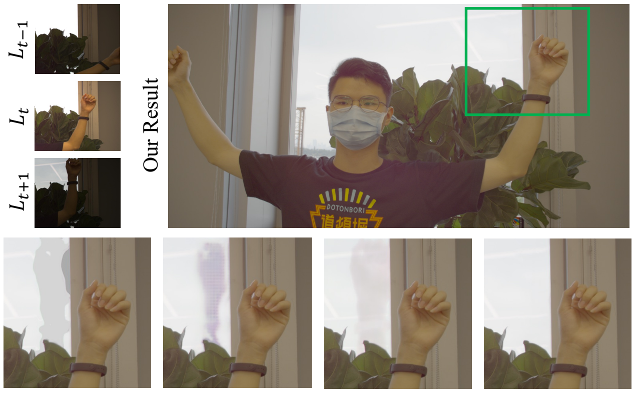
# HDRFlow: Real-Time HDR Video Reconstruction with Large Motions

## Supplementary Material



Kalantari *et al.* Chen *et al.* LAN-HDR Ours

Figure 8. Qualitative comparisons on the DeepHDRVideo dataset (3-Exposure). Compared to previous methods [2, 6, 14], our approach produces ghosting-free results under large motions.

## 6. More Experimental Results

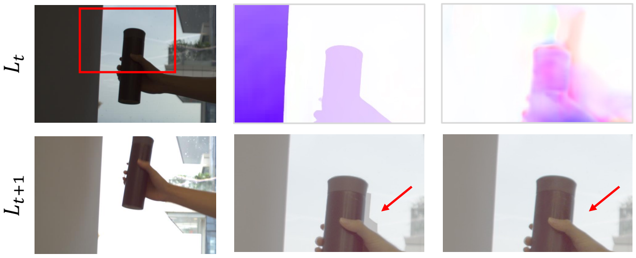
### 6.1. More Comparisons with Previous Methods

As shown in Fig. 8, we provide more visual comparisons with previous methods, Kalantari19 [14], Chen21 [2], and LAN-HDR [6]. The previous methods struggle to handle large motions, resulting in ghosting artifacts in the final HDR output. In comparison, our method produces high-quality, ghosting-free HDR results. We also provide more visual comparisons with flow-based methods [2, 14], shown in Fig. 10. The optical flows predicted by the methods of Kalantari *et al*. [14] and Chen *et al*. [2] are discontinuous, lacking smoothness and completeness. As a result, their HDR outputs exhibit ghosting artifacts and noise, and lose details in saturated regions. In comparison, our predicted flows are more accurate and smooth, enabling precise alignment in regions with large motions.

Fig. 9 shows the comparison between our method and RAFT+fusion. RAFT's [34] flow is sub-optimal, and alignment may fail in occluded regions. In contrast, our method effectively handles occluded regions by learning an HDR-oriented flow.

### 6.2. Runtime of Each Module

We benchmark the runtime of each module during inference for the 2-Exposure case. As shown in Tab. 5, our Flow Network only takes 10ms and 15ms for resolutions of 1280 × 720 and 1536 × 813, respectively. This is faster than most existing flow methods [29, 34].



LDRs RAFT+fusion Ours

Figure 9. Comparisons with RAFT+fusion. We construct RAFT+fusion by using a pre-trained RAFT [34] flow network as the flow estimator and employing the same fusion network as in our approach. 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. In comparison, our method effectively handles occluded regions by learning an HDR-oriented flow.

| Module | 1280 × 720 | 1536 × 813 |
|---|---|---|
| Flow Net with MLK | 10 ms | 15 ms |
| Fusion Net | 15 ms | 20 ms |

Table 5. Runtime time analysis of each module for 2-Exposure case. The input resolutions are 1280×720 and 1536×813, respectively.

## 7. Network Details for the Proposed HDRFlow

### 7.1. Details of Flow Network with Multi-size Large Kernel

Fig. 12 shows the architecture of our 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 corresponding channel numbers are 32, 64, 128, and 256, respectively. 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 at the 1/16 resolution.

