LRHDR: Learning Representation-enhanced HDR Video Reconstruction

Chenzhuo Liao[1], Xin Chen[2], Bingchen Li[1], Yu Meng[3], Tao Yue[1] , Xuemei Hu[1*]
[1] Nanjing University, [2] OPPO, [3] Jiangsu Ocean University.
{chenzhuoliao, 502023250092}@smail.nju.edu.cn, chenxin15@oppo.com,
mengyu2614@163.com, {yuetao,xuemeihu}@nju.edu.cn

Abstract

Reconstructing High Dynamic Range (HDR) video from alternately exposed Low Dynamic Range (LDR) frames is challenged by large motion, exposure-induced photometric inconsistency, and information loss in saturated or under-exposed regions. Prior HDR video pipelines typically follow an alignment–reconstruction paradigm, which is limited by the precision of alignment and the performance of the fusion module. We propose a new reconstruction framework called Learning Representation-enhanced HDR Video Reconstruction (LRHDR), which is built around two novel components: an Amalgamated Cross-exposure Consistent Representation (ACCR) network and an Adaptive Pixel-wise Sparse Weighted Fusion (APSWF). The ACCR includes an Exposure-aware Interleaved Context (EIC) encoder and a Representation Mapper (RM). The EIC couples a large-field path with a high-fidelity sub-pixel path and an exposure gate to produce exposure-aware features. The RM avoids explicit cross-exposure alignment by mapping features from different exposures into a unified representation via per-pixel, per-channel linear modulation and decoding into the calibrated linear HDR domain. The APSWF treats fusion as pixel-wise candidate selection, producing sparse weighted masks to form a normalized fusion in the linear HDR domain, thereby suppressing artifacts. Extensive experiments on standard benchmarks demonstrate that our LRHDR outperforms previous methods.

1. Introduction

High dynamic range (HDR) imaging aims to reproduce the true dynamic range of natural scenes with higher luminance contrast and richer details. Standard cameras capture only a narrow dynamic range, yielding low dynamic range (LDR) measurements. While dedicated HDR hardware (e.g., specialized sensors, multi-aperture optics) directly samples a wider range [3, 5, 12, 25, 37, 45], the wide application of these sensors has been constrained by the necessity for specifically tailored hardware components. Thus, multi-exposure HDR fusion stands as a pragmatic alternative for HDR capture, owing to its inherent convenience, versatility across diverse scenarios, and freedom from specialized hardware requirements. By capturing multiple exposures, scene information across different dynamic ranges can be collected, thereby facilitating HDR synthesis.

Nevertheless, reliance on multi-frame fusion leads to degraded reconstruction quality in dynamic scenes. Moreover, misalignment and photometric inconsistency among multi-exposure frames pose additional challenges for multi-frame alignment and fusion. Most existing HDR video methods adopt a reference-centered alignment paradigm. Some approaches perform patch-based matching[15], optical-flow-driven warping[2, 6, 14, 39], or attention-based feature alignment[4] to explicitly register neighboring frames to a reference. Others rely on two-stage alignment[34], feature-level bidirectional alignment[18], or state-space models that propagate information from the reference frame to adjacent frames[19]. In all cases, the reconstruction pipeline is designed either to align frames from different exposure modes into the reference frame or to use the reference frame to attenuate mismatched information in neighboring frames.

The information loss by over-exposure and under-exposure, along with inter-frame photometric inconsistency from varying exposures, renders pixel-level alignment highly ill-posed. Furthermore, pixel misalignment and information loss pose significant challenges to existing fusion and reconstruction methods, which have to deal with the misalignment and uninformative regions during the fusion.

To address these challenges, we tailor all sub-modules for HDR video reconstruction to tackle inter-frame motion and exposure variations, with three optimized novel designs spanning feature extraction, representation, and fusion. The Exposure-aware Interleaved Context (EIC) encoder, optimized for feature extraction, enhances feature quality. The Representation Mapper (RM) refines cross-exposure features by avoiding explicit cross-exposure alignment and mapping them into a unified representation, thereby reduc-

*Corresponding author.

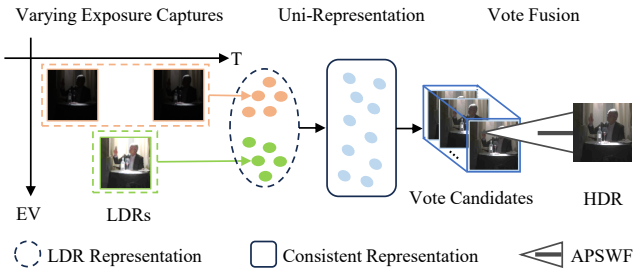

Figure 1. From an alternately exposed LDR stream, we first encode the frames, mapping them from the LDR domain to a cross-exposure-consistent unified representation, which collapses photometric gaps. The calibrated features generate a compact set of HDR candidates. APSWF then performs pixel-wise weighting with sparse masks and carries out normalized fusion in the linear HDR domain to produce the final HDR frame. This pipeline avoids explicit cross-exposure alignment, suppresses ghosting under large motion, and maintains clean edges in both saturated and under-exposed regions.

ing distortion from forced alignment and enabling complementary information from different exposures to be used more reliably. The Adaptive Pixel-wise Sparse Weighted Fusion (APSWF) improves reconstruction reliability by casting fusion as pixel-wise candidate selection. It produces sparse weights and forms a normalized fusion in the linear HDR domain, ensuring physical plausibility while suppressing artifacts. This step-by-step design substantially alleviates the adverse effects of motion and exposure inconsistency in multi-exposure HDR video reconstruction.

Our contributions are concluded as:

- We propose an ACCR network that consists of EIC and RM for extracting exposure-aware features and then mapping the features into a unified representation, and decodes them in the calibrated linear HDR domain.
- We propose APSWF, which improves reconstruction reliability by formulating fusion as pixel-wise candidate selection with sparse weighting. It performs normalized fusion in the linear HDR domain, ensuring physical plausibility while suppressing artifacts.
- Extensive experiments are conducted to demonstrate the effectiveness of the proposed method.

