

---

# Injecting Frame-Event Complementary Fusion into Diffusion for Optical Flow in Challenging Scenes

## *Supplementary Material*

---

Haonan Wang<sup>1</sup>, Hanyu Zhou<sup>2\*</sup>, Haoyue Liu<sup>1</sup>, Luxin Yan<sup>1</sup>

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

<sup>2</sup>School of Computing, National University of Singapore  
{whn\_aurora, yanluxin}@hust.edu.cn, hy.zhou@nus.edu.sg

In this supplementary material, we provide a detailed description of obtaining the Frame-Event Flow Dataset (FEFD) in Sec. A. Then we demonstrate the generalization of our proposed method on various unseen dynamic scenes in Sec. B.1, various unseen illumination scenes in Sec. B.2, and various unseen adverse weather scenes in Sec. B.3. Next, we provide several ablation experiments and discussions about the proposed method, including the impact of three input features in the TVM-MCA module in Sec. C.1, the weight sensitivity of different loss functions in Sec. C.2, the impact of the recurrent times of GRU in the MGDD module in Sec. C.3. Finally, we additionally provide some visual comparisons on synthetic datasets and real datasets in Sec. D.

## A Frame-Event Flow Dataset

The key to effective appearance-boundary fusion is to obtain pixel-level aligned frame data and event data. For our proposed Frame-Event Flow Dataset, we obtain pixel-level aligned frame and event data through two steps: time synchronization and spatial calibration. For time synchronization, we utilize a microcontroller to generate two pulses with different frequencies but consistent timestamps to externally trigger the frame and event camera, including 30 Hz for frame camera and 1 MHz for event camera. For spatial calibration, our strategy is divided into two parts: hardware and software calibration. As shown in Fig. 1, in hardware, we utilize a beam splitter to build a co-axial device for the frame camera and event camera, which makes the field of view of the two cameras generally aligned. In software, we further perform standard stereo calibration on the frame data and event data, and fine-tune the slight calibration error through pixel offset [8]. Through the above methods, we have constructed an extensive frame-event dataset, which covers real complex scenes with various dynamic patterns and various illumination conditions. To obtain the ground truth of optical flow, we introduce LiDAR to obtain the depth information of the scene and project it into the optical flow.

## B Generalization for Various Unseen Scenes

### B.1 Generalization for Various Dynamic Scenes

In Fig. 2, we further verify the generalization of our proposed method for unseen scenes with various dynamic patterns on the proposed dataset. Compared with the multimodal method BFlow [2] and the diffusion-based method FlowDiffuser [6], the proposed method demonstrates stronger robustness to different degrees of dynamic patterns and obtains optical flow with dense appearance saturation and boundary completeness, which verifies that the proposed diffusion-based appearance-boundary fusion framework can adapt to unseen dynamic scenes.

---

\*Corresponding author.

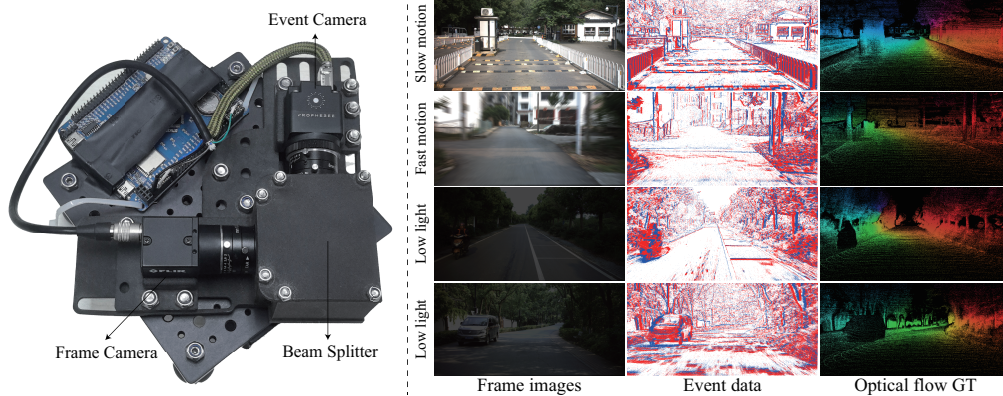


Figure 1: Frame-Event collection device and examples of proposed Frame-Event Flow Dataset

