HDRFlow: Real-Time HDR Video Reconstruction with Large Motions

Gangwei Xu[1,2]*, Yujin Wang[2]*, Jinwei Gu[3], Tianfan Xue[3], Xin Yang[1]†

[1]School of EIC, Huazhong University of Science and Technology
[2]Shanghai AI Laboratory   [3]The Chinese University of Hong Kong

{gwxu, xinyang2014}@hust.edu.cn, wangyujin@pjlab.org.cn

{jwgu@cse, tfxue@ie}.cuhk.edu.hk

Abstract

Reconstructing High Dynamic Range (HDR) video from image sequences captured with alternating exposures is challenging, especially in the presence of large camera or object motion. Existing methods typically align low dynamic range sequences using optical flow or attention mechanism for deghosting. However, they often struggle to handle large complex motions and are computationally expensive. To address these challenges, we propose a robust and efficient flow estimator tailored for real-time HDR video reconstruction, named HDRFlow. HDRFlow has three novel designs: an HDR-domain alignment loss (HALoss), an efficient flow network with a multi-size large kernel (MLK), and a new HDR flow training scheme. The HALoss supervises our flow network to learn an HDR-oriented flow for accurate alignment in saturated and dark regions. The MLK can effectively model large motions at a negligible cost. In addition, we incorporate synthetic data, Sintel, into our training dataset, utilizing both its provided forward flow and backward flow generated by us to supervise our flow network, enhancing our performance in large motion regions. Extensive experiments demonstrate that our HDRFlow outperforms previous methods on standard benchmarks. To the best of our knowledge, HDRFlow is the first real-time HDR video reconstruction method for video sequences captured with alternating exposures, capable of processing 720p resolution inputs at 25ms. Project website: https://openimaginglab.github.io/HDRFlow/.

1. Introduction

Capturing a high dynamic range (HDR) natural scene using a standard digital camera with a limited dynamic range (LDR) often yields undesirable results: either details in
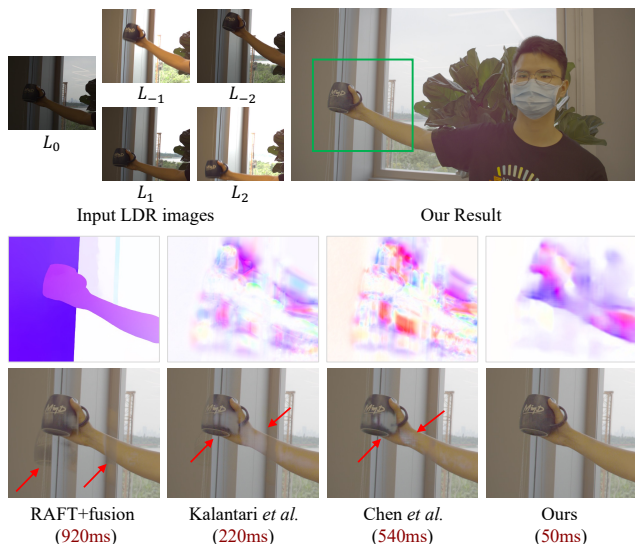

Figure 1. **Row 2:** the optical flow from frame 0 to frame -1. **Row 3:** the resulting HDR images. The methods of Kalantari *et al.* [14] and Chen *et al.* [2] struggle to predict accurate optical flow due to large motions, resulting in ghosting artifacts in the HDR output. In contrast, our HDRFlow predicts HDR-oriented optical flow and exhibits robustness to large motions. We compare our HDR-oriented flow with RAFT's [34] flow. RAFT's flow is sub-optimal for HDR fusion, and alignment may fail in occluded regions, leading to significant ghosting artifacts in the HDR output.

highlights are missing or shadows are too dark. The most prevalent solution to this issue is video HDR fusion, which merges multiple LDR images with varying exposures. With recent advances in deep learning, video HDR fusion has become a popular solution for HDR capturing and has been extensively used on recent mobile cameras.

However, in the presence of large motion, video HDR fusion algorithms still face two main challenges: achieving robust artifact-free merging and ensuring efficient processing for real-time applications. Large motions are common in amateur captures, caused by either camera motion during hand-held capture, or by object motion, such as peo-

---

*Equal contribution. This work was done when Gangwei Xu interned at Shanghai AI Laboratory.
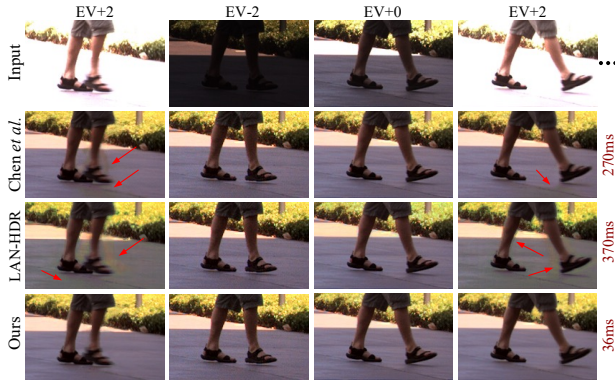
†Corresponding author.

