduce the ghosting artifacts [52, 20, 49, 37]. Recent methods apply deep neural networks to merge multi-exposure images [23, 6, 58, 60, 61, 48]. However, these methods rely on a fixed reference exposure (*e.g.*, the middle exposure) and cannot be directly applied to reconstruct HDR videos from sequences with alternating exposures. Burst denoising technique [36, 18, 34] can also be applied to produce HDR images by denoising the low-exposure images. However, this technique cannot make use of the cleaner details that exist in high-exposure images and have difficulty in handling extremely dark scenes.

There are methods for HDR reconstruction from a single LDR image. Traditional methods expand the dynamic range of the LDR images by applying image processing operations (*e.g.*, function mapping, and filtering) [1, 2, 3, 4, 21, 30]. These methods generally cannot recover the missing details in the clipped regions. Recent methods proposed to adopt CNNs for single image reconstruction [10, 11, 32, 62, 45, 42, 35, 51]. However, these methods focus on hallucinating the saturated regions and cannot deal with the noise in the dark regions of a low-exposure image.

Recently, Kim *et al*. [27, 28] proposed to tackle the problem of joint super-resolution and inverse tone-mapping. Instead of reconstructing the linear luminance image like previous HDR reconstruction methods, their goal was to convert a standard dynamic range (SDR) image to HDR display format (*i.e.*, from BT.709 to BT.2020).

HDR video reconstruction Many existing HDR video reconstruction methods rely on specialized hardware. For example, per-pixel exposure [47], scanline exposure/ISO [16, 19, 7], internal [56, 31] or external [43] beam splitter that can split the light to different sensors, modulo camera [63], and neuromorphic camera [17]. The requirement of specialized hardware limits the widespread application of these methods. Recent methods also explore the problem of joint optimization of the optical encoder and CNN-based decoder for HDR imaging [44, 54].

There are works for HDR video reconstruction from sequences with alternating exposures. Kang *et al*. [26] introduced the first algorithm of this approach by first aligning neighboring frames to the reference frame using optical flow, and then merging the aligned images to an HDR image. Mangiat and Gibson improved this method by a block-based motion estimation and refinement stage [38, 39]. Kalantari *et al*. [25] introduced a patch-based optimization method that synthesizes the missing exposures at each image and then reconstructs the final HDR image. Gryaditskaya *et al*. [15] improved [25] by introducing an adaptive metering algorithm that can adjust the exposures to reduce artifacts caused by motion. Li *et al*. [33] formulated this problem as a maximum a posteriori estimation. Recently, Kalantari and Ramamoorthi [24] introduced an end-to-end deep learning framework that contains a flow network for alignment and a weight network for pixel blending in image space. Different from [24], our coarse-to-fine network performs alignment and fusion sequentially in the image space and feature space for better reconstruction.

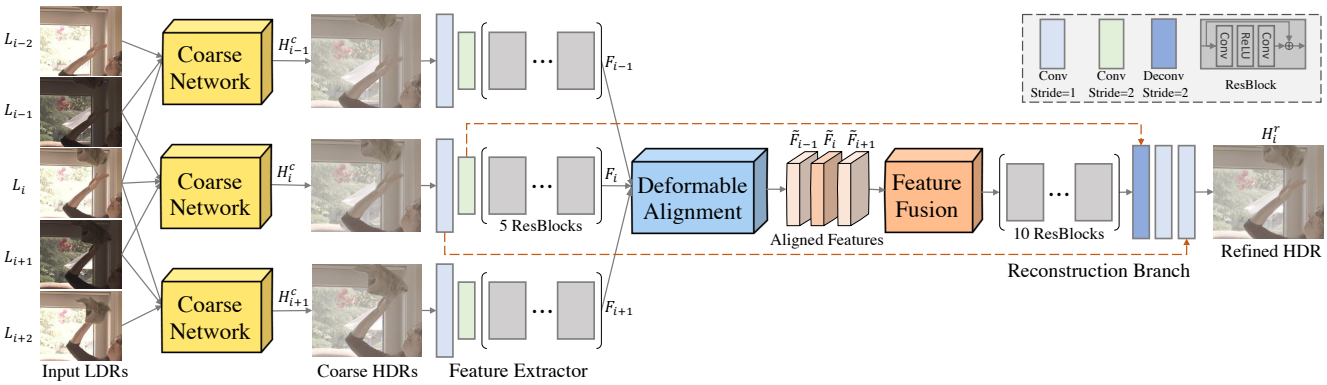


Figure 2. Network architecture of the proposed coarse-to-fine framework for videos captured with two alternating exposures.

3. The Proposed Coarse-to-fine Framework

3.1. Overview

Given an input LDR video $\{\tilde{L}_i | i = 1, \ldots, n\}$ captured with alternating exposures $\{t_i | i = 1, \ldots, n\}$[1], our goal is to reconstruct the corresponding HDR video $\{H_i | i = 1, \ldots, n\}$, as shown in Fig. 1.

