HARP: HIGH DYNAMIC RANGE PHOTOSEQUENCING THROUGH DUAL REVERSED SHUTTER SCANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The adoption of CMOS sensors in mobile photography is frequently compromised by the rolling shutter (RS) effect, which introduces geometric distortions and motion artifacts. Particularly, recent rolling shutter with global reset (RSGR) mode, while mitigating some RS issues, also incurs major limitations, including reduced capture speed and compressed dynamic range. To address these problems, we propose a novel dual reversed scanning setup utilizing both RSGR and inverted RSGR views. This solution not only handles the inherent flaws of RSGR by synchronizing complementary exposures to balance the dynamic range across the frames but also introduces an effective method for HDR photosequencing under highly dynamic scenes. Our proposed network first accommodates row-wise complementarity and manages visual shifts by row-adaptive feature alignment. Subsequently, the hallucination module, built upon a correlation-guided mix-attention block, integrates the mutually reinforced features to recover missing details. In addition, we construct a coaxial imaging system to collect a real-world dataset, enabling robust training and evaluation beyond numerical simulation. Experimental results demonstrate the twofold benefits of our solution in mitigating RSGR limitations and advancing HDR reconstruction techniques.

1 INTRODUCTION

With the widespread use of digital cameras and smartphones, increasing attention has been paid to choosing imaging sensors that balance performance and cost. Compared with Charge-Coupled Devices (CCD), Complementary Metal-Oxide Semiconductor (CMOS) sensors dominate the consumer market due to their low power consumption, high data rate, and ease of integration (Vasu et al., 2018; Schubert et al., 2019), and have become fundamental to visual computing (Kirillov et al., 2023; Ji et al., 2018; Ye et al., 2020; Zhao et al., 2025). However, under dynamic objects or ego-motion, their row-wise scanning mechanism produces rolling shutter (RS) effects, degrading both appearance and geometry. As a result, mitigating RS effects has become an essential problem for robust scene understanding and reconstruction.

Traditional RS correction methods (Albl et al., 2020; Zhuang et al., 2017; Lao & Ait-Aider, 2018; Purkait et al., 2017; Rengarajan et al., 2016) rely on velocity assumptions or geometric constraints, which struggle to capture real camera motion and scene geometry. Deep learning-based approaches have achieved impressive performance for single (Zhuang et al., 2019; Yan et al., 2023) and multi-frame inputs (Liu et al., 2020a; Ringaby & Forssén, 2012; Fan et al., 2023; Qu et al., 2023; Niu et al., 2024). Recent works (Fan & Dai, 2021; Fan et al., 2022; Zhong et al., 2022a) further attempt to recover multiple latent frames from RS inputs, but the ill-posed nature of the problem still leads to ghosting and motion artifacts. To handle this issue, the global-reset feature of RS sensors (RSGR) (Wang et al., 2022; Ji et al., 2023a;b) has been exploited to reformulate RS correction as a more tractable deblurring problem, achieving superior results. But meanwhile, the RSGR mode introduces two overlooked limitations: (1) the merit of RS mode over global shutter (GS) in capture speed is sacrificed because early scanning of the next frame is not allowed anymore (Figure 1a); and (2) row-variant exposure causes dynamic-range compression (Figure 1b). Unlike conventional RS with constant row exposure, RSGR increases exposure duration row by row, leading to under-exposure at the top and over-exposure at the bottom when scanning downward.

Existing HDR approaches cannot be directly applied, because they depend on alternating-exposure LDR sequences combined with motion compensation (Kang et al., 2003; Chen et al., 2021) or feature

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

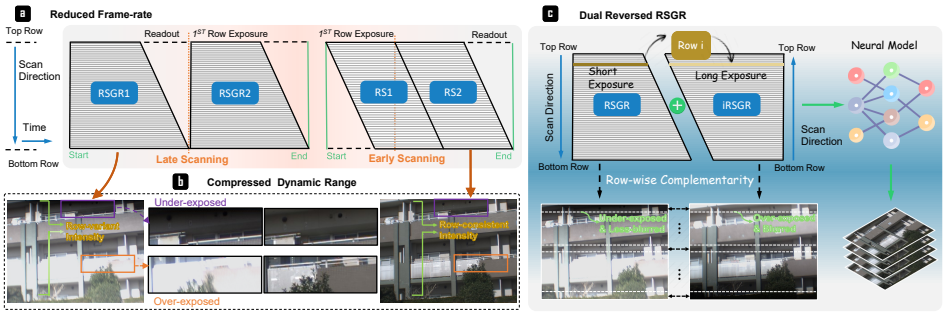


Figure 1: Limitations of RSGR and our dual reversed setting. (a) shows capturing process of two consecutive frames from RSGR and RS sensors. Enabling the global reset deprives RS mode of early scanning and causes delay of a readout time. (b) presents acquired frames from two types of sensors. The dynamic range of RSGR is lower than that of RS. In (c), we reconstruct an HDR video clip from two row-wise complementary observations captured by spatiotemporally aligned RSGR and iRSGR cameras, which addresses issues of RSGR and provides a novel solution for HDR reconstruction.

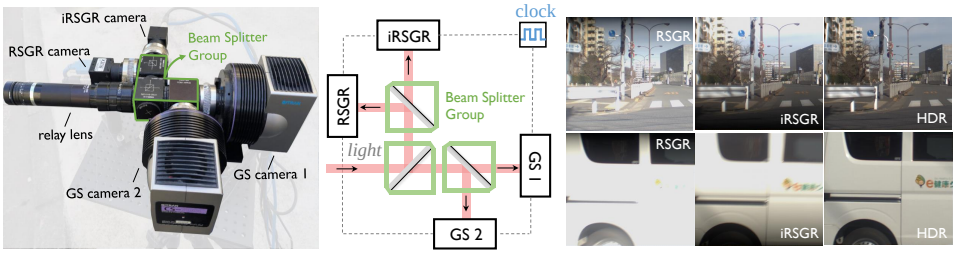
alignment (Chung & Cho, 2023; Liu et al., 2022). Such methods remain vulnerable to distortions and ghosting under fast motion, making them ineffective for handling the limitations. Inspired by Albl et al. (2020), we propose a dual reversed setting simultaneously capturing two fully aligned RSGR and inverted RSGR (iRSGR) views to reconstruct a sequence of clear frames with high speed and high dynamic range (HDR photosequencing), which not only well-addresses two aforementioned issues but also provides a novel solution for HDR reconstruction in dynamic scenarios. As shown in Figure 1c, the corresponding rows in the two views provide complementary exposures for the same visual content under static conditions. Although the middle scanlines may converge to similar intensity levels, missing contrast can be compensated by the hallucination capability of our model. In dynamic scenes, visual shifts between corresponding rows may weaken this complementarity. However, such shifts are typically minor due to the short exposure time, and our method remains robust to them. Meanwhile, these shifts also alleviate intensity convergence in the middle rows. Our experiments in Section 4.3 confirm these conclusions. Since this row-wise complementarity holds in both static and dynamic scenes, our dual reversed RSGR setup ultimately enables high-frame-rate acquisition, mitigates motion degradation, and makes dynamic-range expansion more tractable.

Following the hardware design of Zhong et al. (2022a); Ji et al. (2023b; 2024; 2026), we construct our optical system to capture aligned high-speed sharp HDR videos and low-speed RSGR-iRSGR pairs. Assisted by the collected dataset, we further propose an end-to-end framework for HDR photosequencing. Firstly, the Row-adaptive Feature Alignment module (RFA) conducts multi-scale, row-shaped attention to better adapt the row-wise complementarity and simultaneously manage the visual shifts under highly dynamic scenes. Subsequently, the hallucination module generates sharp details, especially for under/over-exposed rows. To this end, we introduce a residual correlation-guided mix-attention block (RCMB), built upon parallel streams, which aggregates intra-stream self-similar features and supports cross-stream mutual compensation. Finally, a tri-pivot temporal loss on the luminance channel emphasizes geometric consistency and suppresses color variations, improving temporal coherence in the reconstructed video. **In summary, the main contributions of our paper** are: (1) We design a dual-reversed RSGR configuration that inherently accommodates fast motions, addressing intrinsic limitations of RSGR mode, and simultaneously offering a competitive solution to robustify HDR reconstruction. (2) We construct a special imaging system that simultaneously captures RSGR-iRSGR pairs along with their corresponding high-speed HDR videos, and collect a high-quality real dataset of outdoor dynamic scenes. (3) We introduce a novel neural network architecture consisting of a row-adaptive feature alignment and a hallucination module based on correlation-guided mix-attention block. Extensive experiments have validated the effectiveness of our setting and model.

2 Related Work

2.1 Rolling Shutter Correction and Interpolation

The causing effects of rolling shutters under ego-motion or dynamic scenes are often non-negligible. Early works (Meingast et al., 2005) model RS geometry to improve SfM and correct distortions



(a) Imaging System (b) Optical Diagram (c) Data Samples

Figure 2: (a) shows the constructed data capturing platform that is spatially aligned and temporally synchronized through a beam splitter group and a signal controller, respectively. (b) is the corresponding optical diagram. We also present some data samples in (c).

using discrete epipolar constraints (Dai et al., 2016) or differentiable warping (Zhuang et al., 2017), but they rely on strong assumptions about scene structure or camera motion, such as Manhattan-world layouts (Purkait et al., 2017), straight-line preservation (Rengarajan et al., 2016), purely rotational (Lao & Ait-Aider, 2018; Ringaby & Forssén, 2012) or translational motion (Liang et al., 2008; Baker et al., 2010). These priors limit their applicability. Albl et al. (2020) later introduce a dual reversed RS configuration that provides additional geometric constraints from sparse correspondences.

Learning-based approaches (Naor et al., 2022; Zhou et al., 2022; Fan et al., 2021; Zhuang et al., 2019) instead estimate motion fields to warp RS images toward their GS counterparts. Liu et al. (2020a) propose a shutter-unrolling network with differentiable warping, while Qu et al. (2023) solve complex nonlinear motion with a quadratic RS model. Fan et al. (2023) jointly recover appearance and motion in a single-stage design, and the global-reset mode has also been exploited for video (Wang et al., 2022) and single-image correction (Ji et al., 2023a). More broadly, RS interpolation or temporal super-resolution (Fan & Dai, 2021; Fan et al., 2022; Zhong et al., 2022a) reconstructs multiple latent sharp frames within a single RS exposure, and Ji et al. (2023b) provide a unified comparison across degradations. Nevertheless, RS correction fundamentally requires shifting each scanline to a virtual GS coordinate frame based on estimated flow or hallucinated context, making existing methods fragile in the presence of large or complex motions.

