Deep Event Visual Odometry

Simon Klenk*, Marvin Motzet*, Lukas Koestler, Daniel Cremers

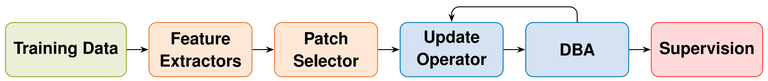
Technical University of Munich (TUM) & Munich Center for Machine Learning (MCML), Germany

3DV 2024

Motivation

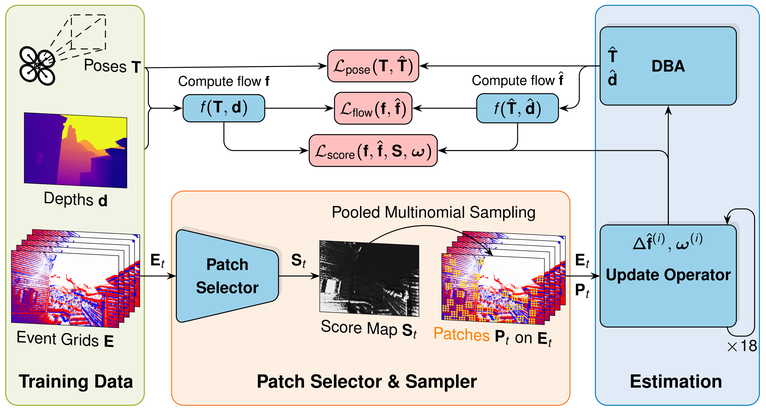
Event cameras offer the exciting properties of high temporal resolution, high dynamic range, and low latency. This makes them ideally suited for camera pose tracking during high-speed motion and in HDR lighting conditions. However, existing event-based monocular visual odometry approaches demonstrate limited performance on recent benchmarks.

We propose **Deep Event VO (DEVO)**, the first monocular event-only VO system with strong performance on seven real-world benchmarks.

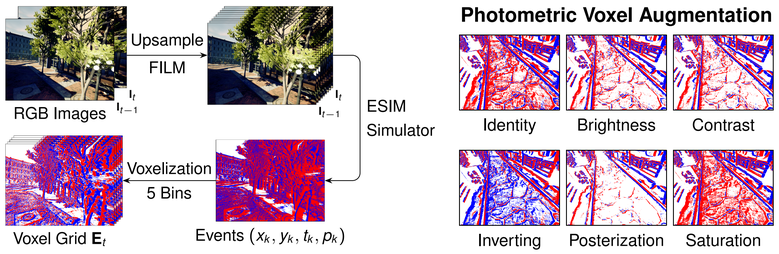


- **Supervised training** using camera poses and depths.
- Novel **patch selector** predicts a score map to highlight optimal coordinates for pose and optical flow estimation.
- Recurrent **update operator** refines optical flow estimates and updates poses and depths using **differentiable bundle adjustment (DBA)**.

Method

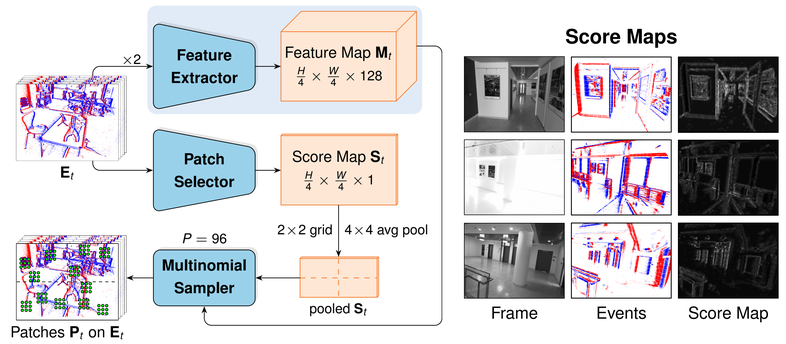


Training Data Generation



Photometric Voxel Augmentation

Patch Selector & Sampler



Loss Functions

Pose Loss

$$\mathcal{L}_{\text{pose}} = \sum_{(i,j)\, i \neq j} \left\| \log_{\text{SE3}} \left(\left(\mathbf{T}_j^{-1} \mathbf{T}_i \right)^{-1} \left(\hat{\mathbf{T}}_j^{-1} \hat{\mathbf{T}}_i \right) \right) \right\|$$

