Unsupervised 3d Motion Estimation Using Event Camera

Supplementary Material

6. Error Analysis

To quantify the approximation introduced in deriving the relation between optical flow divergence and motion in depth, we analyze both the spatial and temporal sources of error.

We first consider the continuous approximation that neglects the spatial derivative term in Eq. (7). Defining the true deviation between $\alpha = \dot{Z}/Z$ and its divergence-based approximation as

$$e_c(x, y, t) \coloneqq \alpha(x, y, t) + \frac{1}{2} \operatorname{div} \mathbf{f}(x, y, t), \tag{16}$$

the exact expression of this quantity follows directly from the analytical form of the divergence:

$$e_c(x, y, t) = \frac{1}{2}\big(x\, \partial_x \alpha(x, y, t) + y\, \partial_y \alpha(x, y, t)\big). \tag{17}$$

Within a local neighborhood of radius R, this yields the bound

$$\big|e_c(x, y, t)\big| \leq \frac{1}{2} R\, \|\nabla \alpha(\cdot, t)\|_\infty, \tag{18}$$

where

$$\|\nabla \alpha(\cdot, t)\|_\infty = \sup_{(x,y)\in \text{patch}} \sqrt{(\partial_x \alpha)^2 + (\partial_y \alpha)^2}. \tag{19}$$

Thus, the spatial approximation error remains small whenever the depth variation is locally smooth or when the patch radius is small.

Next, we analyze the temporal discretization error arising when approximating the depth ratio

$$\tau = \frac{Z(t + \Delta t)}{Z(t)} = \exp\left(\int_t^{t+\Delta t} \alpha(s)\, ds\right). \tag{20}$$

A first-order Taylor expansion around t gives

$$\tau = 1 + \alpha(t)\Delta t + \frac{1}{2}\big(\alpha(t)^2 + \dot{\alpha}(\eta)\big)\Delta t^2 + O(\Delta t^3), \quad \eta \in [t, t + \Delta t]. \tag{21}$$

leading to the temporal approximation bound

$$\big|\tau - (1 + \alpha(t)\Delta t)\big| \leq \frac{1}{2}\big(\alpha_{\max}^2 + \|\dot{\alpha}\|_\infty\big)\Delta t^2 + O(\Delta t^3), \tag{22}$$

where

$$\alpha_{\max} = \sup_{s\in[t,t+\Delta t]} |\alpha(s)|, \qquad \|\dot{\alpha}\|_\infty = \sup_{s\in[t,t+\Delta t]} |\dot{\alpha}(s)|. \tag{23}$$

Combining the spatial and temporal approximations, let

$$\hat{\tau}(t) = 1 - \frac{1}{2} \operatorname{div} \mathbf{f}(t)\, \Delta t \tag{24}$$

be the discrete estimator derived from optical flow divergence. The total deviation between τ and $\hat{\tau}$ decomposes as

$$E = \tau - \hat{\tau} = \big(\tau - (1 + \alpha(t)\Delta t)\big) + \big(\alpha(t) + \tfrac{1}{2}\operatorname{div} \mathbf{f}(t)\big)\Delta t. \tag{25}$$

Substituting the bounds above yields the overall error estimate

$$|E| \leq \frac{1}{2}\big(\alpha_{\max}^2 + \|\dot{\alpha}\|_\infty\big)\Delta t^2 + \frac{1}{2} R\, \|\nabla \alpha(\cdot, t)\|_\infty\, \Delta t + O(\Delta t^3). \tag{26}$$

Equation (26) shows that the discrete approximation remains accurate when (i) local depth variation is spatially smooth, and (ii) the temporal sampling interval Δt is sufficiently small. These insights clarify its limitations under strong perspective distortion or rapidly varying scene depth.

7. Runtime

Our model contains 33.52M parameters and 211.39G MACs. When processing a 100 ms event sequence at a resolution of 320×960, it achieves an inference latency of 13.13 ms on a single RTX 4090 GPU. Under the same setting, Optical Expansion has 12.13M parameters and 81.20G MACs and requires 168.28 ms, while EMoTive has 5.61M parameters and 125.45G MACs and requires 58.86 ms. Although our approach entails higher model capacity and computational cost, it provides substantially lower inference latency.

8. Experiments on Real Dataset

Table 3 reports the performance of three methods on the DSEC dataset when trained on either CarlaEvent3D or DSEC. The evaluation follows the dataset split and protocol of [33, 35], where the ground truth is temporally sparse (0.3,s intervals), making the benchmark more challenging and consistent with prior evaluations. Since the ground-truth annotations of the DSEC test set are not publicly available, [33, 35] further split the training set into training and validation subsets and compute the ground truth motion in depth from the available annotations for evaluation purposes. The same evaluation strategy is adopted in this work.

The results show that the supervised EMoTive model experiences a notable performance drop when evaluated on unseen data, which reflects the impact of domain shift. In contrast, unsupervised methods exhibit better cross-domain

	EMoTive		Expansion		Ours	
EPE	28.74	0.43	5.14	4.07	4.08	3.38
F1	90.92	0.74	61.46	51.26	50.24	40.77
lm	1032.35	152.72	561.31	400.42	431.28	380.36

Left: trained on CarlaEvent3D; Right: trained on DSEC. *lm* = log-mid.

generalization. The proposed method consistently achieves better performance than Expansion on DSEC under both zero-shot and retrained settings.

