Motion Blur Decomposition with Cross-shutter Guidance

Xiang Ji Haiyang Jiang Yinqiang Zheng[†]
The University of Tokyo, Japan
{jixiang,jiang-haiyang777}@g.ecc.u-tokyo.ac.jp,
yqzheng@ai.u-tokyo.ac.jp

Abstract

Motion blur is a frequently observed image artifact, especially under insufficient illumination where exposure time has to be prolonged so as to collect more photons for a bright enough image. Rather than simply removing such blurring effects, recent researches have aimed at decomposing a blurry image into multiple sharp images with spatial and temporal coherence. Since motion blur decomposition itself is highly ambiguous, priors from neighbouring frames or human annotation are usually needed for motion disambiguation. In this paper, inspired by the complementary exposure characteristics of a global shutter (GS) camera and a rolling shutter (RS) camera, we propose to utilize the ordered scanline-wise delay in a rolling shutter image to robustify motion decomposition of a single blurry image. To evaluate this novel dual imaging setting, we construct a triaxial system to collect realistic data, as well as a deep network architecture that explicitly addresses temporal and contextual information through reciprocal branches for cross-shutter motion blur decomposition. Experiment results have verified the effectiveness of our proposed algorithm, as well as the validity of our dual imaging setting.

1. Introduction

Photo capture usually needs sufficient exposure time to collect photons. If ego-motion of camera or dynamic objects are presented during this period, the resultant image will suffer from motion blur. This kind of degradation makes visual content less interpretable. Hence, extensive research has been devoted to reverse this process and produce sharper details.

Generally, the deblurring course is formulated as image-to-image transition by extracting a single latent frame from a blurry input [7, 17, 20, 21, 28, 29]. Recently, researchers step froward to a more ambitious task, retrieving a sharp image sequence instead of just one frame, dubbed as blur decomposition [22]. Unfortunately, averaging effects of motion blur have severely destroyed the temporal ordering of latent frames, which cannot be restored solely by reconstruction loss [11]. To make matters worse, motion ambiguity resides in each dynamic object, thus leading to numerous plausible solutions, yet many of which are physically infeasible. Figure 1 (a) and (b) illustrate this ambiguity by using two dynamic objects. For a given blur observation, there exist four motion sequences that could be interpreted as its decomposition. Models only trained on this data usually have issues of instability and low performance [36].

Handling the loss of temporal order is a problem far from being well-studied in the blur decomposition. Current solutions mainly fall into two categories: (a) introducing ordering-invariant loss [11] and (b) approximating latent temporal order within exposure by motion of consecutive blur frames [34, 36]. The former one is easily caught in sub-optimal solutions owing to the weak supervision propagated from the loss. Similarly, the latter solution suffers from severe degeneration when presented with long exposure time or fast motion. Moreover, motion estimation among blurry frames is inherently nontrivial and time-consuming. Recently, it has been recognized that rolling shutter (RS) images encode the canceled motion due to its row-by-row exposure mode [4, 6], according to which RS effects can be mitigated. But restoring a sequence from single RS input still remains unfinished because of lacking complete global content. In contrast, Blur images contain adequate contextual information but without temporal ordering. On the other hand, dual camera system has been widely exploited in RS correction (RS-RS, RS-Event) [1, 35, 37], deblurring (GS-Event) [27, 31], even vibration sensing (GS-RS) [25].

Therefore, considering the complementarity of Blur and RS images, we propose dual Blur-RS setting to solve the motion ambiguity of blur decomposition. As shown in Figure 1 (c), the RS view not only provides local details but also implicitly captures temporal order of latent frames. Meanwhile, GS view could be exploited to mitigate the initial-state ambiguity from RS counterpart (as discussed in Section 4.4). Inspired by the hardware design of [24, 33], we devised our triaxial imaging system to capture strictly aligned high-speed sharp videos and low-speed Blur-RS pair videos.

[†] Corresponding author
https://github.com/jixiang2016/dualBR

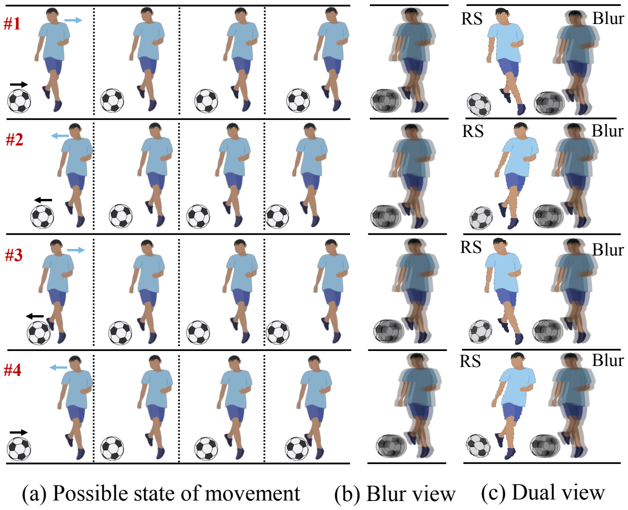
Facilitated by the collected dataset, we further proposed a novel two-staged model, containing motion interpretation and blur decomposition modules, to reconstruct a sharp video sequence from cross shutter views: blur and RS observations. The motion interpretation module firstly disentangles the bilateral motion fields as complementary dual stream: Blur and RS branches, to actively address the contextual characterization and temporal abstraction, respectively. Shutter alignment and aggregation enables mutual boosting of two branches by propagating aligned and aggregated feature to each other. Besides, the temporal positional encoding further enhances model's ability to disambiguate motion direction of latent frames. Subsequently, estimated motion fields along with blur-RS inputs will be warped and refined through blur decomposition module to generate a sharp video clip. At last, we also deeply explored the advantages of our proposed Blur-RS combination over existing settings for blur and RS decomposition by providing experimental results. In summary, our main contributions are:

- We present a new setting of dual Blur-RS combination to address the motion ambiguity of blur decomposition, and demonstrate its superiority to pure RS or blur setting.
- Rather than a biaxial system for image-to-image deblurring, we develop a triaxial imaging system that simultaneously captures Blur-RS pairs along with high-speed ground truth, and collect a real dataset named RealBR.
- We introduce a novel neural network architecture that actively address the contextual characterization and temporal abstraction by dual stream motion interpretation module. Extensive experiments have validated the effectiveness of our setting and model.

