

Bridge Frame and Event: Common Spatiotemporal Fusion for High-Dynamic Scene Optical Flow *Supplementary Material*

Hanyu Zhou^{1,2*}, Haonan Wang¹, Haoyue Liu¹, Yuxing Duan¹, Yi Chang¹, Luxin Yan^{1†}

¹ National Key Lab of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology

² School of Computing, National University of Singapore

hy.zhou@nus.edu.sg yanluxin@hust.edu.cn

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, sensitivity to frame-event calibration error in Sec. 3.2, sensitivity to image blur in Sec. 3.3, complexity of each component in Sec. 3.4, inference time in Sec. 3.5, and weight sensitivity in Sec. 3.6. 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

*The author is currently with National University of Singapore, but this work was finished at Huazhong University of Science and Technology.

†Corresponding author.

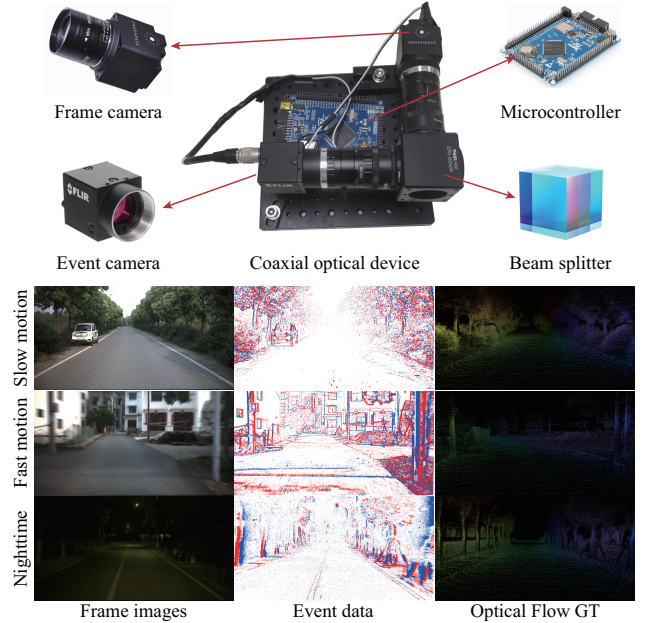


Figure 1. Collection device and examples of the proposed pixel-aligned frame-event dataset. A coaxial optical device is built to collect frame and event data, covering various dynamic patterns and illumination conditions.

[1]. 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. It is worth mentioning that this dataset is the first high dynamic scene optical flow benchmark, promoting the optical flow community.

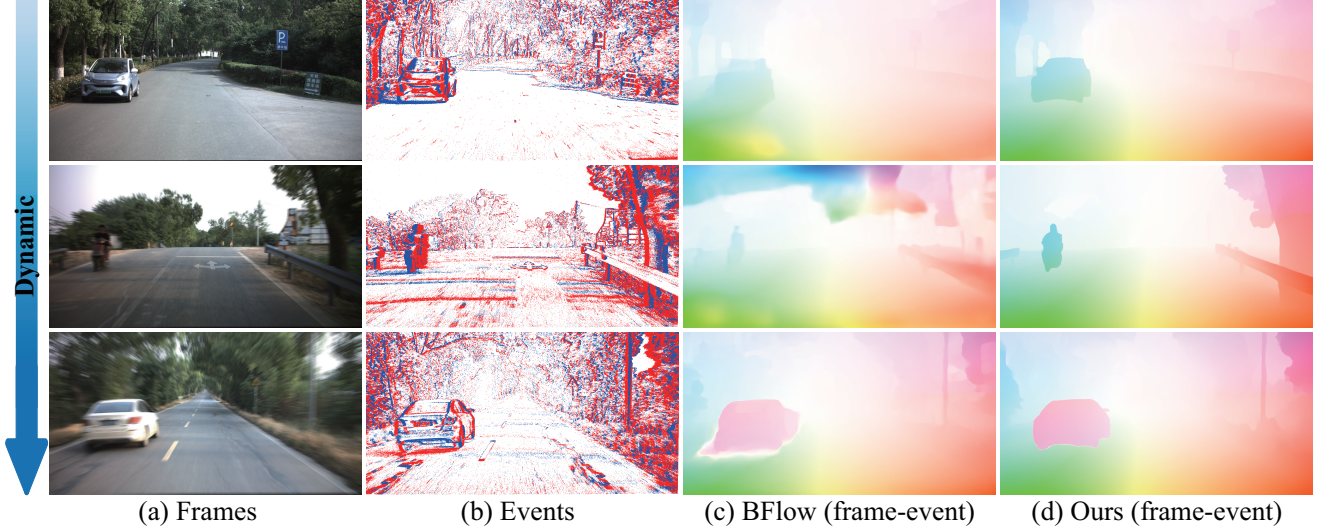


Figure 2. Visual comparison of optical flows on unseen scenes with various dynamic patterns.

