Towards Real-World HDR Video Reconstruction: A Large-Scale Benchmark Dataset and A Two-Stage Alignment Network

Yong Shu, Liquan Shen*, Xiangyu Hu, Mengyao Li, Zihao Zhou
Shanghai University, China
{yungsyu,jsslq,arhu314,sdlmy,yi_yuan}@shu.edu.cn

Abstract

*As an important and practical way to obtain high dynamic range (HDR) video, HDR video reconstruction from sequences with alternating exposures is still less explored, mainly due to the lack of large-scale real-world datasets. Existing methods are mostly trained on synthetic datasets, which perform poorly in real scenes. In this work, to facilitate the development of real-world HDR video reconstruction, we present **Real-HDRV**, a large-scale real-world benchmark dataset for HDR video reconstruction, featuring various scenes, diverse motion patterns, and high-quality labels. Specifically, our dataset contains 500 LDRs-HDRs video pairs, comprising about 28,000 LDR frames and 4,000 HDR labels, covering daytime, nighttime, indoor, and outdoor scenes. To our best knowledge, our dataset is the largest real-world HDR video reconstruction dataset. Correspondingly, we propose an end-to-end network for HDR video reconstruction, where a novel two-stage strategy is designed to perform alignment sequentially. Specifically, the first stage performs global alignment with the adaptively estimated global offsets, reducing the difficulty of subsequent alignment. The second stage implicitly performs local alignment in a coarse-to-fine manner at the feature level using the adaptive separable convolution. Extensive experiments demonstrate that: (1) models trained on our dataset can achieve better performance on real scenes than those trained on synthetic datasets; (2) our method outperforms previous state-of-the-art methods. Our dataset is available at https://github.com/yungsyu99/Real-HDRV.*

1. Introduction

The demands for high dynamic range (HDR) video have drastically increased in recent years since it can bring a better visual experience for users [2, 10, 19, 40]. However, most cameras cannot capture HDR videos directly due to the limitations of sensors. Therefore, some specialized hardware devices [12, 17, 27, 30, 32] are developed to capture HDR videos. However, these devices are typically bulky and expensive, which are not widely adopted [6].

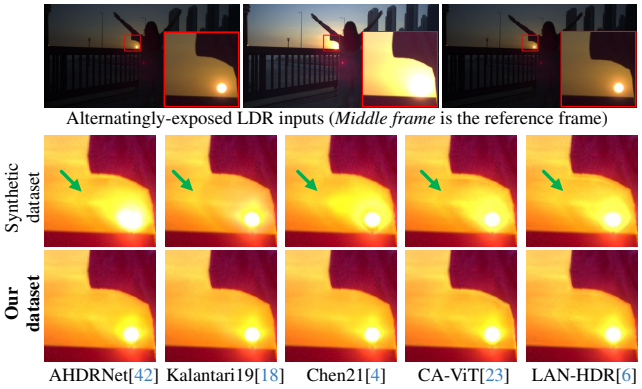

Figure 1. Row 1 shows a real-world sample from the Chen21 dataset[4]. Row 2-3 show the HDR frames reconstructed by models trained on the synthetic dataset [4] and our Real-HDRV, respectively. Obviously, models trained on our dataset are able to recover more and better details of the over-exposed regions.

In contrast, the computational-based HDR video reconstruction [20, 24] is more practical and affordable for obtaining HDR videos. It captures low dynamic range (LDR) sequences with alternating exposures (*e.g.*, sequences with exposure values of {-3,0,-3,0,...}), which are then used to reconstruct the corresponding HDR video. The common reconstruction pipeline is to align the input frames and then merge the aligned inputs to reconstruct the HDR videos. Before the era of deep learning, some optimization-based reconstruction methods [19, 20, 24] are proposed. Recently, learning based methods [4, 6, 18] have shown their effectiveness on HDR video reconstruction, which significantly improve the performance over optimization-based methods.

Despite remarkable progress, the development of deep models for HDR video reconstruction is relatively slow, mainly due to the lack of suitable training datasets. The only publicly accessible labeled real-world dataset of HDR

*Corresponding author. This work was partially supported by China NSFC grant (no.61931022 and no.62271301), Shanghai SSTP grant (22511105200), Shanghai SEALP grant (23XD1401400).

video reconstruction [4] is built for evaluating HDR video reconstruction methods. The number of distinct scenes and the motion patterns in their dataset are limited, making it unsuitable for supervised training. Therefore, existing models are still trained on synthetic datasets. However, the synthetic datasets are not well suited for the study of real-world HDR video reconstruction. Models trained on synthetic datasets are hard to generalize to real scenes (see Fig. 1 Row 2) since the synthetic degradations are far different from the real degradations (*e.g.*, the noise in under-exposed areas, the saturation in over-exposed areas). It is highly desired for a large-scale real-world dataset to facilitate the development of real-world HDR video reconstruction.

Besides the datasets, another key issue of HDR video reconstruction lies in the alignment of input frames. Previous methods [18–20] usually use optical flow or both optical flow and deformable convolution [4, 45] to align the inputs. However, the estimated flows are prone to be inaccurate due to the noise and saturation in alternatingly-exposed inputs, resulting in ghosting artifacts. Recently, Chung *et al*. [6] proposed a luminance-based attention module to align the inputs. However, it cannot properly deal with the under-exposed and over-exposed areas of inputs, resulting in unpleasing artifacts. Additionally, the global motion (caused by camera movements) is not properly modeled in most existing methods [6, 45], which further increases the difficulty of alignment, leading to inferior performance.

Based on the above observations, to facilitate the development of real-world HDR video reconstruction, we build a large-scale real-world dataset, named **Real-HDRV**. In order to get LDRs-HDRs video pairs and ensure the quality of HDR labels, we capture the scene in a frame-by-frame manner using a camera with high continuous shooting speed (up to 40 frames/sec). Specifically, we carefully select the relatively static scenes and manually create different types of motion between neighboring frames. For each static frame, we capture a multi-exposure image stack (7 differently-exposed LDR images guarantee the quality of HDR labels). The images in each multi-exposure stack are then used to synthesize the corresponding HDR label. We collected 500 LDRs-HDRs video pairs, comprising about 28,000 LDR frames and 4,000 HDR labels. Our Real-HDRV cannot only serve as a benchmark for HDR video reconstruction but also be applied to other HDR tasks (*e.g.*, HDR Deghosting [42], single-image HDR reconstruction [46]).