9. Additional Ablation Experiments

In addition to the ablation studies reported in the main paper, experiments are conducted on the input event representation and the length of event window.

Input representation. Different event representations can significantly influence model performance. We evaluate several common forms, with results shown in Tab. 4. Event count accumulates events per pixel over a fixed window. Voxel grid encodes them into a spatiotemporal histogram. Time surface records the latest event timestamp per pixel.

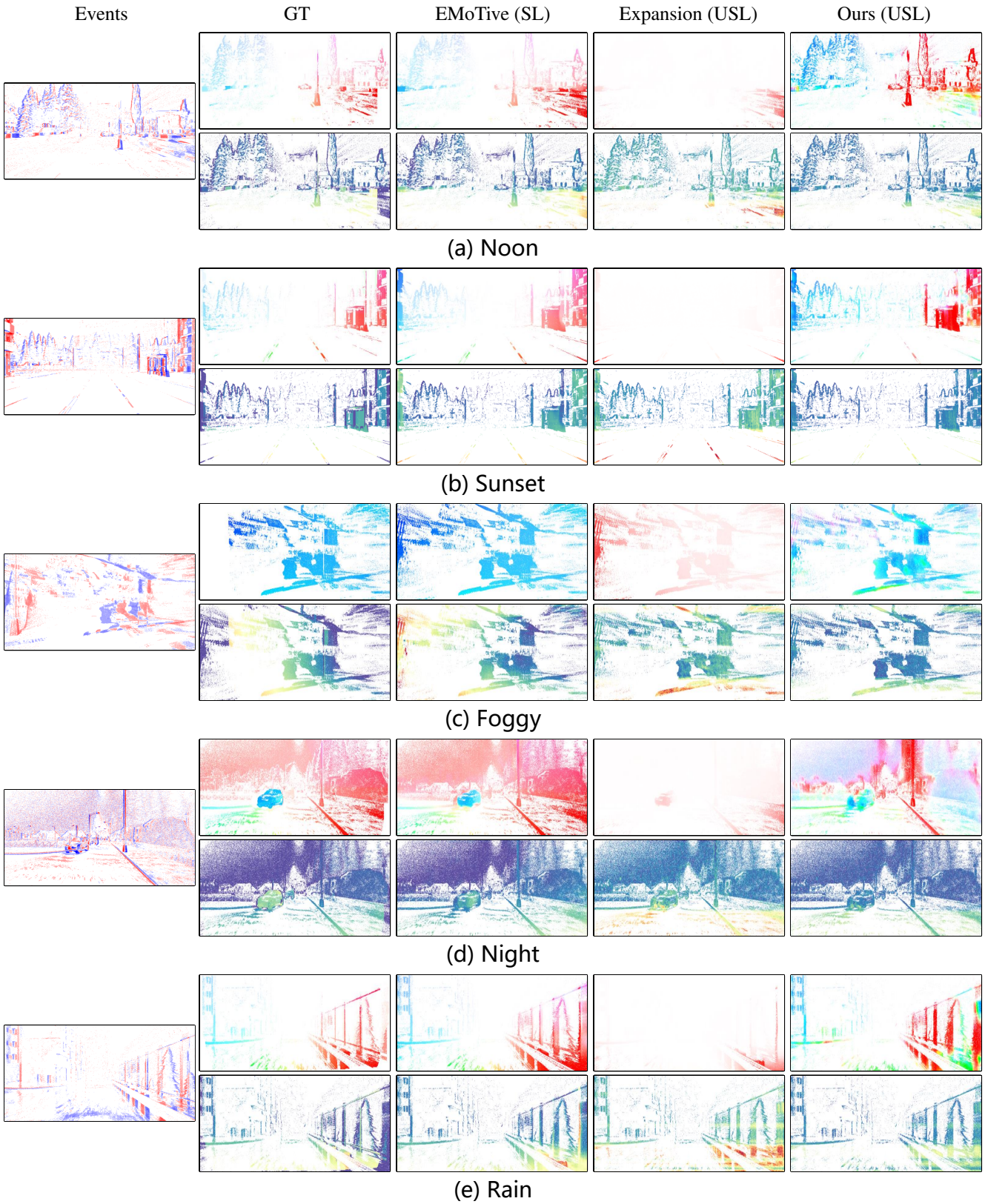
	EPE	F1	log-mid
Event count	3.38	41.07	**361.31**
Time surface	3.40	44.64	397.53
Voxel grid	**3.29**	**39.14**	364.01

Event window. We further evaluated how different numbers of input events affect the results, see Tab. 5.

	2000	4000	6000	8000
EPE	3.31	3.35	**3.29**	3.42
F1	39.76	39.89	**39.14**	41.22
log-mid	375.31	**359.05**	364.01	386.42

10. More Qualitative Results

Additional qualitative results under different weather conditions are shown below.


Events
GT
EMoTive (SL)
Expansion (USL)
Ours (USL)
(a) Noon
(b) Sunset
(c) Foggy
(d) Night
(e) Rain