

# BRIDGE FRAME AND EVENT: COMMON SPATIOTEMPORAL FUSION FOR HIGH-DYNAMIC OPTICAL FLOW

## Supplementary Material

**Anonymous authors**

Paper under double-blind review

In this supplementary material, we provide the detailed description of obtaining the pixel-aligned frame-event data in Sec. 1. Then, we further present the generalization of the proposed method for unseen dynamic scenes in Sec. 2.1 and unseen illumination scenes in Sec. 2.2 using the proposed dataset. Next, we provide several analysis experiments about the proposed method, including impact of boundary class number in Sec. 3.1, weight sensitivity in Sec. 3.2, and inference time in Sec. 3.3. Finally, we provide the qualitative comparison on various datasets from Sec. 4.1 to Sec. 4.3.

## 1 PIXEL-ALIGNED FRAME-EVENT DATASET

The prerequisite for the spatiotemporal motion fusion is to obtain the pixel-aligned frame and event data. To this end, we collect the paired frame-event data via two steps, including time synchronization and spatial calibration. Regarding the issue of time synchronization, we utilize microcontroller to generate two pulses with different frequencies but same timestamp as external trigger to synchronize the time between frame and event cameras, including 30 Hz for frame camera and 1M Hz for event camera. Regarding the issue of spatial calibration, we divide this step into two sub-steps, *i.e.*, hardware calibration and software calibration. As shown in Fig. 1, in hardware, we set up a physically coaxial optical device with a beam splitter for frame and event cameras, which allows the same light to pass through the same lens and enter different cameras, thus achieving the overall field of view alignment. In software calibration, we further perform a standard stereo rectification between frame data and event data, and then fine tune the slight calibration errors via pixel offset (Tulyakov et al., 2022). In this way, we can obtain the spatiotemporal pixel-aligned frame images and event stream. Furthermore, we utilize the coaxial optical device to collect the pixel-aligned frame-event dataset, which covers real complex scenes with various dynamic patterns and various illumination conditions. Regarding the issue of optical flow GT, we further introduce LiDAR to obtain accurate scene depth, which is projected to optical flow.

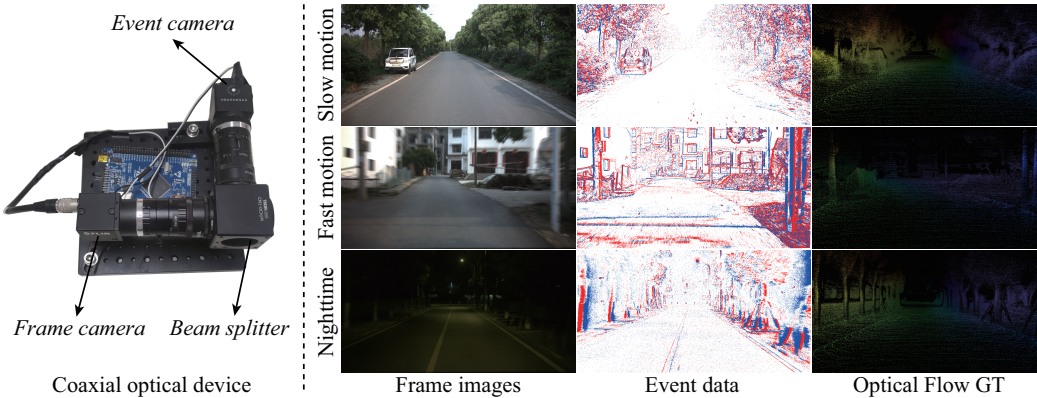


Figure 1: Collection device and examples of the proposed pixel-aligned frame-event dataset.

## 2 GENERALIZATION FOR VARIOUS UNSEEN SCENES

### 2.1 GENERALIZATION FOR VARIOUS DYNAMIC SCENES

In Fig. 2, we further verify the generalization of the proposed method for unseen scene with various dynamic patterns using the proposed dataset. Compared with the competing multimodal method BFlow (Gehrig et al., 2024), the proposed method is more robust to different degrees of dynamic patterns, and the optical flow performance performs better with clear motion boundary. This demonstrates that the proposed common spatiotemporal fusion framework is more adaptable to unseen dynamic scenes.