Preprocessing Following previous methods [25, 33, 24], we assume the camera response function (CRF) [14] $\mathcal{F}$ of the original input images $\tilde{L}_i$ is known. In practice, the CRF of a camera can be robustly estimated using a linear method [9]. As in [24], we replace the CRF of the input images with a fixed gamma curve as $L_i = (\mathcal{F}^{-1}(\tilde{L}_i))^{1/\gamma}$, where $\gamma = 2.2$. This can unify input videos captured under different cameras or configurations. Global alignment is then performed using a similarity transformation to compensate camera motions among neighboring frames.

Pipeline Due to the existence of noise and missing details, accurate image alignment between images with different exposures is difficult. To overcome these challenges, we introduce a two-stage framework for more accurate image alignment and fusion (see Fig. 2). For simplicity, we illustrate our method for handling videos captured with *two* alternating exposures in this paper, and describe how to extend our method for handling *three* exposures in the supplementary material.

The first stage, named *CoarseNet*, aligns images using optical flow and performs HDR fusion in the image space. It takes three frames as input and estimates a 3-channel HDR image for the reference (*i.e.*, center) frame. This stage can recover/remove a large part of the missing details/noise for the reference LDR image. Given five consecutive LDR frames $\{L_i | i = i-2, \ldots, i+2\}$ with two alternating exposures, our CoarseNet can sequentially reconstruct the coarse HDR images for the middle three frames (*i.e.*, H^c_{i-1}, H^c_i, and H^c_{i+1}). The second stage, named *RefineNet*, takes these three coarse HDR images as input to produce a better HDR reconstruction for the reference frame (*i.e.*, H^r_i). It performs a more sophisticated alignment using deformable convolution and temporal fusion in the feature space.

[1]For example, the exposure can be alternated periodically in the order of {EV-3, EV+3, EV-3, . . . } or {EV-2, EV+0, EV+2, EV-2, . . . }.

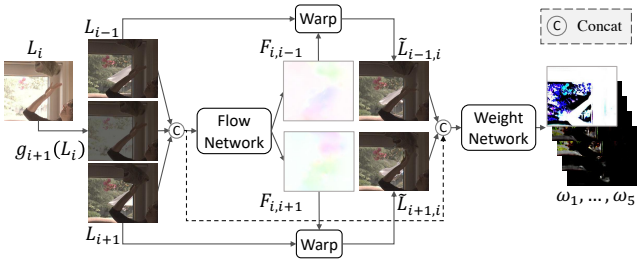


Figure 3. Overview of the CoarseNet.

3.2. Coarse Reconstruction in the Image Space

The CoarseNet follows the design of [24], containing an optical flow estimation network, named *flow network*, and a blending weight estimation network, named *weight network* (see Fig. 3). It first warps two neighboring frames to the center frame using optical flows, and then reconstructs the HDR image by blending the aligned images. The network details can be found in the supplementary materials.

Optical flow for image alignment Since the reference frame L_i and neighboring frames (L_{i-1}, L_{i+1}) have different exposures, we adjust the exposure of the reference frame to be the same as the neighboring frame. The reference LDR image is first converted to the linear radiance domain using its exposure t_i:

$$I_i = h(L_i, t_i) = L_i^{\gamma} / t_i. \tag{1}$$

It is then converted to the LDR domain using the neighboring exposure t_{i+1} as $g_{i+1}(I_i) = \text{clip}[(I_i t_{i+1})^{1/\gamma}]$, where the clip function clips the values to the range of $[0, 1]$.

Traditional flow estimation method takes two images as input and estimates a flow map. However, in our problem,

the center frame has a different exposure as the neighboring frames, such that the adjusted reference frame $g_{i+1}(I_i)$ often contains missing contents or noise. We therefore take three consecutive images, *i.e.*, $\{L_{i-1}, g_{i+1}(I_i), L_{i+1}\}$, as input and estimate two flow maps $\{F_{i,i-1}, F_{i,i+1}\}$) as in [24]. Two neighboring frames can then be aligned to the reference frame ($\hat{L}_{i-1,i}, \hat{L}_{i+1,i}$) using backward warping with bilinear sampling [22].

Pixel-blending for HDR reconstruction The HDR image can be computed as a weighted average of the pixels in the aligned images [9, 23]. Note that the two original neighboring frames are also taken into account for pixel blending, as it is reported to be helpful for reducing artifacts in the background regions [24].

Specifically, the input image number for weight network is 5, *i.e.*, $\{L_{i-1}, \hat{L}_{i-1,i}, L_i, \hat{L}_{i+1,i}, L_{i+1}\}$. We provide these five images as input in both the LDR and linear radiance domain, resulting in a stack of 10 images as inputs. The network predicts five per-pixel weighted maps, *i.e.*, $\{\omega_k | k = 1, \ldots, 5\}$. The coarse HDR at frame i can then be reconstructed as the weighted average of five input images in the linear radiance domain:

$$H_i^c = \frac{\omega_1 I_{i-1} + \omega_2 \hat{I}_{i-1,i} + \omega_3 I_i + \omega_4 \hat{I}_{i+1,i} + \omega_5 I_{i+1}}{\sum_{k=1}^{5} \omega_k}. \tag{2}$$

Similar to [24], we adopt an encoder-decoder architecture to estimate the blending weights.