Correspondingly, we propose an end-to-end network for HDR video reconstruction, in which we design a two-stage strategy to align the inputs sequentially. Specifically, the first stage performs global alignment with the designed global alignment module (GAM), which can effectively handle the global motion and reduce the difficulty of subsequent alignment. The second stage implicitly performs local alignment at the feature level in a coarse-to-fine manner with the designed local alignment module (LAM). The pyramid structure of LAM facilitates the feature alignment under large motion. The adaptive separable convolution [31] used in LAM enables flexibly integrating the useful information in neighboring frames to compensate for the missing content in the reference frame, which facilitates the feature alignment under noise and saturation. Then, a reconstruction module is applied to reconstruct the HDR video from the aligned features. Our two-stage alignment network can effectively handle complex motion and reconstruct high-quality HDR video. In summary, our contributions are as follows:

- We propose a large real-world HDR video reconstruction dataset, featuring various scenes, diverse motion patterns, and high-quality labels. Our dataset cannot only serve as a benchmark for HDR video reconstruction but also be applied to other HDR imaging tasks.
- We propose an end-to-end network for HDR video reconstruction, in which we design a two-stage strategy to perform alignment sequentially. Our network can effectively handle complex motion and achieve high-quality HDR video reconstruction.
- Extensive experiments demonstrate the superiority of our dataset and our method. Our work provides a new platform for researchers to explore real-world HDR video reconstruction techniques.

2. Related Work

HDR Image Reconstruction Many methods[5, 16, 22, 33, 34] attempt to perform HDR reconstruction from a single LDR image. However, these methods cannot effectively handle the noise and saturation due to the limited information in a single image. There are methods[15, 35] for HDR reconstruction from multi-exposure LDR images. Although these methods work well for static scenes, they generally suffer from ghosting artifacts when tackling dynamic scenes. Therefore, many HDR deghosting methods [3, 9, 23, 42, 43] are proposed to alleviate this issue.

HDR Video Reconstruction Datasets Kalantari *et al*. [19] captured 5 LDR sequences with two alternating exposures. To quantitatively evaluate HDR video reconstruction methods, Chen *et al*. [4] collected 76 dynamic image pairs, 49 static image pairs, and 50 unlabeled LDR sequences with two alternating exposures. However, the number of distinct scenes and the motion patterns in their dataset are limited, making it unsuitable for supervised training. Recently, Yue *et al*. [45] collected 85 real-world LDRs-HDRs video pairs using a mobile phone, but, until now, they are not publicly accessible. In addition, they use two images with different exposures to generate an HDR label, which may not cover the full dynamic range of the scene, resulting in limited-quality HDR labels. Due to the lack of publicly accessi-

Table 1. Comparison between different datasets.

Dataset	GT	Numbers	Motion patterns	Scenes
Chen21 (Dynamic) [4]	central frame	76	LM	ID
Chen21 (Static) [4]	only one static frame	49	Static	ID, IN, OD, ON
Ours	per frame	500	GM, LM, FM	ID, IN, OD, ON

1. Our dataset contains per-frame HDR labels, while the Chen21 dataset only contains the HDR labels for the center frames.
2. GM and LM denote global motion (where only the camera is moving) and local motion (where only the foreground is moving), respectively. FM denotes full motion (where both foreground and camera are moving).
3. OD, ON, ID and IN denote outdoor daytime, outdoor nighttime, indoor daytime and indoor nighttime, respectively.

ble large-scale real-world datasets, existing models are still trained on the synthetic dataset [4], which hinders the development of real-world HDR video reconstruction.

HDR Video Reconstruction There are mainly two types of methods to obtain HDR videos: hardware-based methods and computational-based methods. The hardware-based methods [12, 27, 30, 32] typically rely on specialized hardware systems (*e.g.*, beam splitter), which are typically too expensive to be widely adopted.

The computational-based methods reconstruct the HDR video from alternatingly-exposed sequences. Kang *et al*. [20] proposed the first method in this direction, which used optical flow to align the input frames and then merged the aligned frames to generate HDR videos. Mangiat *et al*. [24] improved [20] by introducing a block-based motion estimation method with a refinement stage for ghost removal. Kalantari *et al*. [19] proposed a patch-based method to synthesize the missing exposures at each frame, and these synthesized images are then fused into an HDR frame.

Recently, learning based methods have shown their effectiveness on HDR video reconstruction. Kalantari *et al*. [18] proposed a flow-based framework, which consists of an optical flow network for alignment and a weight network for merging images. Chen *et al*. [4] and Yue *et al*. [45] used both optical flow and deformable convolution to perform alignment for reconstructing HDR videos. Unfortunately, these methods typically generate ghosting artifacts since the estimated flows are prone to be inaccurate due to the noise and saturation. More recently, Chung *et al*. [6] proposed a luminance-based alignment network for HDR video reconstruction. However, it cannot properly deal with the under-exposed and over-exposed areas of inputs, resulting in unpleasing artifacts.

3. Proposed Dataset

To favor the development of real-world HDR video reconstruction, we construct a large-scale real-world HDR video reconstruction dataset, named Real-HDRV. Actually, it is extremely challenging to simultaneously capture alternatingly-exposed LDR sequences and the corresponding HDR sequences for dynamic scenes. One may use a beam splitter and two cameras to build a complex optical system to capture two videos with different exposures simultaneously and then generate the HDR video using the captured two videos. However, the amount of light is halved by the beam splitter[8, 38], which limits the quality of HDR labels. Instead of relying on complex optical systems, to get LDRs-HDRs video pairs and ensure the quality of the HDR labels, we capture LDRs-HDRs video pairs in a frame-by-frame manner. We manually create motions between neighboring frames and capture a multi-exposure image stack (7 LDR images) for each static frame.

One crucial problem that needs to be addressed is how to ensure that the high-quality HDR label can be obtained after capturing each multi-exposure image stack. The details are as follows: (1) We use a camera with high continuous shooting speed (up to 40 frames/sec), the Canon R6 Mark2, which enables us to capture a multi-exposure image stack in a very short period of time with one depression of the wireless shutter. During the capturing, the camera can almost avoid introducing extra motion, such as camera shake, object movement, etc. (2) We carefully select the relatively static scenes and manually create different types of motion (*i.e.*, global motion, local motion, and full motion) between neighboring frames to make the motion controllable and diverse. In addition, the camera is mounted on a tripod, and a wireless remote controller is used to avoid introducing extra motion caused by the shutter release.