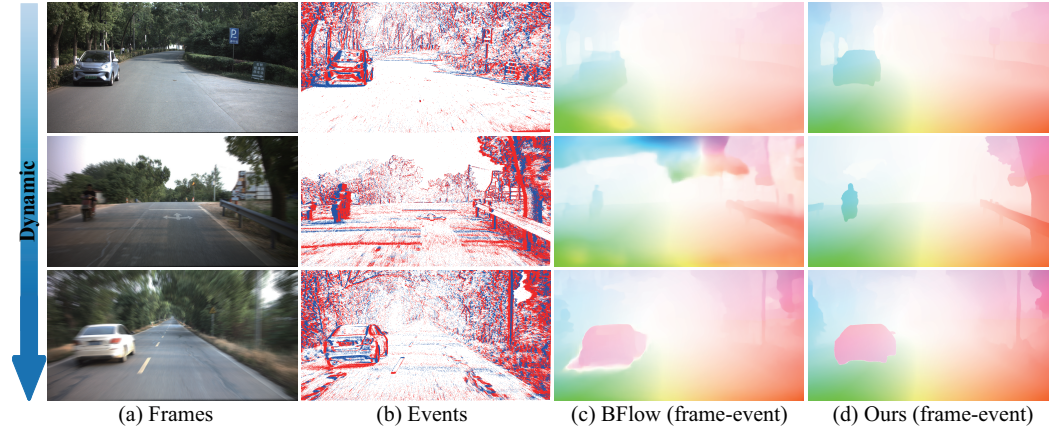


Figure 2: Visual comparison of optical flows on unseen scenes with various dynamic patterns.

### 2.2 GENERALIZATION FOR VARIOUS ILLUMINATION SCENES

In Fig. 3, we further verify the generalization of the proposed method for unseen scene with various illumination conditions using the proposed dataset. As the luminance becomes lower, the optical flows of competing methods (e.g., Selfflow (Liu et al., 2019) and BFlow (Gehrig et al., 2024)) becomes worse, while the proposed method can still perform well.

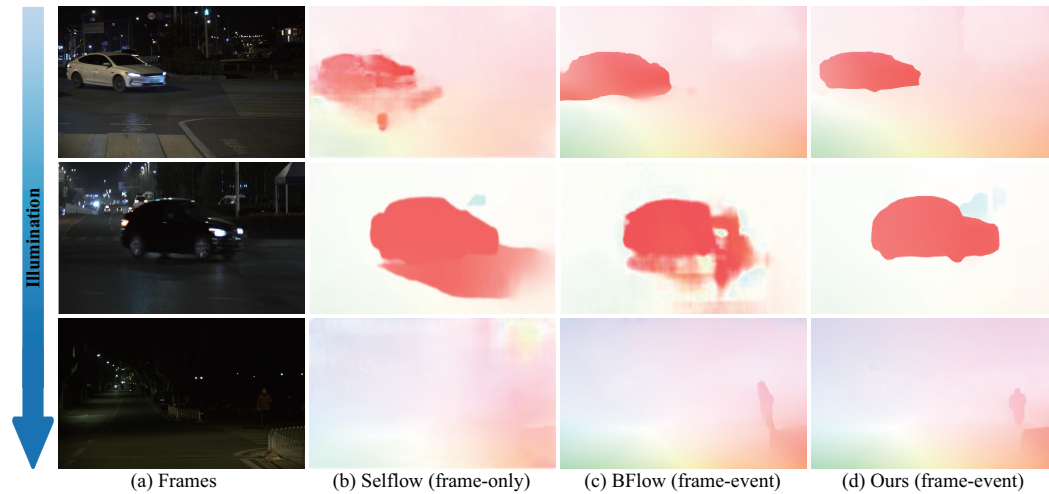


Figure 3: Visual comparison of optical flows on unseen scenes with various illumination conditions.

### 3 DISCUSSION

#### 3.1 IMPACT OF BOUNDARY CLASS NUMBER ON OPTICAL FLOW

Boundary class number  $K$  is a parameter that measures the degree of motion boundary degradation. As shown in Table 1, the boundary class number is not as more as possible, but there is a balance, namely 10. The reason is that motion boundary classification depends on the probability threshold, the larger the motion class number value, the larger the probability threshold corresponding to the normal boundary feature, increasing the risk of misclassification of abnormal boundary features. Therefore, an appropriate boundary class number is important to the final optical flow result.