2. Related Work

2.1. Blur Decomposition

Compared with traditional deblurring task, reconstructing an image sequence from single blurred input is much more challenging because the average effects have destroyed the temporal ordering of latent frames. Jin *et al*. [11] raise this problem and address the temporal ordering ambiguity for the first time. From two aspects, they design a network with large receptive field to tackle the inherent ill-posedness of deblurring and ordering-invariant loss for motion ambiguity. [22] specially proposes a surrogate task to learn motion representation from sharp videos in an unsupervised manner, and then employs it as a guidance of training motion encoder for blurred images. [2] takes advantages of spatial transformer network modules to restore a video sequence and its underlying motion in an end-to-end manner. In order to avoid the directional ambiguity, BiT [34] takes three consecutive blurry frames as input to extract the motion prior. They propose a blur intra-interpolation transformer based on novel multi-scale Swin transformer blocks along with dual-end temporal supervision and symmetric ensembling strategies to effectively unravel the underlying motion within the exposure time and recover arbitrary sharp frames from blur. Zhong *et al*. [36] also emphasize on the motion ambiguity of blur decomposition by introducing motion guidance representation. They provide three interfaces to acquire the motion guidance: learning by network, approximating from blur video and user input. The learning one is also supervised by optical flow of blur frames. So essentially, solutions proposed in [22, 34, 36] all try to represent the latent motion by using consecutive blurry inputs.



(a) Possible state of movement (b) Blur view (c) Dual view

Figure 1. **Motion ambiguity of blur observation.** In this toy example, we show two objects: a soccer and a player, both moving horizontally. (a) shows four possible motion states (both moving right, both moving left, one moving left and the other is towards right.) during the exposure time. (b) presents corresponding motion blurred observations. They are all identical due to averaging effects, which brings about motion ambiguity to blur decomposition. (c) In our dual Blur-RS setting, rolling shutter (RS) view implicitly encoded temporal ordering of latent frames.

Different from blur decomposition, another line of related work inherits video frame interpolation setting by substituting inputs as blurry frames to reconstruct clear images within deadtime. Most of them take as input an image sequence and interpolate a sharp one at the middle of two blurry frames [12, 26, 32]. The recent study has started to generate interpolated frame at arbitrary time [10, 19]. In this paper, we mainly focus on the motion ambiguity of blur decomposition. So, we do not expand the discussion of these works in detail.

2.2. Dual Camera System

Recently, significant progress has been made by constructing dual camera view to handle different vision tasks. According to combination manner, we briefly divide them as: RS-RS [1, 35], RS-Event [37], GS-Event [27, 31] and RS-GS [25].

Due to the ill-posedness of single RS correction, Albl *et*

Table 1. **Specifications of our triaxial imaging system**. The deadtime between two adjacent high speed frames is extremely short and thus can be ignored.

Device	RS camera	GS camera	HS camera
Resolution	800×800	800×800	800×800
Frame rate	20 fps	20 fps	500 fps
Exp. per Row	2 ms	18 ms	2 ms
Delay. per Row	20 μs	0 μs	0 μs
Exp. per Frame	18 ms	18 ms	2 ms
Deadtime	32 ms	32 ms	0 ms

al. [1] resort to a camera configuration: two RS cameras with reversed scanning direction. They further proved that the setup possesses geometric constraints needed to correct rolling shutter distortion using only a sparse set of point correspondences between the two images. Lately, Zhong *et al*. [35] extend this setup to learning based method. Instead of correcting RS geometrically, they develop an end-to-end model, to generate dual optical flow sequence through iterative learning. In contrast, considering the high-speed characteristic of the event camera, Zhou *et al*. [37] introduce a novel computational imaging setup consisting of an RS sensor and an event sensor to correct RS effects.

Similarly, by exploring the high-temporal resolution property of events, [27, 31] use a cross-modal set up: GS-Event to solve the deblurring problem. Sun *et al*. [27] unfold blurring process into an end-to-end two stage restoration network by effectively fusing event and image features. They design an event-image cross-modal attention module, which allows network to focus on relevant features from the event branch and filter out noise. Xu *et al*. [31] aim at data inconsistency, constructing a piece-wise linear motion model taking into account motion non-linearities, to achieve accurate deblurring in a self-supervised manner.

Most interestingly, Sheinin *et al*. [25] simultaneously capture the vibration with two cameras equipped with rolling and global shutter sensors, respectively. The RS camera captures distorted speckle images that encode the high-speed object vibrations, while GS camera captures undistorted reference images of the speckle pattern, helping to decode the source vibrations.

3. Methodology

3.1. Proposed Setup of Blur Decomposition

Usually, Blur accumulation can be formulated as an averaging process presented in a linear space by using inverse camera response function (CRF) on RGB images [17, 18]. While blur decomposition aims at extracting uniformly distributed sharp frames from single blurred image.

As we discussed in Section 1, this process is highly ill-posed because of motion ambiguity residing in accumulation of photons. Considering the characteristic of RS exposure that inherently encodes temporal ordering of latent frames and provide local details as supplementary to global content of blur, we additionally capture an RS view R of each blurred frame B so as to better address the indeterminacy. As a result, the decomposition problem is formulated as:

$$(B, R) \mapsto \mathbf{I} = \{S^t, t \in 0, \cdots, N-1\} \quad (1)$$

3.2. Optical System and Dataset

Optical System In order to capture aligned training inputs (RS, GS) and ground truth sequences that can be recorded by a high-speed camera (HS), we constructed a triaxial optical system as depicted in Figure 2. Similar to [38], the system comprises 2 beam-splitters that partition incident light into 3 identical beams, and 3 cameras for RS (FLIR BFS-U3-63S4C with 2x2 binning), GS (FLIR GS3-U3-23S6C), HS (a high-speed GS camera, BITRAN CS-700C, with forced cooling) separately. Other specifications for 3 cameras are detailed in Table 1.The usage of a neutral density (ND) filter with roughly 20% transmittance is for the purpose of counterbalancing excessive brightness in blurry GS images brought about by their relatively long exposure duration. While beam-splitters feeding light signals of the same scene to 3 cameras, and the other ND-filter equalizing illumination magnitude of blur and RS views, we further incorporated geometrical transforms for pixel-level alignment and synchronization signal control for simultaneity. For more details, please refer to our supplementary materials.

RealBR Dataset Applying the optical system to capture real world image sequences, we established a RealBR (GS Blur & RS) dataset by recording 54 distinct street scenes containing ample amount of objects, like vehicles and pedestrians, and various camera motions. In each scene, we have 56 pairs of consecutive RS and GS blur frames, and 1400 corresponding sharp HS images. As can be seen from Figure 2(c), capturing of single pair of Blur-RS frames has a period of 50ms, with GS and RS cameras finishing their exposure in 18ms leaving the rest 32ms in one period as deadtime, while HS camera exposing within 2ms at 500fps taking 25 frames in total within one period, 9 of which temporally located inside GS/RS exposure duration and the rest 16 inside deadtime. After necessary preprocessing, we reorganized entire dataset and split it into 40, 4, and 10 scenes for training, validation and test. All captured images are in both RGB and RAW format, and will be made publicly available, facilitating related potential exploration in the area.