2.2 High Dynamic Range Imaging

Regardless of HDR video reconstruction or multi-exposure imaging, the goal is to merge motion-varying LDR frames into a clean HDR result. Traditional HDR pipelines focus on alignment to avoid ghosting: motion rejection methods (Khan et al., 2006; Heo et al., 2010; Zhang & Cham, 2011; Oh et al., 2014; Lee et al., 2014) first register LDR inputs and then discard misaligned pixels, often suppressing useful information and degrading HDR quality. More robust techniques use optical flow (Kang et al., 2003; Zimmer et al., 2011) or energy minimization (Sen et al., 2012; Hu et al., 2013), but they remain time-consuming and sub-optimal due to complex optimization.

Deep learning has significantly advanced HDR imaging. CNN- and Transformer-based methods (Kalantari et al., 2017; Cai et al., 2018; Wu et al., 2018; Yan et al., 2019; 2020; Niu et al., 2021) alleviate misalignment and reduce ghosting, even enabling single-image HDR reconstruction (Chen et al., 2023; Marnerides et al., 2018; Liu et al., 2020b). Kalantari et al. (2017) pioneer optical-flow-guided HDR fusion; Wu et al. (2018) avoid explicit motion estimation via image translation. Chen et al. (2021) develop a coarse-to-fine alignment and fusion framework using their collected real dynamic HDR data. More recent works include context-aware Transformers (Liu et al., 2022) and luminance-based alignment (Chung & Cho, 2023) for enhanced detail recovery.

3 Methodology

3.1 Optical System and RSGR-HDR Dataset

To train our model, we require spatially aligned and temporally synchronized RSGR-iRSGR observations together with corresponding high-frame-rate HDR ground truths. Therefore, we develop a novel quad-axis imaging system and collect a real-world dataset, RSGR-HDR. Each capture provides a dual reversed RSGR pair and two sets of high-speed GS frames for composing HDR GTs. To

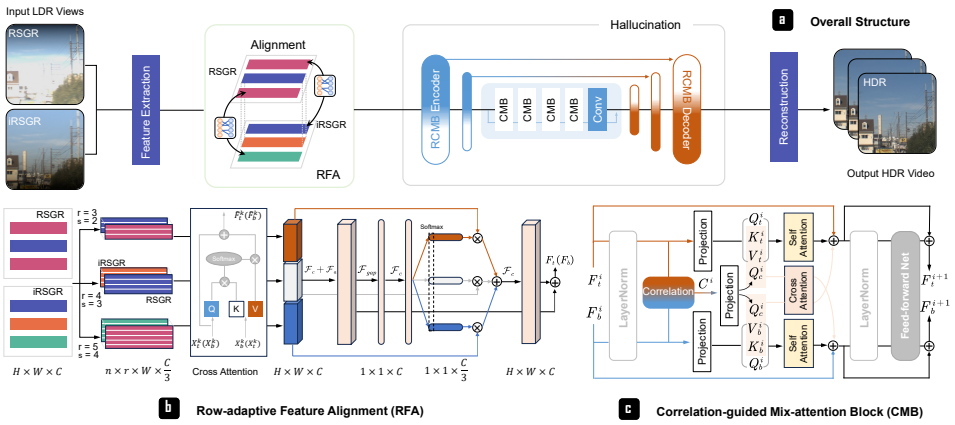


Figure 3: (a) shows the overall architecture containing two parts: alignment module registers motions between reversed RSGR views, and hallucination module generates sharp details, especially for under/over-exposed rows. Instead of simply associating corresponding rows between views, we exploit row-adaptive alignment strategy with multiple scales to handle the visual shift under highly dynamic scenes as shown in (b). (c) presents our correlation-guided mix-attention block that further facilitates the generation of local high frequencies.

the best of our knowledge, such real-world data have not been explored in prior work. Our entire hardware prototype is illustrated in Figure 2. Two RSGR cameras (FLIR BFS-U3-63S4C) with reversed scanning direction share the identical objective scenes with two high-speed GS cameras (BITRAN CS-700C) through a beam splitter group. The RSGR sensors enlarge pitch size from 2.4 μm to 9.6 μm through 4×4 binning, while GS cameras incorporate SONY IMX426 sensor with air-forced cooling. We follow the merging strategy in Lin et al. (2023); Zou et al. (2021) to combine the GS pairs with different scene irradiance into HDR measurements.

Consequently, the RSGR-HDR dataset is constructed using our optical system, which consists of three modes: RSGR, inverted RSGR (iRSGR) and corresponding HDR frames, acquired under highly dynamic urban scenes. In total, we collect 90 videos with resolution of 640×480, each of which contains 70×2 degraded frames, accompanied by 1,400 aligned high-speed HDR frames. Notably, the characteristic differences of reversed RSGR views are quite obvious (Figure 2c). More details about imaging system and dataset are provided in Section A.1 and Section A.2.

3.2 Network Architecture

To perform HDR photosequencing from dual reversed RSGR views, we propose a two-stage framework comprising feature alignment and hallucination modules, as shown in Figure 3a. Given strictly aligned RSGR–iRSGR pair, shallow convolutions first extract their feature maps X_t and X_b. Row-shaped attention is then applied at multiple scales to handle visual shifts under large motion and to promote mutual compensation between the two views. Features aligned with row sizes $\boldsymbol{r}$ and strides $\boldsymbol{s}$ are dynamically fused to accommodate varying motion magnitudes. The hallucination module further aggregates these aligned features using a Transformer-based encoder–decoder to recover fine details in under- or over-exposed regions. Finally, a reconstruction block outputs a high-speed HDR sequence $\boldsymbol{H} = \{H^i, i = 0, 1, \cdots, N-1\}$ within the exposure duration of input. N is the length of reconstructed latent video.

Row-adaptive Feature Alignment Normally, attention-based alignment will attend to all tokens without selection or adhering to epipolar constraints. However, to effectively adapt and fully leverage our dual reversed RSGR configuration, several considerations need to be addressed. As discussed in Ji et al. (2023a); Wang et al. (2022), the unique distortion induced by RSGR mode is naturally row-dependent. Indiscriminate processing tends to corrupt the tempo-

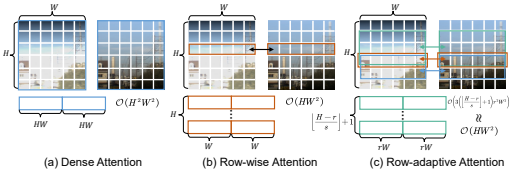


Figure 4: Different alignment strategies based on attention mechanism. We compared the full image, row-wise with single scale and row-adaptive with multiple scales alignment methods. Row size r and stride s are constants.

ral cues implicitly embedded by sequential readout mechanism. Second, the observed RSGR and iRSGR views are inherently row-wise corresponding, facilitating robust alignment even under large motions. Furthermore, applying dense attention across the entire feature space (Figure 4a) leads to high computational demands and potential ambiguities in motion registration. Therefore, we propose row-shaped attention to highlight these concerns. Nevertheless, direct use of row-wise attention with fixed scale (Figure 4b), e.g. simply associating corresponding rows between views (Shi et al., 2021; Li et al., 2024) is only suitable for ideal static scenes. To handle the varying visual shifts under dynamic cases, we propose row-adaptive feature alignments (Figure 4c) that effectively blend multiple scales. For clarity, we proceed to explain with $\boldsymbol{r} = [3, 4, 5]$ and $\boldsymbol{s} = [2, 3, 4]$.

We first compute the bilateral alignments for dual views on multiple scales as shown in Figure 3b and then dynamically fuse them in a learnable manner. By taking the RSGR branch as an example, the extracted features of RSGR and iRSGR views are denoted as $X_t, X_b \in \mathbb{R}^{H\times W\times C}$. On each scale level k, the features are accordingly converted to $X_t^k, X_b^k \in \mathbb{R}^{n\times r^k\times W\times C/3}$ by following row size r^k and stride s^k, where $n = \lfloor \frac{H-r^k}{s^k} + 1 \rfloor$. We generate the query Q_t^k for X_t^k, and the key K_b^k, value V_b^k for X_b^k as follows:

$$Q_t^k, K_b^k, V_b^k = X_t^k * K_1, X_b^k * K_2, X_b^k * K_3, \tag{1}$$

where K_1, K_2, K_3 denote 1×1 kernels and $*$ is convolutional operator. Thus, the aligned feature under corresponding scale $\tilde{F}_t^k$ is formulated as:

$$\tilde{F}_t^k = X_t^k + (\sigma(Q_t^k (K_b^k)^T) V_b^k) * K_4. \tag{2}$$

$*K_4$ represents 1×1 kernel convolution to match the number of channels with that of X_t^k. $\sigma(\cdot)$ is the softmax function.

The aligned features under all scales naturally retain motions with different magnitudes, which can be dynamically fused to make robust alignments. In more detail, we use the concatenated information of all scales to produce respective weights for each branch, then integrate them via weighting. Specifically, we first compute the fused feature $\tilde{F}_t^{fuse}$ using a fully connected layer $\mathcal{F}_c$ and an activation layer $\mathcal{F}_a$ as:

$$\tilde{F}_t^{fuse} = \mathcal{F}_a(\mathcal{F}_c(\oplus_{k=1,2,3} \tilde{F}_t^k)), \tag{3}$$

where $\oplus$ means the concatenation of feature maps along the channel dimension. Subsequently, the weights are formulated as:

$$\tilde{F}_t^w = \mathcal{F}_c(\mathcal{F}_{gap}(\tilde{F}_t^{fuse})), \qquad \alpha_k = \frac{e^{\delta_k(\tilde{F}_t^w)}}{\sum_{j=1}^{3} e^{\delta_j(\tilde{F}_t^w)}}, k = 1, 2, 3, \tag{4}$$

where the $\mathcal{F}_{gap}$ denotes global average pooling and δ_k is an operation to uniformly partition channel dimension according to scale levels and return the k^{th} part. Finally, the aligned feature for RSGR branch is computed by:

$$F_t = \tilde{F}_t^{fuse} + \mathcal{F}_c(\sum_{k=1}^{3} \alpha_k \times \tilde{F}_t^k) \tag{5}$$

The generation of aligned feature F_b for iRSGR branch is similar to the aforementioned process.