Then, we perform the multi-size large kernel convolutions to increase the receptive field and model large motions. 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. We utilize depth-wise convolutions, which almost do not increase

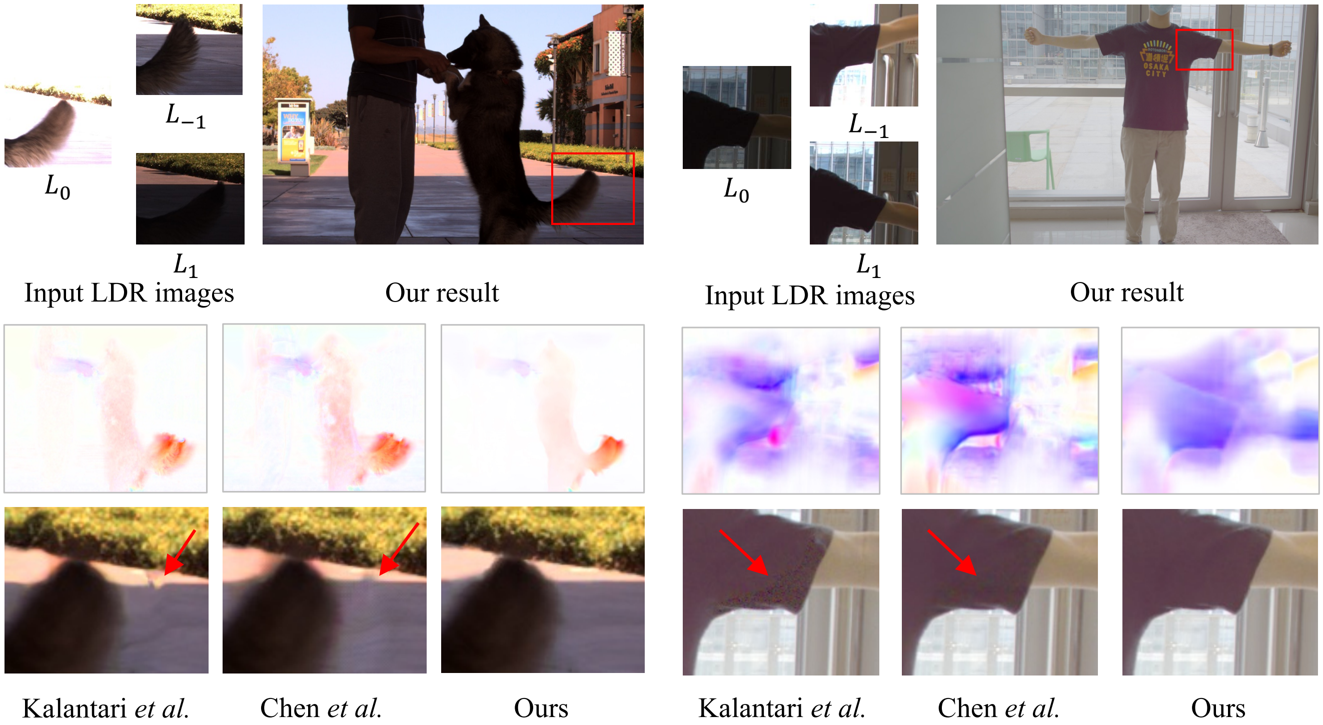


Figure 10. Comparisons with the state-of-the-art methods. The optical flows predicted by the methods of Kalantari *et al*. [14] and Chen *et al*. [2] are discontinuous, lacking smoothness and completeness. As a result, their HDR outputs exhibit ghosting artifacts and noise, and lose details in saturated regions. In comparison, our predicted flows are more accurate and smooth, enabling precise alignment in regions with large motions. Thus, our method produces high-quality HDR output.

the computational costs, shown in Fig. 12.

The decoder of the flow network consists of two upsampling blocks and a flow head. Each upsampling block has a $4 \times 4$ kernel deconvolution with a stride of 2. After each upsampling block, features are concatenated with a skip-connection, and a $1 \times 1$ convolution followed by a $3 \times 3$ convolution is applied to merge the skipped and upsampled features for the current resolution. The upsampled flow features are at resolutions of 1/8 and 1/4, with channel numbers of 128 and 64, respectively. After upsampling the flow feature to 1/4 resolution, the flow head is applied to predict the bidirectional optical flows. The flow head consists of three $5 \times 5$ kernel convolutions.