Figure 2. HDR video reconstruction from sequences (image size: 1280×720) captured with three alternating exposures. Row 1 displays four input LDR frames. Rows 2-4 are the reconstructed HDR frames using methods Chen *et al.* [2], LAN-HDR [6] and ours.

ple walking. Without accurate alignment of neighboring frames, ghosting artifacts will appear in the fused HDR video, as demonstrated in Fig. 1 and Fig. 2. Researchers have attempted to enhance alignment robustness using advanced optical flow methods, such as SpyNet [29], employed by Kalantari *et al.* [14] and Chen *et al.* [2]. However, these flow networks are not specifically designed for HDR fusion and may struggle with inputs of varying exposure. Alternatively, other researchers [6] have attempted direct fusion of multiple frames using an attention module, but this approach is susceptible to aggregating irrelevant information due to inaccurate attention maps, leading to local inconsistencies (Fig. 2, Fig. 6). Furthermore, these alignment modules tend to be computationally intensive, especially in the presence of substantial motion. Thus, off-the-shelf flow or attention modules may be sub-optimal for HDR fusion.

For robust and efficient alignment, we design a flow estimation algorithm that is tailored for real-time HDR video reconstruction, named *HDRFlow*. The proposed HDRFlow has included three novel designs: an HDR-domain alignment loss, an efficient flow network with a multi-size large kernel, and a new HDR flow training scheme.

Firstly, to train robust alignment tailored for HDR fusion, we propose a novel HDR-domain alignment loss (HALoss). A simple solution to align input frames is to use a pre-trained model on extensive optical flow datasets, such as RAFT [34]. However, RAFT's flow is sub-optimal, and alignment may fail in occluded regions (Fig. 1). A better solution for handling occlusion and large motion is to use a task-oriented flow trained with an unsupervised approach, as illustrated by Xue *et al.* [42]. A typical unsupervised flow training calculates the photometric loss between the aligned frames, but this photometric loss relies on the brightness consistency assumption between input frames, which does not hold for HDR fusion, where input frames have different exposures. Therefore, we propose a novel HDR-domain

alignment loss that is robust to varying exposures in input and predicts an HDR-oriented flow for precise alignment.

Secondly, we introduce a novel alignment network using multi-size large kernel convolutions, which can efficiently handle large motions. To manage extensive movement, most existing optical flow methods utilize deep iterative structure [12, 13, 31–34, 37–39] or Transformer [22, 40, 41], capable of estimating flow with large motions, albeit at a high computational cost. In contrast, we present a simple flow network with multi-size large kernel convolutions that only operate on low-resolution input, proficient in handling large motions with minimal computational expense. The flow boundaries may not be as sharp as those produced by computationally expensive methods (Fig. 1), but the fused HDR image retains clarity and sharpness, showing a sharp flow boundary is not imperative for HDR fusion. Consequently, our flow network predicts bidirectional optical flows while requiring only 10 ms for 720p resolution inputs.

Thirdly, we introduce a new HDR flow training scheme that integrates both synthetic and real videos for video HDR training. Most existing video HDR networks are mostly trained on real videos, such as Vimeo-90K [42], which often lack instances of large motion. In this work, we introduced the synthetic data, Sintel [1], into our training dataset, utilizing both its provided ground-truth forward flow and backward flow generated by us to supervise our flow network. Compared to networks trained solely on real videos, it improves the flow's robustness against large motions.

Extensive experiments demonstrate that our approach surpasses state-of-the-art methods [2, 6, 14] on public benchmarks [2, 9]. Specifically, our approach exhibits superior performance in handling large motions (Fig. 1, Fig. 7). In summary, our main contributions are as follows:

- We introduce a novel HDR-domain alignment loss to supervise the flow network, enabling accurate alignment in saturated and dark regions.
- We propose a lightweight flow network with multi-size large kernel convolutions to efficiently model large motions. It predicts bidirectional optical flows with a runtime of 10ms for 720p resolution. The total HDR fusion only takes 25ms, 10× faster than the current state-of-the-art methods [2, 6].
- We propose a novel training scheme that incorporates both synthetic and real videos for training, further boosting the robustness of our network under large motions.

2. Related Work

HDR image reconstruction. The most popular way for HDR imaging is to merge multi-exposure LDR images, which is similar to HDR video reconstruction from alternating-exposure LDR frames. Early works [11, 23, 27, 30] utilize image alignment to reduce ghosting arti-
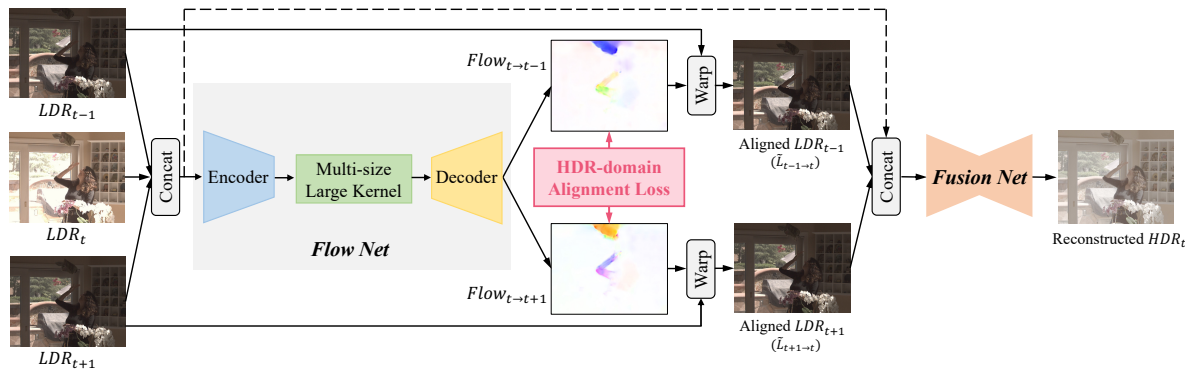
Figure 3. Network architecture of the proposed HDRFlow. We first estimate bidirectional optical flows through the proposed flow network. Then, we align the neighboring frames to the reference frame t based on these estimated flows. To achieve accurate alignment, we introduce a novel HDR-domain alignment loss to supervise our flow network. Finally, the aligned frames and the original frames are fused together through the fusion network to reconstruct a high-quality and ghost-free HDR image for the reference frame.