Correlation-guided Mix-attention Block To enhance the generation ability of local details, we further propose novel correlation-guided mix-attention Transformer layer with dual streams. Therefore, the aggregated features of both views can be mutually improved more effectively from perspective of global context. The calculated correlation volume (Sun et al., 2018) between two views explicitly represents the similarities, which has merits of driving the features in each stream to become more compatible with those of the other. To this end, we build the cross-attention module aided by the correlation volume as the major supplementary to original self-attention. Overall, our proposed mix-attention effectively adapts to self-similar features within a single stream while also enabling reciprocal interactions across different streams.

As shown in Figure 3c, we present the specific operations for the i^{th} CMB layer. Given input features F_t^i and F_b^i after layer normalization, we compute their correlation volumes C^i following the method in Sun et al. (2018); Fan et al. (2021) by setting the search range as 3×3 to balance between complexity and accuracy. Note that, to ensure input order invariance, we consider the bidirectional

nature of cost volumes and project concatenation of them into the same dimension as input features. For the self-attention, F_t^i and F_b^i are linearly projected to obtain their corresponding queries, keys and values: $\{Q_t^i, K_t^i, V_t^i\}$ and $\{Q_b^i, K_b^i, V_b^i\}$. The correlation volume C^i is also projected as a query Q_c^i serving as guidance for cross-attention. Formally, we define the computational process as:

$$F_{t_agg}^i = \sigma(\frac{Q_t^i(K_t^i)^T}{\sqrt{d_k}})V_t^i + \sigma(\frac{Q_c^i(K_b^i)^T}{\sqrt{d_k}})V_b^i, \tag{6}$$

where d_k is the channel length of projected keys for both K_t^i and K_b^i. Similarly, We could calculate the aggregated feature $F_{b_agg}^i$ for the other stream. Thus, final outputs of the i^{th} CMB layer are formulated as:

$$\begin{aligned} F_t^{i+1} &= \mathcal{FFN}(\mathcal{LN}(F_t^i + F_{t_agg}^i)) + (F_t^i + F_{t_agg}^i), \\ F_b^{i+1} &= \mathcal{FFN}(\mathcal{LN}(F_b^i + F_{b_agg}^i)) + (F_b^i + F_{b_agg}^i), \end{aligned} \tag{7}$$

where $\mathcal{LN}$ is layer normalization and $\mathcal{FFN}$ denotes multilayer perceptron. In experiments, we implement the hallucination module with 6 RCMBs with depths $[6, 8, 8, 8, 8, 6]$ and number of heads $[6, 6, 6, 6, 6, 6]$, using an initial channel dimension of 96.

Tri-pivot Temporal Loss To generate perceptually natural video, avoiding the flickering is very essential. Since the correlation of reconstructed HDR frames is weakly constrained, we present tri-pivot temporal loss $\mathcal{L}_{temp}$ to improve structural coherence and temporal consistency. In specific, we sample three frame pivots that are temporally located at $i = \{0, \frac{N-1}{2}, N-1\}$ from the reconstructed HDR video and enforce the difference maps among them to be similar to these of corresponding GTs using the Charbonnier function. On the other hand, we observe that discontinuity or inconsistency is normally caused by wrongly estimated edges and geometric structures. Therefore, instead of calculating temporal loss in sRGB space without distinguishing colors and contents, we conduct it on Y channel of YCrCb space in order to pay more attention to structures and details weakening the effects of color information. The temporal loss for this enforcement is defined as:

$$\begin{aligned} &\Delta_{a,b} I \triangleq [I^a]_Y - [I^b]_Y, \\ &\mathcal{L}_{temp} = \sum_{(a,b)\in\mathcal{P}} \sqrt{\|\Delta_{a,b} H - \Delta_{a,b} G\|^2 + \epsilon^2}. \end{aligned} \tag{8}$$

where $\mathcal{P} = \{(N-1, 0), (N-1, \frac{N-1}{2}), (\frac{N-1}{2}, 0)\}$. $[\cdot]_Y$ denotes extraction of the Y channel of YCrCb color space and ϵ is set as 10^{-3}. G represents groundtruth frame. While the reconstruction loss, $\mathcal{L}_{rec}$ is defined as:

$$\mathcal{L}_{rec} = \sum_{i=0}^{N-1} \sqrt{\|H^i - G^i\|^2 + \epsilon^2}. \tag{9}$$

In total, our training loss is $\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{temp}$, where λ is set as 0.1 based on experiments to balance the reconstruction and temporal losses.

4 Experiments

4.1 Comparison with SOTA methods

We aim to reconstruct a sequence of HDR frames from dual reversed RSGR views with intermediate degradation between RS effects and blur. Therefore, we compare our model with related SOTA methods from the following categories:

- Blur interpolation and decomposition: BiT (Zhong et al., 2023), AfB (Zhong et al., 2022b) and DeMFI (Oh & Kim, 2022) approximate motions based on temporally adjacent blurry frames. We also include the typical interpolation method, RIFE (Huang et al., 2022) and more competitive setting by using the combination of blur frame and corresponding event streams, EBFI (Weng et al., 2023) and UniINR (Lu et al., 2024).
- RS correction and interpolation: CVR (Fan et al., 2022), RSSR (Fan & Dai, 2021), DSUN (Liu et al., 2020a) and JAMNet (Fan et al., 2023) take as input multiple RS frames while IFED (Zhong et al., 2022a) depends on dual reversed RS views and EvUnroll (Zhou et al., 2022) requires extra event information.



Figure 5: Qualitative comparisons on RSGR-HDR. Our model outperforms the existing approaches with different settings in details recovery and dynamic range expansion.

Table 1: Quantitative comparisons on RSGR-HDR. We present experimental results on recovering a single latent HDR frame (HDR reconstruction) and an HDR sequence of length 3, 5, and 9 (HDR photosequencing), both extracted within the exposure duration of the input frame. Performance is evaluated using mean PSNR/SSIM/LPIPS/TMQI. Dual reversed RSGR inputs are denoted as 'RG-iRG'. 'n·RG' is the setting using n neighboring RSGR frames and 'RG-E' combines RSGR with event streams. Note that, LAN takes 5 consecutive LDR frames with alternate long and short exposures (denoted as '5·LS'). We also compute initial performance of inputs as RSGR and iRSGR.

Method	Input	HDR Reconstruction	HDR Photosequencing			Time (s)	Params (M)	FLOPs (G)
			×3	×5	×9			
RSGR	–	12.30/0.5450/0.2774/0.7157	–/–/–	–/–/–	–/–/–	–	–	–
iRSGR		14.94/0.5961/0.2572/0.6993	–/–/–	–/–/–	–/–/–	–	–	–
DSUN	*2·RG*	19.44/0.7278/0.1663/0.7860	–/–/–	–/–/–	–/–/–	0.847	3.900	225.0
JAMNet		21.39/0.7796/0.1397/0.8310	–/–/–	–/–/–	–/–/–	0.690	4.730	51.13
RSSR		12.82/0.6099/0.2427/0.7924	12.75/0.5975/0.2566/0.7871	12.76/0.5970/0.2592/0.7895	12.72/0.5929/0.2639/0.7906	1.469	26.03	42.67
CVR		13.04/0.5623/0.2898/0.7769	13.03/0.5609/0.2946/0.7565	13.04/0.5643/0.2963/0.7573	13.03/0.5621/0.3000/0.7578	1.497	42.70	101.1
RIFE		19.80/0.7542/0.1911/0.8227	19.64/0.7486/0.1970/0.7998	19.67/0.7492/0.1981/0.8005	19.68/0.7496/0.1984/0.8008	0.358	54.82	71.15
AfB		17.56/0.7285/0.2969/0.8342	17.41/0.7217/0.3060/0.8104	17.43/0.7228/0.3056/0.8112	17.44/0.7232/0.3054/0.8115	0.420	129.2	793.7
IFED		12.30/0.4635/0.2723/0.7377	12.24/0.4579/0.2847/0.6831	12.25/0.4597/0.2835/0.6867	12.25/0.4607/0.2828/0.6883	0.394	10.81	**29.58**
PMBNet		20.05/0.7593/0.1819/0.7961	19.98/0.7560/0.1838/0.7970	19.97/0.7556/0.1845/0.7980	19.97/0.7554/0.1850/0.7984	1.497	30.10	121.2
BiT	*3·RG*	18.60/0.7437/0.1783/0.8111	18.59/0.7447/0.1884/0.7938	18.60/0.7452/0.1852/0.7944	18.61/0.7456/0.1833/0.7946	0.167	11.33	57.48
ExpandNet	*1·RG*	16.26/0.6405/0.2956/0.7279	–/–/–	–/–/–	–/–/–	0.587	**0.457**	63.17
RSS-T		19.46/0.7076/0.2024/0.7903	–/–/–	–/–/–	–/–/–	0.136	11.34	176.6
DeMFI	*4·RG*	14.54/0.3247/0.6156/0.6194	14.79/0.3894/0.5207/0.6404	14.73/0.3835/0.5282/0.6390	14.72/0.3763/0.5355/0.6477	0.388	7.410	420.0
NGS	*5·RG*	19.17/0.6662/0.1804/0.7405	–/–/–	–/–/–	–/–/–	1.555	8.670	1266
LAN	*5·LS*	23.66/0.7122/0.3552/0.6582	–/–/–	–/–/–	–/–/–	0.243	7.320	949.1
HDRFlow		22.07/0.4464/0.3425/0.6712	–/–/–	–/–/–	–/–/–	**0.124**	3.270	83.37
EBFI	*RG-E*	18.25/0.7518/0.1464/0.8539	18.21/0.7490/0.1481/**0.8480**	18.21/0.7498/0.1474/0.8485	18.21/0.7503 /0.1469/0.8488	5.253	5.680	432.2
EvUnroll		14.39/0.3929/0.3285/0.4815	14.38/0.3903/0.3299/0.4837	14.39/0.3909/0.3296/0.4841	14.39/0.3911/0.3294/0.4845	2.594	27.47	890.0
EFI		18.61/0.8023/0.1664/0.8408	18.56/0.8006/0.1675/0.8443	18.57/0.8011/0.1671/0.8446	18.57/0.8014/0.1668/0.8447	5.534	16.60	977.2
UniINR		20.75/0.8008/0.2173/0.8337	20.58/0.7880/0.2350/0.8299	20.60/0.7890/0.2335/0.8304	20.61/0.7898/0.2322/0.8307	1.487	1.824	1366
RIFE$_m$	*RG-iRG*	22.37/0.8445/0.1262/0.8486	22.12/0.8341 /0.1322/0.8441	22.15/0.8361/0.1322/0.8449	22.17/0.8371/0.1317/0.8453	0.358	54.82	71.15
IFED$_m$		21.79/0.8166/0.1072/0.8509	21.49/0.8003/0.1172/0.8398	21.54/0.8033/0.1146/0.8415	21.57/0.8050/0.1131/0.8423	0.394	10.81	**29.58**
BiT$_m$		22.51/0.8341/0.1279/0.8512	22.32/0.8289/0.1319/0.8426	22.38/0.8303/0.1305/0.8434	22.40/0.8311/0.1298/0.8443	0.167	11.33	57.48
Ours		**24.24/0.8679/0.0994/0.8543**	**23.84/0.8562/0.1071/0.8480**	**23.91/0.8584/0.1058/0.8487**	**23.95/0.8595/0.1053/0.8490**	0.598	35.05	183.1