Thanks to the high-speed shooting performance of the camera and the careful shooting procedure, we can ensure there is almost no motion between the images in each multi-exposure image stack. Also, each multi-exposure image stack can provide enough images with different exposures (7 images with different exposures guarantee to cover the full dynamic range of a scene). Then, the per-frame high-quality HDR label can be generated from images in each multi-exposure image stack by using the method in [7], and the LDR images can be arranged in a periodic exposure to generate sequences with alternating exposures[1]. Therefore, we can collect LDRs-HDRs video pairs in a frame-by-frame manner. Specifically, for each frame, we capture a multi-exposure image stack of 7 LDR images spaced by $\pm i$-EV difference where $i \in 1, 2, 3$ around the reference exposure. Then, we manually create different types of motion between neighboring frames to capture the following image stacks. Finally, all the image stacks are grouped in their temporal order to generate the LDRs-HDRs video pairs.

In total, we collected 500 LDRs-HDRs video pairs, each containing 7 to 10 frames. For each frame, 7 differently-exposed LDR images and a high-quality HDR label can be

[1] Since we focus on HDR video reconstruction from sequences with two alternating exposures, we selected LDR frames in a periodic exposure (*i.e.*, {EV-3, EV0, EV-3, EV0, ...}, {EV-2, EV+1, EV-2, EV+1, ...} or {EV-1, EV+2, EV-1, EV+2, ...}.) to generate the sequences with two alternating exposures. Note that the LDR frames can be selected in different exposure orders for HDR video reconstruction.

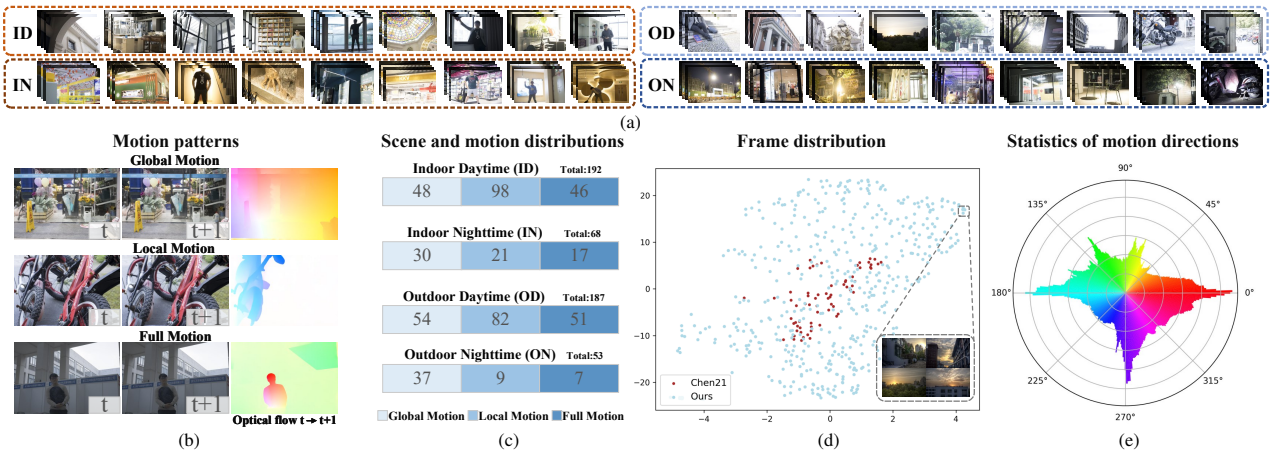

Figure 2. (a) Some typical scenes in our dataset, which can be categorized into 4 categories: indoor daytime (ID), indoor nighttime (IN), outdoor daytime (OD), and outdoor nighttime (ON) scenes. (b) Our dataset contains three kinds of motion: global motion (where only the camera is moving), local motion (where only the foreground is moving), and full motion (where both foreground and camera are moving). (c) Scene and motion distributions of our dataset. (d) Diversity comparison: our dataset *vs.* the Chen21 dataset [4]. (e) Statistics of motion directions in our dataset. We plot a circular histogram, where the color of each bin represents the direction of motion, and the height of the bar represents the proportion of specific directions to all the directions. The per-pixel flow in each frame is computed via RAFT[36].

Table 2. Metrics to assess the diversity of different datasets.

Metrics on the extent of HDR	
FHLP	Fraction of HighLight Pixel: defined in [11]
EHL	Extent of HighLight: defined in [11]
Metrics on intra-frame diversity	
SI	Spatial Information: defined in [1]
CF	ColorFulness: defined in [14]
stdL	standard deviation of Luminance: defined in [11]
Metrics on the overall-style	
ALL	Average Luminance Level: defined in [11]
DR	Dynamic Range [16]: calculated as the log10 differences between the highest 2% luminance and the lowest 2% luminance.

Among these aspects, greater extent of HDR represents more probability for the network to learn pixel in advanced HDR volume beyond LDR's capability, higher intra-frame diversity means that the network may learn better generalization capability. We use these metrics to verify the diversity of our dataset

provided. All the images are captured in RAW format with resolution of 6000×4000. We performed the demosaicing, white balancing, color correction, and gamma compression ($\gamma = 2.2$) to convert the raw data to RGB data. In this work, we rescaled the images to 1500×1000 for training and testing. Figure 2 shows some typical scenes in our dataset and the statistical indicators of our dataset. In addition, since our dataset provides data in RAW format, the data processing pipeline is highly flexible. Therefore, our dataset can be easily adjusted to make training data for different HDR tasks for future research.

Analysis of Our Dataset To quantitatively evaluate the superiority of our dataset, we analyzed the diversity of the Chen21 dataset[4] and our dataset. Following [11, 16], the 7 metrics defined in Table 2 are utilized to assess the diversity of different datasets from 3 aspects, including the extent of HDR, the intra-frame diversity and the overall style of HDR. For each HDR label, 7 different metrics are calculated according to Table 2. Then, we use the t-SNE [37]

Table 3. Statistics of HDR labels in different datasets. Besides the DR, all numbers are in percentage.