2. Related Work

Multi-exposure HDR image reconstruction. HDR image reconstruction has been widely studied, and most methods consider multiple images with different exposures over time to reconstruct an HDR image, without specialized sensors. To avoid artifacts of ghosting and blurring, most methods take alignment into account. Early works [9, 33] reduce artifacts in a patch-based way. Learning-based multi-frame HDR then replaced hand-crafted rules with deep fusion. Kalantari *et al*. [13] introduced the first deep-learning method for HDR fusion of temporally multi-exposed images using optical flow. Wu *et al*. [38] propose to formulate HDR imaging as an image translation problem without optical flow based on an encoder-decoder network. Yan *et al*. [40, 42] adopt a spatial attention module instead of optical flow to discard the interference from motion. Besides, Yan *et al*. [41] attempt to handle the large motion, via extracting global information through non-local module. Further, [20, 35, 36, 43] use transformer-based mechanisms instead of optical flow to enhance alignment. Hu *et al*. [10] use frequency domain processing to reconstruct the HDR image. Kong *et al*. [16] guide fusion with exposure priors. Li *et al*. [17] use deep unfolding paradigm for HDR image reconstruction. Related candidate-selection and reliability-aware fusion ideas have also been explored in HDR deghosting, including hard-voting based fusion and uncertainty-aware HDR merging [1, 29]. However, most HDR image rendering pipelines assume a fixed-exposure reference frame, which does not carry over to HDR video where the reference exposure alternates over time.

Multi-exposure HDR video reconstruction. HDR video reconstruction has received relatively less attention and exploration, compared with HDR image reconstruction. Kalantari *et al*. [14] adopt a multiscale optical flow network to align neighboring frames and U-Net [32] to fuse multi-exposure frames, which is the first work for video HDR reconstruction based on deep learning. Chen *et al*. [2] propose to align by stages, which first coarsely aligns via optical flow and then utilizes the deformable convolution [44] for fine alignment. Chung *et al*. [4] propose a Luminance-based Alignment Network that combines the attention and hallucination modules for the alignment. Xu *et al*. [39] propose a lightweight flow network with multi-size large kernel convolutions to efficiently model large motions. Shu *et al*. [34] propose a two-stage strategy to perform alignment sequentially to handle complex motion. Cui *et al*. [6] propose an exposure completing method which completes the missing exposure information by interpolating LDR frames to render HDR videos. Lin *et al*. [19] propose a state-space–model method, modeling long-range dependencies across both spatial and temporal dimensions to handle ghosting artifacts. Lin *et al*. [18] propose a lightweight method that employs a bidirectional progressive propagation scheme, combined with adaptive alignment and attention-based fusion. Different from them, we propose a framework that maps inputs from different exposures into a unified representation to make full use of each other's information instead of explicit alignment.

3. Learning Representation HDR Network

In this paper, we propose an algorithm for reconstructing an HDR video $\{H_t|t = 1, 2, ...\}$ from an LDR video $\{L_t|t = 1, 2, ...\}$ captured with a conventional camera operating in alternating exposure modes. Let the exposure

modes be $\{e_i\}_{i\in\Theta}$, where e_i is the i-th mode exposure time, $\Theta = \{0, 1, \ldots, N_e - 1\}$ and N_e denotes the number of exposure levels in each cycle. Each LDR frame L_t is acquired under one exposure mode, and the modes repeat cyclically throughout the sequence.

In practice we use $N_e \in \{2, 3\}$ to cover high- and low-luminance regions. For $N_e = 2$, the cycle alternates between EV-3 and EV+0, and for $N_e = 3$, the cycle is EV-2, EV+0, EV+2. For clarity and brevity, this paper details the algorithm for two-exposure alternating mode and defers the treatment of the three-exposure case to the **Supplementary Material**.

Overview. The overall architecture of the proposed framework is shown in Fig. 2. The framework consists of a pretrained Interpolator FiLM [31], a trainable ACCR, a trainable APSWF. At time t, a three-frame sliding window $L_{t-1}^{e_0}, L_t^{e_1}, L_{t+1}^{e_0}$ is fed into the framework, where $L_t^{e_1}$ serves as the reference frame. We obtain a non-reference exposure intermediate $\hat{L}_t^{e_0}$ from $L_{t-1}^{e_0}, L_{t+1}^{e_0}$ using the pretrained Interpolator, providing reference-free motion information. Then $\{L_t^{e_1}, \hat{L}_t^{e_0}\}$ are fed into ACCR to obtain the unified representation linear HDR frames $\{\tilde{H}_t^{e_1}, \tilde{\hat{H}}_t^{e_0}\}$. Then APSWF fuses the original inputs after transforming them to the linear HDR domain. The transformation is defined as

$$\Gamma(L) = L^{\gamma}/e, \tag{1}$$

where L is the LDR input, e is the exposure time, and $\gamma = 2.2$. APSWF then performs fusion in a voting manner on these linear HDR inputs.

3.1. Amalgamated Cross-exposure Consistent Representation.

Forcing inputs from different exposure modes to explicitly align to a single reference frame can intensify ghosting artifacts and place greater strain on subsequent reconstruction modules. To avoid explicit cross-exposure alignment, we propose ACCR, which maps each exposure stream into an exposure-independent unified representation domain.

EIC Encoder. EIC is an encoder tailored for HDR video reconstruction from alternately exposed frames. Its goal is to extract exposure-aware features that retain useful fine structures and provide stable spatial context. Given a pair of differently exposed frames $\{L_t^{e_1}, \hat{L}_t^{e_0}\}$ containing a reference frame and an interpolated frame, the EIC encoder extracts features via two branches and a scalar exposure gate $\alpha(e_i)$, where $i \in \{0, 1\}$ indexes the exposure mode. The fused feature output is

$$F_t^{e_i} = \mathrm{LF}(L_t^{e_i}) + \alpha(e_i) \cdot \mathrm{HF}(L_t^{e_i}). \tag{2}$$

As shown in Fig. 2 (a), the Large-Field (LF) branch applies a typical convolutional layer with stride 2. The High-Fidelity (HF) branch is composed of the pixel unshuffle layer, channel splitting operation, the dot product and a convolutional layer with stride 1. Furthermore, we design an Exposure Gate (EG) as

$$\alpha(e_i) = \sigma\big(w \log(e_i + \varepsilon) + b\big), \tag{3}$$

where $\varepsilon = 10^{-8}$ avoids numerical issues, w, b are learned scalars and $\sigma(\cdot)$ is the sigmoid function. A global exposure-dependent gate modulates per-pixel features, refined via spatial convolutions for exposure-aware, spatially adaptive fusion. LF and HF with EG jointly introduce an exposure-aware multi-scale information exchange mechanism.