- RSGR undistortion and interpolation: NGS (Wang et al., 2022) and RSS-T (Ji et al., 2023a) reconstruct single sharp output while PMBNet (Ji et al., 2023b) can interpolate at arbitrary time.
- HDR reconstruction: LAN (Chung & Cho, 2023), HDRFlow (Xu et al., 2024) and ExpandNet (Marnerides et al., 2018) follow convention by alternating exposures of input LDR frames. EFI (Lin et al., 2023) absorbs complementary information from the RS image and events to retrieve a HDR sequence.

To better demonstrate the superiority of our setup, we also adapt RIFE, IFED and BiT to dual RSGR inputs, which are denoted as RIFE$_m$, IFED$_m$, and BiT$_m$, respectively. As a contrast, RIFE, IFED and BiT represent models using consecutive RSGR frames. Importantly, following prior work Mertens et al. (2007); Lin et al. (2023); Chen et al. (2025), we adopt an exposure-fusion strategy for HDR composition. Unlike conventional pipelines that reconstruct a linear HDR image followed by tone mapping, we directly output tone-mapped LDR results, avoiding error accumulation (Chen et al.,

Table 2: Model architecture ablation. RFA_s and RFA_m denote row-wise alignment with single scale and row-adaptive alignment with multiple scale. STB is Swin Transformer block. R, T_R and T_Y are reconstruction loss, temporal loss calculated in sRGB and YCrCb space. T_W is warping-based temporal constraint.

	Alignment			Transformer		Loss				PSNR / SSIM / LPIPS
	Dense	RFA_s	RFA_m	STB	CMB	R	R+T_R	R+T_W	R+T_Y	
V_1					✓				✓	22.29 / 0.8428 / 0.1254
V_2	✓				✓				✓	22.05 / 0.8299 / 0.1234
V_3		✓			✓				✓	22.79 / 0.8383 / 0.1214
V_4			✓	✓					✓	22.48 / 0.8372 / 0.1208
V_5			✓		✓	✓				23.37 / 0.8500 / 0.1218
V_6			✓		✓		✓			23.52 / 0.8525 / 0.1204
V_7			✓		✓			✓		23.59 / 0.8426 / 0.1483
Ours			✓		✓				✓	**23.95 / 0.8595 / 0.1053**

Table 3: Row-shaped attention ablation. We conduct the comparisons to decide the optimal row shape stride for feature alignment. The results demonstrate the superiority of our proposed row-adaptive strategy compared with row-wise alignment.

Row Size	Stride	Attention Type	PSNR / SSIM / LPIPS
$r = [2]$	$s = [1]$	Row-wise (Single Scale)	22.10 / 0.8316 / 0.1353
$r = [3]$	$s = [2]$		22.61 / 0.8385 / 0.1276
$r = [5]$	$s = [4]$		22.79 / 0.8383 / 0.1214
$r = [2, 3]$	$s = [1, 2]$	Row-adaptive (Multi Scale)	23.68 / 0.8542 / 0.1159
$r = [3, 5]$	$s = [3, 5]$		23.22 / 0.8567 / 0.1097
$r = [2, 5]$	$s = [1, 4]$		**23.95 / 0.8595 / 0.1053**
$r = [2, 3, 5]$	$s = [2, 3, 5]$		23.65 / 0.8576 / 0.1056
$r = [2, 3, 5]$	$s = [1, 2, 4]$		23.90 / 0.8556 / 0.1088

2025). Accordingly, methods are evaluated using PSNR, SSIM, LPIPS and TMQI (Yeganeh & Wang, 2012), rather than linear-HDR metrics such as PSNR-PU, SSIM-μ or HDR-VDP.

After retraining all models on RSGR-HDR, we present experimental results in Table 1. Our method outperforms all existing approaches across diverse setups by a large margin (at least 1.5 dB on PSNR). Overall, algorithms designed for motion degradation fail under extreme overexposure or underexposure conditions, while HDR methods cannot handle motion distortions in dynamic scenes. The significant improvements of RIFE_m, IFED_m, and BiT_m over their corresponding baselines (RIFE, IFED, and BiT) clearly demonstrate the effectiveness and generalizability of our dual reversed setup. The qualitative comparison in Figure 5 further validates that our method can simultaneously expand dynamic range and reconstruct visual details from motion degradation. We further present the inference-time computational cost of all models in Table 1, measured on a single NVIDIA GeForce RTX 4090 with an input size of 640×480. The results demonstrate that our model attains the highest perceptual and fidelity scores while remaining competitively efficient.

To better substantiate the ability of our model in temporal consistency and local detail recovery, we also provide consecutive recovered frames in Section B.3 and corresponding video demos in the supplemental material, respectively, which allow a more direct evaluation of temporal stability.

4.2 Ablation Study

To assess the efficacy of proposed structure in our model, we conduct ablation studies as depicted in Table 2. We first remove feature alignment (V_1), and then successively replace our row-adaptive strategy with the dense and row-wise counterparts (V_2 and V_3). The performance significantly drops in all metrics. We also observe that the Swin Transformer block (V_4) fails to capture the complementarity between RSGR-iRSGR views. Although incorporating temporal loss in the sRGB space leads to certain improvements (V_5 vs. V_6), the gains remain limited. In contrast, our proposed tri-pivot loss contributes substantially to the final performance. We also compared with frequently used warping-based temporal constraint (Lai et al., 2018). Figure 6 demonstrates consistent findings from qualitative aspect. Moreover, we perform ablation study on row shapes used in feature alignment to explore the optimal hyper-parameters. As shown in Table 3, our model is retrained with different combinations of row size r and stride s, which further proves the effectiveness of our tailored design.



Figure 6: Visual results of model architecture ablation. We can observe that each key component contributes to the final results, especially in preserving local structures.

4.3 Visual Shifts and Convergent Exposures

As discussed in Section 1, the exposure convergence in the middle rows of RSGR-iRSGR views may complicate HDR expansion, but missing contrast can be compensated by our model. On the other hand,

Table 5: Quantitative comparisons on GOPRO-Dual. We follow the same evaluation protocol used in our experiments on RSGR-HDR. Performance is evaluated using mean PSNR/SSIM/LPIPS/TMQI.

Method	Input	HDR Reconstruction	HDR Photosequencing		
			×3	×5	×9
DSUN	*2·RG*	19.92 / 0.7850 / 0.1347 / 0.8691	– / – / –	– / – / –	– / – / –
JAMNet		25.16 / 0.8797 / 0.0940 / 0.9186	– / – / –	– / – / –	– / – / –
RSSR		12.21 / 0.6535 / 0.2394 / 0.8164	12.86 / 0.6860 / 0.2054 / 0.8033	12.94 / 0.6918 / 0.2079 / 0.8052	13.02 / 0.6908 / 0.1956 / 0.8060
CVR		13.29 / 0.6886 / 0.2291 / 0.8211	13.08 / 0.6768 / 0.2121 / 0.8069	13.14 / 0.6797 / 0.2164 / 0.8005	12.98 / 0.6699 / 0.2062 / 0.7978
RIFE		20.97 / 0.8327 / 0.1646 / 0.9412	20.91 / 0.8319 / 0.1756 / 0.8868	20.94 / 0.8330 / 0.1728 / 0.8896	20.96 / 0.8335 / 0.1710 / 0.8910
AfB		20.64 / 0.7021 / 0.3080 / 0.9354	20.85 / 0.7251 / 0.3165 / 0.8518	20.84 / 0.7212 / 0.3118 / 0.8540	20.84 / 0.7204 / 0.3092 / 0.8553
IFED		23.96 / 0.8447 / 0.0949 / 0.9431	24.27 / 0.8404 / 0.1036 / 0.8943	24.24 / 0.8424 / 0.1011 / 0.8974	24.24 / 0.8434 / 0.1000 / 0.8989
PMBNet		25.42 / 0.9009 / 0.1061 / 0.9252	24.94 / 0.8870 / 0.1198 / 0.9207	25.07 / 0.8907 / 0.1167 / 0.9191	25.11 / 0.8923 / 0.1160 / 0.9180
BiT	*3·RG*	23.96 / 0.8673 / 0.1034 / 0.9392	24.48 / 0.9011 / 0.0837 / 0.9302	24.40 / 0.8952 / 0.0863 / 0.9287	24.35 / 0.8921 / 0.0882 / 0.9273
ExpandNet	*1·RG*	17.59 / 0.7410 / 0.2702 / 0.8274	– / – / –	– / – / –	– / – / –
RSS-T		23.00 / 0.9229 / 0.0756 / 0.9316	– / – / –	– / – / –	– / – / –
DeMFI	*4·RG*	21.06 / 0.8520 / 0.1324 / 0.9203	20.75 / 0.8288 / 0.1847 / 0.8997	20.84 / 0.8348 / 0.1763 / 0.8940	20.59 / 0.8191 / 0.2014 / 0.8910
NGS	*5·RG*	31.04 / 0.9645 / 0.0343 / 0.9457	– / – / –	– / – / –	– / – / –
LAN	*5·LS*	34.38 / 0.9416 / 0.0589 / 0.7003	– / – / –	– / – / –	– / – / –
HDRFlow		21.17 / 0.4041 / 0.4663 / 0.6520	– / – / –	– / – / –	– / – / –
EBFI	*RG-E*	20.82 / 0.8864 / 0.1034 / 0.9354	20.73 / 0.8824 / 0.1040 / 0.9230	20.75 / 0.8837 / 0.1036 / 0.9232	20.77 / 0.8844 / 0.1034 / 0.9233
EvUnroll		29.27 / 0.9372 / 0.0496 / 0.9446	29.07 / 0.9321 / 0.0522 / 0.9348	29.13 / 0.9336 / 0.0513 / 0.9350	29.17 / 0.9345 / 0.0509 / 0.9351
EFI		20.45 / 0.8905 / 0.1005 / 0.9427	20.38 / 0.8861 / 0.1047 / 0.9371	20.40 / 0.8872 / 0.1032 / 0.9373	20.41 / 0.8874 / 0.1027 / 0.9373
UniINR		24.21 / 0.9124 / 0.1090 / 0.9335	23.31 / 0.8495 / 0.1731 / 0.9120	23.30 / 0.8468 / 0.1763 / 0.9114	23.31 / 0.8448 / 0.1794 / 0.9111
RIFE_m	*RG-iRG*	32.39 / 0.9591 / 0.0659 / 0.9544	33.56 / 0.9475 / 0.0710 / 0.9399	32.89 / 0.9502 / 0.0712 / 0.9407	32.49 / 0.9516 / 0.0712 / 0.9412
IFED_m		32.34 / 0.9625 / **0.0291** / 0.9542	32.64 / 0.9437 / **0.0341** / 0.9404	32.25 / 0.9486 / **0.0331** / 0.9420	32.05 / 0.9509 / **0.0327** / 0.9427
BiT_m		34.17 / 0.9733 / 0.0400 / 0.9538	33.68 / 0.9569 / 0.0477 / 0.9428	33.58 / 0.9614 / 0.0458 / 0.9439	33.53 / 0.9637 / 0.0449 / 0.9444
Ours		**35.88** / **0.9795** / 0.0293 / **0.9546**	**36.41** / **0.9708** / 0.0378 / **0.9471**	**35.88** / **0.9729** / 0.0369 / **0.9473**	**35.55** / **0.9738** / 0.0364 / **0.9474**