facts in dynamic scenes. With the rise of deep neural networks, many works [36, 43, 44] directly learn the complicated mapping between LDR and HDR using CNNs. Wu *et al.* [36] formulated HDR imaging as an image translation problem and proposed the first non-flow-based network for HDR imaging. Yan *et al.* [43, 44] proposed a spatial attention module to suppress undesired content from non-reference images and then used a non-local network to merge them. Based on this spatial attention, several methods [3, 19–21] have been proposed to remove ghosting artifacts. Unluckily, spatial attention produces unsatisfactory results when motion occurs in over-exposed regions or under-exposed regions. To mitigate this, Yan *et al.* [45] proposed to integrate similar content from other non-reference images by patch aggregation. However, these methods rely on a fixed-exposure reference frame, such as medium exposure, making it challenging to apply them to HDR video reconstruction with alternating-exposure reference frames.

HDR video reconstruction. Some existing methods rely on dedicated hardware solutions for direct HDR video acquisition, such as scanline exposure/ISO [5, 10] and internal/external beam splitter [18, 26, 35]. However, these methods require sophisticated designs and are costly.

A practical solution for HDR video reconstruction is to merge multiple LDR images with varying exposures. Kang *et al.* [16] introduced the first algorithm of this category by first aligning neighboring frames to the reference frame using global and local registration and then merging the aligned images to an HDR image. Mangiat *et al.* [24] improved this method by block-based motion estimation and refined the motion vectors. Kalantari *et al.* [15] adopted patch-based optimization to reconstruct missing images with different exposures. Recently, Kalantari *et al.* [14] presented the first end-to-end CNN-base framework that consists of a flow network for alignment and a weight network for merging the aligned LDR images. Fol-

low this, Chen *et al.* [2] performed a more sophisticated alignment by incorporating deformable convolution [7] after the coarse alignment using optical flow. Chung *et al.* [6] aligned adjacent frames to the reference frame by computing luminance-based attention score. However, these methods suffer from flow estimation errors and attention calculation errors when large motions occur. Different from them, we propose a robust HDR-oriented flow estimator to achieve precise alignment. Meanwhile, our method is significantly faster than theirs.

3. HDRFlow

In HDR fusion, the input LDR video consists of LDR frames $\{L_t\}$ captured under different exposures $\{e_t\}$, where $t = 1, \ldots, n$. Our goal is to efficiently reconstruct high-quality HDR video, consisting of HDR frames $\{H_t\}$. Following the convention [14], we evaluate two different types of input: three frames with two alternating exposures {EV-3, EV+0, EV-3, ... }, and five frames for three alternating exposures {EV-2, EV+0, EV+2, EV-2, EV+0, ... }. In both cases, the middle frame is selected as the reference, and the rest frames are fused to the reference frame to generate HDR output. For simplicity, we introduce our algorithm for handling videos captured with two alternating exposures in this paper, and discuss the extension to three exposures in the **supplementary material**.

The overview of the proposed network framework is illustrated in Fig. 3. Our framework is composed of the flow network and the fusion network. Firstly, we input three LDR frames into the flow network to estimate bidirectional optical flows, denoted as $F_{t \to t-1}$ and $F_{t \to t+1}$. These flows are utilized to warp the two neighboring frames to the reference frame. Subsequently, we feed the following information to the fusion network: the warped neighboring frames, the reference frame, the original neighboring frames in the LDR domain, and their counterparts in the linear HDR do-
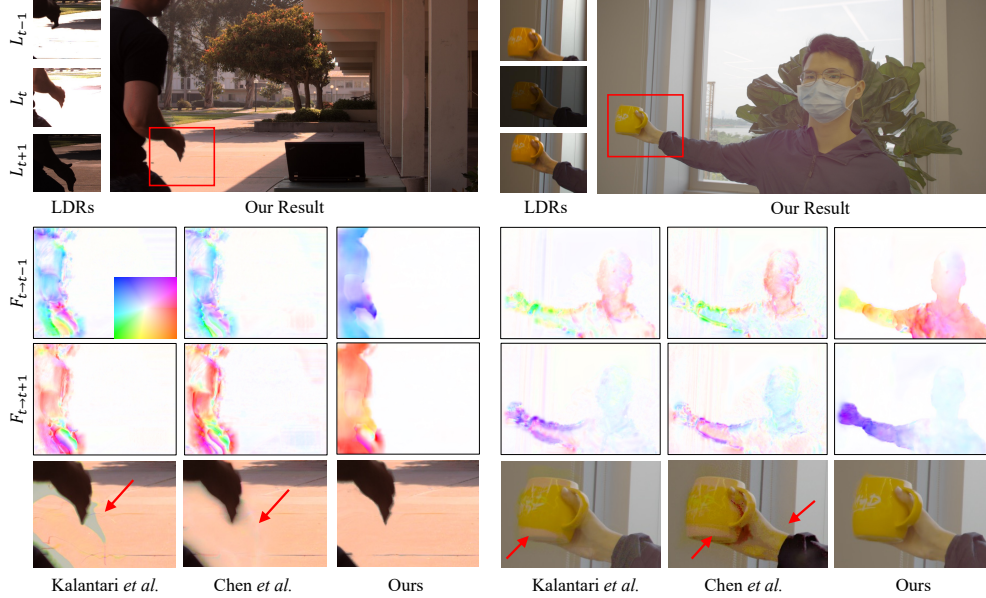
Figure 4. Qualitative comparisons with flow-based methods. Left: 3-Exposure scene from the Kalantari13 dataset [15]. Right: 2-Exposure scene from the DeepHDRVideo dataset [2]. Flow visualization is based on the color wheel shown on the corner of the first flow map.