	Extent of HDR		**Intra-frame Diversity**			**Overall-style**	
Dataset	**FHLP**	**EHL**	**SI**	**CF**	**stdL**	**ALL**	**DR**
Chen21 [4]	8.85	2.46	8.93	2.77	11.35	5.29	2.54
Ours	**13.75**	**2.72**	**9.16**	**4.30**	**12.13**	5.05	**2.73**

to project the single frame's 7-D vector (consisting of 7 metrics from Table 2) to the corresponding 2D-coordinate for plotting the frame distribution of our dataset and the Chen21 dataset. As shown in Fig. 2 (d), our dataset contains wider frame distribution than the Chen21 dataset, indicating that the networks trained with our dataset may be better generalized to different scenarios. And the statistics of different datasets are shown in Table 3. In addition, our dataset contains more diverse motion patterns (see Table 1 and Fig. 2 (c)). The diversity in both scenes and motion patterns makes that our dataset can naturally be used for training deep networks and assessing the generalization capability of the networks across different scenes.

4. Proposed Method

Global motion (caused by camera movements) and local motion (caused by object motion) are almost inevitable when capturing videos, which imposes a core issue for HDR video reconstruction: how to perform alignment for the alternatingly-exposed inputs. Without effective alignment, the areas with motion in neighboring frames cannot be properly utilized to reconstruct the HDR frame, leading to severe ghosting artifacts. In this work, considering the differences between the global motion and local motion, we introduce a two-stage alignment network for HDR video reconstruc-

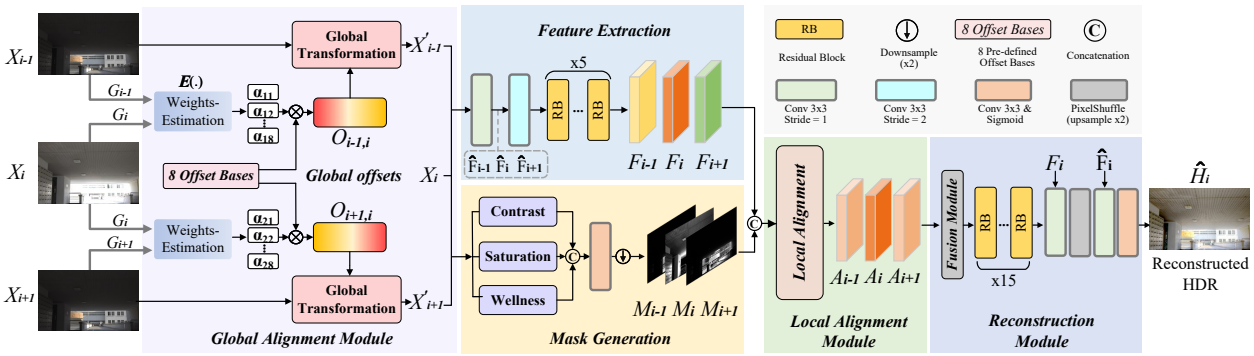

Figure 3. The architecture of our proposed network.

tion, which firstly performs alignment for the inputs (from global to local) and then adaptively fuses the aligned features to reconstruct the HDR video.

Overview Given an input LDR video $\{I_i|i = 1,...,n\}$ with alternating exposures $\{t_i|i = 1,...,n\}$ [2], our target is to reconstruct the corresponding HDR video $\{H_i|i = 1,...,n\}$. Following [4, 18], the input images are firstly mapped into the linear HDR domain by applying gamma correction:

$$\bar{I}_i = I_i^{\gamma}/t_i,\ (\gamma = 2.2)\,. \tag{1}$$

where t_i is the exposure time of I_i. Then, the input image I_i and the linear image $\bar{I}_i$ are concatenated into a 6-channels input X_i. Our network takes three continuous frames $\{X_{i-1}, X_i, X_{i+1}\}$ as input and predicts the HDR frame $\hat{H}_i$ for the center frame. As shown in Fig. 3, our network consists of the global alignment module (GAM) for compensating global motion, the local alignment module (LAM) for compensating local motion, and the reconstruction module for reconstructing the HDR frame.

Global Alignment Module The global motion is relatively simple, which does not need to be modeled by dense pixel-wise optical flow with high Degree-of-Freedoms. Inspired by [39, 44], we use the pre-defined offset bases with 8 Degree-of-Freedoms (with each 2 for translation, rotation, scale, perspective [13]) to model the global motion. Specifically, we design the GAM to estimate a weighted sum of 8 pre-defined offset bases for generating the global offsets. The global offsets are then used to spatially transform the inputs. Since all the operations in the GAM are differentiable, the GAM can be optimized through end-to-end training. In this way, the GAM can adaptively learn to compensate for the global motion between neighboring frames.

As shown in Fig. 3, given the input $\{X_j|j = i-1, i, i+1\}$, the GAM firstly uses a shared encoding layer to extract feature maps G_j with 16 channels from inputs. Then, the features $\{G_j|j = i-1, i+1\}$ of neighboring frames are fed into the weights estimation module $E(.)$ (see our *supplementary file* for the detailed architecture) along with the feature map G_i of the reference image to obtain the corresponding weights $\{\alpha_{1k}, \alpha_{2k}\}$, generating the global offsets:

$$O_{i-1,i} = \sum_{k=1}^{8} \alpha_{1k} n_k \quad (k = 1, 2, ..., 8)\,, \tag{2a}$$

$$O_{i+1,i} = \sum_{k=1}^{8} \alpha_{2k} n_k \quad (k = 1, 2, ..., 8)\,. \tag{2b}$$

where the pre-defined offset bases n_k are computed with the same settings in [44]. The global offsets are then used to spatially transform the neighboring frames for compensating global motion between neighboring frames.

Local Alignment Module The LAM is designed to perform local alignment, which estimates the kernel weights at multiple scales in a coarse-to-fine manner and then performs a transformation for input features using adaptive separable convolution[31] with the estimated kernel weights. In this way, the LAM can adaptively learn to integrate useful information in neighboring frames to compensate the missing content in reference frame, which facilitates the feature alignment under the noise and saturation.