random visual shifts between corresponding rows are inevitable due to physical misalignments and fast motion. Such shifts, in turn, effectively mitigate the exposure-convergence issue. Meanwhile, our dual setting normally equips some tolerance to these constrained visual shifts (Zhong et al., 2022a; Ji et al., 2024). To prove this, we randomly shift the RSGR view in the image space along the horizontal and vertical axes, with maximal translational offsets as $4, 6, 8$ pixels, respectively. We then retrain our model using these misaligned pairs and present comparisons in Table 4 and Figure 7. The results reveal no significant performance drop, validating that our dual-view setup and customized model remain robust to dynamic scenes, where motion-induced spatial displacement helps alleviate central exposure convergence and limits its impact on reconstruction quality.

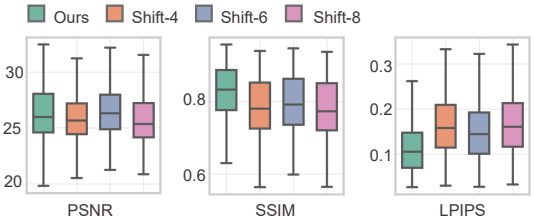


Figure 7: Performance distribution of our method and misaligned variants ('Shift-4', 'Shift-6' and 'Shift-8'). The box-plot further validates robustness to exposure convergence and visual shifts under fast motion.

4.4 Comparison on Synthetic Data

We also perform experiments on the synthetic dataset GOPRO-Dual to further demonstrate the conclusions drawn from real-data RSGR-HDR. Quantitative and qualitative results are presented in Table 5 and Figure B.5. The dataset synthesizing process and more experimental results are presented in Section A.2 and Section B.4.

Table 4: Quantitative comparisons on misaligned RSGR-iRSGR view. 'Shift-n' denotes misalignment with maximal offsets n.

Method	×3			×5			×9		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Shift-4	**23.88**	0.8483	0.1273	**23.94**	0.8501	0.1262	**23.97**	0.8511	0.1255
Shift-6	23.81	0.8536	0.1202	23.88	0.8555	0.1190	23.91	0.8563	0.1185
Shift-8	23.79	0.8464	0.1275	23.86	0.8483	0.1262	23.89	0.8493	0.1256
Ours	23.84	**0.8562**	**0.1071**	23.91	**0.8584**	**0.1058**	23.95	**0.8595**	**0.1053**

5 Conclusion

In this paper, we present a new task of HDR photosequencing from dual reversed RSGR inputs, addressing the inherent limitations of RSGR in frame rate and dynamic range. By exploiting the row-wise complementary property, our framework enables robust feature alignment and dynamic range expansion even under large motions. We also construct an optical system to collect dual reversed RSGR observations with sharp HDR sequences, and develop a network with row-adaptive alignment, mix-attention hallucination, and tri-pivot structural supervision, achieving improved detail recovery and reduced temporal flickering. Extensive experiments on both synthetic and real data validate the effectiveness of our proposed setting and model, demonstrating clear advantages over existing solutions.

AI USE STATEMENT

We used generative AI tools solely for language editing and polishing of the manuscript, including improvements to grammar, wording, clarity, and writing style. No generative AI tools were used for research ideation, methodology development, experiments, data analysis, or generation of research results. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the imaging system, data acquisition process, synthetic data generation, training settings, and additional experimental results in the appendix. These details are intended to facilitate reproduction of both our imaging setup and learning framework.

REFERENCES

Cenek Albl, Zuzana Kukelova, Viktor Larsson, Michal Polic, Tomas Pajdla, and Konrad Schindler. From two rolling shutters to one global shutter. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2505–2513, 2020.

Simon Baker, Eric Bennett, Sing Bing Kang, and Richard Szeliski. Removing rolling shutter wobble. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2392–2399. IEEE, 2010.

Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4):2049–2062, 2018.

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*, pp. 2502–2511, 2021.

Su-Kai Chen, Hung-Lin Yen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Wen-Hsiao Peng, and Yen-Yu Lin. Learning continuous exposure value representations for single-image hdr reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12990–13000, 2023.

Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 16111–16121, 2025.

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*, pp. 12760–12769, 2023.

Yuchao Dai, Hongdong Li, and Laurent Kneip. Rolling shutter camera relative pose: Generalized epipolar geometry. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4132–4140, 2016.

Bin Fan and Yuchao Dai. Inverting a rolling shutter camera: bring rolling shutter images to high framerate global shutter video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4228–4237, 2021.

Bin Fan, Yuchao Dai, and Mingyi He. Sunet: symmetric undistortion network for rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4541–4550, 2021.

Bin Fan, Yuchao Dai, Zhiyuan Zhang, Qi Liu, and Mingyi He. Context-aware video reconstruction for rolling shutter cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17572–17582, 2022.

Bin Fan, Yuxin Mao, Yuchao Dai, Zhexiong Wan, and Qi Liu. Joint appearance and motion learning for efficient rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5671–5681, 2023.

Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3586–3595, 2020.

Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *Asian Conference on Computer Vision*, pp. 486–500. Springer, 2010.

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*, pp. 1163–1170, 2013.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pp. 624–642. Springer, 2022.

Xiang Ji, Xinchen Ye, Hongcan Xu, and Haojie Li. Dense reconstruction from monocular slam with fusion of sparse map-points and cnn-inferred depth. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2018.

Xiang Ji, Zhixiang Wang, Shin'ichi Satoh, and Yinqiang Zheng. Single image deblurring with row-dependent blur magnitude. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12269–12280, 2023a.

Xiang Ji, Zhixiang Wang, Zhihang Zhong, and Yinqiang Zheng. Rethinking video frame interpolation from shutter mode induced degradation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12259–12268, 2023b.

Xiang Ji, Haiyang Jiang, and Yinqiang Zheng. Motion blur decomposition with cross-shutter guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12534–12543, June 2024.

Xiang Ji, Guixu Lin, Zhengwei Yin, Jiancheng Zhao, and Yinqiang Zheng. Moment-reenacting: Inverse motion degradation with cross-shutter guidance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

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

Erum Arif Khan, Ahmet Oguz Akyuz, and Erik Reinhard. Ghost removal in high dynamic range images. In *2006 International Conference on Image Processing*, pp. 2005–2008. IEEE, 2006.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

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.

Wei-Sheng Lai, Jia-Bin Huang, Oliver Wang, Eli Shechtman, Ersin Yumer, and Ming-Hsuan Yang. Learning blind video temporal consistency. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 170–185, 2018.

Yizhen Lao and Omar Ait-Aider. A robust method for strong rolling shutter effects correction using lines with automatic feature selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4795–4803, 2018.

Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014.

Peng Li, Yuan Liu, Xiaoxiao Long, Feihu Zhang, Cheng Lin, Mengfei Li, Xingqun Qi, Shanghang Zhang, Wei Xue, Wenhan Luo, et al. Era3d: High-resolution multiview diffusion using efficient row-wise attention. *Advances in Neural Information Processing Systems*, 37:55975–56000, 2024.

Chia-Kai Liang, Li-Wen Chang, and Homer H Chen. Analysis and compensation of rolling shutter effect. *IEEE transactions on image processing*, 17(8):1323–1330, 2008.

Guixu Lin, Jin Han, Mingdeng Cao, Zhihang Zhong, and Yinqiang Zheng. Event-guided frame interpolation and dynamic range expansion of single rolling shutter image. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3078–3088, 2023.

Peidong Liu, Zhaopeng Cui, Viktor Larsson, and Marc Pollefeys. Deep shutter unrolling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5941–5949, 2020a.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1651–1660, 2020b.

Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In *European Conference on Computer Vision*, pp. 344–360. Springer, 2022.

Yunfan Lu, Guoqiang Liang, Yusheng Wang, Lin Wang, and Hui Xiong. Uniinr: Event-guided unified rolling shutter correction, deblurring, and interpolation. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pp. 37–49. Wiley Online Library, 2018.

Marci Meingast, Christopher Geyer, and Shankar Sastry. Geometric models of rolling-shutter cameras. *arXiv preprint cs/0503076*, 2005.

Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion. In *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, pp. 382–390. IEEE, 2007.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Eyal Naor, Itai Antebi, Shai Bagon, and Michal Irani. Combining internal and external constraints for unrolling shutter in videos. In *European Conference on Computer Vision*, pp. 119–134. Springer, 2022.

Muyao Niu, Tong Chen, Yifan Zhan, Zhuoxiao Li, Xiang Ji, and Yinqiang Zheng. Rs-nerf: Neural radiance fields from rolling shutter images. In *European Conference on Computer Vision*, pp. 163–180. Springer, 2024.

Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021.

Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *European Conference on Computer Vision*, pp. 198–215. Springer, 2022.

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.

Pulak Purkait, Christopher Zach, and Ales Leonardis. Rolling shutter correction in manhattan world. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 882–890, 2017.

Delin Qu, Yizhen Lao, Zhigang Wang, Dong Wang, Bin Zhao, and Xuelong Li. Towards nonlinear-motion-aware and occlusion-robust rolling shutter correction. *arXiv preprint arXiv:2303.18125*, 2023.

Vijay Rengarajan, Ambasamudram N Rajagopalan, and Rangarajan Aravind. From bows to arrows: Rolling shutter rectification of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2773–2781, 2016.

Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European conference on computer vision*, pp. 184–201. Springer, 2020.

Erik Ringaby and Per-Erik Forssén. Efficient video rectification and stabilisation for cell-phones. *International journal of computer vision*, 96:335–352, 2012.

David Schubert, Nikolaus Demmel, Lukas von Stumberg, Vladyslav Usenko, and Daniel Cremers. Rolling-shutter modelling for direct visual-inertial odometry. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2462–2469. IEEE, 2019.

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

Zifan Shi, Na Fan, Dit-Yan Yeung, and Qifeng Chen. Stereo waterdrop removal with row-wise dilated attention. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3829–3836. IEEE, 2021.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 237–246, 2017.

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 conference on computer vision and pattern recognition*, pp. 8934–8943, 2018.

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.

Jonas Unger and Stefan Gustavson. High-dynamic-range video for photometric measurement of illumination. In *Sensors, Cameras, and Systems for Scientific/Industrial Applications VIII*, volume 6501, pp. 106–115. SPIE, 2007.

Subeesh Vasu, AN Rajagopalan, et al. Occlusion-aware rolling shutter rectification of 3d scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 636–645, 2018.

Zhixiang Wang, Xiang Ji, Jia-Bin Huang, Shin'ichi Satoh, Xiao Zhou, and Yinqiang Zheng. Neural global shutter: Learn to restore video from a rolling shutter camera with global reset feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17794–17803, 2022.

Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based blurry frame interpolation under blind exposure. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1588–1598, 2023.

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 (ECCV)*, pp. 117–132, 2018.

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*, pp. 24851–24860, 2024.

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*, pp. 1751–1760, 2019.

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.

Weilong Yan, Robby T Tan, Bing Zeng, and Shuaicheng Liu. Deep homography mixture for single image rolling shutter correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9868–9877, 2023.

Xinchen Ye, Xiang Ji, Baoli Sun, Shenglun Chen, Zhihui Wang, and Haojie Li. Drm-slam: Towards dense reconstruction of monocular slam with scene depth fusion. *Neurocomputing*, 396:76–91, 2020.

Hojatollah Yeganeh and Zhou Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image processing*, 22(2):657–667, 2012.

Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2011.

Jiancheng Zhao, Yifan Zhan, Qingtian Zhu, Mingze Ma, Muyao Niu, Zunian Wan, Xiang Ji, and Yinqiang Zheng. Tree-nerv: Efficient non-uniform sampling for neural video representation via tree-structured feature grids. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15076–15085, 2025.

Zhihang Zhong, Mingdeng Cao, Xiao Sun, Zhirong Wu, Zhongyi Zhou, Yinqiang Zheng, Stephen Lin, and Imari Sato. Bringing rolling shutter images alive with dual reversed distortion. In *European Conference on Computer Vision*, pp. 233–249. Springer, 2022a.

Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from blur: Multi-modal blur decomposition with motion guidance. In *European Conference on Computer Vision*, pp. 599–615. Springer, 2022b.

Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5713–5723, 2023.

Xinyu Zhou, Peiqi Duan, Yi Ma, and Boxin Shi. Evunroll: Neuromorphic events based rolling shutter image correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17775–17784, 2022.

Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Rolling-shutter-aware differential sfm and image rectification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 948–956, 2017.

Bingbing Zhuang, Quoc-Huy Tran, Pan Ji, Loong-Fah Cheong, and Manmohan Chandraker. Learning structure-and-motion-aware rolling shutter correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4551–4560, 2019.

Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pp. 405–414. Wiley Online Library, 2011.

Yunhao Zou, Yinqiang Zheng, Tsuyoshi Takatani, and Ying Fu. Learning to reconstruct high speed and high dynamic range videos from events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2024–2033, 2021.

A Implementation Details

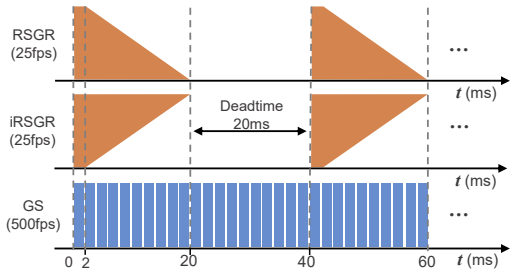
A.1 Construction of Imaging System

To construct a realistic dataset that contains perfectly aligned RSGR, iRSGR, and high-speed HDR video streams, we begin by fixing the RSGR camera and manually tuning the pose of the remaining cameras. This is guided by inspecting checkerboard residuals. However, achieving sub-pixel alignment for all four cameras across every axis proves extremely challenging due to the intricate nature of the multi-sensor rig. As a result, we eventually lock all components of the system in place and perform the alignment computationally. Specifically, we follow strategy in Rim et al. (2020) and estimate three homographies under a closed-loop constraint to synchronize the viewpoints of the four sensors. The detailed specifications and acquisition parameters for each device are listed in Table A.1.

Extended Table A.1: Specifications of our coaxial imaging system. The deadtime between two adjacent high-speed frames of GS1 and GS2 is extremely short and can therefore be ignored.

Device	RSGR	iRSGR	GS 1	GS 2
Exp. per Row	$2 \sim 20$ ms	$2 \sim 20$ ms	2 ms	2 ms
Delay. per Row	37.5 μs	37.5 μs	0 μs	0 μs
Exp. per Frame	20 ms	20 ms	2 ms	2 ms
Deadtime	20 ms	20 ms	0 ms	0 ms
Frame rate	25 fps	25 fps	500 fps	500 fps
Resolution	640×480	640×480	640×480	640×480

The GS camera used in our system—BTRAN CS-700C equipped with a SONY IMX426 sensor—exhibits very low noise. Two factors mainly contribute to this. First, the IMX426 increases its effective pixel size to $9\mu m \times 9\mu m$ by combining four $4.5\mu m \times 4.5\mu m$ subpixels at the circuit level, resulting in high sensitivity and reduced read noise. Second, the camera is actively cooled to $0°$C, which further suppresses thermally induced noise.



Extended Figure A.1: Temporal alignment of sensors in our imaging system.

To ensure that the system remains feasible in practice, we pair the imaging device with a FUJINON HF12.5HA-1S lens (focal length 12.5 mm, supporting 2/3-inch formats). Given the effective pixel pitch of $9\mu m$, the active sensing region of 640×480 corresponds to $5.7mm \times 4.3mm$. Because the relay lens operates at a $1:1$ magnification ratio, the horizontal and vertical fields of view become $26.1°$ and $19.5°$, respectively. At a distance of $10m$, this equates to a view coverage of $4.6m$ horizontally and $3.4m$ vertically, making the setup suitable for real-world capture. Although lens distortion may interact with rolling-shutter effects, selecting a high-grade FUJINON lens helps minimize such artifacts.

We follow the method proposed in Lin et al. (2023); Zou et al. (2021) to merge the two GS frames into HDR image. One of the high-speed sensors is fitted with a Thorlabs ND513B neutral-density filter to attenuate incoming light. ND filters reduce irradiance uniformly in both spatial and spectral domains. This enables us to record two GS frames with different illumination levels without altering exposure duration, which is often difficult to control reliably when capturing at high frame rates. The two complementary exposures provide the necessary variation for synthesizing HDR results.

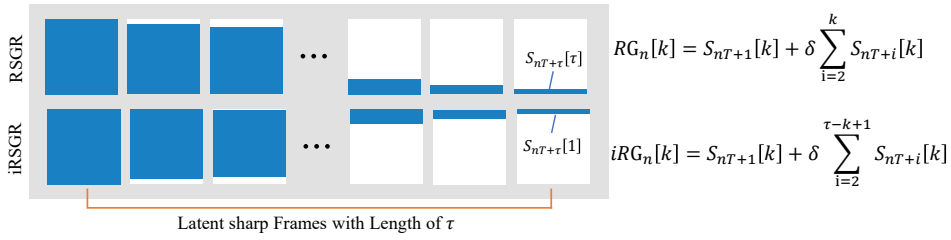
A.2 Experimental Data and Training Details

A.2.1 Real Dataset

The comprehensive methodology for collecting our real data is detailed in Section 3.1. Utilizing the capabilities of our imaging system, we collected the RSGR-HDR dataset, which comprises 90 urban scenes. Each scene includes 70×2 degraded frames and 1,400 aligned high-speed sharp frames. As shown in Table A.1 and Figure A.1, each RSGR–iRSGR pair is acquired over a 40 ms cycle, consisting of a 20 ms exposure followed by a 20 ms dead time. Meanwhile, the two GS cameras operate at 500 fps with a 2 ms exposure and no dead time, producing 20 synchronized GS pairs per cycle. Ten pairs fall within the RSGR–iRSGR exposure window and are fused into 10 temporally aligned HDR frames, while the remaining ten fall within the dead time. Therefore, in RSGR-HDR, each RSGR–iRSGR pair corresponds to 10 HDR GT frames. Representative RSGR-



Extended Figure A.2: Data samples from our RSGR-HDR. RSGR view is exposed downwards while iRSGR scans upwards. We present three groundtruth frames which are temporally located at $T = 0, 0.5, 1$. Here, the time instants are defined as normalized timestamps in $[0, 1]$ over the latent video duration.