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

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

Pixel offset (pix)	0	1	3	5	7	9
EPE	3.78	3.86	4.53	5.89	7.31	10.55

Table 2. Impact of frame-event calibration error on optical flow.

Blur kernel size (pix)	1	3	5	7	9
EPE	0.47	0.49	0.52	0.54	0.58

Table 3. Effect of blur degree on optical flow.

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 [2], 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.

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 [3] and BFlow [2]) becomes worse, while the proposed method can still perform well.

3. Discussion

3.1. Impact of Boundary Class Number

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.

3.2. Sensitivity to Frame-Event Calibration Error

We simulate the frame-event alignment degree via intentional misalignment, and analyze the impact of the alignment degree on optical flow in Table 2. We can observe that the optical flow model is sensitive to pixel calibration errors. To this end, we will apply dynamic networks (*e.g.*, deformed convolution) to align cross-modal data at the feature level.

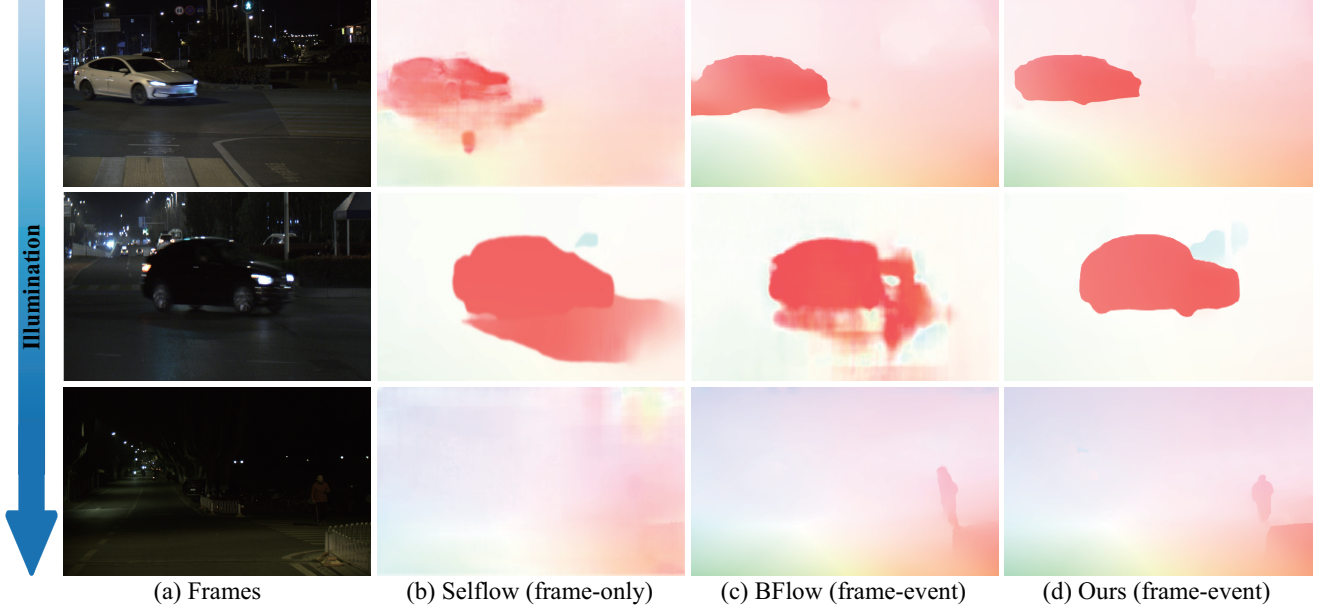


Figure 3. Visual comparison of optical flows on unseen scenes with various illumination conditions.

Metric	GMA	Our framework					
		Pyramid	E_f	E_e	Transformer	GRU	Total
Runtime (ms)	137	19	19	15	80	22	155
Param (M)	5.91	2.19	1.80	1.79	7.53	1.06	14.37
Size (MB)	58.73	5.68	8.83	10.04	52.42	22.27	99.24
EPE	1.24	0.58					

Table 4. Complexity comparison of different components.