3.3. The Proposed Architecture

The overview of our proposed architecture is shown as Figure 3 (a). We mainly focus on reconstructing a clear video sequence from a blurred image with the assistance of its RS view to address the motion ambiguity issues. The inference

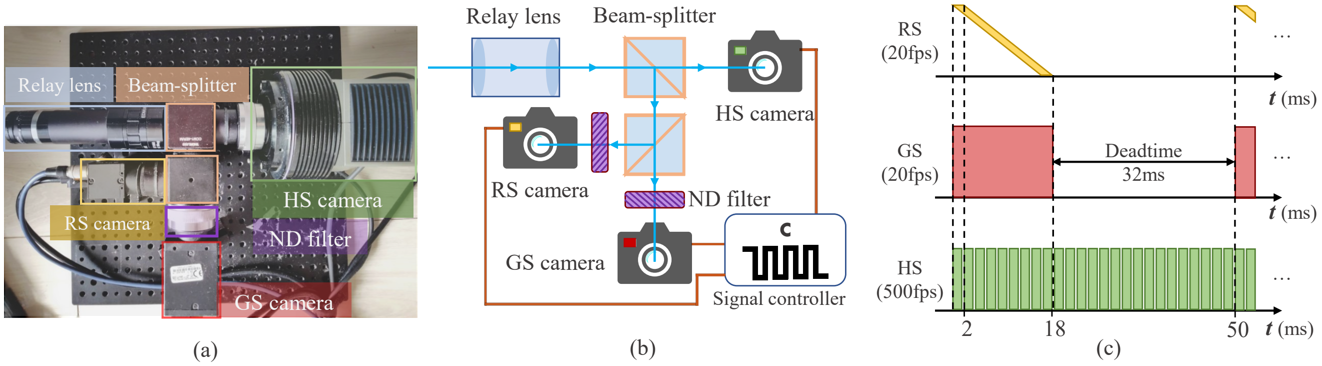


Figure 2. **Our triaxial imaging system.** (a) A photo of the actual system for data gathering; (b) Optical diagram of our system; (c) Illustration of exposure duration for all cameras on temporal axis. In picture (c), its vertical axes can be interpreted as spatial rows of captured images from each camera.

process of our $\mathcal{F}$ is formulated as:

$$\mathbf{S} = \mathcal{F}\left(B, R\right) \tag{2}$$

where $\mathbf{S} = \{S^t, t \in 0, \cdots, N-1\}$ denotes extracted latent sharp video sequence with a length of N. R is the RS view of blurred input B.

Overall, the whole model could be divided into two stages: motion interpretation and blur decomposition. Motion interpretation (MI) is highly attentive to explore the benefits of our Blur-RS combination in an iterative manner with three motion interpretation blocks (MIB). With the guidance of temporal positional encoding, it explicitly emphasizes the contextual characterization and temporal abstraction from disentangled blur and RS streams, respectively. Estimating bidirectional motion fields can be described as bellow:

$$\left(\mathbf{F}_{S\to B}, \mathbf{F}_{S\to R}, M\right) = \mathcal{MI}\left(B, R\right) \tag{3}$$

這裡都是一組！

where $\mathbf{F}_{S\to B} = \{F_{S^t\to B}, t \in 1, \cdots, N\}$ are the intermediate flows from targeted latent frames to the blurry input. Similarly, $\mathbf{F}_{S\to R}$ denotes counterparts from latent frames to the RS view and M is a predicted mask to aggregate warped frames using bilateral motion fields. The blur decomposition part is implemented through *GenNet* in an encoder-decoder architecture to warp and refine the reconstructed latent video sequence, which is formulated as:

$$\mathbf{S} = GenNet\left(B, R, \mathbf{F}_{S\to B}, \mathbf{F}_{S\to R}, M\right) \tag{4}$$

Hereinafter, we highlight core components of the model: motion interpretation block with mutually boosted branches; temporal positional encoding; and shutter alignment and aggregation. As for the structure of *GenNet* in blur decomposition, it is presented in supplemental materials.

Dual Streams with Mutual Incentive As we observed, methods exploiting neighboring frames [34, 36] or dual view from different cameras [35] tend to simply concatenate them to capture motion fields. But considering the fact regarding our cross shutter views that blur and RS inputs play contrasting roles in the blur decomposition task, we separately address them into two parallel branches. As shown in Figure 3 (b), the i^{th} motion interpretation block (MIB$_i$) is implemented as two streams with mutual incentive through a shutter alignment and aggregation module. The RS branch offers local details and disambiguates motion directions. Meanwhile, blur counterpart with full global context will elevate the accuracy of motion magnitude and mitigate initial-state ambiguity residing in RS views.

To promote the interaction of two branches, the aligned and aggregated feature are extracted. Instead of directly concatenating encoded features from each other, we firstly predict bidirectional displacement maps between two input views: $F^i_{B\to R}$ and $F^i_{R\to B}$. Then corresponding aligned features $\phi^i_{alin\_R}$ and $\phi^i_{alin\_B}$ for two streams are warped as:

$$\begin{aligned} \phi^i_{alin\_R} &= \mathcal{W}\left(\phi^i_B, F^i_{R\to B}\right) \\ \phi^i_{alin\_B} &= \mathcal{W}\left(\phi^i_R, F^i_{B\to R}\right) \end{aligned} \tag{5}$$

where $\mathcal{W}$ denotes backward-warping process. ϕ^i_B and ϕ^i_R are represented feature of blur and RS views. This aligning process enables us to adaptively selects helpful features and rejects incorrect ones from the other view. In addition, the aggregated feature ϕ^i_{agg} has also been taken into account as an auxiliary.