Loss function As HDR images are typically displayed after tonemapping, we compute the loss in the tonemapped HDR space. Following [23, 58, 60, 24], we adopt the differentiable μ-law function:

$$T_i^c = \frac{\log(1 + \mu H_i^c)}{\log(1 + \mu)}, \tag{3}$$

where T_i^c is the tonemapped HDR image, and μ is a parameter controlling the compression level and is set to 5000. We train CoarseNet with the L1 loss $\mathcal{L}^c = \| T_i^c - \tilde{T}_i \|_1$, where $\tilde{T}_i$ is the ground-truth tonemapped HDR image. Since both the flow network and weight network are differentiable, the CoarseNet can be trained end-to-end.

3.3. HDR Refinement in the Feature Space

Taking three coarse HDR images (*i.e.*, H_{i-1}^c, H_i^c, and H_{i+1}^c) estimated by the CoarseNet as input, the RefineNet performs alignment and fusion in the feature space to produce better HDR reconstruction for the center frame, as the problem of missing contents or noise has been largely solved in the first stage (see the right part of Fig. 2).

Our RefineNet first extracts a 64-channel feature for each input (*i.e.*, F_{i-1}, F_i, and F_{i+1}) using a share-weight feature extractor. Features of the neighboring frames are then aligned to the center frame using a deformable alignment module [8, 57]. The aligned features are fused using a temporal attention fusion module for the final HDR reconstruction.

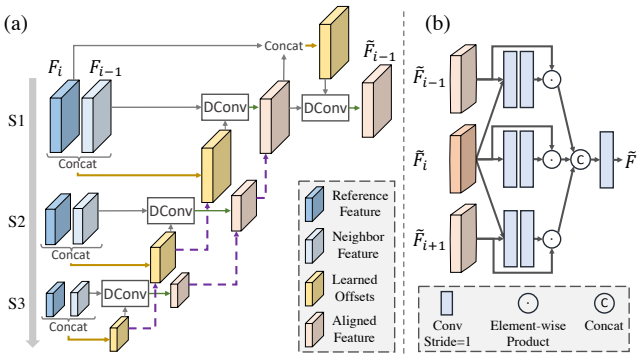


Figure 4. Structure of the (a) deformable alignment module and (b) temporal attention fusion module.

Deformable feature alignment Deformable convolution [8] has recently been successfully applied to feature alignment for the problem of video super-resolution (*e.g.*, EDVR [57] and TDAN [55]). The core idea of deformable feature alignment is as follows. Given two features (*e.g.*, F_{i-1} and F_i) as input, an offset prediction module (can be general convolutional layers) predicts an offset:

$$\Delta p_{i-1} = f([F_{i-1}, F_i]). \tag{4}$$

With the learned offset, the neighboring feature F_{i-1} can be sampled and aligned to the reference frame F_i using deformable convolution [8]:

$$\tilde{F}_{i-1} = \text{DConv}(F_{i-1}, \Delta p_{i-1}). \tag{5}$$

We adopt the pyramid, cascading and deformable (PCD) alignment module [57], which performs deformable alignment in three pyramid levels, as our feature alignment module (see Fig. 4 (a)). This alignment process is implicitly learned to optimize the final HDR reconstruction.

Multi-feature fusion Given the aligned features ($\tilde{F}_{i-1}$, $\tilde{F}_i$, and $\tilde{F}_{i+1}$), we propose a temporal attention fusion module for suppressing the misaligned features and merging complementary information for more accurate HDR reconstruction (see Fig. 4 (b)). Each feature is concatenated with the reference feature as the input for two convolutional layers to estimate an attention map that has the same size as the feature. Each feature is then weighted by their corresponding attention map. Last, three attended features are concatenated and fused using a convolutional layer.

HDR reconstruction The reconstruction branch takes the fused feature as input and regresses the HDR image (H_i^r). Two skip connections are added to concatenate encoder features of the *reference frame* to decoder features that have the same dimensions.

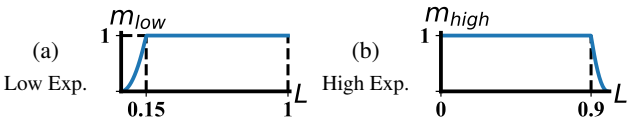


Figure 5. Weight curves for computing the well-exposed regions for (a) low- and (b) high-exposure reference image. L is the pixel value of the reference LDR image.