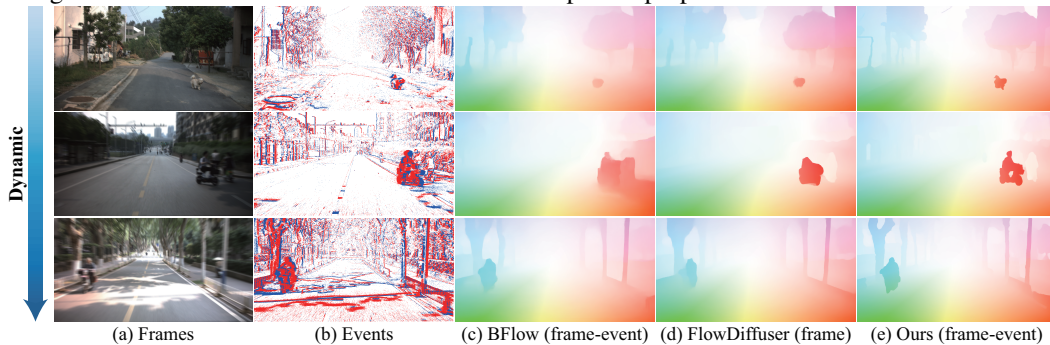


Figure 2: Visual results of optical flows on unseen scenes with various dynamic patterns.

## B.2 Generalization for Various Illumination Scenes

In Fig. 3, we further verify the generalization of our proposed method for unseen scenes with various illumination conditions on the proposed dataset. As the illumination decreases, our proposed method maintains good performance in both appearance and boundary areas, while the optical flow results of competing methods (e.g., BFlow [2] based on frame and event input, and FlowDiffuser [6] based on diffusion models) deteriorate, especially in boundary areas, which demonstrates the strong robustness of the proposed method to low-light scenes.

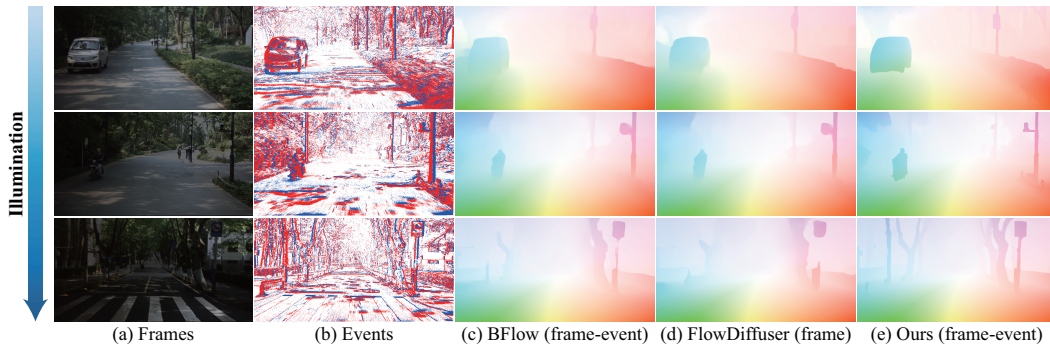


Figure 3: Visual results of optical flows on unseen scenes with various illumination conditions.

## B.3 Generalization for Various Adverse Weather Scenes

Our proposed method has been verified to perform well under various dynamic and illumination conditions. Furthermore, to verify the performance of our method under different adverse weather conditions, we conduct additional experiments on rainy, foggy, large-displacement, and overexposed



Table 1: Quantitative results on the proposed unseen HS-FEFD and LL-FEFD datasets.