main. The mapping function that takes the LDR frame L_t to the linear HDR domain I_t is defined as:

$$I_t = L_t^\gamma / e_t, \qquad (1)$$

where e_t is the exposure time of L_t and $\gamma = 2.2$. The fusion network estimates the blending weights, which are then utilized to fuse the linear HDR frames into the final HDR output.

3.1. Flow Network with Multi-size Large Kernel

To reconstruct the missing content at frame t, aligning neighboring frames with the reference frame is required. To accomplish this, we design an efficient flow network tailored for HDR fusion task to estimate the flow field from the reference frame t to the neighboring frames, that is, previous frame $t - 1$ and next frame $t + 1$.

Network design. The flow network consists of encoder, multi-size large kernel and decoder. Specifically, we incorporate multi-size large kernel at the end of the encoder, which can effectively model large motions at a negligible computational cost. In contrast to previous works [2, 14], our network is a single feed-forward network that does not require any iterative prediction and estimated intermediate warped frames. Thus, our flow network is efficient and can predict bidirectional optical flows within a 10ms runtime for 720p resolution. Flow network details can be found in the **supplementary materials**.

Due to the exposure difference between input frames, we adjust the exposure of the reference frame t to match neighboring frames before injecting it into the flow network,

$$g_{t+1}(L_t) = \text{clip}\Big(((L_t^\gamma / e_t)e_{t+1})^{1/\gamma}\Big), \qquad (2)$$

where $g_{t+1}(L_t)$ is adjusted reference frame. We concatenate L_{t-1}, $g_{t+1}(L_t)$ and L_{t+1} and send them into the flow network. The encoder of the flow network consists of two subnetworks, one builds a feature pyramid and another one builds an image pyramid. The feature pyramid consists of 8 residual blocks, 2 at 1/2 resolution, 2 at 1/4 resolution, 2 at 1/8 resolution, and 2 at 1/16 resolution. The image pyramid is obtained by applying pooling operations on the concatenated LDR frames. We concatenate the feature pyramid and the image pyramid at 1/4, 1/8, and 1/16 resolution. Finally, we obtain the flow feature Z_e ($Z_e \in \mathbb{R}^{H/16 \times W/16 \times 256}$). H and W represent image height and width. At the end of the encoder, we perform the proposed multi-size large kernel to increase the receptive field and model large motions.

The decoder of the flow network consists of two upsampling blocks and a flow head. After upsampling the flow feature to 1/4 resolution, the flow head is applied to predict the bidirectional optical flows. The predicted flows are at 1/4 resolution, and we upsample them to full resolution by bilinear interpolation. The final bidirectional flows are denoted as $F_{t \to t-1}$ and $F_{t \to t+1}$. The two neighboring frames can then be aligned to the reference frame as $\{\tilde{L}_{t-1 \to t}, \tilde{L}_{t+1 \to t}\}$.

Multi-size Large Kernel. In recent literature [4, 8], it has been observed that large kernel convolutions have much larger effective receptive fields, leading to improved performance in segmentation and detection areas. Inspired by

Figure 5. Effectiveness of HALoss.

this, we designed the multi-size large kernel convolutions to increase the receptive field and model large motions. As shown in Fig. 3, we place the multi-size large kernel at the coarsest resolution of the flow network to minimize computational costs. The multi-size large kernel consists of three different-sized large kernel convolutions (*i.e.*, 7×7, 9×9, and 11×11), each modeling different degrees of large motions,

$$Z_{mlk} = \text{Concat}\{\text{DConv}_{7\times7}(Z_e), \text{DConv}_{9\times9}(Z_e), \\ \text{DConv}_{11\times11}(Z_e)\}, \quad (3)$$

where DConv denotes depth-wise convolution. Then, we apply a 1×1 convolution to merge the concatenated feature Z_{mlk}, and add the merged feature to Z_e.

3.2. HDR-domain Alignment Loss

To reconstruct a high-quality, artifact-free HDR video, accurate alignment between the reference frame and neighboring frames is crucial. However, previous works [2, 6, 14] solely compute the loss between the final estimated and ground truth HDR frame, without direct supervision of the intermediate alignment. Therefore, it is challenging to achieve accurate alignment in over-exposed or under-exposed regions (see Fig. 5 and Fig. 4).

Thus, additional supervision on intermediate alignment is important. An intuitive solution is to use photometric consistency loss between the reference frame and the warped neighboring frames. However, the input LDR frames have different brightness, which violates the photometric consistency assumption.

To deal with brightness differences, we propose the HDR-domain Alignment Loss (HALoss) that is robust to brightness change. Since the training LDR frames were generated from clean HDR videos, where there are no brightness changes, we can use HDR frames to calculate photometric consistency loss. Specifically, our flow network predicts the flow field based on the input LDR frames, but the photometric consistency loss is calculated on the warped clean HDR frames, where the predicted flow fields are used for warping. For better perceptual quality, we compute the loss in the tonemapped HDR space. Following previous works [14, 36, 43], we use the differentiable μ-law function as the tonemapping function $\mathcal{T}$:

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + H)}, \quad (4)$$

where μ is set to 5,000.