The fused features $F_t^{e_i}$ are then fed into the RM module, which maps them into a unified representation consistent across exposures.

Representation Mapper. We aim to learn a unified representation $R_t(x)$ that maintains exposure-independence at the position of pixel x at time t, and a normalized mapping Π^e that projects any exposure feature onto that unified representation as

$$\tilde{F}_t^e(x) = \Pi^e\left(F_t^e(x)\right) \approx R_t(x), \quad e \in \{e_0, e_1\}. \tag{4}$$

We aim to map input features from different exposures toward a shared representation $R_t(x)$. To model the form of normalization mapping Π^e, we begin with the fundamental model of imaging, the output LDR frame is controlled by the $s = \log e$. In locally well-exposed regions and under a common monotonically differentiable camera response function, the derivative of the observation X^e with respect to s at pixel x can be approximated as

$$\partial_s X^e = \frac{\partial X^e}{\partial e} e \approx A(x) X^e + B(x), \tag{5}$$

where $A(x)$ is an intensity-dependent term and $B(x)$ absorbs bias terms such as black level and dark current. For encoder $E(\cdot)$, expanding along the direction of s gives

$$\partial_s E(x) \approx a(x, s) E(x) + b(x, s). \tag{6}$$

Integrating along the path from s_a to s_b, the solution is in affine form as $E(X^{e_b})(x) \approx k(x) E(X^{e_a})(x) + b(x)$, which motivates a pixel-wise, channel-wise linear modulation in feature space. Feature-wise affine modulation has appeared in broader conditioning and feature-statistics alignment literature [11, 27], here we use it for cross-exposure representation normalization guided by radiometric cues and supervised in the calibrated HDR domain.

Based on the above theory, we define the unified mapper Π^e as an element-wise linear modulation, thus, we designed the Representation Mapper module as the Π^e shown in Fig. 2,

$$\tilde{F}_t^e(x) = K_t^e(x) \odot F_t^e(x) + B_t^e(x), \tag{7}$$

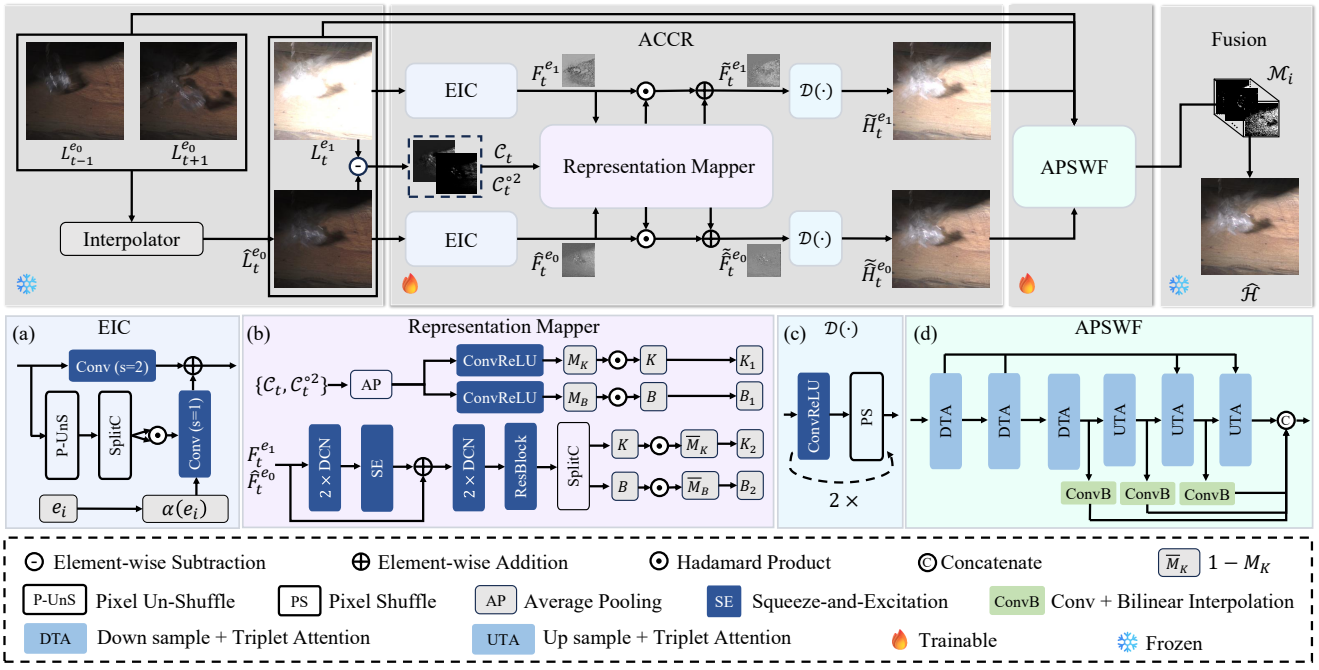

Figure 2. Overview of the proposed LRHDR. We take a three-frame LDR window with alternating exposures and interpolate a same-exposure middle frame using pre-trained FiLM [31]. EIC encodes reference frame and interpolated middle frame with a large-field branch, a high-fidelity sub-pixel branch, and an exposure gate. We compute cross-exposure cues in the linear-radiance domain. RM applies per-pixel, per-channel linear modulation, guided by the cues, to map the features into a cross-exposure-consistent unified representation, and then decodes them into linear HDR outputs. APSWF receives all linear HDR domain candidates and predicts sparse, normalized masks for pixelwise weighting. The final HDR frame is the masked fusion of the candidates in the linear HDR domain.

