


# Generalizing Event-Based Motion Deblurring in Real-World Scenarios

## - Supplementary Material -

Xiang Zhang<sup>1</sup>, Lei Yu<sup>1</sup>, Wen Yang<sup>1</sup>, Jianzhuang Liu<sup>2</sup>, Gui-Song Xia<sup>1</sup>

<sup>1</sup>Wuhan University <sup>2</sup>Shenzhen Institute of Advanced Technology

{xiangz, ly.wd, yangwen, guisong.xia}@whu.edu.cn, jz.liu@siat.ac.cn

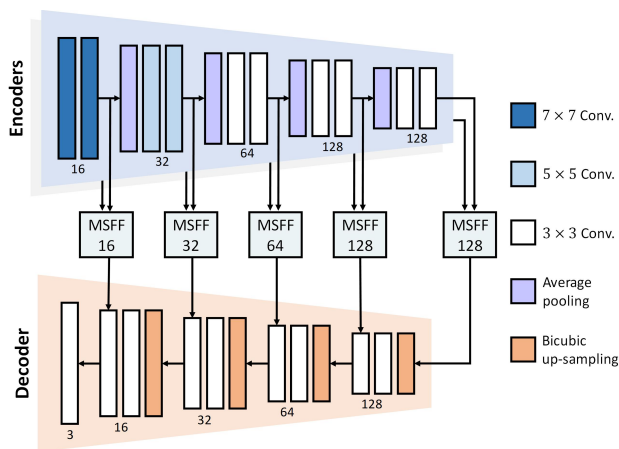


Figure 1: Detailed architecture of our scale-aware network, where MSFF 16, 32, 64, and 128 indicate the MSFF block composed of a 4-layer MLP and a 1-layer DCN with the channel set to 16, 32, 64, and 128, respectively.

## 1. Network Details

The detailed structure of our Scale-Aware Network (SAN) is shown in Fig. 1. The encoders for blurry frames and events are implemented with the same network structure. At each scale, the encoded multi-modal features are fused using the corresponding Multi-Scale Feature Fusion (MSFF) block. Finally, the latent image is restored by decoding the fused features.

## 2. Dataset Details

### 2.1. Multi-scale Real-world Blurry Dataset

This section provides more details on our proposed Multi-scale Real-world Blurry Dataset (MS-RBD), including camera setup, data alignment, and dataset composition.

**Setup.** The MS-RBD is collected with the camera

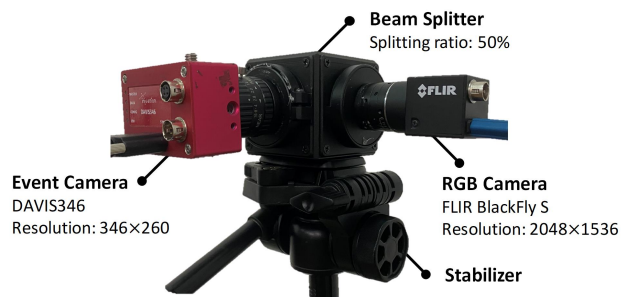



Figure 2: Illustration of our capture system, which is implemented with a beam splitter connecting a DAVIS346 event camera and a FLIR BlackFly S RGB camera. The system is mounted on a stabilizer to control the camera motion.

setup shown in Fig. 2, where a DAVIS346 event camera (346×260) is connected with a FLIR BlackFly S global shutter RGB camera (2048×1536) using a beam splitter (50% splitting). The camera system is mounted on a stabilizer for convenient control of camera motion.

**Alignment.** To ensure the same field of view of events and frames, we perform alignment using a homography estimated by matching SIFT features [1], which is computed using the gray-scale Active Pixel Sensor (APS) frames output from the DAVIS346 camera and the RGB frames output from the FLIR camera. RANSAC is also employed to filter false matches in the homography estimation process. After alignment, the events and the RGB frames are cropped to sizes 288×192 and 1152×768, respectively.

**Composition.** Our MS-RBD contains 32 sequences composed of 22 indoor and 10 outdoor scenes. Each sequence contains 60 RGB 1152×768 blurry frames and the concurrent 288×192 events without ground-truth images. During data collection, the frame rate of the FLIR camera is set to 30 and 15 FPS to imitate the motion blur of different temporal scales, and the motion blur caused by both dynamic targets and camera ego-motion (from simple rotation to complex random motion) is considered in MS-RBD. For self-supervised methods, we choose 5 and 3 sequences

 Corresponding author

The research was partially supported by the National Natural Science Foundation of China under Grants 62271354 and 61871297.

Table 1: Overview of our MS-RBD. #Event indicates the total number of events in the sequence. FPS is the frame rate of the FLIR camera. Dynamic/Static shows whether the target scene is dynamic or static.