### 7.2. Details of Fusion Network

The fusion network adopts a U-Net architecture with skip connections, comprising three downsampling blocks and three upsampling blocks. In more detail, each downsampling block consists of a $3 \times 3$ convolution with a stride of 2, followed by a $3 \times 3$ convolution with a stride of 1. After three downsampling blocks, we obtain features at three different resolutions: 1/2, 1/4, and 1/8 of the original resolution. The corresponding channel numbers for these resolutions are 32, 64, and 128, respectively. Each upsampling block consists of a $4 \times 4$ deconvolution with stride 2, followed by a $3 \times 3$ convolution with stride 1. The fusion network outputs the fusion weights for five LDR frames in the linear domain.

### 7.3. Generation of Aligned Neighboring Frames

We use predicted bidirectional optical flows, $F_{t\to t-1}$ and $F_{t\to t+1}$, to align neighboring frames to reference frame via warping operation,

$$\begin{aligned} \tilde{L}_{t-1\to t} &= \mathcal{W}(L_{t-1}, F_{t\to t-1}), \\ \tilde{L}_{t+1\to t} &= \mathcal{W}(L_{t+1}, F_{t\to t+1}). \end{aligned} \tag{10}$$

The $\tilde{L}_{t-1\to t}$ and $\tilde{L}_{t+1\to t}$ are aligned neighboring frames.

### 7.4. Optical Flow Labels for Sintel

We use the Sintel dataset as our training dataset. As shown in Fig. 11, the Sintel dataset provides ground-truth forward flow ($F_{t\to t+1}$, the second row of Fig. 11). However, the Sintel does not provide backward flow. To train our flow network, we use pre-trained RAFT [34] flow network to generate backward flow ($F_{t\to t-1}$, the third row of Fig. 11) as pseudo-labels.

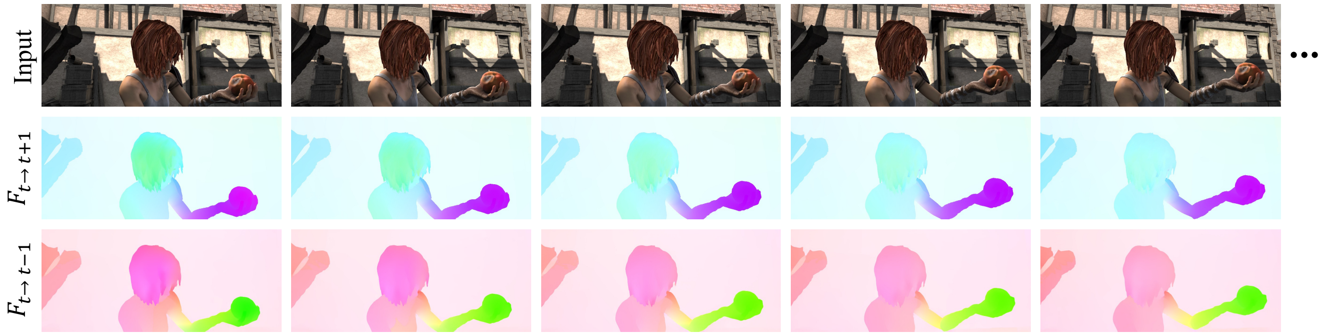


Figure 11. Optical flow labels for Sintel [1] dataset. The first row is video frames of Sintel dataset, and the second row is ground-truth forward optical flow from frame t to t+1. The Sintel does not provide backward flow. Therefore, we use pre-trained RAFT [34] flow network to generate backward optical flow from frame t to t-1 as pseudo-labels, shown in the third row.