where $\odot$ is Hadamard product, x denotes the index of each pixel. Π^e maps features from different exposures toward a coarse shared representation. However, in areas of oversaturation and those overwhelmed by dark noise, it is difficult to reliably map a single path to the correct R_t. Thus, we use the cues $\mathcal{C}_t, \mathcal{C}_t^{\circ 2}$, which are defined as

$$\begin{aligned} \mathcal{C}_t &= \Gamma(\hat{L}_t^{e_0}) - \Gamma(L_t^{e_1}), \\ \mathcal{C}_t^{\circ 2} &= \mathcal{C}_t \odot \mathcal{C}_t, \end{aligned} \tag{8}$$

to guide the estimation of K and B, where $\Gamma(\cdot)$ is defined in Eq. (1) and $\odot$ stands for the Hadamard product. $\mathcal{C}_t$ serves as a signed cue for cross-exposure discrepancy and indicates whether a region tends to require enhancement or suppression, while $\mathcal{C}_t^{\circ 2}$ provides a magnitude and reliability cue that helps estimate the confidence and modulation strength. After completing these operations, we decode the exposure-normalized features to the image domain to provide supervision for the normalization mapping Π^e. Thus, with the input of the features obtained from the EIC encoder $\{F_t^{e_1}, \hat{F}_t^{e_0}\}$ and the cues $\{\mathcal{C}_t, \mathcal{C}_t^{\circ 2}\}$,the total RM module as shown in Fig. 2 (b) is as follows

$$\begin{aligned} \tilde{\hat{F}}_t^{e_0} &= K_t^{e_0} \odot \hat{F}_t^{e_0} + B_t^{e_0}, \\ \tilde{F}_t^{e_1} &= K_t^{e_1} \odot F_t^{e_1} + B_t^{e_1}. \end{aligned} \tag{9}$$

To better supervise ACCR training, we decode the representation information learned by the RM into the linear HDR domain for supervision following

$$\begin{aligned} \tilde{\hat{H}}_t^{e_0} &= \mathcal{D}(\tilde{\hat{F}}_t^{e_0}), \\ \tilde{H}_t^{e_1} &= \mathcal{D}(\tilde{F}_t^{e_1}). \end{aligned} \tag{10}$$

ACCR Loss. To guide the ACCR in learning accurate unified representation, we design a dedicated multi-term loss function. This loss targets the two linear HDR output $\tilde{\hat{H}}_t^{e_0}$ and $\tilde{H}_t^{e_1}$ by ACCR, enforcing consistency with the ground-truth HDR frame $H_t^\star$. The ACCR Loss is defined as

$$\mathcal{L}_{\text{ACCR}} = \lambda_{\text{L1}} \cdot \mathcal{L}_{\text{L1}} + \lambda_{\text{grad}} \cdot \mathcal{L}_{\text{grad}}, \tag{11}$$

where $\lambda_{\text{L1}} = 1.0$ and $\lambda_{\text{grad}} = 0.01$ are hyperparameters selected via cross-validation to balance the two objectives. To enhance perceptual quality [39], we formulate the loss in the tonemapped HDR domain using a simple differentiable μ-law function as the tonemapping function $\mathcal{T}$

$$\mathcal{T}(H) = \frac{\log(1+\mu H)}{\log(1+\mu)}, \tag{12}$$

where μ is set to 5000. The $\mathcal{L}_{\text{L1}}$ is defined as

$$\mathcal{L}_{\text{L1}} = \left\| \mathcal{T}(\tilde{H}_t^{e_1}) - \mathcal{T}(H_t^\star) \right\|_1 + \eta \cdot \left\| \mathcal{T}(\tilde{\hat{H}}_t^{e_0}) - \mathcal{T}(H_t^\star) \right\|_1, \tag{13}$$

Figure 3. Effectiveness of ACCR. Compared to recent methods that rely on explicit cross-exposure frame aligning, our ACCR reconstruction produces intermediate results with significantly less distortion. This enables us to achieve higher-fidelity HDR final outputs. Right three columns: Row 1 shows the aligned intermediate frames from HDRFlow [39], NECHDR [6] exposure completion, and our decoded unified representation, and Row 2 shows the corresponding HDR results.

where $\eta = 0.7$ is a reduced weight for the interpolated frame. This design mitigates over-constraining the interpolated stream. To preserve structural details and avoid artifacts, we define $\mathcal{L}_{\text{grad}}$ as a robust multi-scale gradient fidelity loss, adapted to HDR-specific challenges

$$\mathcal{L}_{\text{grad}} = \mathcal{L}_{\nabla}(\mathcal{T}(\tilde{H}_t^{e_1}), \mathcal{T}(H_t^{\star})) + \omega_s \cdot \mathcal{L}_{\nabla}(\mathcal{T}(\tilde{\hat{H}}_t^{e_0}), \mathcal{T}(H_t^{\star})), \tag{14}$$

where $\mathcal{L}_{\nabla}$ is a top-down multiscale approach gradient loss with the resolution of $1\times, 0.5\times, 0.25\times$, defined as

$$\mathcal{L}_{\nabla} = \mathbb{E}\left[\sqrt{(G_x - G_x^{\star})^2 + (G_y - G_y^{\star})^2 + \epsilon}\right], \tag{15}$$

where $\mathbb{E}$ denotes the spatial average and G stands for gradient, $\epsilon = 10^{-6}$ prevents division by zero. The multi-scale weights ω_s are normalized to ensure balanced contribution across resolutions and set as $\omega_s = (1.0, 0.5, 0.25)$ the same as the resolution scales.

3.2. Adaptive Pixel-wise Sparse Weighted Fusion

APSWF treats fusion as pixel-wise candidate selection. Rather than *asking how much to average from each source*, APSWF learns sparse, exposure-aware weights that adaptively activate only reliable candidates at each pixel. The final estimate is a normalized weighted fusion in the linear HDR domain, which reduces artifacts.

Concretely, based on the ACCR stage, the original LDR frames and the interpolated LDR frame $\{L_{t-1}^{e_0}, L_t^{e_1}, L_{t+1}^{e_0}, \hat{L}_t^{e_0}\}$, together with the unified $\{\tilde{H}_t^{e_1}, \tilde{\hat{H}}_t^{e_0}\}$ obtained by ACCR and the linear-radiance candidates $\{H_{t-1}^{e_0}, H_t^{e_1}, H_{t+1}^{e_0}, \hat{H}_t^{e_0}\}$ obtained via $\Gamma(\cdot)$ in Eq. (1), are fed into APSWF. APSWF learns a sparse mask tuple $\mathcal{M}(x)$ for each pixel x

$$\mathcal{M}(x) = (\mathcal{M}_1(x), \ldots, \mathcal{M}_6(x)), \tag{16}$$

and for every pixel x, with

$$\mathcal{M}_i(x) \geq 0, \quad \sum\nolimits_{i=1}^{6} \mathcal{M}_i(x) = 1, \tag{17}$$

thus, the HDR estimate is fused by

$$\hat{\mathcal{H}}_t = \sum\nolimits_{i=1}^{6} \mathcal{M}_i H_i, \tag{18}$$

and $(H_1, \ldots, H_6) = (H_{t-1}^{e_0}, H_t^{e_1}, H_{t+1}^{e_0}, \hat{H}_t^{e_0}, \tilde{H}_t^{e_1}, \tilde{\hat{H}}_t^{e_0})$.