3.3. Sensitivity to Image Blur

We also study the sensitivity of the proposed method to the degree of image blur. As shown in Table 3, our method performs well on degraded images with different blur kernels. The main reason is that event data from our framework can provide continuous motion knowledge with high temporal resolution, contributing to enhancing the anti-interference ability of the flow model to blur.

3.4. Necessity and Complexity of Each Component

The whole framework only looks complicated, but is a simple end-to-end network during testing phase. Within the proposed framework, common space is built to close the modality gap, and interpretable optimization strategies are introduced to model the spatiotemporal fusion. For end-to-end inference, feature pyramid and transformer are used to learn the above process. Moreover, we compare the performance and efficiency of each component in Table 4, which verifies that our method achieves a trade-off between performance gains and computational costs. Overall, the whole framework looks complex, but is necessary and efficient.

3.5. Inference Time

In Table 5, we choose inference time as the efficiency metric of different competing methods (*e.g.*, Selfflow [3], RAFT [4], GMA [5], E-RAFT [6], BFlow [2]) 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.

3.6. 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

Method	Selflow	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%

Table 5. Discussion on inference time on image 640×480 .

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 $[0.01, 1.0, 1.0, 1.0]$.

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 Selflow [3] with frame-only and multimodal method BFlow [2] 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.

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 Selflow 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 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 [7] and E-RAFT [6]) 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 losses slight motion details in the motion boundaries. Instead, our event optical flow E-ABDA still works well, verifying its superiority.

References

- [1] 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.*, pages 17755–17764, 2022. 1
- [2] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. 2, 3, 4
- [3] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4571–4580, 2019. 2, 3, 4
- [4] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Eur. Conf. Comput. Vis.*, pages 402–419, 2020. 3
- [5] 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.*, pages 9772–9781, 2021. 3
- [6] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *International Conference on 3D Vision*, pages 197–206, 2021. 3, 4
- [7] 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. 4