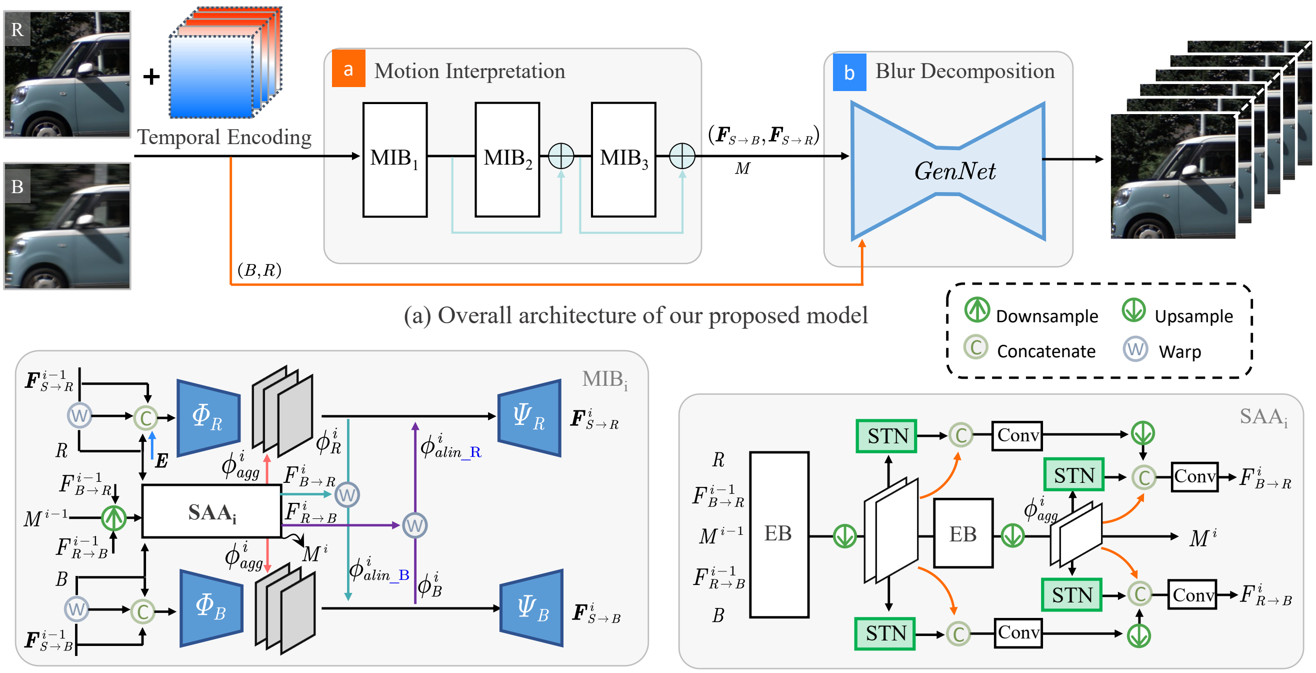
Therefore, taking the blur branch as an example, which can be formulated as:

$$\begin{aligned} \phi^i_B &= \Phi_B\left(B, \mathbf{F}^{i-1}_{S\to B}\right) \\ \mathbf{F}^i_{S\to B} &= \Psi_B\left(\left[\phi^i_B, \phi^i_{alin\_B}, \phi^i_{agg}\right]\right) \end{aligned} \tag{6}$$

i 是 MIB 的編號，單純疊起來

where Φ_B, Ψ_B are encoder and decoder. $[\cdot]$ denotes concatenation. The processing is similar under RS branch except for temporal positional encoding. Overall, the i^{th} MIB can be described as:

$$\mathbf{F}^i, F^i, M^i = \mathcal{MIB}_i\left(B, R, \mathbf{E}, \mathbf{F}^{i-1}, F^{i-1}, M^{i-1}\right) \tag{7}$$



(a) Overall architecture of our proposed model

(b) Motion Interpretation Block (MIB)

(c) Shutter Alignment and Aggregation (SAA)

BRS feature 一起整合
- FSRB 經過 encoder decoder
- 與下一層的 FSRB 用 SAA 的 aligned feature motion field 來整合再一起

BR feature 內部整合，
- 生成出 BR 的 aligned feature motion field
- aggregation feature

Figure 3. **Our proposed model.** (a) shows the overall architecture containing two stages: motion interpretation and blur decomposition. Blur decomposition is implemented through a *GenNet*. Motion interpretation takes as input a blur image B and an RS image R along with its temporal positional encoding E. It consists of three blocks and one of them is unfolded in (b). (c) presents specific details of shutter alignment and aggregation (SAA). Feature extracted by encoder block (EB) will be converted using spatial transformer network (STN), and then enhanced through a Conv. block to accurately predict displacement field between shutters.

where $\mathbf{F} = [\mathbf{F}_{S\to B}, \mathbf{F}_{S\to R}]$, $F = [F_{B\to R}, F_{R\to B}]$ and M is the predicted mask to aggregate warped frames in blur decomposition.

Temporal Positional Encoding To further enhance model's ability to disambiguate motion direction of latent frames, we propose a temporal positional encoding for the RS branch in MIB. Row-by-row exposure of RS cameras inherently carves the latent motion into captured images. [5, 14] approximates this process by copying image rows from corresponding latent frames to synthesize RS effects. So, naturally, the temporal positional encoding for RS input R and latent frame S^t are:

RS image 上，標記每一個 frame 所對應到的 row 們（只有一張）

每一個 frame Est 整張都是同一個數字（共有N張）

$$[E_R]_k = k, k = 0, 1, \cdots, N-1$$
$$E_{S^t} = \frac{H-1}{N-1} t \cdot \mathbb{1} \tag{8}$$

where $[\cdot]_k$ is the operation that extracts k^{th} row and $\mathbb{1}$ denotes a 2-D tensor with all elements being 1. H and N are the image height and length of recovered video clip. Instead of directly using the absolute positional encoding of latent frames, we further compute the relative one to E_R:

$$\mathbf{E} = \{(E_R - E_{S^t}), t = 0, 1, \cdots, N-1\} \tag{9}$$

Finally, the positional encoding map will be concatenated to R and taken as input to RS branch of MIB.

Shutter Alignment and Aggregation SAA module promotes the information propagation across two streams from two perspectives: aggregated and aligned features. The aggregated feature ϕ_{agg} is extracted by feeding the concatenation of bidirectional displacement maps to encoder blocks (EB), which address the correlation between two input views, while aligned features ϕ_{alin_B} and ϕ_{alin_R} focus on absorbing uniquely beneficial parts from each other to stress the contextual characterization and temporal abstraction, respectively.

The SAA module mainly consists of two encoders implemented by convolutional layers with multi-output strategy like MIMO-UNet [3]. Two spatial transformer networks (STN) predict a global transformation conditioned on the output of each encoder to spatially transform features, making the model more robust to visual distortions. We then feed transformed features into their corresponding Conv. blocks to generate bidirectional motion fields. A connection is built through downsampling coarse outputs to next block and the aggregated feature ϕ_{agg} is extracted from output of the second encoder. Inference process with i^{th} SAA is as

Table 2. **Quantitative comparisons** of reconstructed latent frame sequence with lengths of 3, 5 and 9 on RealBR. Subscript of AfB denotes different motion guidance used, while subscripts of RIFE and IFED suggest input settings. '*B-R*' is our proposed dual blur-RS view and '*n·B*' is the setting using *n* neighboring blur frames to tackle motion ambiguity. The performance is measured with mean PSNR, SSIM and LPIPS. We also compute the running time, number of parameters and FLOPs.