As shown in Fig. 2, APSWF adopts a U-Net [32] architecture augmented with triplet attention (TA) blocks [24], thereby providing per-pixel information that reliably guides learning sparse pixel weights. The backbone produces features at four scales $(1/8, 1/4, 1/2, 1\times)$. At each scale, we attach a lightweight six-channel head that predicts per-candidate weighting logits $\ell^{(s)}$. Coarse scales capture long-range context. The $1\times$ head refines the boundaries of the masks. We upsample and fuse the multi-scale logits in a top-down manner to obtain the final weighting logits. A critical measure of APSWF is to map ℓ to sparse probability masks so that only a few candidates remain active at any pixel. We therefore use α-entmax [28] to project logits onto the 6-simplex $\{p \in \mathbb{R}^6_{\geq 0} : \sum_i p_i = 1\}$, and we set $\alpha = 1.75$ for both training and inference to obtain winner-takes-most masks with exact zeros while keeping gradients smooth.

Vote Loss. To supervise APSWF, we derive oracle votes to find the candidate that best matches the ground-truth HDR. We also use Eq. (12) to enhance perceptual quality, based on this, we define the per-candidate error as

$$E_i(x) = \|\mathcal{T}(H_i(x)) - \mathcal{T}(H^{\star}(x))\|_2^2, \tag{19}$$

where $H^{\star}$ is the ground-truth HDR while x is the pixel, and the oracle label is $i^{\star}(x) = \arg\min_{i \in \{1,\ldots,6\}} E_i(x)$. Given weighting logits $\ell(x) \in \mathbb{R}^6$, we map logits to a sparse probability simplex with α-entmax and minimize the corresponding Entmax cross-entropy

$$\mathcal{L}_{\text{vote}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \text{CE}_{\text{entmax}_\alpha}(\ell(x), i^{\star}(x)), \quad \alpha = 1.75, \tag{20}$$

where Ω stands for image domain. This calibrated surrogate aligns training with our winner-takes-most masks such that the dominance of a single candidate leads to an optimal one-hot solution, otherwise, a low-entropy mixture is permitted while avoiding dense averaging.

The total loss consists of ACCR Loss $\mathcal{L}_{\text{ACCR}}$, Vote Loss $\mathcal{L}_{\text{vote}}$ and Frame Reconstruction Loss $\mathcal{L}_{\text{Recon}}$. The $\mathcal{L}_{\text{Recon}}$ uses $\mathcal{L}_1$ loss as

$$\mathcal{L}_{\text{Recon}} = \|\mathcal{T}(\hat{\mathcal{H}}_t) - \mathcal{T}(H_t^{\star})\|_1. \tag{21}$$

And the total loss $\mathcal{L}_{total}$ is defined as

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{\text{Recon}} + \lambda_2 \mathcal{L}_{\text{ACCR}} + \lambda_3 \mathcal{L}_{\text{vote}}, \tag{22}$$

where $\lambda_1 = 1, \lambda_2 = 0.1$, and $\lambda_3 = 0.5$.

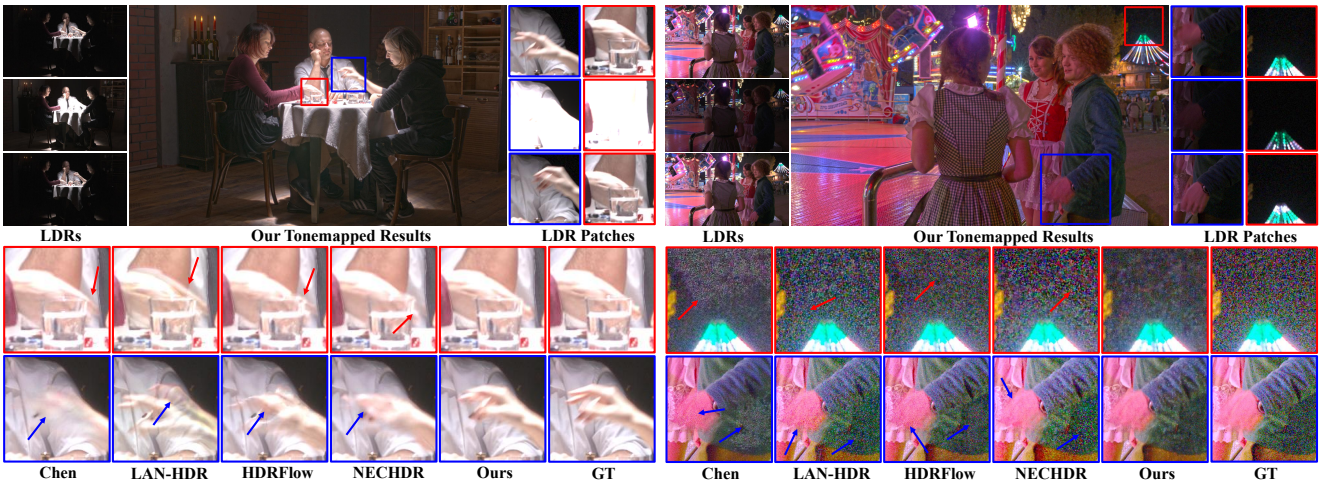

Figure 4. Qualitative comparisons on the Cinematic Video [8] dataset under 2-exposure setting. Left: scene with saturation and motion. Right: scene with complex lighting and severe noise interference.