Given HDR frames H_{t-1}, H_t, H_{t+1} corresponding to LDR frames L_{t-1}, L_t, L_{t+1}, as well as estimated flows $F_{t\rightarrow t-1}$ and $F_{t\rightarrow t+1}$, the HALoss $\mathcal{L}_{HA}$ is expressed as:

$$\mathcal{L}_{t,t-1}^{photo} = \| \mathcal{T}(H_t) - \mathcal{W}(\mathcal{T}(H_{t-1}), F_{t\rightarrow t-1}) \|_1,$$
$$\mathcal{L}_{t,t+1}^{photo} = \| \mathcal{T}(H_t) - \mathcal{W}(\mathcal{T}(H_{t+1}), F_{t\rightarrow t+1}) \|_1, \quad (5)$$
$$\mathcal{L}_{HA} = (1 - M_t) \odot (\mathcal{L}_{t,t-1}^{photo} + \mathcal{L}_{t,t+1}^{photo}),$$

where $\mathcal{W}(\cdot, \cdot)$ denotes the warping of neighboring frames to the reference frame using optical flow. $\mathcal{L}_{t,t-1}^{photo}$ and $\mathcal{L}_{t,t+1}^{photo}$ denote photometric loss. The M_t is a mask indicating the well-exposed regions of the reference frame t. We convert L_t to the YCbCr space to obtain the luminance channel, Y. Then, the M_t is defined as $\delta_{low} < Y < \delta_{high}$. δ_{low} and δ_{high} respectively denote the low and high luminance thresholds. Since our objective is to integrate neighboring frames' information into the reference frame in the not well-exposed regions, we only calculate the HALoss in these regions to learn the HDR-oriented optical flows. As illustrated in Fig. 5, our HALoss is effective.

3.3. Fusion Network

The objective of the fusion network is to generate a high-quality HDR frame from the reference frame, aligned neighboring frames, and original neighboring frames. In the reference LDR frame, both static regions and moving objects can experience over-exposed or under-exposed conditions. Consequently, aligned neighboring frames (*i.e.*, $\tilde{L}_{t-1\rightarrow t}$ and $\tilde{L}_{t+1\rightarrow t}$, see Fig. 3) obtained by warping operation primarily contribute to missing content in dynamic regions, while original neighboring frames (*i.e.*, L_{t-1} and L_{t+1}) mainly provide missing content in static regions.

With this consideration, five LDR frames, along with their corresponding linear HDR domain frames (Eq. (1)) are fed into the fusion network. The fusion network adopts a U-Net architecture with skip connections, comprising three downsampling blocks and three upsampling blocks. Fusion network details can be found in the **supplementary materials**. The fusion network outputs the fusion weights for five linear HDR frames. The final HDR frame $\hat{H}_t$ is computed as a weighted average of the five linear HDR frames using their fusion weights as:

$$\hat{H}_t = \frac{w_0 I_t + w_1 \tilde{I}_{t-1\rightarrow t} + w_2 \tilde{I}_{t+1\rightarrow t} + w_3 I_{t-1} + w_4 I_{t+1}}{\sum_{j=0}^4 w_j}. \quad (6)$$

3.4. Training Scheme and Loss

New training scheme. Existing video HDR networks [2, 6] are mostly trained on real videos, such as Vimeo-90K, which often lack instances of large motion. Synthetic data,

Methods	2-Exposure				3-Exposure			
	$PSNR_T$	$SSIM_T$	HDR-VDP-2	Time (ms)	$PSNR_T$	$SSIM_T$	HDR-VDP-2	Time (ms)
Kalantari13 [15]	37.51	0.9016	60.16	-	30.36	0.8133	57.68	-
Kalantari19 [14]	37.06	0.9053	70.82	230	33.21	0.8402	62.44	260
Yan19 [43]	31.65	0.8757	69.05	460	34.22	0.8604	<u>66.18</u>	-
Prabhakar [28]	34.72	0.8761	68.82	-	34.02	0.8633	65.00	-
Chen [2]	35.65	0.8949	**72.09**	550	34.15	0.8847	**66.81**	570
LAN-HDR [6]	38.22	0.9100	69.15	707	35.07	0.8695	65.42	905
Ours (Vimeo)	<u>39.20</u>	<u>0.9154</u>	70.98	55	<u>36.55</u>	<u>0.9039</u>	65.89	76
Ours (Vimeo+Sintel)	**39.30**	**0.9156**	<u>71.05</u>	55	**36.65**	**0.9055**	66.02	76

Table 1. Quantitative comparisons of our method with other state-of-the-art methods on the Cinematic Video dataset [9]. The time is the inference time for the 1920×1080 resolution dataset. **Bold**: best, <u>underline</u>: second best.