Scene	Sequence	Train/Test	#Event (K)	FPS	Dynamic/Static	Camera Motion
Indoor	Badminton	Train	8096	15	Static	Rotation
	Book	Train	12128	15	Static	Rotation
	Book2	Train	3647	15	Dynamic	No motion
	Card	Train	3351	15	Dynamic	No motion
	Chinese	Train	13584	15	Dynamic	No motion
	Cube	Train	2026	15	Dynamic	No motion
	Cube2	Train	6114	15	Dynamic	No motion
	Cylinders	Train	9119	15	Static	Rotation
	Cylinders2	Train	11389	15	Static	Random
	Desk	Train	7703	15	Static	Rotation
	English	Train	6582	15	Dynamic	No motion
	Game2	Train	7751	15	Static	Rotation
	Printer	Train	6312	15	Static	Rotation
	Printer2	Train	8809	15	Static	Random
	Tools	Train	7193	15	Static	Rotation
	Toys	Train	4765	15	Static	Rotation
	Toys2	Train	6422	15	Static	Random
	Bag	Test	9892	15	Dynamic	No motion
	Balls	Test	5182	15	Static	Rotation
	Balls2	Test	5622	15	Static	Random
Chessboard	Test	5574	15	Dynamic	No motion	
Game	Test	7075	15	Static	Random	
Outdoor	Bike	Train	6810	30	Static	Random
	Poster	Train	1328	30	Static	Rotation
	Poster2	Train	3070	15	Static	Random
	Road	Train	2732	15	Dynamic	Rotation
	Road2	Train	2888	15	Dynamic	Rotation
	Street	Train	3198	30	Dynamic	Random
	Text	Train	1158	15	Static	Rotation
	Building	Test	3066	15	Static	Rotation
	Dog	Test	2026	15	Static	Rotation
	Mall	Test	4403	30	Static	Random

from the indoor and outdoor scenes for testing and leave the rest for training. For supervised approaches, all sequences in our MS-RBD can be used for qualitative evaluation of deblurring performance in real-world scenarios. Finally, our MS-RBD is summarized in Tab. 1, and some examples are illustrated in Fig. 3.

## 2.2. High-Speed Events-RGB Dataset

Since only the test set of the High-Speed Events-RGB (HS-ERGB) dataset [2] is available, we choose 4 sequences (*far-bridge\_lake\_01*, *close-fountain\_schaffhauserplatz\_02*, *close-spinning\_umbrella*, and *close-water\_bomb\_floor\_01*) for testing and leave the rest 11 sequences for model training. Besides, we manually filter the static frames in the