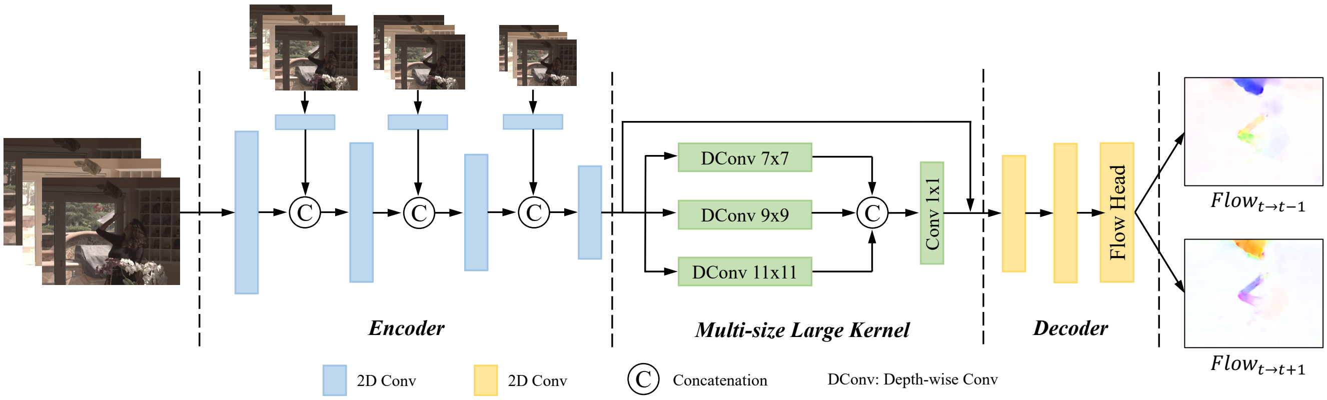


Figure 12. Flow Network with Multi-size Large Kernel. The flow network consists of the encoder, a multi-size large kernel, and a decoder. The flow network takes LDR images as input and outputs bidirectional optical flows.

### 7.5. Extension to Three Exposures

We have illustrated our HDRFlow for handling videos captured with two alternating exposures in the paper. Here we discuss the extension to three exposures.

**Review of two-exposure model** For sequences captured with two alternating exposures (*e.g.*, {EV-3, EV+0, EV-3, ...}), our flow network takes three LDR frames $\{L_{t-1}, L_t, L_{t+1}\}$ as input and estimates the optical flow, $F_{t\rightarrow t-1}$ and $F_{t\rightarrow t+1}$. Then, we align the neighboring frames $\{L_{t-1}, L_{t+1}\}$ to the reference frame $t$ based on these estimated flows. Finally, the aligned frames (2 images) and the original frames (3 images) in the linear domain are fused together through the fusion network to reconstruct a high-quality and ghost-free HDR image for the reference frame.

**HDRFlow for sequences with three exposures** For sequences with three alternating exposures (*e.g.*, {EV-2, EV+0, EV+2, EV-2, EV+0, ...}), our HDRFlow takes five frames $\{L_{t-2}, L_{t-1}, L_t, L_{t+1}, L_{t+2}\}$ as input and estimates the HDR image for the reference frame $t$. Specifically, we adjust the exposure of the reference frame $t$ to match neighboring frames before injecting it into the flow network. Thus, the flow network takes $\{L_{t-2}, g_{t+1}(L_t), L_{t+1}\}$ and $\{L_{t-1}, g_{t+2}(L_t), L_{t+2}\}$ as input and estimates four flow maps, $F_{t\rightarrow t-2}$, $F_{t\rightarrow t-1}$, $F_{t\rightarrow t+1}$, and $F_{t\rightarrow t+2}$. The four neighboring frames can then be aligned to the reference frame as $\{\tilde{L}_{t-2\rightarrow t}, \tilde{L}_{t-1\rightarrow t}, \tilde{L}_{t+1\rightarrow t}, \tilde{L}_{t+2\rightarrow t}\}$ using the estimated flows. The aligned frames (4 images) and the original input frames (5 images) in both the LDR and linear HDR domain are used as the input (54 channels) for the fusion network to estimate 9 fusion weight maps. Then, the HDR image for the reference frame $t$ can be reconstructed as the weighted average of 9 input images in the linear domain. The overall architectures of both the flow network and fusion network remain consistent between sequences with two and three exposures. The sole distinction lies in the channel numbers at the input and output layers.