First, the shallow features $\{F_j|j = i-1, i, i+1\}$ of input images are extracted. Inspired by Mertens *et al*. [28], the contrast maps c_j, exposure wellness maps e_j, and saturation maps s_j are extracted to provide the exposure information of the inputs. All three maps are concatenated together to generate the adaptive masks M_j for input images. To handle large motions, the pyramidal processing is adopted, we generate an L-level pyramid of feature representation $\{F_j^l|j = i-1, i, i+1; l = 1, ..., L\}$ for input images and an L-level pyramid of masks $\{M_j^l|j = i-1, i, i+1; l = 1, ..., L\}$. Then, we concatenate the corresponding masks and features along the channel dimension at each level to obtain the pyramid of tensors $\{K_j^l|j = i-1, i, i+1; l = 1, ..., L\}$, which are then utilized to predict kernel weights W_j^l for neighboring features. With the predicted kernel weights, the aligned features A_j^l can be obtained after performing adap-

[2] For example, the input sequences can alternate between two exposures {EV0, EV+3, EV0, EV+3, ...} or three exposures {EV-2, EV0, EV+2, EV-2, EV0, EV+2, ...}. In this work, we reconstruct the HDR video from sequences with two alternating exposures, while our network can be easily extended to other cases (*e.g.*, three exposures).

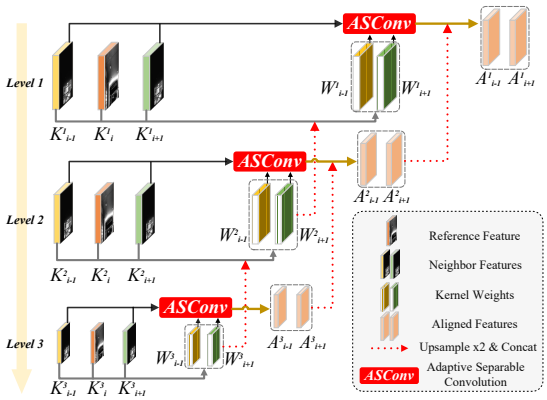

Figure 4. The architecture of local alignment module (LAM).

tive separable convolution for neighboring features. Specifically, at the l-th level, kernel weights and aligned features are predicted also with the $\times 2$ upsampled kernel weights and aligned features from the upper $(l+1)$-th level, respectively (red dot lines in Fig. 4):

$$W_{i-1}^{l}, W_{i+1}^{l} = g\Big([K_{i-1}^{l}, K_{i}^{l}, K_{i+1}^{l}, \left(W_{i-1}^{l+1}\right)^{\uparrow\times 2}, \left(W_{i+1}^{l+1}\right)^{\uparrow\times 2}]\Big), \tag{3a}$$

$$A_{i-1}^{l} = h\left(ASConv\left(F_{i-1}^{l}, W_{i-1}^{l}\right), \left(A_{i-1}^{l+1}\right)^{\uparrow\times 2}\right), \tag{3b}$$

$$A_{i+1}^{l} = h\left(ASConv\left(F_{i+1}^{l}, W_{i+1}^{l}\right), \left(A_{i+1}^{l+1}\right)^{\uparrow\times 2}\right). \tag{3c}$$

where $(.)^{\uparrow\times 2}$ is the upscaling operation with a factor of 2, $[.]$ is the concat operation, $g(.)$ is the kernel weights predictor consisting of several convolution layers, $ASConv(.)$ denotes the adaptive separable convolution, $h(.)$ is the general function with several convolution layers. We use three-level pyramid, *i.e.*, L=3, in LAM. The kernel size is set to 31 in the adaptive separable convolution.

Fusion and Reconstruction The fusion module is used to fuse the aligned features, which can suppress the harmful features from under-exposed and over-exposed areas. As shown in Fig. 5, the aligned features are concatenated as the input of the fusion module, generating fusion masks for fusing the aligned features A_j. Then, the fused feature F_{fusion} is passed through a series of residual blocks. Two skip connections are added to concatenate shallow features of the reference frame. Finally, the HDR frame $\hat{H}$ can be obtained after a convolution layer and a sigmoid activation layer.

Loss Function Since HDR images are typically displayed after tonemapping, following [4, 18, 42], we use the μ-law tonemapping function to the HDR image:

$$T(H) = \frac{log(1+\mu H)}{log(1+\mu)}, \mu = 5000. \tag{4}$$

where $T(H)$ denotes the tonemapped HDR image. We train the network by minimizing the $l1$ loss distance $L = ||T(\hat{H}) - T(H)||$ between the tonemapped estimated $\hat{H}$ and ground truth H.

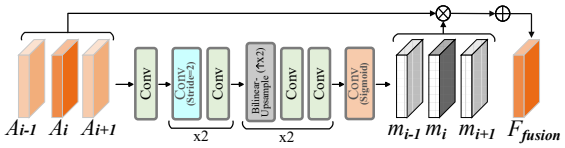

Figure 5. The architecture of fusion module.

5. Experiments

5.1. Experimental Settings

Datasets There are three datasets adopted, including our Real-HDRV, the Chen21 dataset [4] and the synthetic dataset [4]. As for our dataset, it is divided into the training collection (450 videos) and the testing collection (27 videos for indoor daytime and outdoor daytime scenes, 23 videos for indoor nighttime and outdoor nighttime scenes). Each video in the testing collection provides 8 LDR frames with two alternating exposures and the corresponding HDR labels. As for the synthetic dataset, we utilized 21 existing HDR videos from [8, 21] and Vimeo-90K [41] to synthesize the dataset with the same settings as in [4]. The Chen21 dataset contains 76 dynamic image pairs, 49 static image pairs augmented with random global motion, and 50 unlabeled sequences with two alternating exposures.

Implementation Details We generate LDR sequences with two alternate exposures separated by three stops for each video in our training collection. We then sample three LDR frames as input and produce the HDR label for the center frame to generate a training sample. We crop patches of size 256×256 with a stride of 128 from the training set for training. Random rotation and flipping augmentation are applied. We use Adam optimizer, and set the batch size and initial learning rate as 4 and 0.0001, respectively. We implement our model using PyTorch with 6 NVIDIA 3090 GPUs and train for 100 epochs.

Evaluation Metrics We use six common metrics for testing, *i.e.*, HDR-VDP-2 [25], PSNR-μ, SSIM-μ, PU-PSNR, PU-SSIM and HDR-VQM [29]. PSNR-μ and SSIM-μ are computed after tonemapping with μ-law function (in Equation (4)). PU-PSNR and PU-SSIM are computed after perceptually uniform encoding [26]. When computing the HDR-VDP-2, the diagonal display size is set to 30 inches.

5.2. Evaluation of Our Proposed Dataset

To evaluate the effectiveness of our dataset, we compare our dataset with the synthetic dataset [4]. We train representative HDR reconstruction models [4, 6, 18, 23, 42] on our dataset and the synthetic dataset, and evaluate the performance of trained models on the Chen21 dataset [4].