Note that our RefineNet aims to refine the results of CoarseNet in the not well-exposed regions. For a low-exposure image, we empirically define that regions with LDR pixel values smaller than 0.15 are not well-exposed, while for a high-exposure image, regions with pixel values larger than 0.9 are not well-exposed [25]. The final predicted HDR is then computed as

$$H_i = M_i \odot H_i^c + (1 - M_i) \odot H_i^r, \tag{6}$$

where M_i is a mask indicating the well-exposed regions of the reference frame i, and $\odot$ is the element-wise product. Figure 5 shows how M_i is computed for low- and high-exposure reference image. For example, the well-exposed mask of a low-exposure reference image L_i is computed as

$$M_i = \begin{cases} 1, & \text{if } L_i >= 0.15 \\ (L_i/0.15)^2, & \text{if } L_i < 0.15 \end{cases} \tag{7}$$

Loss function We adopt L1 loss and perceptual loss to compute the loss for RefineNet as $\mathcal{L}^r = \mathcal{L}^r_{l1} + \mathcal{L}^r_{\text{perc}}$. The L1 loss is defined as

$$\mathcal{L}^r_{l1} = \| T_i - \tilde{T}_i \|_1 \; / \; \| 1 - M_i \|_1, \tag{8}$$

where T_i is the tonemapped image of H_i. The loss is normalized by the number of not well-exposed pixels. The perceptual loss is defined as $\mathcal{L}^r_{\text{perc}} = \sum_k \| \phi_k(T_i) - \phi_k(\tilde{T}_i) \|_1$, where $\phi_k(\cdot)$ extracts image features from the k^{th} layer of VGG16 network [53]. We use three layers {*relu1_2*, *relu2_2*, *relu3_3*} to compute the loss.

4. Real-world Benchmark Dataset

In this section, we introduce a real-world benchmark dataset for qualitative and quantitative evaluation.

Existing real-world video dataset Currently, there is no benchmark dataset with ground-truth HDR for this problem. The only public real-world dataset is the *Kalantari13* dataset [25], which consists of 9 videos for dynamic scenes in RGB image format. However, due to the lack of ground-truth HDR, previous works can only evaluate their methods qualitatively on this dataset. In addition, this dataset is too small to be used for possible semi-supervised or unsupervised learning in the future.

Table 1. Comparison between our dataset and the *Kalantari13* dataset [25]. Frame number shows the image number. 2-Exp and 3-Exp indicate videos with two and three exposures, respectively.

		Static Scenes w/ GT $6-9$ frames		Dynamic Scenes w/ GT $5-7$ frames		Dynamic Scenes w/o GT $50-200$ frames	
Data	Size	2-Exp	3-Exp	2-Exp	3-Exp	2-Exp	3-Exp
[25]	1280×720	-	-	-	-	5	4
Ours	4096×2168	49	48	76	108	37	38

Dataset overview To facilitate a more comprehensive evaluation on real data, we captured a real-world dataset and generated reliable ground truth HDR for evaluation. We used an off-the-shelf Basler acA4096-30uc camera for capturing videos with alternating exposures (*i.e.*, two and three exposures) in a variety of scenes, including indoor, outdoor, daytime, and nighttime scenes.

Three different types of video data are captured, namely, *static scenes with GT* ($\mathcal{D}_s^{gt}$), *dynamic scenes with GT* ($\mathcal{D}_d^{gt}$), and *dynamic scenes without GT* ($\mathcal{D}_d$).[2] Table 1 compares the statistics between our dataset and *Kalantari13* dataset.

Static scenes with GT For static scenes, we captured 49 two-exposure and 48 three-exposure sequences, each with $15-20$ frames. The ground-truth HDR frames for static scenes were generated by merging multi-exposure images [9]. We first averaged images having the same exposure to reduce noise, and then merged multi-exposure images using a weighting function similar to [23]. For each scene, we will release $6-9$ captured frames and the generated HDR frame.

Dynamic scenes with GT Generating per-frame ground-truth HDR for dynamic videos is very challenging. Following the strategy used for capturing dynamic HDR image [23], we propose to create image pairs consisting of input LDR frames and the HDR of the center frame. We considered static environment and used a human subject to simulate motion in videos.

For each scene, we first asked the subject to stay still for $1-2$ seconds, where we can find 2 consecutive still frames (or 3 frames for three-exposure) without motions for generating the HDR image for this timestamp. We then asked the subject to move back-and-forth (*e.g.*, waving hands or walking). We selected an image sequence whose center frame was the static frame, and arranged this sequence to be the proper LDRs-HDR pairs (see Fig. 6 for an example). For each reference frame with GT HDR, we also created a pair with a larger motion by sampling the neighboring frames in a frame interval of 2, which doubles the number of pairs. In total, we created 76 and 108 pairs for the case of two-exposure (5 input frames) and three-exposure (7 input frames), respectively.

[2] GT is short for the ground-truth HDR.

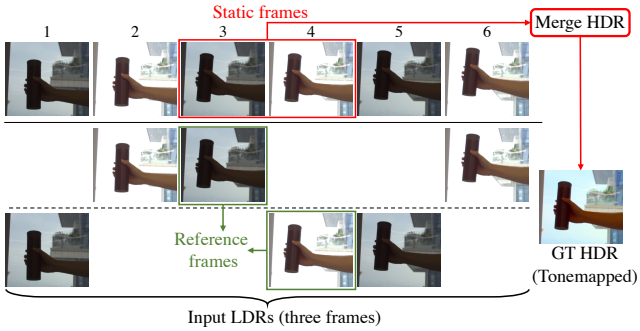


Figure 6. Illustration of generating the LDRs-HDR pairs for a two-exposure scene (3 frames). Row 1 shows the selected image sequence. Rows 2 and 3 are two sample pairs with low-exposure and high-exposure reference frames, respectively.