Method	Input	×3			×5			×9			Time (s)	Params (M)	FLOPs (G)
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS			
LEVS [11]	*1·B*	21.77	0.7042	0.2886	21.62	0.7153	0.2683	21.83	0.7277	0.2535	1.47	15.9	304
AfB$_p$ [36]	*2·B*	21.50	0.7596	0.4102	21.65	0.7648	0.4055	21.82	0.7686	0.4017	0.15	190	839
AfB$_v$ [36]		22.83	0.7877	0.3904	22.96	0.7903	0.3883	23.10	0.7924	0.3860	0.22	129	793
RIFE$_B$ [8]		24.60	0.8172	0.2254	24.73	0.8199	0.2268	24.83	0.8219	0.2268	1.33	54.8	71.1
IFED$_B$ [35]		24.45	0.8105	0.1817	24.62	0.8141	0.1811	24.74	0.8164	0.1798	1.33	10.8	29.5
BiT [34]	*3·B*	21.90	0.7664	0.2583	21.88	0.7694	0.2574	22.02	0.7729	0.2546	0.11	11.3	57.4
DeMFI [19]	*4·B*	25.55	0.8485	0.2247	25.26	0.8466	0.2275	26.20	0.8577	0.2165	4.86	7.41	420
RIFE$_{BR}$ [8]	*B-R*	30.26	0.8983	0.1071	30.53	0.9030	0.1046	30.67	0.9053	0.1042	1.33	54.8	71.1
IFED$_{BR}$ [35]		30.46	0.9030	0.0467	30.70	0.9064	0.0445	30.84	0.9084	0.0434	1.33	10.8	29.5
Ours		30.87	0.9073	0.0696	31.05	0.9103	0.0684	31.15	0.9120	0.0678	1.30	105	183

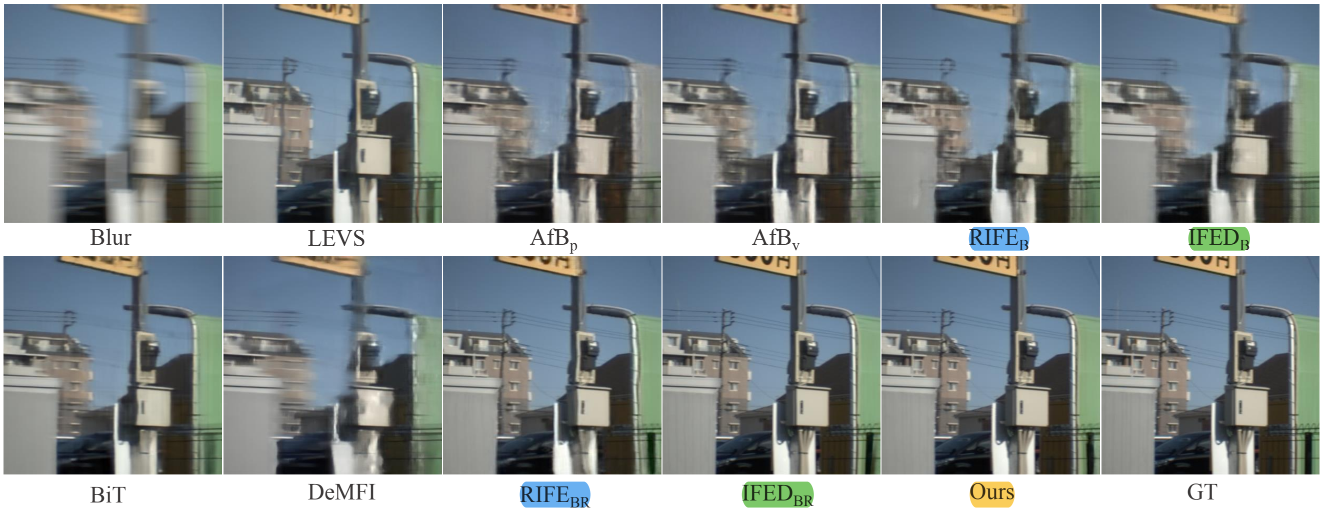


Figure 4. **Qualitative comparison.** Our model outperforms the approaches approximating latent motion fields relying on adjacent blurry inputs. Especially, RIFE$_{BR}$ and IFED$_{BR}$ implemented by dual Blur-RS view reconstruct much sharper details than RIFE$_B$ and IFED$_B$.

follows:

$$F^i, M^i = \mathcal{SAA}_i\left(B, R, F^{i-1}, M^{i-1}\right) \qquad (10)$$

4. Experiments

As explained in Section 3.2, each blur-RS pair corresponds to 9 high-speed sharp frames. So, for training and validation, each sample comprises a paired input (B, R) and groundtruth video clip S' with a length of 9. Our model is trained by using Adam optimizer [13] with epoch of 800. The initial learning rate is set to 10^{-4} and decreases to 10^{-6} through a cosine annealing scheduler. To augment the training data, we first crop the samples into 512 and then conduct random horizontal flipping and channel reverse. Experiments are performed on two GPUs of NVIDIA Tesla V100 with batch size of 8. Besides conducting comparisons on our collected real-world dataset RealBR, we also train all models on synthesized dataset based on GOPRO data [17].

4.1. Comparison with SOTA methods

We compare our model with existing state-of-the-arts to handle motion ambiguity of blur decomposition including LEVS [11], AfB [36] and BiT [34]. Notably, the AfB is implemented by using different motion guidance: learned by a predictor (AfB$_p$) or extracted from neighboring blur frames (AfB$_v$). In addition, considering that the cutting-edge methods RIFE [8] for video frame interpolation and IFED [35] for RS temporal super-resolution can be easily adapted to our blur-RS setting, we therefore combine the two models with our setting (denoted as RIFE$_{BR}$ and IFED$_{BR}$) and conduct the experiments. As a contrast, results of these two models using consecutive blur frames (denoted as RIFE$_B$

and IFED$_B$) are also provided. Although, we focus on blur decomposition, we also compared against a blur frame interpolation method, DeMFI [19] that could be converted to our setting for fair comparison. UTI-VFI [32], TNTI [12] and BIN [26], which cannot distinguish exposure or dead-time when interpolating, are unable to be integrated into our comparison experiments without losing fairness. To better demonstrate the performance of all models, we retrained them on our collected data RealBR.

