EDCFlow: Exploring Temporally Dense Difference Maps for Event-based Optical Flow Estimation

Supplementary Material

6. More comparison with SOTA methods

Compared with frame-based methods. Several frame-based SOTA methods [16, 40, 44] focus on achieving high-resolution optical flow estimation. They primarily perform pixel matching or refinement across multiple spatial resolutions in frames. Unlike frames, event data provides high temporal resolution, necessitating efficient method to capture continuous motion features. In this context, our multi-scale difference layer effectively handles temporal dynamics by leveraging temporal feature differences with low computations and can refine RAFT-like networks. Using publicly available source code under the same training settings, our method outperforms these methods in accuracy and efficiency on DSEC (in Tab. 8), highlighting its importance in event-based flow estimation.

Compared with event-based methods. We present accuracy vs. complexity in Fig. 5 to facilitate comparison. Our EDCFlow achieves higher performance as well as significant reductions in computational overhead over the state-of-the-art methods.

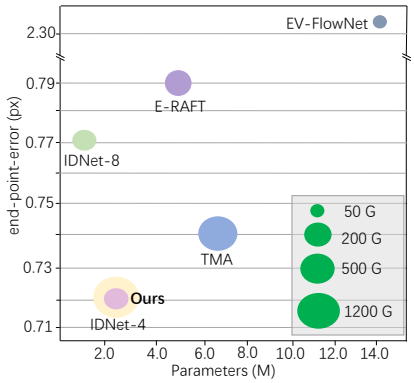

Figure 5. End-point-error (px) on DSEC vs. computational complexity (MACs: G) and model size (Parameters: M). All models are trained on DSEC, and tested with one NVIDIA 4090 GPU. The computational complexity corresponding to the size of the circle is shown in the legend in the lower right corner.

7. More Visualizations

Qualitative Results on MVSEC. Fig. 7 presents a qualitative comparison of our method with other methods on outdoor_day1 sequence of the MVSEC [45]. Compared to DSEC dataset [11], MVSEC has lower resolution and sparser events (especially at $dt = 1$), and it lacks occlusion and moving objects handling for its ground truth, making the data quality relatively poor. Despite this, compared to other methods, our approach holds superior performance at both $dt = 1$ and $dt = 4$, capturing clearer motion boundaries. This confirms that our method demonstrates greater adaptability across diverse data distributions and scenes.

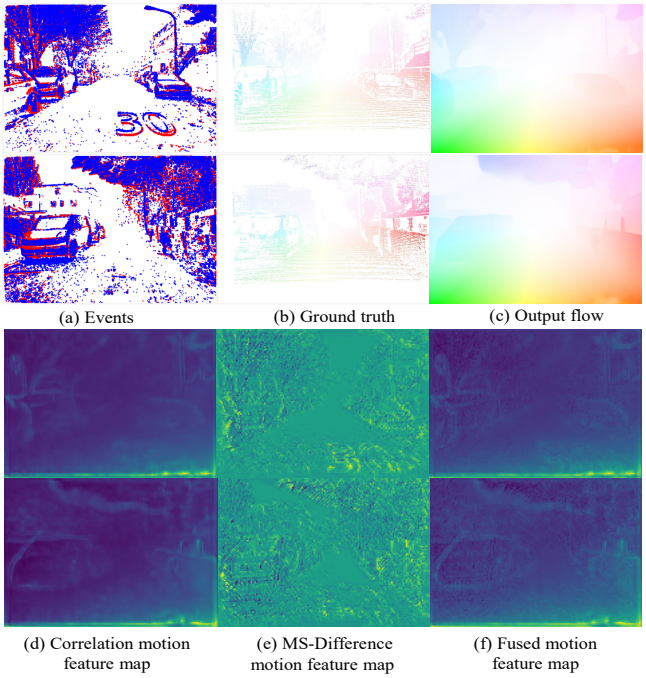

Figure 6. Illustration of motion feature maps. (a) Event data. (b) Optical flow ground truth. (c) Optical flow estimated by our method. (d) Correlation motion feature map. (e) Multi-scale temporal difference motion feature map fused with multi-scale attention. (f) The final motion feature map is aggregated from the difference motion feature and correlation motion feature.