Dynamic scenes without GT We captured a larger scale dataset containing uncontrolled dynamic scenes for qualitative evaluation. Specifically, we captured 37 two-exposure and 38 three-exposure sequences, each contains around 100 frames. This dataset can also be used for semi-supervised or unsupervised training in the future.

Data processing We saved the raw data of the captured videos and performed demosaicing, white balancing, color correction, and gamma compression ($\gamma = 2.2$) to convert the raw data to RGB data using the recorded metadata. In this paper, we rescaled the images to 1536×813 for evaluation. Both the captured raw data and processed images will be released.

5. Experiments

In this section, we conduct experiments on synthetic and real-world datasets to verify the effectiveness of the proposed method. We compared our methods with Kalantari13 [25], Kalantari19 [24], and Yan19 [60]. Kalantari13 [25] is an optimization-based method and we used the publicly available code for testing. Note that Yan19 [60] is a state-of-the-art method for multi-exposure HDR image reconstruction, and we adapted it for video reconstruction by changing the network input. We re-implemented [24, 60] and trained them using the same dataset as our method.

We evaluated the estimated HDR in terms of PSNR (in the μ-law tonemapped domain), HDR-VDP-2 [41], and HDR-VQM [46]. HDR-VQM is designed for evaluating the quality of HDR videos. All visual results in the experiment are tonemapped using Reinhard *et al.*'s method [50] following [24, 25, 26]. In addition, a user study [5] (*i.e.*, pair comparison test) was conducted.

5.1. Training Datasets and Details

Synthetic training dataset Since there is no publicly available real video dataset with alternating exposures and their ground-truth HDR, we resort to synthetic data for training. Following [24], we selected 21 HDR videos [12, 31] to synthesize the training dataset. Since the size of the HDR video dataset is limited, we also adopted the high-quality Vimeo-90K dataset [59] to be the source videos. Please refer to our supplementary material for more details.

Data augmentation As the training data was generated from clean HDR videos, the resulting input sequences lack noise in the low-exposure images. To close this gap, we randomly added zero-mean Gaussian noise ($\sigma = 10^{-3}$) in the linear domain of the inputs. We also perturbed the tone of the reference image using a gamma function ($\gamma = \exp(d), d \in [-0.7, 0.7]$) to simulate the possibly inaccurate CRF [24, 13]. Random horizontal/vertical flipping and rotation were applied. Patches of size 256×256 were cropped out to be the network input.

Implementation details We trained our method using Adam optimizer [29] with default parameters. We first trained the CoarseNet with 10 epochs using a batch size of 16, and then trained the RefineNet with 15 epochs using a batch size of 8. The learning rate were initially set to 0.0001 and halved every 5 epochs for both networks . We then end-to-end finetuned the whole network for 2 epochs using a learning rate of 0.00002.

Table 2. Averaged results on synthetic dataset.

Method	2-Exposure			3-Exposure		
	PSNR	HDR-VDP2	HDR-VQM	PSNR	HDR-VDP2	HDR-VQM
Kalantari13 [25]	37.53	59.07	84.51	30.36	56.56	65.90
Yan19 [60]	39.05	70.61	71.27	36.28	65.47	72.20
Kalantari19 [24]	37.48	70.67	84.57	36.27	65.51	72.58
Ours	**40.34**	**71.79**	**85.71**	**37.04**	**66.44**	**73.38**

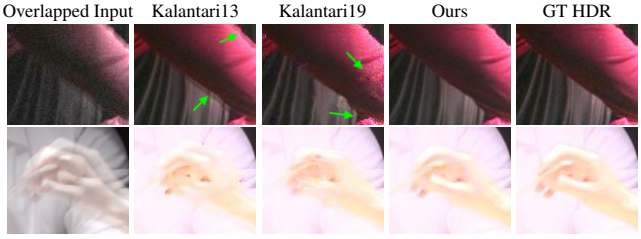


Figure 7. Visual results on the synthetic dataset.

5.2. Evaluation on Synthetic Dataset

We first evaluated our method on a synthetic dataset generated using two HDR videos (*i.e.*, POKER FULLSHOT and CAROUSEL FIREWORKS) [12], which are not used for training. Each video contains 60 frames and has a resolution of 1920×1080. Random Gaussian noise was added on the low-exposure images. Table 2 clearly shows that our method outperforms previous methods in all metrics on the this dataset. Figure 7 visualizes that our method can effectively remove the noise (top row) and ghosting artifacts (bottom row) in the reconstructed HDR.

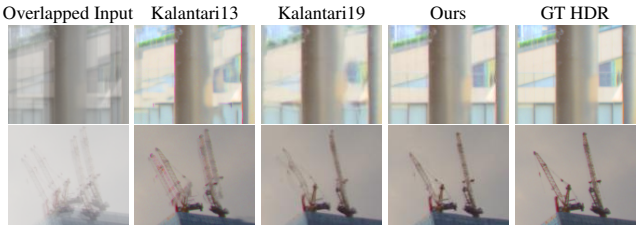
Table 3. Quantitative results on the introduced real dataset. The averaged results for each exposure and all exposures are shown. **Red** text indicates the best and **blue** text indicates the second best result, respectively.