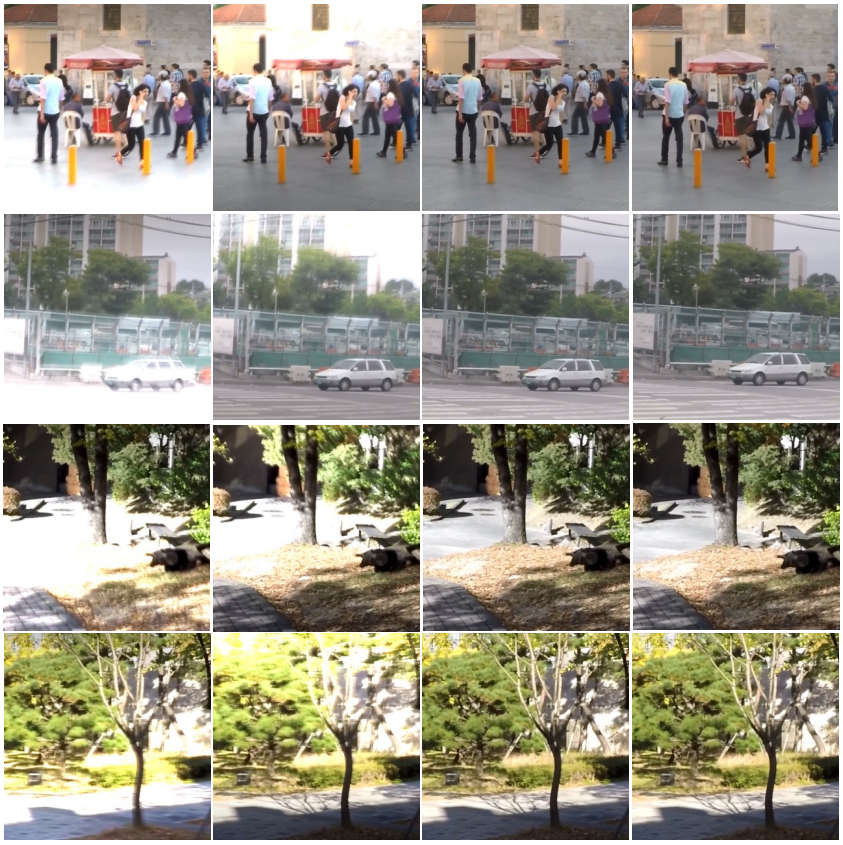


Extended Figure A.3: Synthetic method. The notation $RG[k]$ denotes extracting the k-th row from frame RG. n is the index of RSGR or iRSGR view. T is the number of latent frames that correspond to exposure and deadtime. RG denotes the RSGR images with scanning direction from top to bottom while iRG has the inverted exposure direction. δ is the ratio between constant delay among rows and duration of the first exposed row.

HDR samples are illustrated in Figure A.2. The visual discrepancies between the two views are distinctly noticeable. The mutually inverted scanning manner allows the exposure compensation, which provides supplemental details and cues for robust HDR reconstruction. We further divide them into 54, 12 and 24 sequences as our training, validating, and testing set, respectively.

A.2.2 Synthetic Dataset

To strengthen our study of how dual reversed RSGR setting aids HDR photosequencing, we also developed a synthetic dataset called GOPRO-Dual, adhering to the methodologies specified in Ji et al. (2023a); Wang et al. (2022). While synthetic data cannot fully capture the complexity and diversity of real-world scenes, it is used here only for supplementary and independent validation. Our main training and evaluation are conducted on the real-world RSGR-HDR dataset. Meanwhile, experiments on synthetic data further verify our findings and provide a meaningful reference for specific analyses (e.g., ensuring fair comparison across methods). This dataset draws from the GOPRO data (Nah et al., 2017), which encompasses 33 videos with a resolution of 1280×720. Each video features 1200 consecutive frames shot at 240fps. To render the dataset more lifelike, we significantly enhanced the



Extended Figure A.4: Data samples from our synthetic GOPRO-Dual dataset. For each dual reversed RSGR pair, we present two groundtruth HDR frames here which are temporally located at $T = 0, 1$.

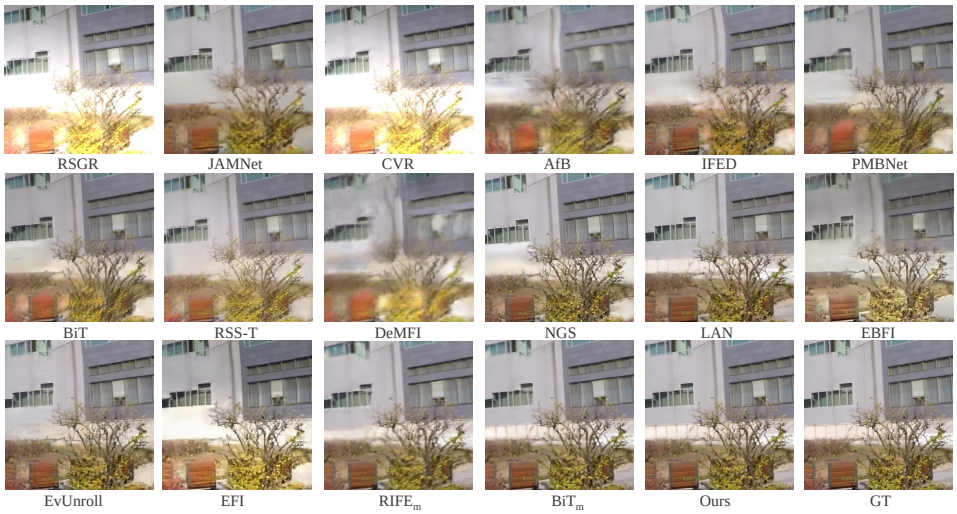
framerate of the GOPRO videos, increasing them by a factor of $\times 64$ through a widely-used video interpolation algorithm (Huang et al., 2022).

Figure A.3 illustrates our synthesizing method. We formulate the imaging process of RSGR as weighted sum of patches extracted from corresponding latent sharp frames within exposure time τ (Ji et al., 2023a; Wang et al., 2022). RSGR view is synthesized with scanning direction from top to bottom, while iRSGR exposes frames upwards. We strictly constrain the generation process as shown in Figure A.3, the two reversed RSGR views are aligned in frame level to ensure they observe the identical visual content of scenes. Specifically, the original GOPRO frames are centrally cropped as 512×512 and we set $T = \tau = 512$, which leads to 512 sharp frames corresponding to an RSGR or iRSGR frame. For enabling supervised learning, we uniformly sampled 9 frames out of them to take as ground truth. Figure A.4 exhibits some examples of synthesized data.

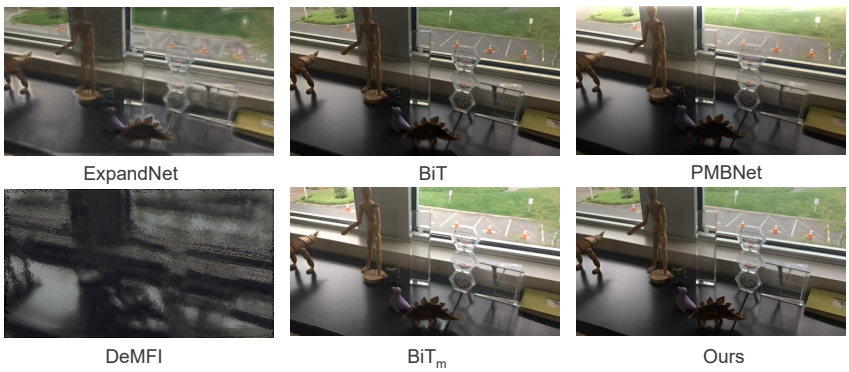
In the experiments, we compared our method with some competitive settings: EFI (Lin et al., 2023), EvUnroll (Zhou et al., 2022) and EBFI (Weng et al., 2023) assisted by *event camera*, LAN (Chung & Cho, 2023) using *alternating exposure* within an input sequence. However, our collected dataset: RSGR-HDR and GOPRO-Dual, exclude the event or multi-exposure counterpart. Thus, we synthesized the event stream from the high frame-rate videos using an event simulator (Gehrig et al., 2020). The alternating exposure sequence could be generated based on the HDR frames through mapping strategy provided in Chen et al. (2021).

A.2.3 Training Details

During learning process, each training sample consists of RSGR-iRSGR pairs and a HDR video clip with frame length of 9. We adopt Adam optimizer (Kingma & Ba, 2014) with initial learning rate of



Extended Figure B.5: Qualitative comparisons on GOPRO-Dual. Our model outperforms the existing approaches with different settings in details recovery and dynamic range expansion. Best viewed with zoom.

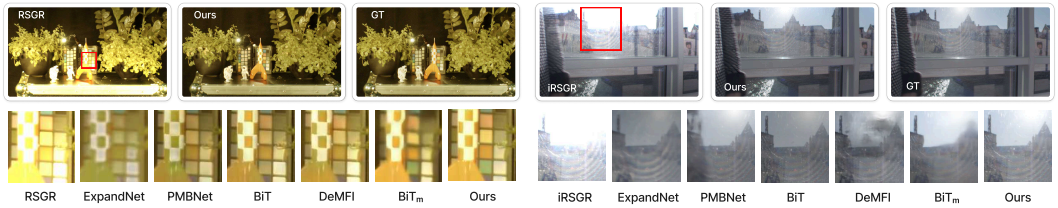


Extended Figure B.6: **Qualitative comparison** with representative SOTAs on Adobe-Dual testset.

Extended Table B.2: **Quantitative comparison** with representative SOTAs on Adobe-Dual testset.

Method	Input	×1			×9		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
ExpandNet (Marnerides et al., 2018)	*1·RG*	15.75	0.7220	0.1877	–	–	–
PMBNet (Ji et al., 2023b)	*2·RG*	16.79	0.7631	0.1450	17.36	0.7550	0.1632
BiT (Zhong et al., 2023)	*3·RG*	18.79	0.8187	0.1291	18.29	0.7676	0.1481
DeMFI (Oh & Kim, 2022)	*4·RG*	13.94	0.4089	0.4709	13.29	0.3717	0.5231
BiT$_m$ (Zhong et al., 2023)	*RG·iRG*	22.48	0.8802	0.0954	21.18	0.8270	0.1169
Ours		23.54	0.8950	0.0775	22.25	0.8463	0.0964

10^{-4} combined with a cosine annealing scheduler. The total training epoch is set as 800. To augment data, training samples are randomly cropped to 256×256, then flipped vertically and horizontally. The whole training process is performed on two GPUs of NVIDIA GeForce RTX 5090 with batch size of 8. Conventionally, we assess the performance of our models using established metrics: PSNR, SSIM, and LPIPS.