Table 2 shows quantitative comparisons of all methods on reconstructing video sequence with lengths of 3, 5 and 9. Overall, retrieving a video from single blurred image is quite challenging. The carefully designed supervision of LEVS is barely useful to solve the ambiguity posed by averaging effects. On the other hand, resorting to predicting motion between neighboring inputs, models can indeed speculate the temporal order of latent frames to a certain extent. DeMFI obtains the highest performance among those methods but still lower than our model. Benefited from the dual view, the performance gains of RIFE$_{BR}$ and IFED$_{BR}$, compared with RIFE$_B$ and IFED$_B$, are quite remarkable (at least 5.6 dB on PSNR). The improvement sufficiently demonstrate the effectiveness of blur-RS combination. Reconstructed frames at middle time are presented in Figure 4 for visual comparison. Although existing methods offer reduced distortions, they do not fully restore local details and structures, whereas our results are significantly clearer. Table 2 also presents the computed complexity of all algorithms. Specifically, the FLOPs and running time were evaluated by recovering 9 latent frames with a size of 256 × 256 on an NVIDIA Geforce RTX 3090. The results indicate that our model's complexity is in moderate level.

To better substantiate the ability of our model in motion direction disambiguation and local details recovery, we apply all models to generating 9 consecutive latent frames, whose visual results are given in supplemental materials. We also provide the video demo for a comprehensive comparisons.

Table 3. **Model ablation** on RealBR dataset. 'T' denotes temporal encoding and 'Single' indicates using single branch.

Variants	**PSNR** (↑)	**SSIM** (↑)	**LPIPS** (↓)
W/o T (v1)	31.06	0.9104	0.0690
W/o SAA (v2)	27.96	0.8645	0.1442
Single (v3)	30.33	0.9013	0.0929
Ours (full)	31.15	0.9120	0.0678

4.2. Model Ablation

To assess the efficacy of the components in our proposed model, we conducted an ablation study as depicted in Table 3. The experiments demonstrate performance of three model variants: v_1 (without temporal positional encoding), v_2 (without shutter alignment and aggregation), and v_3 (using single motion interpretation branch that takes concatenation of blur and RS views as input). Our findings reveal that dual stream with mutual incentive through SAA module is effective to handle blur decomposition and the temporal positional encoding further improves the performance.

Table 4. Quantitative comparisons on misaligned RS-Blur view. 'Shift-n' denotes misalignment with maximal offsets n and 'Noise-m' is an experiment conducted on synthesized low-light RS view under peak value m.

Method	×3			×5			×9		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Shift-4	30.95	0.9077	0.0705	31.18	0.9113	0.0696	31.32	0.9135	0.0690
Shift-6	30.82	0.9062	0.0747	31.05	0.9098	0.0727	31.17	0.9117	0.0724
Shift-8	30.56	0.9011	0.0907	30.73	0.9038	0.0915	30.87	0.9065	0.0887
Noise-300	30.56	0.8979	0.0849	30.77	0.9028	0.0841	30.88	0.9053	0.0844
Noise-500	30.92	0.9012	0.0849	30.98	0.9048	0.0848	30.99	0.9064	0.0848
Noise-800	30.95	0.9072	0.0823	31.04	0.9083	0.0805	31.04	0.9084	0.0805
Ours	30.87	0.9073	0.0696	31.05	0.9103	0.0684	31.15	0.9120	0.0678

4.3. Challenging Scenarios

Misaligned Views RS view is taken as a motion guidance and complement of local details to reconstruct multiple latent frames. As discussed in [35], the ambiguity mainly lies in the forward and backward directions of the motions. Hence the motion guidance has no need to be very precise and tolerates some spatial misalignment to blur view. To validate this, we randomly shift RS view in image space along horizontal and vertical axes. The maximal translational offsets are set as 4, 6, 8 pixels, respectively. Then we retrained our model on this misaligned pairs and provide comparisons in Table 4 and Figure 5. There is no obvious drop, which verifies the robustness of our dual-view setting to misalignment.

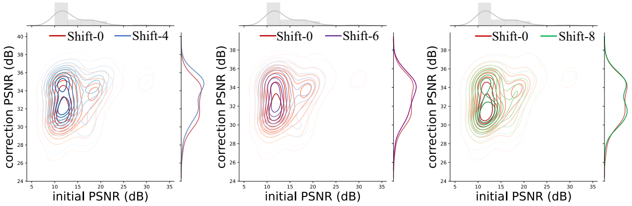


Figure 5. **PSNR distribution** of our method with one aligned ('Shift-0') and three misaligned views('Shift-4', 'Shift-6' and 'Shift-8') under a selected sequence. The horizontal axis is initial PSNR computed by blur view and the first latent frame while the vertical axis denotes PSNR computed between corrected blur view and its ground truth.

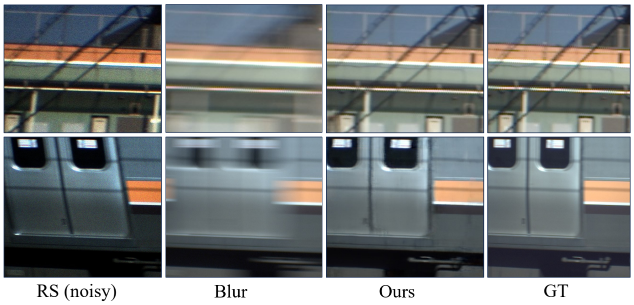
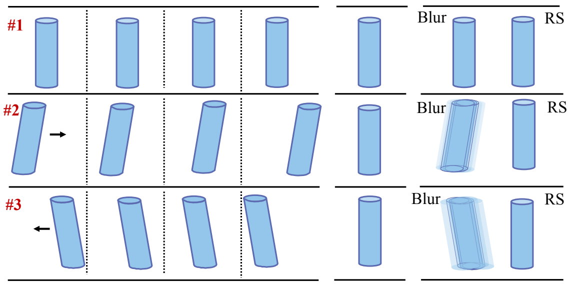


Figure 6. **Visual results** of our proposed method under low-lit scenes with peak value 500. Due to short exposure time of each rows in RS view, it suffers from obvious noise. But our setting is still capable of dealing with this challenge. Best viewed in zoom.



(a) Possible state of movement (b) RS view (c) Dual view

Figure 7. **Initial-state ambiguity of RS observation.** In this toy example, we show one object moving horizontally. (a) shows three possible motion states (static, moving right and moving left) during the exposure time. (b) presents corresponding RS observations. They are all identical due to different initial-state (upright, tilted to the right, and tilted to the left), which brings about ambiguity to RS temporal super-resolution. (c) In our dual Blur-RS setting, blur view sufficiently indicated initial-state of latent frames.