(a) Results on *static scenes with GT* ($\mathcal{D}_s^{gt}$) augmented with random global motion.

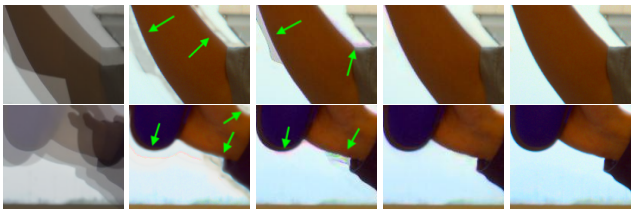
	2-Exposure							3-Exposure									
	Low-Exposure		High-Exposure		All-Exposure			Low-Exposure		Middle-Exposure		High-Exposure		All-Exposure			
Method	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	HDR-VQM	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	HDR-VQM
Kalantari13 [25]	**40.00**	73.70	**40.04**	**70.08**	**40.02**	71.89	**76.22**	**39.61**	73.24	**39.67**	**73.24**	**40.01**	67.90	**39.77**	70.37	79.55
Yan19 [60]	34.54	80.22	39.25	65.96	36.90	73.09	65.33	36.51	77.78	37.45	69.79	39.02	64.57	37.66	70.71	70.13
Kalantari19 [24]	39.79	**81.02**	39.96	67.25	39.88	**74.13**	73.84	39.48	**78.13**	38.43	70.08	39.60	**67.94**	39.17	**72.05**	**80.70**
Ours	**41.95**	**81.03**	**40.41**	**71.27**	**41.18**	**76.15**	**78.84**	**40.00**	**78.66**	**39.27**	**73.10**	**39.99**	**69.99**	**39.75**	**73.92**	**82.87**

(b) Results on *dynamic scenes with GT* ($\mathcal{D}_d^{gt}$).

	2-Exposure							3-Exposure									
	Low-Exposure		High-Exposure		All-Exposure			Low-Exposure		Middle-Exposure		High-Exposure		All-Exposure			
Method	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	HDR-VQM	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2	HDR-VQM
Kalantari13 [25]	37.73	74.05	45.71	66.67	41.72	70.36	85.33	37.53	72.03	36.38	65.37	34.73	62.24	36.21	66.55	84.43
Yan19 [60]	36.41	85.68	**49.89**	**69.90**	43.15	77.79	78.92	36.43	77.74	39.80	**67.88**	**43.03**	**64.74**	39.75	**70.12**	87.93
Kalantari19 [24]	**39.94**	**86.77**	49.49	69.04	**44.72**	**77.91**	**87.16**	**38.34**	**78.04**	**41.21**	66.07	42.66	64.01	**40.74**	69.37	**89.36**
Ours	**40.83**	**86.84**	**50.10**	**71.33**	**45.46**	**79.09**	**87.40**	**38.77**	**78.11**	**41.47**	**68.49**	**43.24**	**65.08**	**41.16**	**70.56**	**89.56**



(a) Results on *static scenes* augmented with random global motion.



(b) Results on *dynamic scenes with GT*.

Figure 8. Visual results on the captured dataset. For each dataset, row 1 is for two-exposure scene and row 2 is for three-exposure.

5.3. Evaluation on Real-world Dataset

To validate the generalization ability of our method on real data, we then evaluated the proposed method on the introduced real-world dataset and *Kalantari13* dataset [25].

Evaluation on static scenes We evaluated our method on $\mathcal{D}_s^{gt}$ augmented with random global motions (*i.e.*, random translation for each frame in the range of $[0, 5]$ pixels). We did not pre-align the input frames for all methods to investigate their robustness against input with inaccurate global alignment. Table 3 (a) shows that our method achieves the best results for two-exposure scenes and the most robust results for three-exposure scenes. Although Kalantari13 [25] shows slightly better averaged PSNR values for three-exposure scenes (*i.e.*, 39.77 vs. 39.75), it suffers from the ghosting artifacts for over-exposed regions (see Fig. 8 (a)).

Evaluation on dynamic scenes Table 3 (b) summarizes the results on $\mathcal{D}_d^{gt}$, where our method performs the best in all metrics. Compared with our method, the performance of Kalantari13 [25] drops quickly for dynamic scenes, as this dataset contains the more challenging local motions. Figure 8 (b) shows that the results of methods performing alignment and fusion in the image space [25, 24] produce unpleasing artifacts around the motion boundaries. In contrast, our two-stage coarse-to-fine framework enables more accurate alignment and fusion, and is therefore robust to regions with large motion and produces ghost-free reconstructions for scenes with two and three exposures.