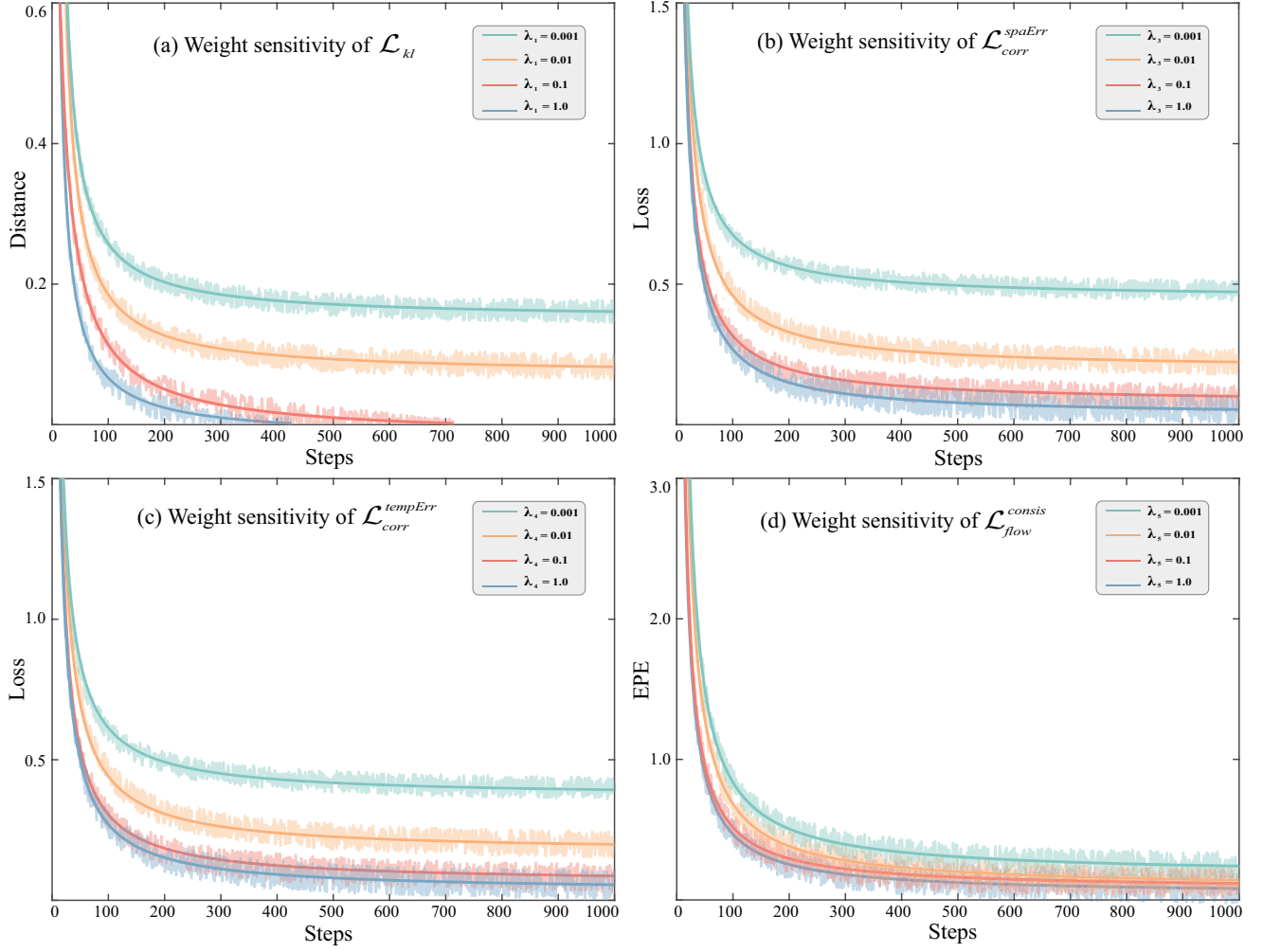


Figure 4. The weight sensitivity of model fusion losses.

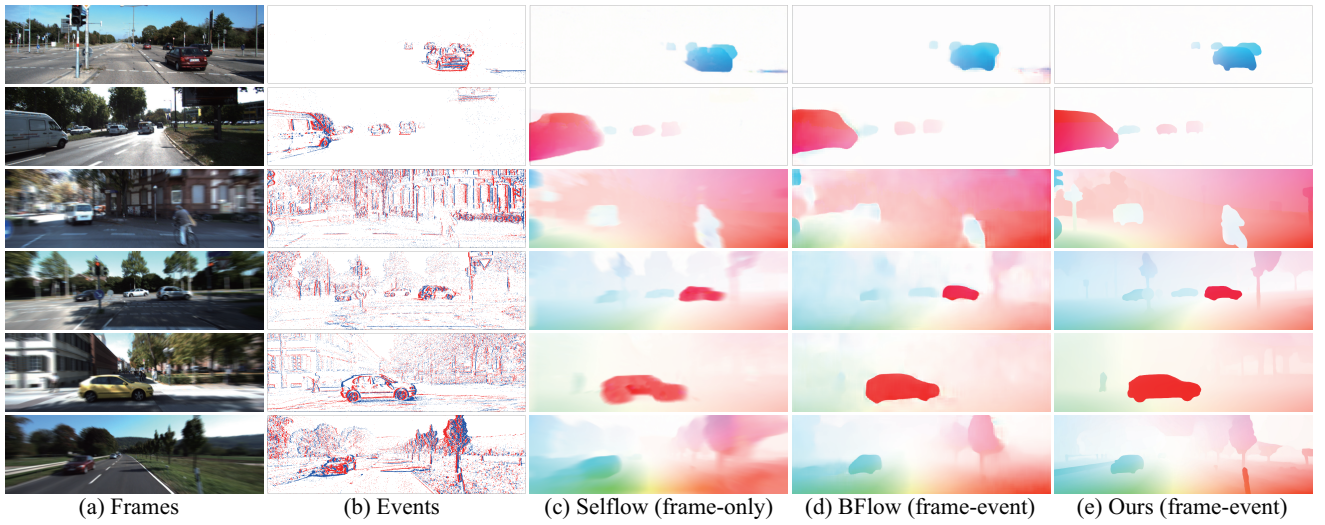


Figure 5. Comparison of optical flows on synthetic Event-KITTI dataset.