Table 1: Discussion on the choice of boundary class number.

Boundary class number $K$	EPE	F1-all
2	0.65	2.24%
5	0.60	2.03%
10	<b>0.58</b>	<b>1.96%</b>
15	0.61	2.11%

#### 3.2 WEIGHT SENSITIVITY OF MODEL LOSSES

To choose the optimal weight parameters, we conduct the study on the weight sensitivity of the typical fusion losses in Fig. 4, such as  $\mathcal{L}_{kl}$ ,  $\mathcal{L}_{corr}^{spaErr}$ ,  $\mathcal{L}_{corr}^{tempErr}$  and  $\mathcal{L}_{flow}^{consis}$ . In Fig. 4 (a), the K-L divergence loss  $\mathcal{L}_{kl}$  is sensitive to the training of the proposed fusion framework. If the weight is too large, the backpropagation gradient will disappear, making the training curve coverage to zero. In Fig. 4 (b) and (c), the larger the weights of  $\mathcal{L}_{corr}^{spaErr}$  and  $\mathcal{L}_{corr}^{tempErr}$ , the more rapidly the fusion framework coverages. In Fig. 4 (d), the flow consistency loss  $\mathcal{L}_{flow}^{consis}$  is robust to the framework training. Therefore, we set the main fusion losses weights as  $[\lambda_1, \lambda_3, \lambda_4, \lambda_5]$  as  $[0.01, 1.0, 1.0, 1.0]$ .

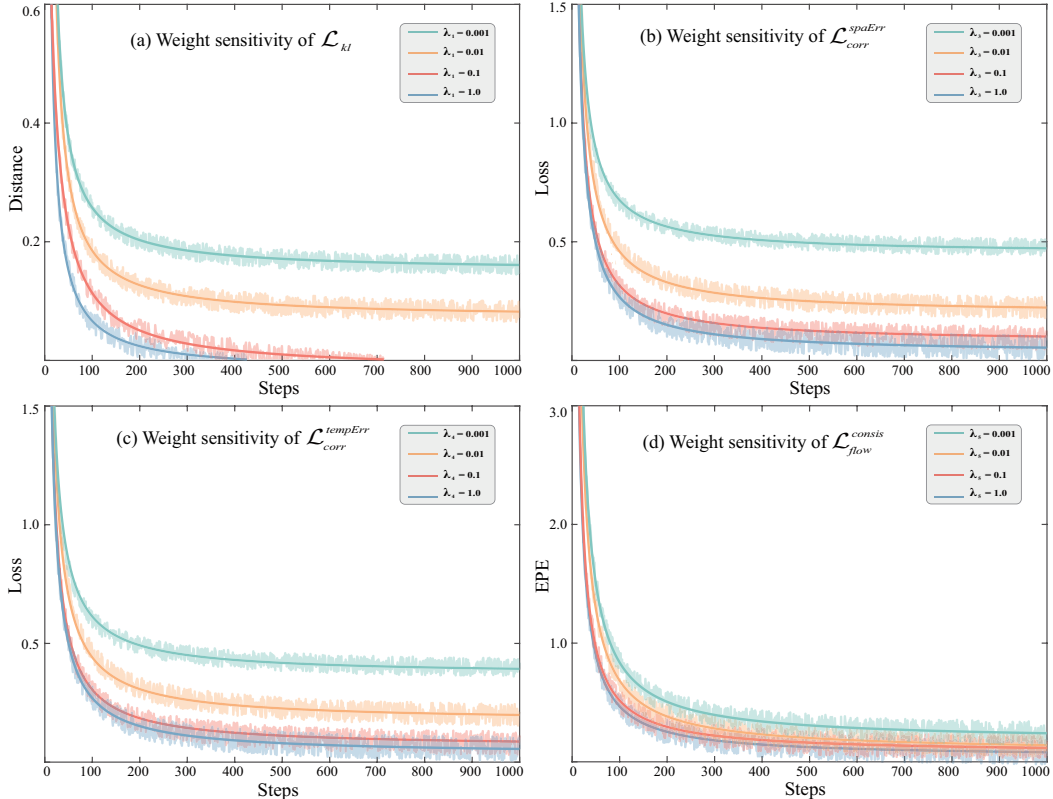


Figure 4: The weight sensitivity of model fusion losses.