Methods	2-Exposure				3-Exposure			
	$PSNR_T$	$SSIM_T$	HDR-VDP-2	Time (ms)	$PSNR_T$	$SSIM_T$	HDR-VDP-2	Time (ms)
Kalantari13 [15]	40.33	0.9409	66.11	-	38.45	0.9489	57.31	-
Kalantari19 [14]	39.91	0.9329	71.11	200	38.78	0.9331	65.73	220
Yan19 [43]	40.54	0.9452	69.67	280	40.20	0.9531	68.23	-
Prabhakar [28]	40.21	0.9414	70.27	-	39.48	0.9453	65.93	-
Chen [2]	42.48	**0.9620**	74.80	522	39.44	**0.9569**	67.76	540
LAN-HDR [6]	41.59	0.9472	71.34	415	40.48	0.9504	68.61	525
Ours (Vimeo)	<u>43.18</u>	0.9510	<u>77.11</u>	35	<u>40.45</u>	0.9530	<u>72.30</u>	50
Ours (Vimeo+Sintel)	**43.25**	<u>0.9520</u>	**77.29**	35	**40.56**	<u>0.9535</u>	**72.42**	50

Table 2. Quantitative comparisons of our method with other state-of-the-art methods on the DeepHDRVideo dataset. The time is the inference time for the 1536×813 resolution dataset.

such as Sintel [1], include examples of large motion. In this paper, we propose to incorporate both synthetic and real videos for training. Specially, the Sintel data provides ground-truth forward flow ($F_{t \to t+1}^{gt}$), and we generate backward flow ($F_{t \to t-1}^{gt}$) using the pre-trained RAFT [34] flow network. Then, we use the forward and backward flow supervision to further enhance the accuracy and robustness of optical flow.

Total loss. The total loss consists of reconstruction loss, alignment loss, and flow loss. We compute the reconstruction loss $\mathcal{L}_{rec}$ between the predicted $\hat{H}_t$ and ground-truth H_t^{gt} using $\mathcal{L}_1$ loss:

$$\mathcal{L}_{rec} = \| \mathcal{T}(\hat{H}_t) - \mathcal{T}(H_t^{gt}) \|_1 . \quad (7)$$

For Sintel data, we additionally compute the loss between predicted flow and ground truth flow to further improve the flow's robustness against large motions. The flow loss $\mathcal{L}_{flow}$ is defined as:

$$\mathcal{L}_{flow} = \| F_{t \to t-1} - F_{t \to t-1}^{gt} \|_1 + \| F_{t \to t+1} - F_{t \to t+1}^{gt} \|_1 . \quad (8)$$

The total loss $\mathcal{L}_{total}$ is represented as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{HA} + \lambda_3 \mathcal{L}_{flow}, \quad (9)$$

where $\lambda_1 = 1$, $\lambda_2 = 0.5$, and $\lambda_3 = 0.001$.

4. Experiments

4.1. Experimental Setup

Datasets. We utilize the high-quality Vimeo-90K [42] and Sintel [1] datasets as our training sets. As the Vimeo-90K and Sintel datasets are not tailored for HDR video reconstruction, we convert the original data to LDR sequences with alternating exposures following the previous work[2, 14]. During training, we first apply random horizontal/vertical flipping and rotation, and then randomly crop the resulting images to obtain patches of size 256×256 as inputs to the network. We evaluate our method on two synthetic videos (POKER FULLSHOT and CAROUSEL FIREWORKS) from the Cinematic Video dataset [9] and DeepHDRVideo dataset [2]. The resolution of Cinematic Video dataset is 1920×1080, and the resolution of Deep-HDRVideo dataset is 1536×813. The DeepHDRVideo dataset consists of both real-world dynamic scenes and static scenes that have been augmented with random global motion. We also utilize the HDRVideo dataset [15], which has a resolution of 1280×720, for qualitative evaluation.

Model	FN	MLK	HALoss	Mask for HALoss	DeepHDRVideo-D		Cinematic Video		Time
					$PSNR_T$	$SSIM_T$	$PSNR_T$	$SSIM_T$	(ms)
Base (SPyNet)					44.82	0.9650	38.78	0.9137	73
FN	✓				45.10	0.9655	38.92	0.9133	34
FN+MLK	✓	✓			45.30	0.9661	39.08	0.9142	35
FN+MLK+HA	✓	✓	✓		45.41	0.9669	39.21	0.9148	35
Full model (HDRFlow)	✓	✓	✓	✓	**45.50**	**0.9683**	**39.30**	**0.9156**	35

Table 3. Ablation study of HDRFlow on the dynamic scenes of DeepHDRVideo dataset and Cinematic Video dataset. FN denotes our flow network, and HALoss denotes HDR-domain alignment loss. Following the previous works [2, 14], the Base (SPyNet) utilizes SPyNet [29] as the optical flow estimator and does not incorporate HALoss. The time is the inference time for 1536×813 resolution. **Bold**: Best.

Implementation details. We implement our approach with PyTorch and perform our experiments using an NVIDIA 3090 GPU. We adopt AdamW optimizer [17] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train our network with 40 epochs using a batch size of 16. The learning rate was initially set to 0.0001 and halved after epochs 20 and 30. In our experiments, we empirically set δ_{low} to 0.2 and δ_{high} to 0.8.

Evaluation metrics We adopt $PSNR_T$, $SSIM_T$ and HDR-VDP-2 [25] as the evaluation metris. $PSNR_T$, $SSIM_T$ are computed in the μ-law tonemapped domain.

4.2. Comparisons with State-of-the-art

Quantitative evaluations. Tab. 1 and Tab. 2 show our quantitative results on the Cinematic Video [9] and Deep-HDRVideo [2] datasets. Ours (Vimeo) denotes that we only use Vimeo-90K [42] as our training dataset, which is consistent with other methods. Ours (Vimeo+Sintel) denotes that we use both Vimeo-90K and Sintel [1] as our training datasets. The performance can be further improved by incorporating Sintel into the training dataset. Our method achieves superior or comparable results to state-of-the-art methods. Specifically, on the Cinematic Video dataset, our method outperforms the second-best method by up to 1.08dB and 1.58dB in terms of $PSNR_T$ for the 2-Exposure and 3-Exposure cases, respectively. On the DeepHDRVideo dataset, our method has also achieved the best results among all methods in terms of $PSNR_T$ and HDR-VDP-2.