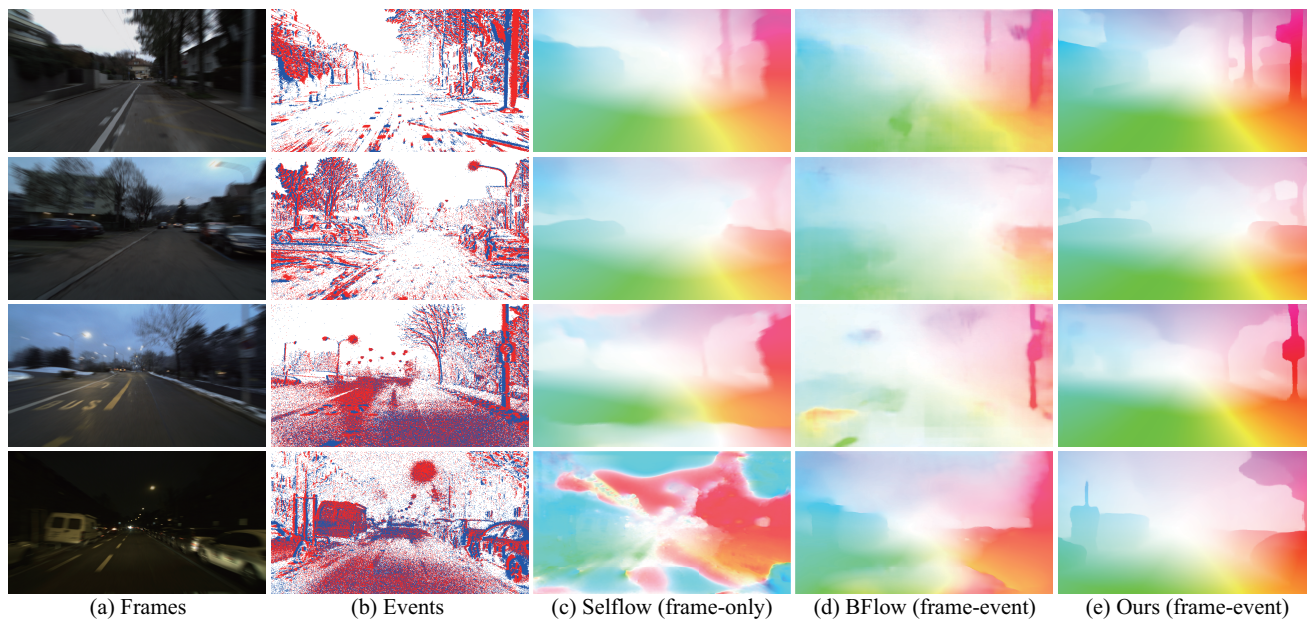


Figure 6. Comparison of optical flows on real DSEC dataset.

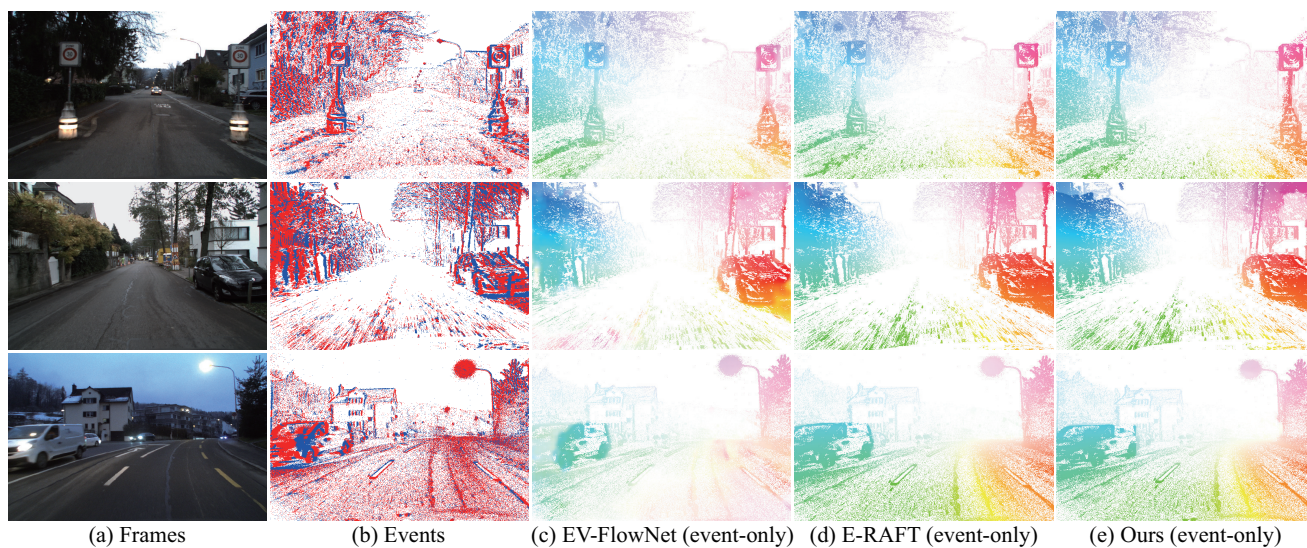


Figure 7. Comparison of event-based optical flows on event stream from DSEC dataset.