### 3.3 INFERENCE TIME

In Table 2, we choose inference time as the efficiency metric of different competing methods (*e.g.*, Selfflow (Liu et al., 2019), RAFT (Teed & Deng, 2020), GMA (Jiang et al., 2021), E-RAFT (Gehrig et al., 2021), BFlow (Gehrig et al., 2024)) for optical flow estimation, and RTX 3090 as the inference platform. We can observe that the multimodal methods do take a little more time to infer than the unimodal methods, but the performance is significantly improved. The main reason is that the multimodal methods need to process the data representation of more modalities and fuse the cross-modal complementary motion knowledge, causing the more computing resources. Moreover, compared with other competing methods, the proposed method can achieve state-of-the-art results within the reasonable inference time.

Table 2: Discussion on inference time on image  $640 \times 480$ .

Method	Selfflow	RAFT	GMA	E-RAFT	BFlow	ComST-Flow
Runtime (ms)	53.3	114.7	137.4	107.4	141.6	155.5
EPE	16.16	1.35	1.24	0.95	0.87	0.58
F1-all	78.07%	6.26%	5.12%	3.65%	2.89%	1.96%

## 4 COMPARISON EXPERIMENTS

### 4.1 COMPARISON ON SYNTHETIC DATASET

The visual results of optical flow predicted by the proposed multimodal method and the competing methods on the synthetic Event-KITTI dataset are presented in Fig. 5. The competing methods include unimodal method Selfflow (Liu et al., 2019) with frame-only and multimodal method BFlow (Gehrig et al., 2024) with frame-event. We have two conclusion. First, the multimodal methods are superior to the unimodal method. This is because these multimodal methods can fuse the complementary knowledge between different modalities to improve optical flow. Second, compared to the multimodal method BFlow with direct fusion, the proposed method with common fusion performs better.

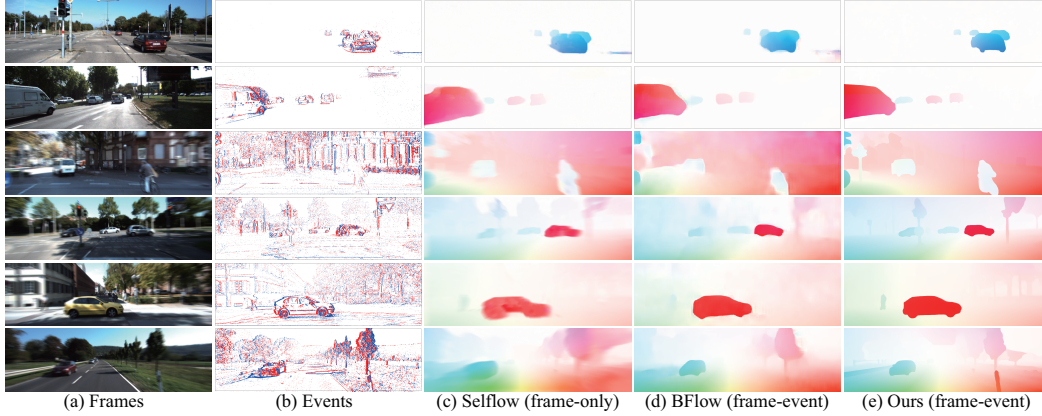


Figure 5: Comparison of optical flows on synthetic Event-KITTI dataset.

### 4.2 COMPARISON ON REAL DATASET

We also show the visual results of the proposed method ComST-Flow and the competing methods on the real DSEC dataset with various illumination conditions in Fig. 6, where we perform blurry effect and frame extraction on images to simulate the spatiotemporal degradation. We have two observations. First, the frame-based method Selfflow almost cannot work normally in nighttime scenes, while the event-based methods can still perform well. This is because event camera has the advantage of high dynamic range to model the motion even in nighttime scenes. Second, the proposed method is superior to other multimodal method BFlow in real scenarios. The main reason is that other multimodal methods suffer the large gap between frame and event modalities, while