Low-lit Scenes Following the conventional setting, we select proper exposure time of rows for avoiding saturation in GS view and reducing blur in RS view. But it is likely that RS observation will suffer from noise when presenting low-lit scenes. Hence, we further explore effects of noisy RS observations to our method. Following [15, 16], we apply a random gamma adjustment and Poisson noise to clear RS view to synthesize low-light captures. Different peak values are chosen to simulate noise of different intensities. The quantitative results are presented in Table 4 showing that low-lit scene causes about 0.27 dB drop on PSNR, which still outperforms the second-best approach in Table 2. The visual results are also presented in Figure 6.

Table 5. **Quantitative comparisons** on RS image temporal super-resolution. Subscripts of RIFE and IFED suggests input settings. '*n·R*' is the setting using *n* neighboring RS frames to tackle initial-state ambiguity.

Method	Input	×5			×9		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
RSSR [4]	*2·R*	21.92	0.7633	0.1526	22.82	0.7833	0.1362
CVR [6]		21.70	0.7620	0.1717	22.26	0.7761	0.1564
RIFE$_R$ [8]		24.14	0.8124	0.1875	24.36	0.8181	0.1813
IFED$_R$ [35]		24.33	0.8033	0.1032	24.54	0.8085	0.0989
RIFE$_{BR}$ [8]	*B-R*	30.53	0.9030	0.1046	30.67	0.9053	0.1042
IFED$_{BR}$ [35]		30.70	0.9064	0.0445	30.84	0.9084	0.0434
Ours		31.05	0.9103	0.0684	31.15	0.9120	0.0678

4.4. Justification for Dual Blur-RS Setting

Our proposed dual blur-RS setting requires extra view compared with existing methods using neighboring frames, which may increase the cost of device. But considering the performance gains, it is worthwhile. The reason is that, on one hand, blur decomposition entails motion ambiguity as shown in Figure 1. The quantitative and qualitative comparisons with SOTA methods corroborate that RS view can guide blur input to infer a temporally plausible video sequence with more local details.

Table 6. **Quantitative comparisons** with more competitive settings under task of blur decomposition and RS temporal super resolution based on synthetic data. Dual reserved RS setting is denoted as '*R-iR*'. '*B-Event*' and '*R-Event*' are blur decomposition and RS temporal super resolution assisted by event camera, respectively. '*B-SL*' means photosequencing from blur using short long exposure.

Method	Input	×3			×7		
		PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
LEVS [11]	*1·B*	17.27	0.6063	0.3410	16.64	0.58	0.3811
AfB$_p$ [36]	*2·B*	23.38	0.7411	0.2271	23.41	0.7517	0.2183
AfB$_v$ [36]		28.10	0.8760	0.1496	28.39	0.8815	0.1461
RIFE$_B$ [8]		31.26	0.9410	0.0896	31.49	0.9430	0.0892
IFED$_B$ [35]		29.46	0.9193	0.0897	29.75	0.9225	0.0874
BiT [34]	*3·B*	32.31	0.9234	0.0708	32.56	0.9266	0.0691
DeMFI [19]	*4·B*	27.57	0.9002	0.1332	27.44	0.8984	0.1304
PMB [23]	*B-SL*	35.48	0.9723	0.0349	35.11	0.9715	0.0324
EBFI [30]	*B-Event*	33.21	0.9568	0.0703	33.51	0.9591	0.0685
RSSR [4]	*2·R*	22.73	0.8116	0.1039	22.65	0.8090	0.1154
CVR [6]		23.50	0.8342	0.0818	23.47	0.8332	0.0815
RIFE$_R$ [8]		24.16	0.8318	0.1697	24.32	0.8365	0.1618
IFED$_R$ [35]		28.30	0.9122	0.0475	28.63	0.9181	0.0446
IFED [35]	*R-iR*	30.89	0.9417	0.0372	31.96	0.9530	0.0307
EvUnroll [37]	*R-Event*	33.06	0.9558	0.0737	33.48	0.9587	0.0699
RIFE$_{BR}$ [8]	*B-R*	34.49	0.9701	0.0398	35.02	0.9733	0.0366
IFED$_{BR}$ [35]		33.03	0.9627	0.0332	33.72	0.9675	0.0304
Ours		34.92	0.9732	0.0310	35.51	0.9764	0.0305

On the other hand, we delve deeper into this blur-RS setting under the context of RS temporal super resolution to well justify its superiority. Although human's visual perception to RS effects is not that sensitive like blur, the high ill-posedness makes the correction process barely tractable even harder than debluring [9]. One possible reason is the initial-sate ambiguity as illustrated in Figure 7. Given an RS observation of upright cylinder, due to the unknown of its initial state, there exists three possible motion patterns within exposure. Similarly, [4, 6] exploit consecutive RS inputs to mitigate the ambiguity, while [35] resorts to dual RS view with reversed exposure direction. In Table 5, we compare our solution with corresponding SOTA methods from RS temporal super resolution. The RIFE and IFED are also implemented by taking neighboring RS frames as input, denoted as RIFE$_R$ and IFED$_R$ respectively. All experiments validated that blur view provide cues of initial state and global context to RS counterpart that effectively boost reconstruction performance.

目標是從 RS 滾動快門圖像中重建高時序分辨率（temporal high-resolution）的清晰幀序列。

Notably, because of lacking different modal data in RealBR, comparisons between more competitive settings: IFED [35] with *dual reversed RS views*, EvUnroll [37] and EBFI [30] assisted by *event camera*, PMB [23] using *short and long exposure*, are conducted on synthetic data. The details of data synthesis are included in supplementary. From Table 6 and Figure 8, we further validate that our setting and method have superiority against existing solutions.