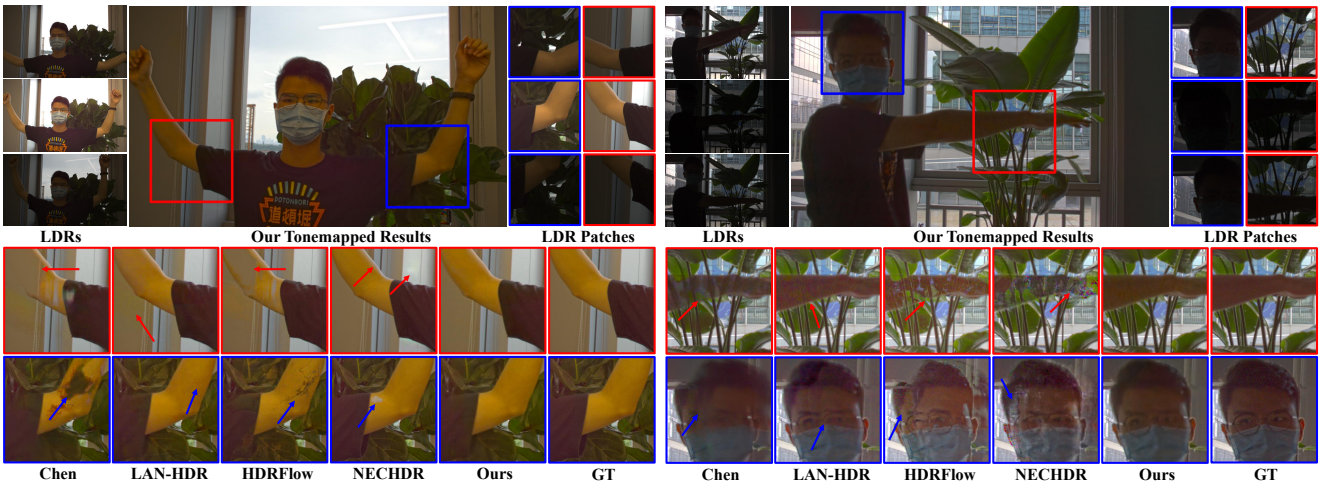

Figure 5. Qualitative comparisons on the DeepHDRVideo [2] dataset under 3-exposure setting in the scenes with large motion.

4. Experiments

4.1. Experimental Setups

Datasets. We use the training data from the Cinematic Video dataset [8], which is generated following the same procedure as [2]. The proposed framework is evaluated on two synthetic video sequences, POKER FULLSHOT and CAROUSEL FIREWORKS, from the Cinematic Video dataset, as well as on the DeepHDRVideo dataset [2]. The Cinematic Video dataset includes diverse indoor and outdoor scenes, while the DeepHDRVideo dataset consists of real-world dynamic scenes and static scenes augmented with simulated global motion. Because the Kalantari13 dataset [15] lacks ground truth, it is only used for qualitative evaluation as [4].

Training details. We use the AdamW optimizer [22] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 10^{-4} for APSWF and 10^{-5} for the ACCR. A cosine annealing schedule [21] is used to reduce both learning rates for ACCR and APSWF to 10^{-6} over 300 epochs. The mini-batch size is 8. All experiments are implemented in PyTorch [26] and executed on four NVIDIA RTX 4090 GPUs with 24 GB of memory each. And we train LRHDR once and evaluate the same frozen weights on all datasets.

Evaluation metrics. We use PSNR_T, SSIM_T, and HDR-VDP-2 [23] as evaluation metrics. PSNR_T and SSIM_T are computed in the μ law tone mapped domain. For HDR-VDP-2, the angular resolution is set to 30 pixels per visual degree for all compared methods, following [4, 6].

4.2. Quantitative and Qualitative Comparisons

Quantitative comparisons. As shown in Tab. 1, we provide quantitative comparisons between our method and prior state-of-the-art methods on the Cinematic Video [8]

Table 1. Quantitative comparisons of our method with other state-of-the-art methods on both Cinematic Video [8] and DeepHDRVideo [2] datasets under both 2-exposure setting and 3-exposure setting. The best and the second best results are highlighted in **red** and **blue**.

	Dataset	Metric	Method							
			Kalantari [14]	Yan [40]	Prabhakar [30]	Chen [2]	LAN-HDR [4]	HDRFlow [39]	NECHDR [6]	Ours
2 Exposure	[8]	$PSNR_T$	37.06	31.65	34.72	35.65	38.22	39.30	**40.59**	**41.11**
		$SSIM_T$	0.9053	0.8757	0.8761	0.8949	0.9100	0.9156	**0.9241**	**0.9274**
		HDR-VDP-2	70.82	69.05	68.82	72.09	69.15	71.05	**73.31**	**75.23**
	[2]	$PSNR_T$	39.91	40.54	40.21	42.48	41.59	43.25	**43.44**	**43.49**
		$SSIM_T$	0.9329	0.9452	0.9414	**0.9620**	0.9472	0.9520	0.9558	**0.9630**
		HDR-VDP-2	71.11	69.67	70.27	74.80	71.34	77.29	**79.20**	**80.68**
3 Exposure	[8]	$PSNR_T$	33.21	34.22	34.02	34.15	35.07	36.65	**37.24**	**37.64**
		$SSIM_T$	0.8402	0.8604	0.8633	0.8847	0.8695	0.9055	**0.9102**	**0.9120**
		HDR-VDP-2	62.44	66.18	65.00	66.81	65.42	66.02	**68.36**	**71.01**
	[2]	$PSNR_T$	38.78	40.20	39.48	39.44	40.48	**40.56**	40.13	**40.93**
		$SSIM_T$	0.9331	0.9531	0.9453	**0.9569**	0.9504	0.9535	0.9550	**0.9646**
		HDR-VDP-2	65.73	68.23	65.93	67.76	68.61	72.42	**76.98**	**76.01**

and DeepHDRVideo [2] test sets. Across both datasets and both exposure settings, our method attains the highest $PSNR_T$ and $SSIM_T$ and delivers competitive or superior HDR-VDP-2. And on the Cinematic Video dataset [8], our method improves HDR-VDP-2 over the second best by 1.92 in the two exposure setting and by 2.65 in the three exposure setting.

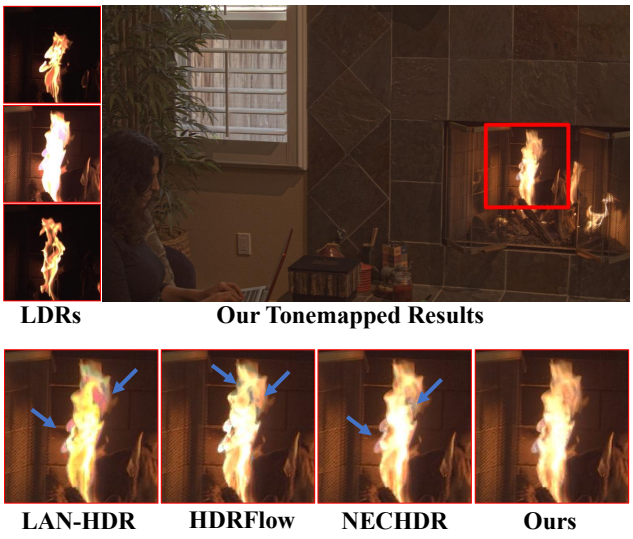

Figure 6. Comparisons with previous state-of-the-art methods [4, 6, 39] on the Kalantari13 dataset [15].