Qualitative evaluations. We compared the visual results of our approach with flow-based methods [2, 14] in Fig. 4, and with attention-based methods [6] in Fig. 6. As shown in Fig. 4, the optical flows predicted by the methods of Kalantari *et al*. and Chen *et al*. are relatively discontinuous, lacking smoothness and completeness. Consequently, their methods exhibit ghosting artifacts in regions with large motions and lose details in saturated regions. In contrast, our predicted optical flow is more accurate and smooth, enabling precise alignment in regions with large motions. Without estimating optical flow, LAN-HDR [6] performs alignment using an attention module. However, using attention easily leads to the aggregation of irrelevant information, resulting in local inconsistency. As shown in Fig. 6,

LAN-HDR [6] introduces color distortion due to the aggregation of grass content onto the ground. Additionally, LAN-HDR struggles to handle noise in extremely dark areas. On the contrary, our method can produce pleasing results in these regions.

Inference time. To demonstrate the high efficiency of our method, we compare the inference time of our method with other HDR video reconstruction methods in Tab. 1 and Tab. 2. For fair comparisons, the inference times of all methods are tested on a single NVIDIA 3090 GPU. Our method is approximately $10\times$ faster than both Chen [2] and LAN-HDR [6]. To the best of our knowledge, our method is the first real-time HDR video reconstruction method for video sequences captured with alternating exposures.

4.3. Ablation Study

We conduct ablation studies to validate the effectiveness of the proposed components with the example of sequences having two alternating exposures. All the quantitative evaluations are conducted on the Cinematic Video dataset and the dynamic scenes of DeepHDRVideo dataset, which contains large motions. We take Base (SPyNet) as the baseline. The Base (SPyNet) uses SPyNet [29] as flow estimator, and the fusion network is consistent with ours. SPyNet [29] utilizes a hierarchical coarse-to-fine architecture that needs iterative prediction and estimated intermediate warped frames. This process is computationally intensive and cannot achieve real-time performance. Furthermore, saturated and dark regions in LDR images lack texture information to the extent that at finer resolutions, SPyNet [29] struggles to accurately estimate optical flow relying solely on image information. In contrast, our flow network (FN) efficiently leverages multi-resolution feature information to estimate accurate optical flow. Compared to SPyNet (Tab. 3), our FN obviously improves performance and greatly reduces inference time. When incorporating the proposed MLK into FN, our model better models large motions, leading to further improvements in the results.

To achieve precise alignment, we introduce a novel HDR-domain Alignment Loss (HALoss). As shown in Tab. 3, HALoss significantly improves performance on
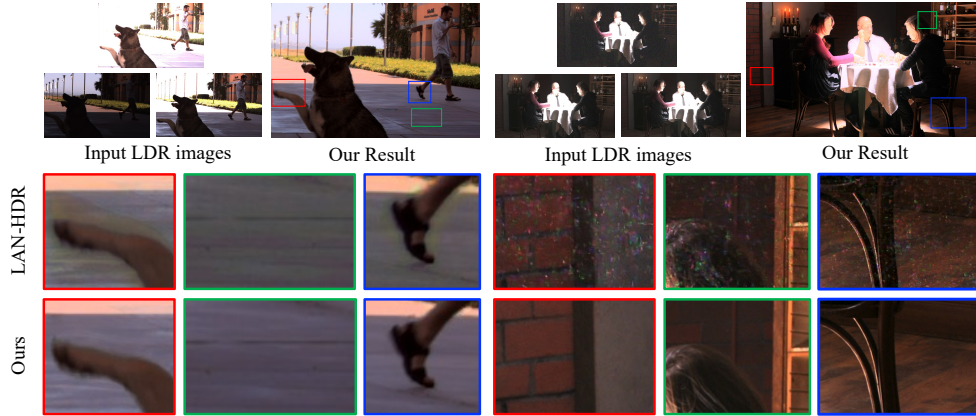
Figure 6. Qualitative comparisons with attention-based method [6].

| Method | DeepHDRVideo-D | | Cinematic Video | | Time |
	$PSNR_T$	$SSIM_T$	$PSNR_T$	$SSIM_T$	(ms)
RAFT+fusion	44.85	0.9634	38.63	0.9131	532
Ours	**45.50**	**0.9683**	**39.30**	**0.9156**	**35**

Table 4. Quantitative comparison with RAFT+fusion.

benchmarks and does not increase the inference time. Since our objective is to integrate neighboring frames' information into the reference frame in the not-well-exposed regions, we define a luminance mask that indicates over-exposed and under-exposed regions, and only compute HALoss within these regions to learn the HDR-oriented optical flow. This mask further improves the results on benchmarks. Visual comparisons are shown in Fig. 5.

4.4. Analysis

Comparison with state-of-the-art flow method. We compare our method with the RAFT+fusion method, which is constructed by using a pre-trained RAFT flow network as the flow estimator and employing the same fusion network as in our approach. For training RAFT+fusion, we freeze the optical flow network and only update the fusion network parameters. Quantitative comparisons are shown in Tab. 4. Our HDRFlow not only outperforms RAFT+fusion but is also about $15\times$ faster than it. Visual comparisons are shown in Fig. 1, RAFT can effectively match visible objects, exhibiting clear flow boundaries, but this flow is incapable of handling occlusions during the alignment. As a result, the fused HDR image exhibits ghosting artifacts in occluded regions. Furthermore, RAFT performs poorly in large textureless areas due to limited receptive field.