Evaluation on *Kalantari13* dataset We then evaluated our method on *Kalantari13* dataset. Note that the result of Kalantari19 [24] for this dataset is provided by the authors. Figure 9 compares the results for three consecutive frames from THROWING TOWEL 2EXP scene, where our method achieves significantly better visual results. For a high-exposure reference frame, our method can recover the fine details of the over-exposed regions without introducing artifacts (see rows 1 and 3). In comparison, methods based on optical flow alignment and image blending [25, 24] suffers from artifacts for the over-exposed regions. For a low-exposure reference frame, compared with Kalantari13 [25], our method can remove the noise and preserve the structure for the dark regions (see row 2).

User study We also conducted a user study on the dynamic scene dataset (3-Exp) to further demonstrate the visual quality of our results (see Fig. 10). 33 participants were invited to give preference on 36 pairs of image. Note that the GT HDR was also shown for reference. Overall, 76% and 78% of the users preferred results of our

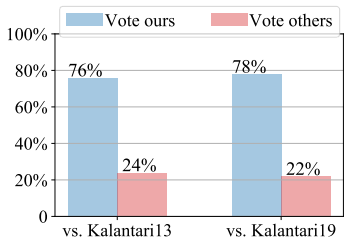


Figure 10. User study results.

Figure 9. Visual comparison on THROWING TOWEL 2EXP scene from *Kalantari13* dataset.

Table 4. Model parameter and runtime for producing an HDR frame of different resolutions.

Method	# Parameter	2-Exposure		3-Exposure	
		1280 × 720	1920 × 1080	1280 × 720	1920 × 1080
Kalantari13 [25]	-	125s	185s	300s	520s
Kalantari19 [24]	$9.0M$	0.35s	0.59s	0.42	0.64
Ours	$6.1M$	0.51s	0.97s	0.64	1.09s

method over Kalantari13 [25] and Kalantari19 [24], reiterating the effectiveness of our method.

5.4. Network Analysis

We first discussed the network parameter and runtime, and then conducted ablation study for the proposed method.

Parameters and runtime Table 4 compares the parameter and runtime of three methods. Note that Kalantari19 [24] and our method were run on a NVIDIA V100 GPU, while Kalantari13 [25] was run on CPUs. Our model contains 6.1 million parameters, including $3.1M$ parameters for CoarseNet and $3.0M$ for RefineNet. It takes around 1 second for our method to produce an HDR frame with a resolution of 1920×1080, which is comparable to Kalantari19 [24] and significantly faster than Kalantari13 [25].

Coarse-to-fine architecture To verify the design of our coarse-to-fine architecture, we compared our method with two baselines. The first one was CoarseNet, which performs optical flow alignment and fusion in the image space (similar to [24]). The second one was RefineNet† that directly takes the LDR frames as input and performs alignment and fusion in the feature space. Experiments with IDs 0-2 in Table 5 show that our method achieves the best results on three datasets, demonstrating the effectiveness of our coarse-to-fine architecture.

Table 5. Ablation study on three datasets with two alternating exposures. CNet and RNet are short for CoarseNet and RefineNet.

ID	Method	Synthetic Dataset		$\mathcal{D}_s^{gt}$		$\mathcal{D}_d^{gt}$	
		PSNR	HDR-VDP2	PSNR	HDR-VDP2	PSNR	HDR-VDP2
0	CNet	39.25	70.81	40.62	74.51	44.43	77.74
1	RefineNet†	39.69	70.95	37.61	75.30	43.70	78.97
2	CNet + RNet	**40.34**	**71.79**	**41.18**	**76.15**	**45.46**	**79.09**
3	CNet + RNet w/o DA	39.72	71.38	40.52	74.79	45.09	78.24
4	CNet + RNet w/o TAF	40.03	71.66	40.80	76.12	45.17	78.99

Network design of the RefineNet To investigate the effect of deformable alignment (DA) module and temporal attention fusion (TAF) module, we trained two variant models, one without DA module and one replacing DAF module with a convolution after feature concatenation. Experiments with IDs 2-4 in Table 5 show that removing either component will result in decreased performance, verifying the network design of the RefineNet.

6. Conclusion

We have introduced a coarse-to-fine deep learning framework for HDR video reconstruction from sequences with alternating exposures. Our method first performs coarse HDR video reconstruction in the image space and then refines the coarse predictions in the feature space to remove the ghosting artifacts. To enable more comprehensive evaluation on real data, we created a real-world benchmark dataset for this problem. Extensive experiments on synthetic and real datasets show that our method significantly outperforms previous methods.

Currently, our method was trained on synthetic data. Since we have captured a large-scale dynamic scene dataset, we will investigate self-supervised training or finetuning using real-world videos in the future.