Quantitative Results As shown in Table 4, the models trained on our dataset can achieve better performance on the real-world dataset [4] than the models trained on the synthetic dataset, demonstrating the effectiveness of our

Table 4. Quantitative comparison for training on the synthetic dataset [4] or our dataset, while evaluating on the Chen21 dataset [4]. The better results are highlighted in bold. Among these evaluation metrics, the higher quality of the tested HDR image leads to the higher score.

Methods	Training dataset	Evaluation on the dynamic set						Evaluation on the static set					
		PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM	HDR-VDP-2	HDR-VQM	PSNR-μ	SSIM-μ	PU-PSNR	PU-SSIM	HDR-VDP-2	HDR-VQM
AHDRNet [42]	Synthetic	44.34	0.9668	38.48	0.9718	62.05	84.56	38.38	0.9329	32.99	0.9422	58.19	71.40
	Our	**45.02**	**0.9741**	**39.17**	**0.9808**	**62.51**	**89.60**	**40.16**	**0.9589**	**34.69**	**0.9638**	**59.53**	**77.84**
Kalantari19 [18]	Synthetic	44.15	0.9637	38.39	0.9728	59.44	86.56	40.59	0.9316	35.22	0.9429	57.53	74.60
	Our	**45.31**	**0.9689**	**39.37**	**0.9757**	**61.39**	**86.95**	**41.19**	**0.9336**	**36.03**	0.9429	**59.69**	**81.83**
Chen21 [4]	Synthetic	45.46	0.9706	39.46	0.9760	61.26	87.40	41.21	0.9412	35.81	0.9483	59.19	78.98
	Our	**45.65**	**0.9716**	**39.79**	**0.9768**	**61.39**	**90.33**	**41.37**	**0.9419**	**36.20**	**0.9516**	**59.46**	**81.43**
CA-ViT [23]	Synthetic	44.76	0.9664	38.81	0.9714	61.95	88.78	38.26	0.9252	32.84	0.9368	58.27	71.90
	Our	**45.19**	**0.9744**	**39.33**	**0.9814**	**62.43**	**90.41**	**39.91**	**0.9570**	**34.30**	**0.9609**	**59.05**	**77.69**
LAN-HDR [6]	Synthetic	44.81	0.9714	38.64	0.9773	61.64	88.86	39.34	0.9424	33.87	0.9490	57.12	70.47
	Our	**45.38**	**0.9743**	**39.53**	**0.9804**	**62.83**	**88.96**	**40.09**	**0.9565**	**34.63**	**0.9595**	**59.78**	**79.29**

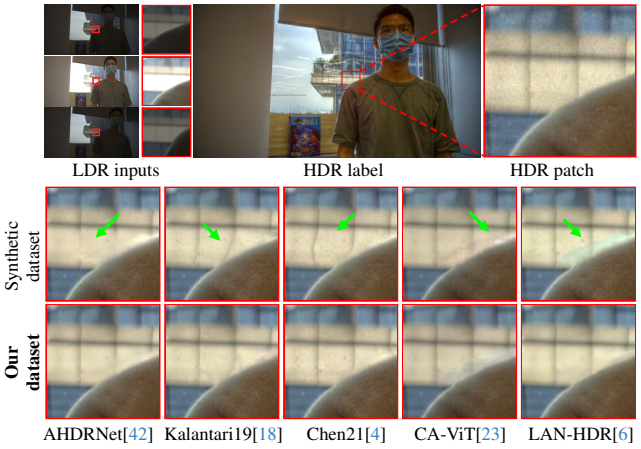

Figure 6. Visual comparison of different models trained on the synthetic dataset [4] and our dataset.

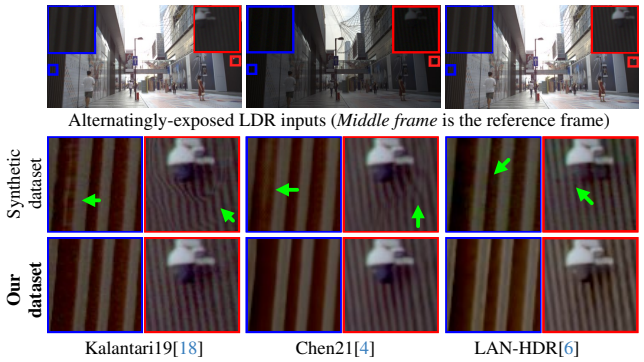

Figure 7. Visual example of models trained on different datasets.

dataset. For example, compared with Kalantari19 trained on the synthetic dataset, the same model trained on our dataset can achieve more than 1 dB gain (PSNR-μ) on the dynamic set of Chen21 dataset, which is significant. Similar improvements can also be observed in other methods.

Qualitative Results The visual comparison for the models trained on different datasets is provided in Fig. 6. Obviously, the models trained on our dataset yield better visual quality, while the models trained on the synthetic dataset typically yield severe ghosting artifacts or color distortions. The superior performance of the models trained with our datasets comes from the real degradation distribution in our dataset (more qualitative comparisons are provided in *supplementary file*). In summary, models trained on our Real-HDRV can better handle real-world scenes, demonstrating the effectiveness of our dataset.

Evaluation on Unlabeled Real-World Dataset We also evaluate the varying models on unlabeled sequences of the Chen21 dataset [4]. The visual comparison is provided in Fig. 7. As seen, when the reference frame is low-exposure, the models trained on our dataset can recover clear details, while the models trained on the synthetic dataset generate corrupted details or color distortions. Similar improvements can be observed when the reference frame is high-exposure, please refer to our *supplementary file* for more details.

5.3. Evaluation of Our Proposed Method

We compare our method with prevalent state-of-the-art HDR video reconstruction methods [4, 6, 18, 19, 45] and state-of-the-art HDR deghosting methods [23, 42] on our dataset for a comprehensive evaluation. For a fair comparison, we use their officially released codes, if accessible, otherwise, we re-implement their methods based on their papers. Note that the AHDRNet [42] and CA-ViT [23] are adapted for HDR video reconstruction by changing the network input. In addition, we evaluate our method on the Chen21 dataset[4] to demonstrate the generalization of our method (more details can be found in *supplementary file*).

Quantitative Results The quantitative comparison between our method and other methods is listed in Table 5. Compared to other methods, our method achieves the best average performance in all the evaluation metrics, demonstrating the effectiveness of our method. In addition, evaluated in different scenes, our method can also acquire the best performance, demonstrating that our method can better handle sequences under different real scenes.