Robustness to large motions. We evaluate the performance of our method with previous flow-based methods [2, 14] at different motion magnitudes (Fig. 7). To construct the evaluation dataset, we use RAFT [34] to process dynamic scenes of DeepHDRVideo and obtain optical flow maps. Then, we manually crop out these images with rea-
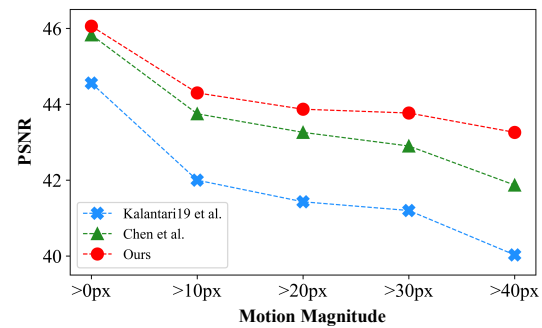

Figure 7. Comparisons with flow-based methods [2, 14] across different motion magnitude ranges.

sonable flow predictions. We divide cropped images into 128×128 blocks and calculate the average motion magnitude within each block. Finally, we evaluate the PSNR of the blocks corresponding to different motion magnitudes. As shown in Fig. 7, our HDRFlow is more robust compared to other methods as the motion magnitude increases.

5. Conclusion

In this paper, we have proposed a robust and efficient flow estimation algorithm tailored for real-time HDR video reconstruction, named HDRFlow. The HDRFlow has introduced three novel designs: an HDR-domain alignment loss, an efficient flow network with a multi-size large kernel, and a new HDR flow training scheme. Extensive experiments demonstrate that our approach surpasses state-of-the-art methods on public benchmarks. In particular, our approach exhibits superior performance in handling large motion regions. To the best of our knowledge, HDRFlow is the first real-time HDR video reconstruction method.

References

[1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 611–625. Springer, 2012. 2, 6, 7, 3

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2502–2511, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[3] Jie Chen, Zaifeng Yang, Tsz Nam Chan, Hui Li, Junhui Hou, and Lap-Pui Chau. Attention-guided progressive neural texture fusion for high dynamic range image restoration. *IEEE Transactions on Image Processing*, 31:2661–2672, 2022. 3

[4] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13488–13498, 2023. 4

[5] Inchang Choi, Seung-Hwan Baek, and Min H Kim. Reconstructing interlaced high-dynamic-range video using joint learning. *IEEE Transactions on Image Processing*, 26(11): 5353–5366, 2017. 3

[6] Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12760–12769, 2023. 2, 3, 5, 6, 7, 8, 1

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[8] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022. 4

[9] 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. SPIE, 2014. 2, 6, 7

[10] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pająk, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, 33(6):1–13, 2014. 3

[11] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1163–1170, 2013. 2

[12] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 668–685. Springer, 2022. 2

[13] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. 2

[14] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, pages 193–205. Wiley Online Library, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[15] 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 (TOG)*, 32(6):202–1, 2013. 3, 4, 6

[16] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 3

[17] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, page 6. San Diego, California;, 2015. 7

[18] 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. 3

[19] Shuaizheng Liu, Xindong Zhang, Lingchen Sun, Zhetong Liang, Hui Zeng, and Lei Zhang. Joint hdr denoising and fusion: A real-world mobile hdr image dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13966–13975, 2023. 3

[20] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. Adnet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 463–470, 2021.

[21] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *Proceedings of the European Conference on Computer Vision*, pages 344–360. Springer, 2022. 3

[22] Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, and Dongfang Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18063–18073, 2023. 2

[23] Kede Ma, Hui Li, Hongwei Yong, Zhou Wang, Deyu Meng, and Lei Zhang. Robust multi-exposure image fusion: a structural patch decomposition approach. *IEEE Transactions on Image Processing*, 26(5):2519–2532, 2017. 2

[24] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*, pages 307–314. SPIE, 2010. 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 (TOG)*, 30(4):1–14, 2011. 7

[26] 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. 3

[27] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1219–1232, 2014. 2

[28] K Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4875–4885, 2021. 6

[29] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017. 2, 7, 1

[30] 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 (TOG)*, 31(6):203–1, 2012. 2

[31] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023. 2

[32] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023.

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018.

[34] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 2, 6, 8, 3

[35] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *ACM Transactions on Graphics (TOG)*, 30(4):1–10, 2011. 3

[36] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision*, pages 117–132, 2018. 3, 5

[37] Gangwei Xu, Shujun Chen, Hao Jia, Miaojie Feng, and Xin Yang. Memory-efficient optical flow via radius-distribution orthogonal cost volume. *arXiv preprint arXiv:2312.03790*, 2023. 2

[38] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21919–21928, 2023.

[39] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[40] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8121–8130, 2022. 2

[41] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[42] 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. 2, 6, 7

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 3, 5, 6

[44] Qingsen Yan, Dong Gong, Javen Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Dual-attention-guided network for ghost-free high dynamic range imaging. *International Journal of Computer Vision*, pages 1–19, 2022. 3

[45] Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, and Yanning Zhang. A unified hdr imaging method with pixel and patch level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22211–22220, 2023. 3