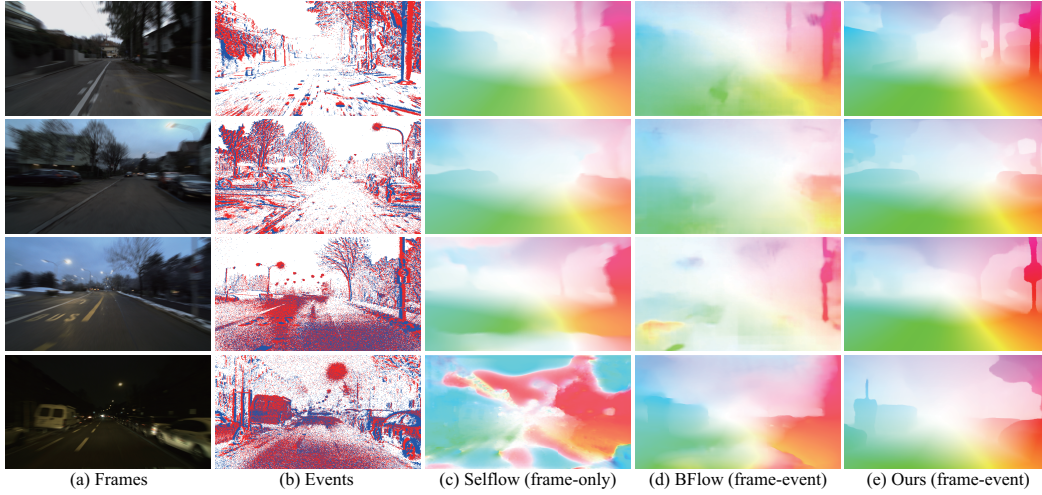


Figure 6: Comparison of optical flows on real DSEC dataset.

the common-latent space of the proposed method bridges the modality gap, thus promoting the spatiotemporal fusion of motion features for optical flow.

#### 4.3 COMPARISON ON EVENT OPTICAL FLOW

In Fig. 7, we compare the state-of-the-art event optical flow models (EV-FlowNet (Zhu et al., 2018) and E-RAFT (Gehrig et al. (2021))) with our event model on the real event stream from DSEC dataset. We can observe that the optical flow estimated by EV-FlowNet is over-smooth, and E-RAFT loses slight motion details in the motion boundaries. Instead, our event optical flow E-ABDA still works well, verifying its superiority.

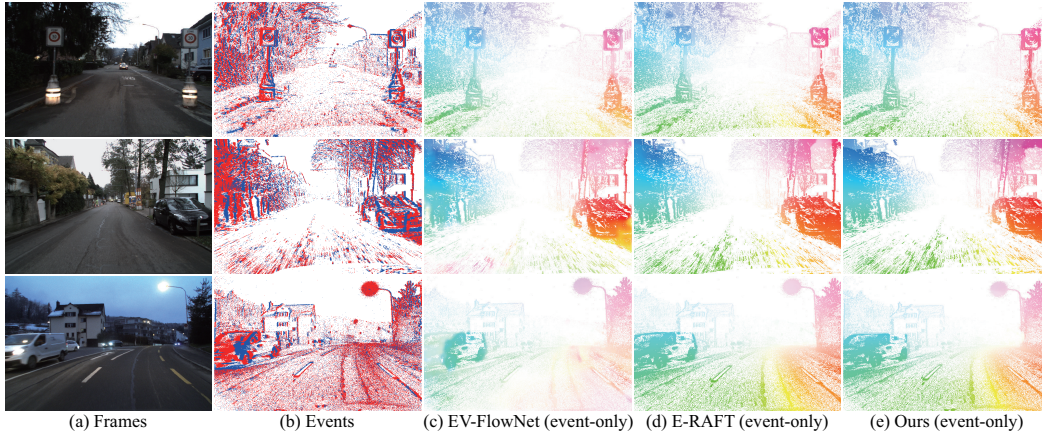


Figure 7: Comparison of event-based optical flows on event stream from DSEC dataset.

## REFERENCES

- Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *International Conference on 3D Vision*, pp. 197–206, 2021. 4, 5
- Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 4
- Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Int. Conf. Comput. Vis.*, pp. 9772–9781, 2021. 4

- Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selfflow: Self-supervised learning of optical flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4571–4580, 2019. 2, 4
- Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pp. 402–419, 2020. 4
- Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 17755–17764, 2022. 1
- Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Ev-flownet: Self-supervised optical flow estimation for event-based cameras. *Robotics: Science and Systems*, 2018. 5