Method		Discriminative model					Generative model	
		GMA [4]	FF [7]	E-RAFT [1]	BFlow [2]	CH <sup>2</sup> DA-Flow [9]	FD [6]	Ours
	Input	Frame	Frame	Event	Frame-event	Frame-event	Frame	Frame-event
Rain-KITTI	EPE ↓	8.02	7.54	18.73	6.27	5.78	4.97	<b>4.23</b>
	Fl-all ↓	46.31	41.35	75.72	37.55	33.59	26.72	<b>24.92</b>
Rain-GOF	EPE ↓	6.03	4.97	14.57	4.22	3.54	3.28	<b>3.19</b>
	Fl-all ↓	35.73	31.62	57.35	21.74	22.75	18.94	<b>17.36</b>
Fog-KITTI	EPE ↓	7.59	6.75	17.54	6.19	5.98	5.72	<b>4.89</b>
	Fl-all ↓	43.62	38.91	67.91	37.29	38.94	32.47	<b>28.74</b>
Fog-GOF	EPE ↓	7.14	5.47	15.12	4.68	3.87	3.79	<b>3.65</b>
	Fl-all ↓	44.29	36.84	59.68	26.91	24.52	24.77	<b>22.83</b>
LD-DSEC	EPE ↓	5.45	4.23	8.62	3.52	4.18	3.79	<b>2.67</b>
	Fl-all ↓	34.82	30.52	47.34	21.64	28.96	24.83	<b>12.45</b>
OE-DSEC	EPE ↓	5.68	3.91	9.43	3.04	3.96	3.25	<b>2.38</b>
	Fl-all ↓	37.04	28.79	55.86	17.28	27.43	18.42	<b>10.94</b>

scenes. For rainy and foggy scenes, we use the Weather-KITTI2015 (Rain-KITTI, Fog-KITTI) and Weather-GOF (Rain-GOF, Fog-GOF) introduced by Zhou et al. [9], with synthetic events generated using the v2e model [3]. For large displacements, we constructed the LD-DSEC dataset by sampling one frame every three frames and re-segmenting events from the original DSEC, using the method of Ce Liu et al. [5] to generate optical flow ground truth. For overexposed scenes, we built the OE-DSEC dataset by selecting frames with intense car headlights or streetlights from DSEC. Each dataset includes 800 randomly selected images for generalization testing. We additionally include CH<sup>2</sup>DA-Flow [9], an unsupervised method for optical flow in adverse weather, as a comparison baseline. As shown in Table 1, our proposed method performs well in large displacement and overexposed scenes, but perform not so well in the rainy and foggy scenes. This is because raindrops and fog not only blur the frame image but also introduce a large number of noise points in the event stream, making it difficult for our method to obtain good visual features.

## C Ablation Study and Discussion

### C.1 Ablation Study on TVM-MCA Module

Our proposed TVM-MCA module takes temporal embedding, visual features, and motion features as input. To verify the effects of the three inputs, we conducted an ablation experiment on the three input features. As shown in Table 2, it is clear that the three inputs can significantly improve the optical flow results. Moreover, we can conclude that visual features have the greatest impact on the evaluation metrics, followed by motion features, and temporal embedding has the least impact.

Time Embedding	Visual Feature	Motion Feature	EPE	Fl-all
✗	✓	✗	1.72	6.63
✗	✗	✓	1.87	7.48
✓	✓	✗	1.54	5.32
✓	✗	✓	1.68	6.14
✗	✓	✓	1.35	4.57
✓	✓	✓	<b>1.09</b>	<b>3.83</b>

Table 2: Ablation study on input features of TVM-MCA Module.

## C.2 Weight Sensitivity of Model Loss

Three loss functions are used in our model training process, among which the smoothness loss  $\mathcal{L}_{smooth}$  and the event consistency loss  $\mathcal{L}_{event}$  have weights that need to be set manually. In order to choose the optimal weight parameters, we conduct an experiment on the weight sensitivity as shown in Fig. 4. Based on the experimental results, we set the weights  $[\lambda_{smooth}, \lambda_{event}]$  to  $[1.0, 1.0]$ , corresponding to the curve with the fastest loss decrease and the lowest convergence value.

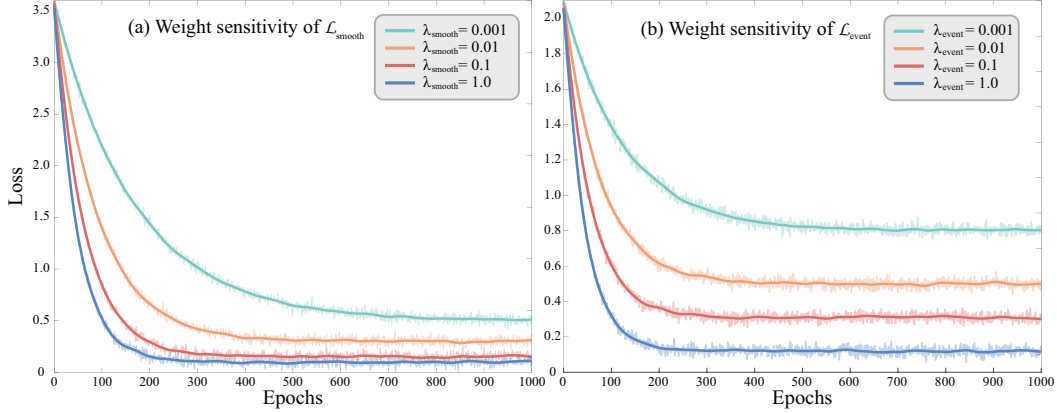


Figure 4: The weight sensitivity of model losses.