Qualitative comparisons. As shown in Figs. 4 and 5, we compare our approach with recent state-of-the-art HDR video reconstruction methods, including Chen *et al.* [2], LAN-HDR [4], HDRFlow [39], and NECHDR [6], on the Cinematic Video [8] and DeepHDRVideo [2] datasets. Specifically, Fig. 4 illustrates results from scenes with saturation and motion, as well as scenes with complex lighting and severe noise interference, on the Cinematic Video dataset under the 2-exposure setting. In scenes with saturation and motion, competing methods are more prone to artifacts and fine-detail loss, while LRHDR produces cleaner boundaries and more stable highlight reconstruction. Under complex lighting and severe noise interference, competing methods tend to show stronger noise amplification and detail degradation, while LRHDR yields more stable reconstructions. As shown in Fig. 5, we further present results under the three exposure setting on the DeepHDRVideo [2] dataset, which contains scenes with large motion. In these scenes, large displacements increase the difficulty of cross-frame matching, and visible ghosting is more likely to appear in prior methods [2, 4, 6, 39]. By comparison, LRHDR preserves cleaner contours and finer textures in this example. Furthermore, as shown in Fig. 6, we also demonstrate that our model outperforms existing models in reconstructing non-rigid motion.

4.3. Analysis

Ablation study. We conduct ablation studies on sequences with two alternating exposures. The experiments assess the contribution of each component on the Cinematic Video [8] dataset and on the dynamic sub-dataset of DeepHDRVideo [2], which contains large motion. Quantitative results are reported in Tab. 2. We build a baseline that uses FiLM [31] to synthesize the non-reference intermediate and employs APSWF for fusion. Training for this baseline uses the frame reconstruction loss with our Vote Loss. Adding the EIC encoder, which is tailored for alternately exposed inputs, yields clear gains by providing exposure-aware features. Introducing the RM module further improves all metrics by mapping cross-exposure features into a unified representation and by complementing unreliable regions across streams, which reduces mismatches. To ensure that RM learns the correct representation, we add the ACCR loss, and this constraint delivers additional, consistent improvements. Finally, we supervise candidate selection with a Vote Loss that aligns the predicted masks with oracle choices, this supervision outperforms the corresponding variant without Vote Loss, giving the best overall results in Tab. 2.

Table 2. Ablation study of LRHDR on the dynamic scenes of DeepHDRVideo (DeepHDRVideo-D in the table) [2] and Cinematic Video [8]. The Base uses FiLM [31] with our APSWF and Vote Loss. And ACCR has been decomposed into two parts: EIC and RM. **Bold** indicates the best score.

Model	ACCR		APSWF	ACCRLoss	Vote Loss	DeepHDRVideo-D		Cinematic Video	
	EIC	RM				$PSNR_T$	$SSIM_T$	$PSNR_T$	$SSIM_T$
Base						44.51	0.9699	40.01	0.9157
EIC	✓		✓		✓	44.86	0.9714	40.27	0.9220
RM		✓	✓		✓	45.11	0.9724	40.47	0.9232
RM+ACCRLoss		✓	✓	✓	✓	45.46	0.9725	40.61	0.9260
ACCR+APSWF	✓	✓	✓		✓	45.57	0.9722	40.99	0.9263
ALL w/o. Vote Loss	✓	✓	✓	✓		45.52	0.9733	40.97	0.9246
ALL (Ours)	✓	✓	✓	✓	✓	**45.89**	**0.9753**	**41.11**	**0.9274**

ACCR. We further analyze ACCR by comparing it with explicit alignment-based reconstruction methods, as shown in Fig. 3, explicit optical-flow alignment becomes less reliable for HDR video with alternating exposures, leading to more noticeable distortions in the intermediate results. Such intermediate distortions propagate to the final HDR reconstruction and are especially noticeable in dynamic regions. And in Tab. 3, we use Fusion Net in previous works [6, 39], further supports this observation quantitatively. These results suggest that ACCR provides a more robust reconstruction strategy than motion-aligned alternatives under alternating exposures.

APSWF. To study the effect of APSWF, we compare our sparse weighted fusion with the dense fusion networks used in prior methods [6, 39]. APSWF performs normalized fusion in the linear HDR domain using sparse weighted masks. As shown in Tab. 3, ACCR with APSWF consistently achieves higher $PSNR_T$ and $SSIM_T$ scores on both the Cinematic Video [8] dataset and the dynamic scenes of DeepHDRVideo [2] than ACCR equipped with the dense fusion network used in [6, 39].

Robustness to large motions. We evaluate our method against recent state-of-the-art [4, 6, 39] across different motion magnitudes, as shown in Fig. 7. To build the evaluation set, we apply [7] by using OpenCV on the dynamic scenes from DeepHDRVideo [2] to estimate optical flow. We then manually crop images with reliable flow and partition each crop into 128×128 blocks. For each block, we compute the average motion magnitude and measure blockwise PSNR. We finally report PSNR for blocks grouped by motion magnitude. As shown in Fig. 7, LRHDR retains stronger PSNR across different motion-magnitude ranges.

Table 3. Effectiveness of ACCR and APSWF in LRHDR. **Bold** indicates the best score, and underline indicates the second score.

Model	DeepHDRVideo-D		Cinematic Video	
	$PSNR_T$	$SSIM_T$	$PSNR_T$	$SSIM_T$
HDRFlow [39]	45.50	0.9683	39.30	0.9156
NECHDR [6]	45.35	0.9713	40.59	0.9241
Ours†	<u>45.61</u>	<u>0.9729</u>	<u>40.90</u>	<u>0.9264</u>
Ours	**45.89**	**0.9753**	**41.11**	**0.9274**

† We use Fusion Net in [6, 39] instead of our APSWF.