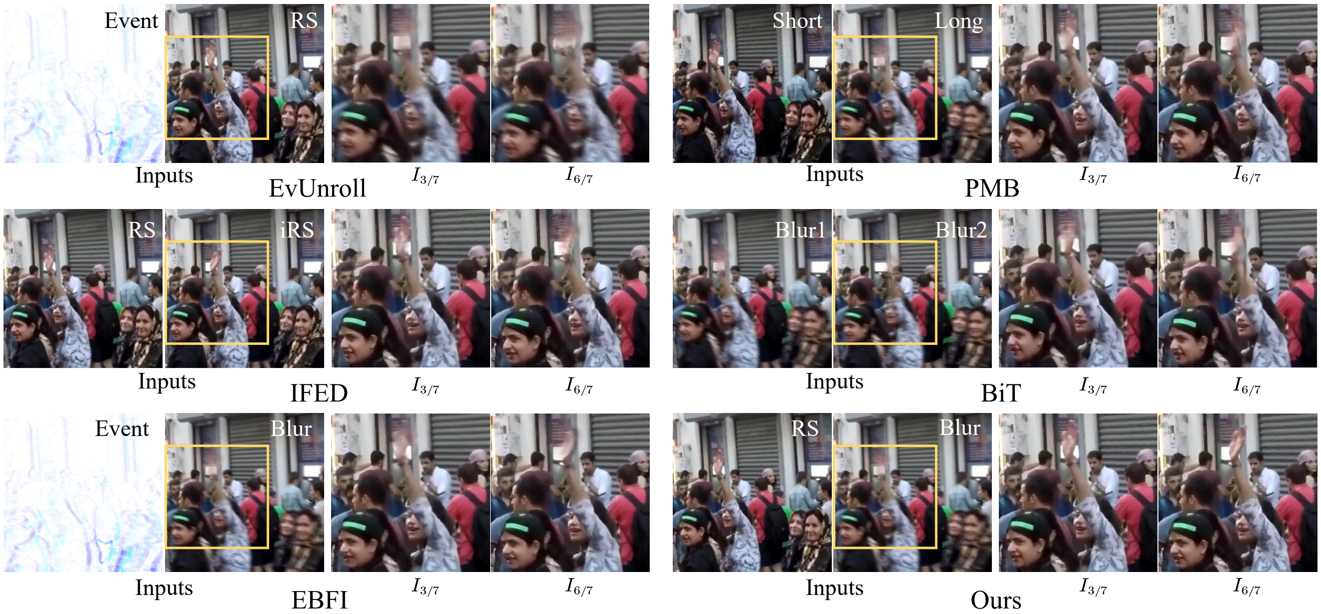


Figure 8. **Visual results** of comparisons with competitive settings on synthetic data. Best viewed in zoom.

5. Conclusion

In this paper, we have proposed a novel cross-shutter setting for motion decomposition of a single blurry image, inspired by the complementary exposure characteristics of GS and RS cameras. Since this setting is new, we first developed a triaxial image capture system to collect triplets of blurry image, rolling shutter image and consecutive sharp frames at higher frame rate. In the arithmetic aspect, we proposed a novel network architecture that actively addresses the contextual characterization and temporal abstraction in a mutual incentive manner. Experiments on our real dataset have verified the effectiveness of proposed algorithm. With synthetic data, we further demonstrated the superiority of our global-shutter/rolling-shutter dual imaging setting. Our current implementation requires a beamsplitter to align two different shutters, which is demanding for compact mobile devices, and our future work is to explore the feasibility of using synchronized sensors placed in parallel.

References

[1] 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*, pages 2505–2513, 2020. 1, 2, 3

[2] Dawit Mureja Argaw, Junsik Kim, Francois Rameau, Chaoning Zhang, and In So Kweon. Restoration of video frames from a single blurred image with motion understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 701–710, 2021. 2

[3] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 5

[4] 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*, pages 4228–4237, 2021. 1, 8

[5] 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*, pages 4541–4550, 2021. 5

[6] 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*, pages 17572–17582, 2022. 1, 8

[7] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3848–3856, 2019. 1

[8] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*, pages 624–642. Springer, 2022. 6, 8

[9] 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 (ICCV)*, pages 12269–12280, 2023. 8

[10] 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*, pages 12259–12268, 2023. 2

[11] Meiguang Jin, Givi Meishvili, and Paolo Favaro. Learning to extract a video sequence from a single motion-blurred image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6334–6342, 2018. 1, 2, 6, 8

[12] Meiguang Jin, Zhe Hu, and Paolo Favaro. Learning to extract flawless slow motion from blurry videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8112–8121, 2019. 2, 7

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] 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*, pages 5941–5949, 2020. 5

[15] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017. 8

[16] Feifan Lv, Feng Lu, Jianhua Wu, and Chongsoon Lim. Mbllen: Low-light image/video enhancement using cnns. In *BMVC*, page 4, 2018. 8

[17] 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*, pages 3883–3891, 2017. 1, 3, 6

[18] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[19] Jihyong Oh and Munchurl Kim. Demfi: deep joint deblurring and multi-frame interpolation with flow-guided attentive correlation and recursive boosting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 198–215. Springer, 2022. 2, 6, 7, 8

[20] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 327–343. Springer, 2020. 1

[21] Kuldeep Purohit and AN Rajagopalan. Region-adaptive dense network for efficient motion deblurring. arxiv eprints, page. *arXiv preprint arXiv:1903.11394*, 2(5), 2019. 1

[22] Kuldeep Purohit, Anshul Shah, and AN Rajagopalan. Bringing alive blurred moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2019. 1, 2

[23] Vijay Rengarajan, Shuo Zhao, Ruiwen Zhen, John Glotzbach, Hamid Sheikh, and Aswin C Sankaranarayanan. Photosequencing of motion blur using short and long exposures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 510–511, 2020. 8

[24] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 184–201. Springer, 2020. 1

[25] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G Narasimhan. Dual-shutter optical vibration sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16324–16333, 2022. 1, 2, 3

[26] Wang Shen, Wenbo Bao, Guangtao Zhai, Li Chen, Xiongkuo Min, and Zhiyong Gao. Blurry video frame interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5114–5123, 2020. 2, 7

[27] Lei Sun, Christos Sakaridis, Jingyun Liang, Qi Jiang, Kailun Yang, Peng Sun, Yaozu Ye, Kaiwei Wang, and Luc Van Gool. Event-based fusion for motion deblurring with cross-modal attention. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 412–428. Springer, 2022. 1, 2, 3

[28] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 1

[29] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 1

[30] 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*, pages 1588–1598, 2023. 8

[31] Fang Xu, Lei Yu, Bishan Wang, Wen Yang, Gui-Song Xia, Xu Jia, Zhendong Qiao, and Jianzhuang Liu. Motion deblurring with real events. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2583–2592, 2021. 1, 2, 3

[32] Youjian Zhang, Chaoyue Wang, and Dacheng Tao. Video frame interpolation without temporal priors. *Advances in Neural Information Processing Systems*, 33:13308–13318, 2020. 2, 7

[33] Zhihang Zhong, Yinqiang Zheng, and Imari Sato. Towards rolling shutter correction and deblurring in dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9219–9228, 2021. 1

[34] Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, and Imari Sato. Blur interpolation transformer for real-world motion from blur. *arXiv preprint arXiv:2211.11423*, 2022. 1, 2, 4, 6, 8

[35] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 233–249. Springer, 2022. 1, 2, 3, 4, 6, 7, 8

[36] Zhihang Zhong, Xiao Sun, Zhirong Wu, Yinqiang Zheng, Stephen Lin, and Imari Sato. Animation from blur: Multimodal blur decomposition with motion guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIX*, pages 599–615. Springer, 2022. 1, 2, 4, 6, 8

[37] 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*, pages 17775–17784, 2022. 1, 2, 3, 8

[38] 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*, pages 2024–2033, 2021. 3