Qualitative Results The visual comparison of varying methods on our dataset is shown in Fig. 8. Obviously, our

Table 5. Quantitative comparison of our method with state-of-the-art methods on our dataset. Red text indicates the best and blue text indicates the second best result, respectively. ID&OD denotes indoor daytime and outdoor daytime scenes. IN&ON denotes indoor nighttime and outdoor nighttime scenes.

	PSNR-μ			SSIM-μ			PU-PSNR			PU-SSIM			HDR-VDP-2			HDR-VQM		
Methods	ID&OD	IN&ON	Avg	ID&OD	IN&ON	Avg	ID&OD	IN&ON	Avg	ID&OD	IN&ON	Avg	ID&OD	IN&ON	Avg	ID&OD	IN&ON	Avg
Kalantari13[19]	41.27	37.24	39.41	0.9697	0.9246	0.9490	35.11	31.76	33.57	0.9738	0.9439	0.9601	58.03	56.42	57.30	90.67	78.80	85.21
AHDRNet[42]	45.02	40.72	43.06	0.9840	0.9613	0.9736	38.96	35.08	37.17	0.9831	0.9666	0.9755	60.53	57.78	59.27	90.20	77.60	84.40
Kalantari19[18]	43.90	39.23	41.75	0.9769	0.9437	0.9616	38.07	33.99	36.19	0.9797	0.9576	0.9695	61.61	59.05	60.43	92.13	82.16	87.54
Chen21[4]	44.42	39.43	42.12	0.9789	0.9466	0.9640	38.70	34.29	36.67	0.9832	0.9612	0.9731	63.89	60.70	62.41	94.22	83.93	89.49
CA-ViT [23]	44.65	40.16	42.58	0.9834	0.9603	0.9728	38.44	34.34	36.55	0.9820	0.9651	0.9743	60.62	57.38	59.13	90.38	77.63	84.51
Yue23 [45]	45.42	41.20	43.48	0.9849	0.9626	0.9746	39.37	35.65	37.66	0.9845	0.9682	0.9770	62.93	59.70	61.44	93.23	82.99	88.52
LAN-HDR[6]	44.23	40.54	42.53	0.9836	0.9607	0.9731	38.16	34.93	36.67	0.9823	0.9646	0.9742	60.89	58.38	59.73	91.28	78.73	85.51
Ours	45.81	41.58	43.86	0.9856	0.9643	0.9758	39.83	36.10	38.11	0.9860	0.9711	0.9792	65.34	61.40	63.53	94.47	84.13	89.71

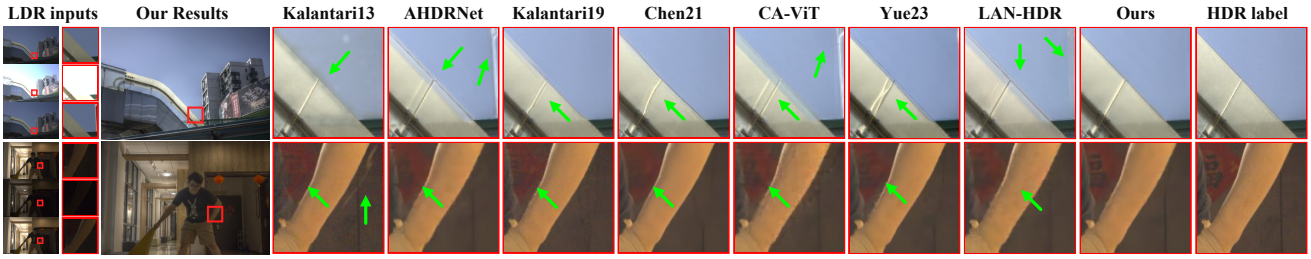

Figure 8. Visual comparison of different networks trained on our dataset. Please zoom in for more details.

Table 6. Computation complexity comparison. All methods are executed on a Nvidia 3090 GPU.

Methods	Kalantari19[18]	Chen21[4]	Yue23[45]	LAN-HDR[6]	Ours
Params. (M)	10.39	6.44	3.5	7.31	5.98
Flops (T)	2.13	5.29	3.96	1.24	2.07
Time (s)	0.24	0.87	0.76	0.55	0.34

method achieves more excellent visual quality than other methods, which can recover the missing content of the over-exposed areas without introducing artifacts when the reference frame is high-exposure (see the 1^{st} row in Fig. 8). Also, our method can better remove the noise and faithfully preserve the structure of the under-exposed areas when the reference frame is low-exposure (see the 2^{nd} row in Fig. 8). In contrast, due to the inaccurate-prone flow, the flow-based methods [4, 18, 19, 45] usually suffer from unpleasing artifacts for the over-exposed areas, and they cannot faithfully recover the details in the under-exposed areas. Additionally, due to the lack of effective alignment, the attention-based methods [23, 42] can easily introduce ghosting artifacts.

Complexity Comparisons For each test method, we record the quantity of the model parameters, the execution time and the floating point operations (flops) of generating an HDR frame with the size of 1500 $\times$ 1000. As shown in Table 6, our model can achieve the best trade-off between the performance and the computational cost.

Ablation Study To analyze the effectiveness of each component in our network, we conduct comprehensive ablation studies on our dataset. As shown in Table 7, the GAM and LAM both improve the performance, demonstrating the effectiveness of the GAM and the LAM. On the one hand, the model (Baseline) performs poorly when directly conducting

Table 7. Quantitative results of the ablation studies. Our baseline network uses the same architecture as our full model (Model4), but with the GAM and the LAM removed

Models	Baseline	GAM	LAM	HDR-VDP-2	PSNR-μ	HDR-VQM
Model1	✓	×	×	59.49	42.67	85.34
Model2	✓	✓	×	61.72	43.07	87.62
Model3	✓	×	✓	62.58	43.61	88.45
Model4	✓	✓	✓	63.53	43.86	89.71

HDR video reconstruction without performing alignment, demonstrating that the alignment is very critical to HDR video reconstruction. On the other hand, by using the GAM to perform global alignment, our full model can more effectively handle the complex motion, obtaining the higher HDR-VDP-2 score and the higher HDR-VQM score than ours (w/o GAM). Also, our full model can achieve better performance than ours (w/o LAM).