Motion Feature Maps. To better illustrate the complementarity between the difference maps and correlation maps, we present in Fig. 6 the multi-scale temporal feature differences of adjacent feature maps and the motion features encoded from the cost volume on the DSEC dataset. The difference motion features exhibit strong responses in textured areas but are noisy, while the correlation motion features may produce blurred boundaries due to matching ambiguities. By adaptively fusing these two features, the response at motion boundaries can be enhanced, resulting in high-quality optical flow.

Flow Error Maps. To analyze the advantages and lim-

Method	EPE	AE	1PE	2PE	3PE	Param (M)	MACs (G)	Runt. (ms)
DIP [44]	3.06	10.95	77.8	46.6	27.5	5.4	590	92
GMFlow+ [40]	6.55	7.40	94.1	77.9	63.6	4.6	223	141
CCMR [16]	GPU out of memory (> 40GB)					11.6	2255	-
Ours	**0.72**	**2.65**	**10.0**	**3.6**	**2.1**	2.5	247	39

Table 8. Compared with frame-based SOTA methods.

itations of the model, we visualize the EPE distributions in Fig. 8, where the error for each pixel is represented as the square root error between the estimated and ground truth flows, shown using heatmaps. Since the ground truth of the DSEC dataset's test set is not publicly available, we use zurich_city_05_b and zurich_city_11_c from the training set as the test set for error analysis, while the remaining training samples are used to train the model. Our model can estimate accurate flow in most scenarios, particularly for complex textured objects like trees. In failure cases, however, our model encounters significant estimation errors in the spatial edge regions. This may be due to two reasons: first, the sparse events in these regions lack sufficient texture information, making it difficult for feature differences to encode motion boundaries and for the cost volume to resolve matching ambiguities; second, pixels at the spatial edges cannot aggregate enough contextual information to encode accurate motion features. These issues could be addressed by fusing images and leveraging multiple preceding and subsequent event streams.

8. More Ablation Studies

More ablation studies are also conducted on the DSEC dataset [11].

Iterations. Some existing methods [5, 35, 39] achieve better optical flow results through iterative refinement strategies, particularly for small objects with large displacements. The results in Tab. 9 show that as the number of iterations increases, the flow results become stable and reach a convergent state. The performance stabilizes when the number of iterations reaches 6. Moreover, excessive iterations may cause overfitting or oscillation around a local optimum, such as ours-8_6/12 in Tab. 7.

iterations	EPE	AE	1PE	2PE	3PE
2	0.86	3.11	13.8	5.3	3.1
4	0.77	2.85	11.5	4.2	2.4
<u>6</u>	**0.72**	2.65	**10.0**	**3.6**	**2.1**
8	**0.72**	**2.64**	10.1	**3.6**	**2.1**

Table 9. Iterations. The number of iterative refinements.

Event splitting. The time windows and temporal bins determine the input's temporal resolution and the number of input channels, respectively. The results in Tab. 10 show that a small value of g leads to performance degradation due to the loss of intermediate motion information. When the g is set to 5 and B to 3, sufficient continuous motion features are captured, achieving good performance. Increasing the window count further achieves comparable performance but introduces additional computational overhead.

g, B	EPE	AE	1PE	2PE	3PE	Param (M)	MACs (G)
1, 15	0.79	2.80	11.9	4.4	2.9	2.2	182
3, 5	0.75	2.74	10.7	3.9	2.3	2.3	209
<u>5, 3</u>	**0.72**	**2.65**	**10.0**	3.6	2.1	2.5	247
15, 1	**0.72**	2.68	**10.0**	**3.5**	**2.0**	2.5	322

Table 10. Event splitting. g represents the number of time windows. B denotes the number of time bins.

Optical flow estimation resolution. Different resolutions capture varying degrees of detail, offering distinct levels of granularity in the information provided. As shown in Tab. 11, flow estimation performs best at 1/4 resolution. We attribute this to the fact that the feature difference strategy is more effective in capturing local detail, making it more accurate at higher resolutions, while at 1/8 resolution, some detail is inevitably lost. Although the 1/2 resolution retains more detail, its tolerance to noise decreases, and the correlation features upsampled by two times remain too coarse to robustly enhance the final motion feature representation. Additionally, this resolution struggles to handle large displacements effectively.