## C.3 Discussion on GRU Recurrent Times in MGDD Module

In the proposed MGDD module, we recurrently execute the GRU with the flow head and the DDIM module to perform iterative denoising. In order to obtain the best optical flow results in the shortest possible inference time, we conduct an experiment to evaluate the impact of the GRU recurrent times on the optical flow evaluation metrics and inference time. According to the results shown in Table 3, we set the GRU recurrent times  $N$  to 6.

Recurrent Times of GRU	EPE	Fl-all	Inference Time (ms)
2	1.45	5.82	141.5
4	1.21	4.65	169.4
<b>6</b>	<b>1.09</b>	<b>3.83</b>	<b>203.9</b>
8	1.08	3.79	232.5
10	1.09	3.84	264.7

Table 3: Discussion on the choice of GRU recurrent times.

## D Comparison Experiments

### D.1 Comparison on Synthetic Dataset

In Fig. 5, we present the visual results of optical flow estimated by the proposed method Diff-ABFlow and the competing methods on the synthetic HS-KITTI and LL-KITTI datasets with various dynamic patterns and illumination conditions. The competing methods include multi-modal method BFlow [2] with frame-event input and diffusion-based uni-modal method FlowDiffuser [6] with only frame input. We can conclude that in high-speed and low-light conditions, our proposed method performs much better in both appearance and boundary areas than the competing methods since we effectively fuse the frame and event data utilizing the appearance-boundary complementarity, and introduce the paradigm of diffusion models that is robust to degraded input features.

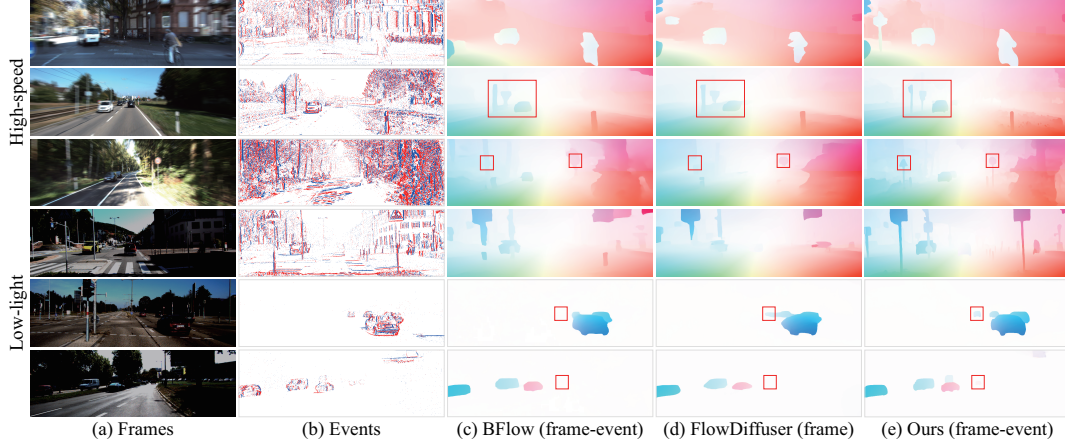


Figure 5: Comparison of optical flows on synthetic HS-KITTI and LL-KITTI datasets.

## D.2 Comparison on Real Dataset

In Fig. 6, we demonstrate the visual results of our proposed Diff-ABFlow and the competing methods on the real HS-DSEC and LL-DSEC datasets with various dynamic patterns and illumination conditions. According to the results, we have two conclusions. First, the multi-modal method BFlow [2] performs slightly better than the diffusion-based uni-modal method FlowDiffuser [6], especially in boundary areas, since the event data provides extra boundary information. However, BFlow fails to adapt to the degradation in the appearance areas. Second, our proposed method is superior to competing methods in both appearance and boundary areas because we effectively utilize the appearance-boundary complementarity of the frame and event data. Moreover, we introduce diffusion models as the backbone of optical flow, to adapt to the degraded input features. Thus, the proposed method achieves state-of-the-art performance under both high-speed and low-light conditions.