Extended Figure B.7: **Qualitative comparison** with representative SOTAs on LiU-HDRv-Dual testset. Best viewed with zoom.

Extended Table B.3: **Quantitative comparison** with representative SOTAs on LiU-HDRv-Dual testset.

Method	Input	×1			×9		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
ExpandNet (Marnerides et al., 2018)	*1·RG*	17.45	0.7274	0.2987	–	–	–
PMBNet (Ji et al., 2023b)	*2·RG*	19.41	0.7753	0.2374	19.39	0.7753	0.2544
BiT (Zhong et al., 2023)	*3·RG*	19.94	0.7687	0.2460	20.36	0.7964	0.1861
DeMFI (Oh & Kim, 2022)	*4·RG*	17.42	0.7257	0.2582	17.66	0.7254	0.2969
BiT$_m$ (Zhong et al., 2023)	*RG·iRG*	22.14	0.8323	0.2165	22.80	0.8282	0.2111
Ours		24.24	0.8604	0.1538	24.72	0.8567	0.1512

B Additional Results

B.1 Qualitative Results on Synthetic Data

In Section 4.4, we compare our model with SOTA methods and report quantitative results. Here, we further present the corresponding qualitative comparisons in Figure B.5. These visual results complement the quantitative evaluation and demonstrate the advantages of our method in recovering fine details and handling regions with challenging dynamic ranges under different baseline settings.

B.2 Third-party Evaluation

To further evaluate the generalization ability of our framework, we additionally construct two third-party testsets, Adobe-Dual and LiU-HDRv-Dual, based on Adobe240 (Su et al., 2017) and the LiU HDRv (Kronander et al., 2014), respectively. Adobe-Dual contains diverse scene contents, motion patterns, illumination conditions, and synthesis parameters that differ from those of our training data, while LiU-HDRv-Dual is introduced to evaluate the performance of our method on conventional HDR scenes with pronounced illumination variations, dark regions, and saturated highlights. Following the same dual reversed RSGR imaging pipeline described in Section A.2.2, we synthesize the corresponding RSGR-iRSGR observations and HDR targets for both testsets. In all experiments, the pretrained model is directly evaluated without any finetuning or domain adaptation.

Table B.2 and Figure B.6 present the quantitative and visual comparisons on Adobe-Dual, demonstrating the robustness of our method to unseen scenes, motions, synthesis parameters, and illumination conditions. Meanwhile, the results on LiU-HDRv-Dual in Table B.3 and Figure B.7 further verify its generalization to conventional HDR scenes with challenging underexposed and overexposed regions. On both testsets, our method achieves the best overall performance and consistently reconstructs fine temporal details with a wide dynamic range.

These results suggest that the superiority of our framework does not mainly originate from overfitting to a specific dataset distribution, but rather from the intrinsic row-wise complementary property of dual reversed RSGR inputs (BiT *vs.* BiT$_m$) together with the proposed row-adaptive alignment and correlation-guided hallucination modules (Ours *vs.* BiT$_m$).

B.3 Video Reconstruction Results

To better substantiate the temporal consistency of our model in HDR photosequencing, we apply our model and competitive SOTAs to generate multiple consecutive latent HDR frames, whose

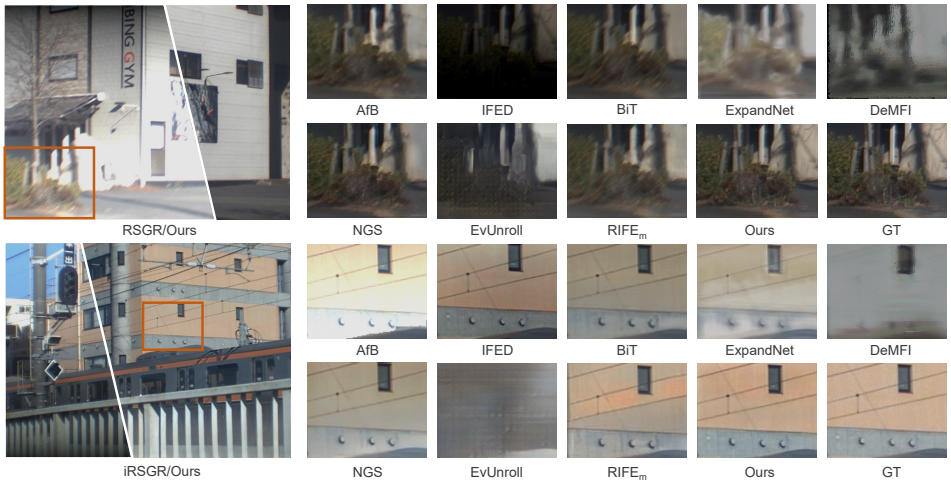


Extended Figure B.8: Qualitative comparison on HDR photosequencing of GOPRO-Dual. We present the multiple intermediate frames (5 out of 9) at different time generated by different models. Each row denotes reconstructed latent frames temporally located at $t = [0, 2/8, 4/8, 6/8, 1]$. Best viewed with zoom.

visual results on GOPRO-Dual are given in Figure B.8. In terms of motion prediction, EFI fails to distinguish the movement of waving hand throughout all latent frames. The other approaches, such as PMBNet, EvUnroll and DeMFI, are able to disambiguate temporal direction but also introduce significant artifacts and distortions. Compared with BiTm, our results are clearer and closer to ground truths. Consistent observation can be found in the real data presented in Figure B.9.



Extended Figure B.9: Qualitative comparison on HDR photosequencing of RSGR-HDR. We present the multiple intermediate frames (5 out of 9) at different time generated by different models. Each row denotes reconstructed latent frames temporally located at $t = [0, 2/8, 4/8, 6/8, 1]$. Best viewed with zoom.



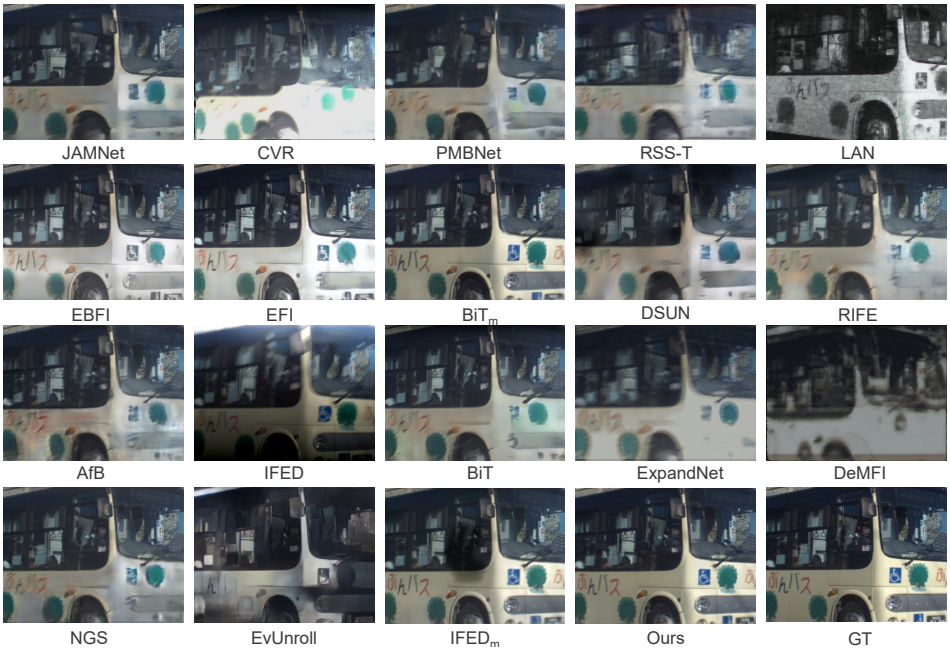
Extended Figure B.10: Qualitative comparisons on RSGR-HDR. Supplementary qualitative results corresponding to Figure 5 in the main manuscript.



Extended Figure B.11: Additional qualitative comparisons on RSGR-HDR.

B.4 Additional Qualitative Results

We present additional qualitative comparisons on RSGR-HDR and GOPRO-Dual in Figure B.10 – Figure B.16.



Extended Figure B.12: Additional qualitative comparisons on RSGR-HDR.



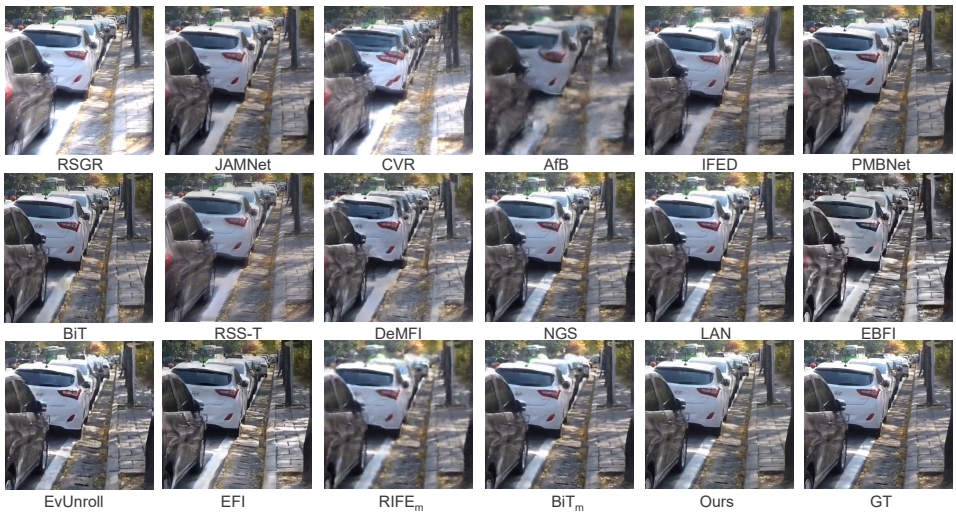
Extended Figure B.13: Additional qualitative comparisons on RSGR-HDR.



Extended Figure B.14: Additional qualitative comparisons on GOPRO-Dual.



Extended Figure B.15: Additional qualitative comparisons on GOPRO-Dual.



Extended Figure B.16: Additional qualitative comparisons on GOPRO-Dual.