6. Conclusion

We constructed a novel dataset for HDR video reconstruction, which contains various scenes, diverse motion patterns, and high-quality labels. Our dataset can also be applied to other HDR tasks for future research. Then, we proposed a novel framework for HDR video reconstruction, which considers the differences between global motion and local motion. The designed GAM enables our framework to better handle global motion. And the designed LAM can adaptively integrate the useful information in neighboring frames to help reconstruct the reference HDR frame, effectively decreasing the ghosting artifacts caused by large local motion. Extensive experiments demonstrate the superiority of our dataset and our method.

References

[1] ITU, Geneva, Switzerland, Recommendation ITU-R BT.500-14: Methodologies for the subjective assessment of the quality of television images, 2019. 4

[2] Nabajeet Barman and Maria G Martini. User generated hdr gaming video streaming: Dataset, codec comparison, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1236–1249, 2021. 1

[3] Sibi Catley-Chandar, Thomas Tanay, Lucas Vandroux, Aleš Leonardis, Gregory Slabaugh, and Eduardo Pérez-Pellitero. Flexhdr: Modeling alignment and exposure uncertainties for flexible hdr imaging. *IEEE Transactions on Image Processing*, 31:5923–5935, 2022. 2

[4] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2502–2511, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[5] Xiangyu Chen, Yihao Liu, Zhengwen Zhang, Yu Qiao, and Chao Dong. Hdrunet: Single image hdr reconstruction with denoising and dequantization. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 354–363, 2021. 2

[6] Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12760–12769, 2023. 1, 2, 3, 6, 7, 8

[7] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM International Conferenceon Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 369–378, 1997. 3

[8] 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*, pages 279–288, 2014. 3, 6

[9] Orazio Gallo, Alejandro Troccoli, Jun Hu, Kari Pulli, and Jan Kautz. Locally non-rigid registration for mobile hdr photography. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 48–55, 2015. 2

[10] Yulia Gryaditskaya, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel. Motion aware exposure bracketing for hdr video. *Computer Graphics Forum*, 34(4):119–130, 2015. 1

[11] Cheng Guo, Leidong Fan, Ziyu Xue, and Xiuhua Jiang. Learning a practical sdr-to-hdrtv up-conversion using new dataset and degradation models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22231–22241, 2023. 4

[12] Jin Han, Yixin Yang, Peiqi Duan, Chu Zhou, Lei Ma, Chao Xu, Tiejun Huang, Imari Sato, and Boxin Shi. Hybrid high dynamic range imaging fusing neuromorphic and conventional images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 3

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[14] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, pages 87–95, 2003. 4

[15] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1163–1170, 2013. 2

[16] Xiangyu Hu, Liquan Shen, Mingxing Jiang, Ran Ma, and Ping An. La-hdr: Light adaptive hdr reconstruction framework for single ldr image considering varied light conditions. *IEEE Transactions on Multimedia*, pages 1–16, 2022. 2, 4

[17] Anthony Huggett, Chris Silsby, Sergi Cami, and Jeff Beck. A dual-conversion-gain video sensor with dewarping and overlay on a single chip. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 52–53, 2009. 1

[18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. *Computer Graphics Forum*, 38(2):193–205, 2019. 1, 2, 3, 5, 6, 7, 8

[19] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Transactions on Graphics*, 32(202):1–8, 2013. 1, 2, 3, 7, 8

[20] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. In *ACM International Conferenceon Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 319—-325, 2003. 1, 2, 3

[21] Joel Kronander, Stefan Gustavson, Gerhard Bonnet, Anders Ynnerman, and Jonas Unger. A unified framework for multi-sensor hdr video reconstruction. *Signal Processing: Image Communication*, 29(2):203–215, 2014. 6

[22] Phuoc-Hieu Le, Quynh Le, Rang Nguyen, and Binh-Son Hua. Single-image hdr reconstruction by multi-exposure generation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 4052–4061, 2023. 2

[23] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision (ECCV)*, pages 344–360, 2022. 1, 2, 6, 7, 8

[24] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*, pages 307–314, 2010. 1, 3

[25] Rafał 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. *ACM Transactions on graphics*, 30(4):1–14, 2011. 6

[26] Rafał K. Mantiuk and Maryam Azimi. Pu21: A novel perceptually uniform encoding for adapting existing quality metrics for hdr. In *2021 Picture Coding Symposium (PCS)*, pages 1–5, 2021. 6

[27] 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*, 27(2):32–42, 2007. 1, 3

[28] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *Proceedings of 15th Pacific Conference on Computer Graphics and Applications (PG)*, pages 382–390, 2007. 5

[29] 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*, 35:46–60, 2015. 6

[30] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 472–479, 2000. 1, 3

[31] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 261–270, 2017. 2, 5

[32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1964–1980, 2021. 1, 3

[33] Erik Reinhard, Michael Stark, Peter Shirley, and James Ferwerda. Photographic tone reproduction for digital images. *ACM Transactions on Graphics*, 21(3):267–276, 2002. 2

[34] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. In *ACM International Conferenceon Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 39–es, 2007. 2

[35] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics*, 31(6):203–1, 2012. 2

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 4

[37] Laurens vd Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11): 2579–2605, 2008. 4

[38] Ruixing Wang, Xiaogang Xu, Chi-Wing Fu, Jiangbo Lu, Bei Yu, and Jiaya Jia. Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *IEEE International Conference on Computer Vision (ICCV)*, pages 9700–9709, 2021. 3

[39] Jonas Wulff and Michael J. Black. Efficient sparse-to-dense optical flow estimation using a learned basis and layers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 120–130, 2015. 5

[40] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions, 2024. 1

[41] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 6

[42] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1751–1760, 2019. 1, 2, 6, 7, 8

[43] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22211–22220, 2023. 2

[44] Nianjin Ye, Chuan Wang, Haoqiang Fan, and Shuaicheng Liu. Motion basis learning for unsupervised deep homography estimation with subspace projection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 13097–13105, 2021. 5

[45] Huanjing Yue, Yubo Peng, Biting Yu, Xuanwu Yin, Zhenyu Zhou, and Jingyu Yang. Hdr video reconstruction with a large dynamic dataset in raw and srgb domains, 2023. 2, 3, 7, 8

[46] Yunhao Zou, Chenggang Yan, and Ying Fu. Rawhdr: High dynamic range image reconstruction from a single raw image. In *IEEE International Conference on Computer Vision (ICCV)*, pages 12334–12344, 2023. 2