References

[1] Ahmet Oğuz Akyüz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bülthoff. Do HDR displays support LDR content? a psychophysical evaluation. *TOG*, 2007. 2

[2] Francesco Banterle, Kurt Debattista, Alessandro Artusi, Sumanta Pattanaik, Karol Myszkowski, Patrick Ledda, and Alan Chalmers. High dynamic range imaging and low dynamic range expansion for generating HDR content. In *Computer Graphics Forum*, 2009. 2

[3] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, 2006. 2

[4] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Expanding low dynamic range videos for high dynamic range applications. In *Proceedings of the 24th Spring Conference on Computer Graphics*, 2008. 2

[5] Marcelo Bertalmío. *Vision models for high dynamic range and wide colour gamut imaging: techniques and applications*. Academic Press, 2019. 6

[6] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *TIP*, 2018. 2

[7] Inchang Choi, Seung-Hwan Baek, and Min H Kim. Reconstructing interlaced high-dynamic-range video using joint learning. *TIP*, 2017. 2

[8] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *ICCV*, 2017. 2, 4

[9] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*, 1997. 2, 3, 4, 5

[10] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. HDR image reconstruction from a single exposure using deep cnns. *TOG*, 2017. 2

[11] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *TOG*, 2017. 2

[12] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. In *Digital Photography X*, 2014. 6

[13] Raquel Gil Rodríguez, Javier Vazquez-Corral, and Marcelo Bertalmío. Issues with common assumptions about the camera pipeline and their impact in hdr imaging from multiple exposures. *SIAM Journal on Imaging Sciences*, 2019. 6

[14] Michael D Grossberg and Shree K Nayar. What is the space of camera response functions? In *CVPR*, 2003. 3

[15] Yulia Gryaditskaya, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel. Motion aware exposure bracketing for HDR video. In *Computer Graphics Forum*, 2015. 2

[16] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualiso HDR reconstruction. *EURASIP Journal on Image and Video Processing*, 2015, 2015. 2

[17] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *CVPR*, 2020. 2

[18] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *TOG*, 2016. 2

[19] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. FlexISP: A flexible camera image processing framework. *TOG*, 2014. 2

[20] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *CVPR*, 2013. 2

[21] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 2014. 2

[22] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 4

[23] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *TOG*, 2017. 1, 2, 4, 5

[24] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR video from sequences with alternating exposures. In *Computer Graphics Forum*, 2019. 1, 2, 3, 4, 6, 7, 8

[25] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *TOG*, 2013. 1, 2, 3, 5, 6, 7, 8

[26] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. In *TOG*, 2003. 1, 2, 6

[27] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. Deep SR-ITM: Joint learning of super-resolution and inverse tone-mapping for 4K UHD HDR applications. 2019. 2

[28] Soo Ye Kim, Jihyong Oh, and Munchurl Kim. JSI-GAN: Gan-based joint super-resolution and inverse tone-mapping with pixel-wise task-specific filters for UHD HDR video. In *AAAI*, 2020. 2

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[30] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, 2014. 2

[31] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, Anders Ynnerman, and Jonas Unger. A unified framework for multi-sensor HDR video reconstruction. *Signal Processing: Image Communication*, 2014. 1, 2, 6

[32] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[33] Yuelong Li, Chul Lee, and Vishal Monga. A maximum a posteriori estimation framework for robust high dynamic range video synthesis. *TIP*, 2016. 1, 2, 3

[34] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T Barron, Dillon

Sharlet, Ryan Geiss, et al. Handheld mobile photography in very low light. *TOG*, 2019. 2

[35] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *CVPR*, 2020. 2

[36] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *TOG*, 2014. 2

[37] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *TIP*, 2017. 2

[38] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing*, 2010. 2

[39] Stephen Mangiat and Jerry Gibson. Spatially adaptive filtering for registration artifact removal in HDR video. In *ICIP*. IEEE, 2011. 1, 2

[40] Steve Mann and Rosalind Picard. On being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures. In *IS&T*, 1995. 2

[41] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *TOG*, 2011. 6

[42] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, 2018. 2

[43] Morgan McGuire, Wojciech Matusik, Hanspeter Pfister, Billy Chen, John F Hughes, and Shree K Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 2007. 2

[44] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. In *CVPR*, 2020. 2

[45] Kenta Moriwaki, Ryota Yoshihashi, Rei Kawakami, Shaodi You, and Takeshi Naemura. Hybrid loss for learning single-image-based HDR reconstruction. *arXiv preprint arXiv:1812.07134*, 2018. 2

[46] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. HDR-VQM: An objective quality measure for high dynamic range video. *Signal Processing: Image Communication*, 2015. 6

[47] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *CVPR*, 2000. 2

[48] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *arXiv preprint arXiv:2007.01628*, 2020. 2

[49] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *TPAMI*, 2014. 2

[50] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. In *TOG*, 2002. 6

[51] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image HDR reconstruction using a cnn with masked features and perceptual loss. In *SIGGRAPH*, 2020. 2

[52] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *TOG*, 2012. 2

[53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[54] Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In *CVPR*, 2020. 2

[55] Y. Tian, Y. Zhang, Y. Fu, and C. Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2, 4

[56] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile HDR video production system. In *TOG*, 2011. 1, 2

[57] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: Video restoration with enhanced deformable convolutional networks. In *CVPR Workshops*, 2019. 2, 4

[58] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*, 2018. 1, 2, 4

[59] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 2019. 6

[60] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *CVPR*, 2019. 1, 2, 4, 6, 7

[61] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep HDR imaging via a non-local network. *TIP*, 2020. 2

[62] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. In *ICCV*, 2017. 2

[63] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. In *ICCP*, 2015. 1, 2