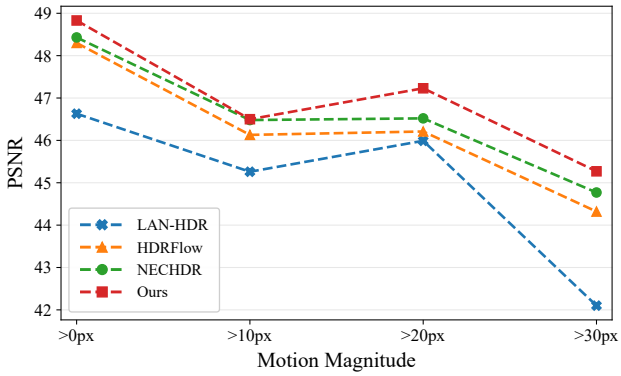

Figure 7. Comparisons with previous state-of-the-art methods [4, 6, 39] across different motion magnitude ranges.

5. Conclusion

In this paper, we present LRHDR, an end-to-end framework for HDR video reconstruction from alternately exposed LDR inputs. ACCR learns a cross-exposure consistent representation without explicit cross-exposure alignment, while APSWF performs sparse pixel-wise fusion using α-entmax masks and vote supervision. Together, these components enable more effective use of complementary information across input frames. Extensive experiments on public benchmarks show that LRHDR consistently improves reconstruction quality under alternating exposures, especially in scenes with saturation and large motion.

Acknowledgement. This work was supported by the National Natural Science Foundation of China No. 62522113 and the National Key Research and Development Program of China No. 2022YFA1207200.

References

[1] 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

[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. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2, 6, 7, 8

[3] 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. 1

[4] 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. 1, 2, 6, 7, 8

[5] Ugur Cogalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Learning hdr video reconstruction for dual-exposure sensors with temporally-alternating exposures. *Computers & Graphics*, 105:57–72, 2022. 1

[6] Jiahao Cui, Wei Jiang, Zhan Peng, Zhiyu Pan, and Zhiguo Cao. Exposure completing for temporally consistent neural high dynamic range video rendering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 10027–10035, New York, NY, USA, 2024. Association for Computing Machinery. 1, 2, 5, 6, 7, 8

[7] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis, 13th Scandinavian Conference, 2003, Halmstad, Sweden, June 29 - July 2, 2003, Proceedings*, pages 363–370. Springer, 2003. 8

[8] Jan Fröhlich, 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, part of the IS&T-SPIE Electronic Imaging Symposium, San Francisco, California, USA, February 2, 2014, Proceedings*, page 90230X. SPIE/IS&T, 2014. 6, 7, 8

[9] 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

[10] Tao Hu, Qingsen Yan, Yuankai Qi, and Yanning Zhang. Generating content for hdr deghosting from frequency view. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 25732–25741, 2024. 2

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, pages 1510–1519, 2017. 3

[12] Yitong Jiang, Inchang Choi, Jun Jiang, and Jinwei Gu. HDR video reconstruction with tri-exposure quad-bayer sensors. *CoRR*, abs/2103.10982, 2021. 1

[13] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4), 2017. 2

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

[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*, 32(6), 2013. 1, 6, 7

[16] Lingtong Kong, Bo Li, Yike Xiong, Hao Zhang, Hong Gu, and Jinwei Chen. Safnet: Selective alignment fusion network for efficient hdr imaging. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[17] Xinyue Li, Zhangkai Ni, and Wenhan Yang. Afunet: Cross-iterative alignment-fusion synergy for hdr reconstruction via deep unfolding paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10666–10675, 2025. 2

[18] Yifeng Lin, Gongning Yang, Rouqi Zhang, Jinhui Lai, Nian He, and Yiwen Xu. Bim-net: Lightweight hdr video reconstruction with multi-sensorial ldr videos. In *Proceedings of the 1st International Workshop on Multi-Sensorial Media and Applications*, page 27–33, New York, NY, USA, 2025. Association for Computing Machinery. 1, 2

[19] Yu-Shan Lin and Jui-Chiu Chiang. Hdrmamba: High dynamic range video reconstruction based on state space models. In *International Computer Symposium*, pages 219–223, 2024. 1, 2

[20] 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. 2

[21] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*. OpenReview.net, 2017. 6

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[23] Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions on Graphics*, 30(4), 2011. 6

[24] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3139–3148, 2021. 5

[25] S.K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 472–479 vol.1, 2000. 1

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner,

Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 6

[27] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2018. 3

[28] Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019. 5

[29] K. Ram Prabhakar, Rajat Arora, Adhitya Swaminathan, Kunal Pratap Singh, and R. Venkatesh Babu. A fast, scalable, and reliable deghosting method for extreme exposure fusion. In *IEEE International Conference on Computational Photography*, pages 1–8, 2019. 2

[30] 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 Conference on Computer Vision and Pattern Recognition*, pages 4873–4883, 2021. 7

[31] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *Proceedings of the European Conference on Computer Vision*, 2022. 3, 4, 7, 8

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, Cham, 2015. Springer International Publishing. 2, 5

[33] 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), 2012. 2

[34] Yong Shu, Liquan Shen, Xiangyu Hu, Mengyao Li, and Zihao Zhou. Towards real-world hdr video reconstruction: A large-scale benchmark dataset and a two-stage alignment network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2879–2888, 2024. 1, 2

[35] Jou Won Song, Ye-In Park, Kyeongbo Kong, Jaeho Kwak, and Suk-Ju Kang. Selective transhdr: Transformer-based selective hdr imaging using ghost region mask. In *Proceedings of the European Conference on Computer Vision*, pages 288–304, Cham, 2022. Springer Nature Switzerland. 2

[36] Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, and Dominique Ginhac. Alignment-free hdr deghosting with semantics consistent transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 2

[37] Sung-Min Woo, Je-Ho Ryu, and Jong-Ok Kim. Ghost-free deep high-dynamic-range imaging using focus pixels for complex motion scenes. *IEEE Transactions on Image Processing*, 30:5001–5016, 2021. 1

[38] 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*, 2018. 2

[39] Gangwei Xu, Yujin Wang, Jinwei Gu, Tianfan Xue, and Xin Yang. Hdrflow: Real-time hdr video reconstruction with large motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24851–24860, 2024. 1, 2, 4, 5, 7, 8

[40] 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. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 2, 7

[41] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29: 4308–4322, 2020. 2

[42] 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, 2021. 2

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22211–22220, 2023. 2

[44] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9300–9308, 2019. 2

[45] Uğur çoğalan and Ahmet Oğuz Akyüz. Deep joint deinterlacing and denoising for single shot dual-iso hdr reconstruction. *IEEE Transactions on Image Processing*, 29:7511–7524, 2020. 1