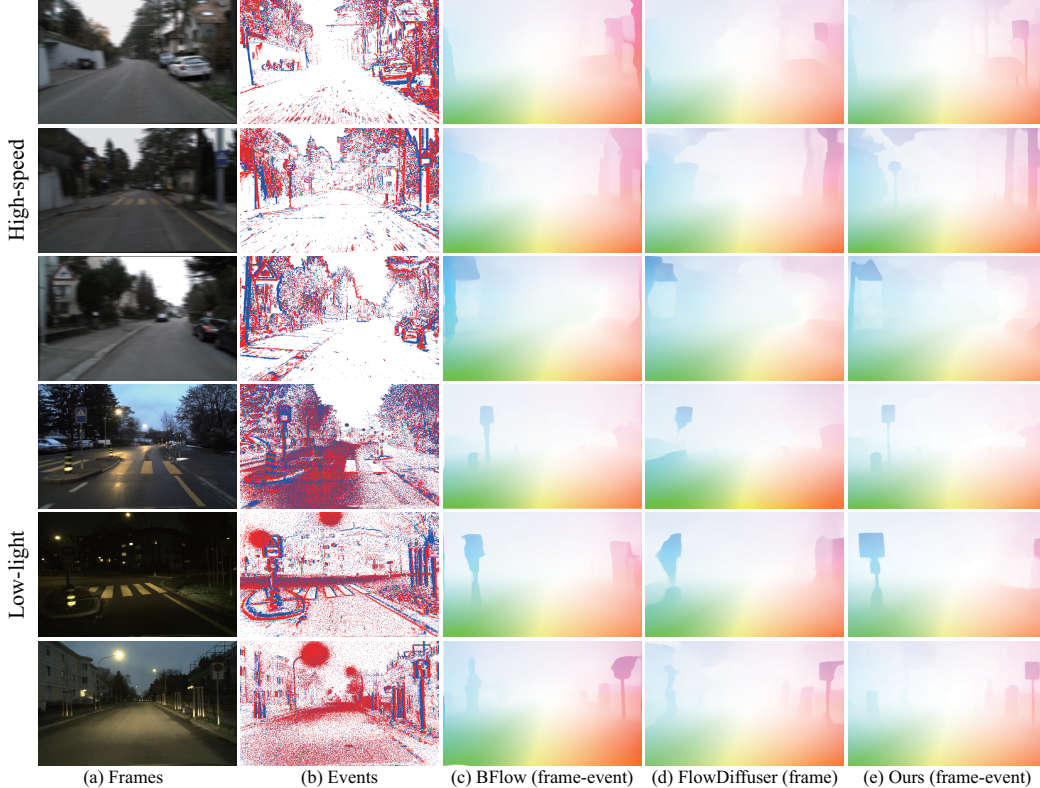


Figure 6: Comparison of optical flows on real HS-DSEC and LL-DSEC datasets.



## References

- [1] Mathias Gehrig, Mario Millhäusler, Daniel Gehrig, and Davide Scaramuzza. E-raft: Dense optical flow from event cameras. In *2021 International Conference on 3D Vision (3DV)*, pages 197–206. IEEE, 2021.
- [2] Mathias Gehrig, Manasi Muglikar, and Davide Scaramuzza. Dense continuous-time optical flow from event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4736–4746, 2024.
- [3] Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1312–1321, 2021.
- [4] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9772–9781, 2021.
- [5] Ce Liu, William T Freeman, Edward H Adelson, and Yair Weiss. Human-assisted motion annotation. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.
- [6] Ao Luo, Xin Li, Fan Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Flowdiffuser: Advancing optical flow estimation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19167–19176, 2024.
- [7] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1599–1610, 2023.
- [8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17755–17764, 2022.
- [9] Hanyu Zhou, Yi Chang, Zhiwei Shi, Wending Yan, Gang Chen, Yonghong Tian, and Luxin Yan. Adverse weather optical flow: Cumulative homogeneous-heterogeneous adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.