Resolution	EPE	AE	1PE	3PE	MACs (G)	Param (M)
8	0.78	2.82	12.7	2.5	216	6.8
<u>4</u>	**0.72**	**2.65**	**10.0**	**2.1**	247	2.5
2	0.77	2.89	10.8	2.5	288	1.1

Table 11. Optical flow estimation resolution. The resolutions of 8, 4, and 2 represent flow computed at 1/8, 1/4, and 1/2 resolution of the input, respectively.

r	EPE	AE	1PE	3PE	MACs (G)	Param (M)
<u>1</u>	**0.72**	**2.65**	**10.0**	**2.1**	247	2.5
2	0.74	2.65	10.4	2.2	244	2.5
8	0.75	2.73	11.0	2.2	241	2.5
w/o	0.73	2.67	10.2	2.2	244	2.5

Table 12. Reduction ratio in the multi-scale temporal difference layer.

Feature dimension reduction ratio. In Tab. 12, we explore

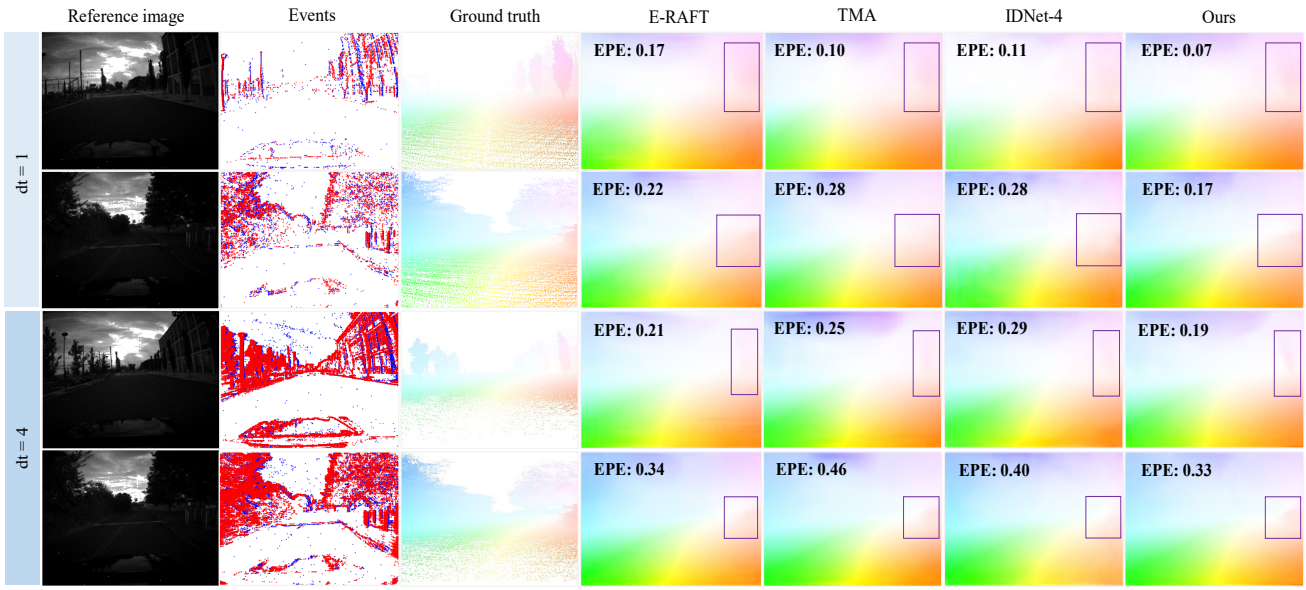

Figure 7. Qualitative results on the outdoor_day1 sequence on MVSEC [45]. Please zoom in for details.