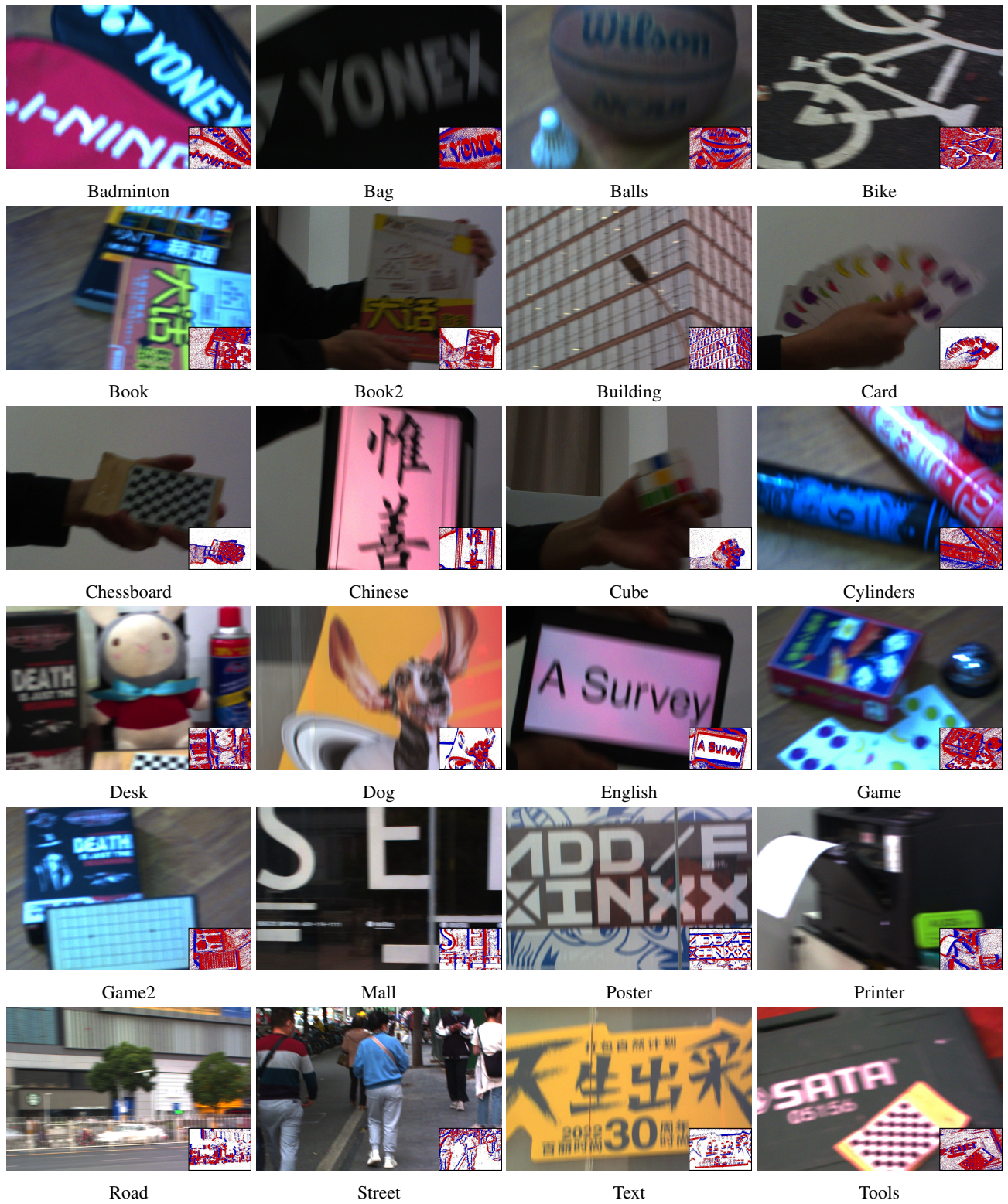


Figure 3: Examples in our MS-RBD, where frames are at size  $1152 \times 768$  and events are at size  $288 \times 192$ . The events accumulated over the exposure time of blurry frames are shown at the bottom right of the corresponding frames (red/blue dots denote positive/negative events).

Table 2: Efficiency comparisons on the FLOPs required to infer a  $160 \times 320$  image and the training hours needed on the Ev-REDS dataset. Best results are **bolded**.

Method	FLOPs	Training time
EVDI [3]	13.45 G	48 hr
Ours	<b>11.15 G</b>	<b>22 hr</b>

Table 3: Results under  $\mathcal{R}(B_T, \mathcal{E}_T) = 1$  on the Ev-REDS dataset. \* means finetuning with  $\#S = 49$  (normal blur) and 97 (large blur), where  $\#S$  denotes the number of sharp images used to synthesize one blurry frame. † means finetuning using our proposed  $\mathcal{L}_{TC}$ . Best and second-best results are **bolded** and underlined, respectively.

Method	Normal blur		Large blur	
	PSNR $\uparrow$	SSIM $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
EVDI [3]	23.88	0.7789	23.02	0.7148
EVDI*	23.41	0.7569	22.55	0.7157
EVDI†	<u>24.01</u>	<u>0.7827</u>	<u>23.29</u>	<u>0.7530</u>
Ours	<b>24.12</b>	<b>0.7898</b>	<b>23.57</b>	<b>0.7663</b>

HS-ERGB dataset as motion blur does not occur in such cases, and only use the dynamic scenes in our experiments to ensure valid evaluation of the deblurring performance.

### 3. Comparisons with EVDI

Although a self-supervised deblurring approach is proposed in the previous work EVDI [3], our method shows advantages over EVDI in terms of both efficiency and effectiveness.

**Efficiency.** EVDI and the first-stage training of our method both aim to achieve self-supervised learning by constraining the brightness and structure consistencies, but the methodology is fundamentally different. Our  $\mathcal{L}_{BC}$  directly ensures brightness consistency by learning blur2blur conversion instead of the reblurring method used in EVDI. In addition, our  $\mathcal{L}_{SC}$  guarantees structure recovery by transferring the knowledge learned from blur2blur to the blur2sharp case, while EVDI employs cross-modal signals for supervision. Thus, unlike EVDI which requires restoring a large number of latent frames per input for training, our approach shows better efficiency (especially for the training hours) as shown in Tab. 2.

**Effectiveness.** Our second-stage training designs a self-distillation technique to handle the varying blurriness levels and different spatial scales of motion blur, which are not considered in EVDI. One possible solution for EVDI

to tackle different blurriness levels is to fit the distribution of motion blur with varying temporal scales. However, directly training with different blurriness levels (normal and large blur) degrades the performance of EVDI (see EVDI\* in Tab. 3) as EVDI gains supervision from input frames and events, which tend to be deteriorated by severer blur and more event noise under the case of large blur. By contrasting different blurriness levels, our learning method provides pseudo-ground-truth as strong supervision and progressively teaches models to handle large blur. Thus, our method better utilizes the different blurriness levels of motion blur for performance generalization, which is validated by the PSNR/SSIM gain of EVDI† over EVDI\* in Tab. 3.

### References

- [1] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1
- [2] Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *CVPR*, pages 16155–16164, 2021. 2
- [3] Xiang Zhang and Lei Yu. Unifying motion deblurring and frame interpolation with events. In *CVPR*, pages 17765–17774, 2022. 4