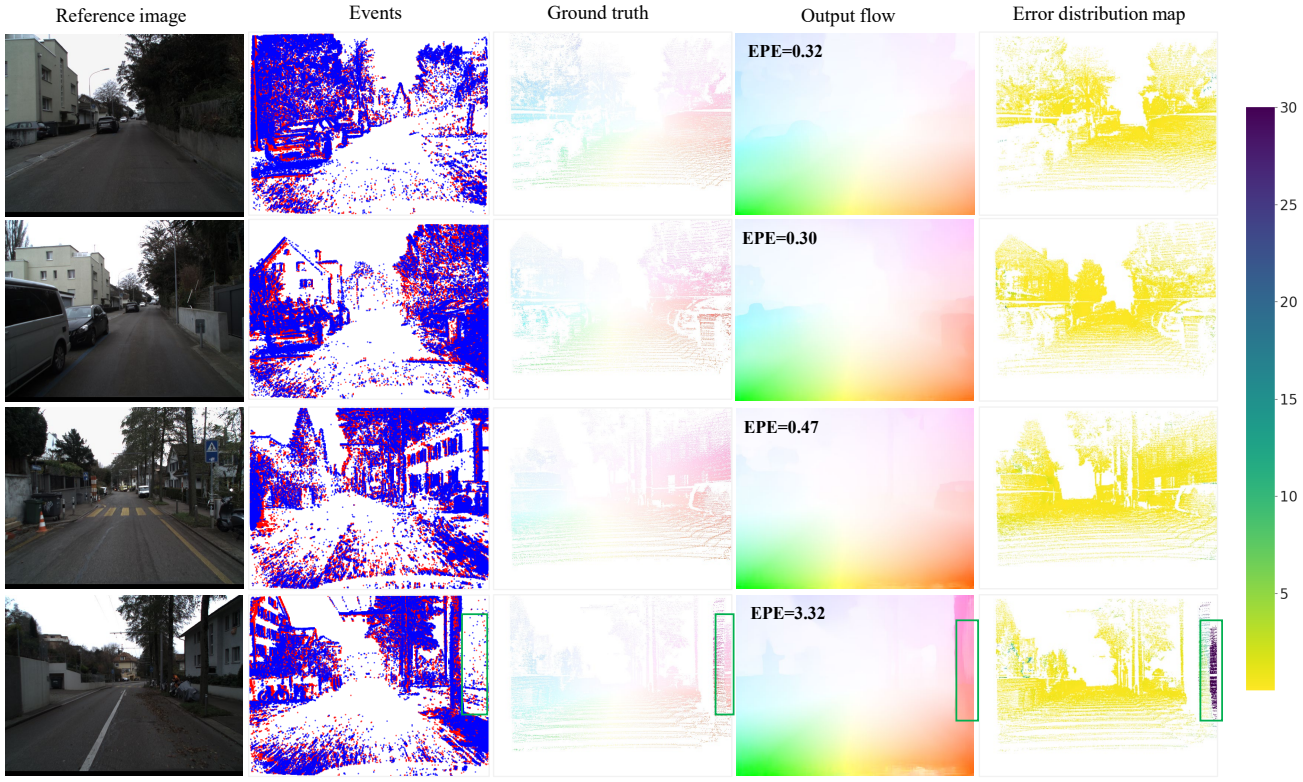

Figure 8. Visualization of error distribution maps. We present three high-quality flow estimation results with smaller EPE (the first three columns) and one failure case with larger EPE (the last column). The color bar indicates the per-pixel square root error magnitude, with darker colors representing larger errors.

the impact of r in the multi-scale temporal difference layer, with "w/o" indicating conv1 removed. We set $r = 1$ to balance accuracy and computations. When compared to "w/o", setting $r = 1$ brings imporvements in accuracy by leverag-

ing conv1 to enhance feature interaction. Since the channel number is 64 in our experiment, the effect of r on computation is marginal. For larger channel numbers, $r > 1$ can be used to balance accuracy and computation.

9. Future Work

In our work, we assume linear motion within short time windows (20 ms for DSEC and 10/40 ms for MVSEC), which shows good empirical performance and lower computational complexity. However, investigating alternative motion models, such as estimating higher temporal resolution intermediate flows to capture complex trajectories, could be an interesting direction for future research. Furthermore, event-image fusion-based optical flow estimation represents a promising research direction. By effectively combining the high temporal resolution and motion-capturing capability of event streams with the rich appearance and texture information provided by images, the accuracy and robustness of optical flow estimation can be significantly enhanced.