ground truth pose $\mathbf{T}$
estimated pose $\hat{\mathbf{T}}$

Flow Loss

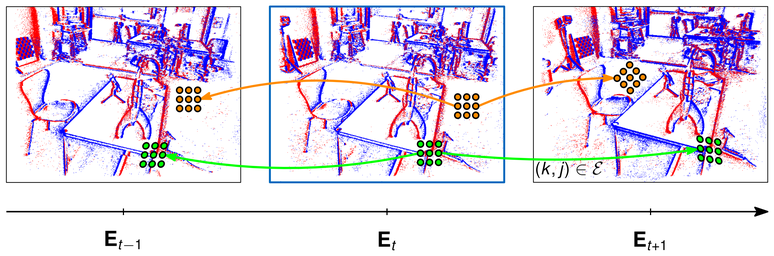
$$\mathcal{L}_{\text{flow}} = \sum_{(k,j) \in \mathcal{E}} \underbrace{\min \left\| \left(\mathbf{f}_{kj} - \hat{\mathbf{f}}_{kj} \right) \right\|}_{=: r_{kj} \text{ (flow residual)}}$$

ground truth optical flow $\mathbf{f}$
estimated optical flow $\hat{\mathbf{f}}$

Score Loss

$$\mathcal{L}_{\text{score}} = \frac{1}{|\mathcal{E}|} \sum_{(k,j) \in \mathcal{E}} s_k r_{kj} \underbrace{\left(1 - \alpha \ln \omega_{kj} \right)}_{\text{“inverted” weight}} - \ln \mathbf{S}[\mathbf{P}]$$

score value s_k
flow residual r_{kj}
sampled values $\mathbf{S}[\mathbf{P}]$



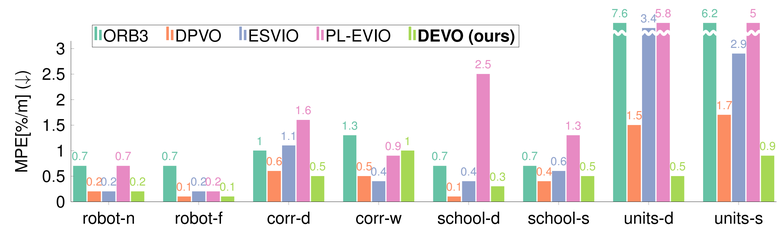
Results

Method	Modality	indoor forward 3	5	6	7	9	10	indoor 45 deg 2	4	9
ORB-SLAM3	Stereo VIO	0.55	1.19	–	**0.36**	0.77	1.02	2.18	1.53	**0.49**
VINS-Fusion	Stereo VIO	0.84	–	1.45	0.61	2.87	4.48	–	–	–
VINS-Mono	Mono VIO	0.65	1.07	**0.25**	0.37	0.51	0.92	**0.53**	1.72	1.25
DPVO	Mono VO	–	–	–	–	–	–	–	–	–
USLAM	Mono EVIO	–	–	–	–	–	–	–	9.79	4.74
PL-EVIO	Mono EVIO	0.38	0.90	0.30	0.55	**0.44**	1.06	0.55	1.30	0.76
EVO	Mono EO	–	–	–	–	–	–	–	–	–
DPVO†	Mono EO†	0.52	0.42	0.55	–	0.45	0.54	–	1.21	–
DEVO (ours)	Mono EO	**0.37**	**0.40**	0.31	0.50	0.61	**0.52**	0.72	**0.45**	0.89

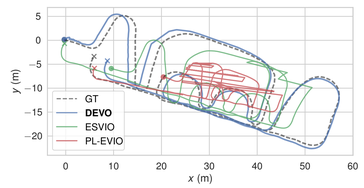
Evaluation on UZH-FPV Drone Racing Dataset with MPE[%/m] (↓), highlighting **Top1** and Top2. DPVO† is a re-trained DPVO model on E2VID reconstructions.

Results

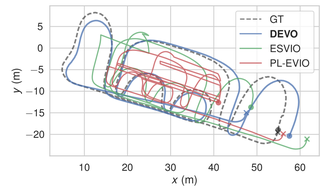
Quantitative Evaluation on VECtor Benchmark



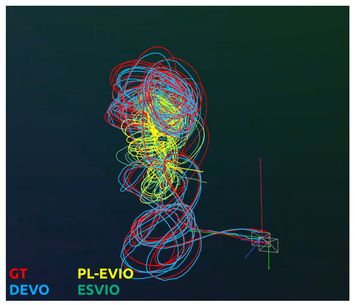
VECtor units-dolly



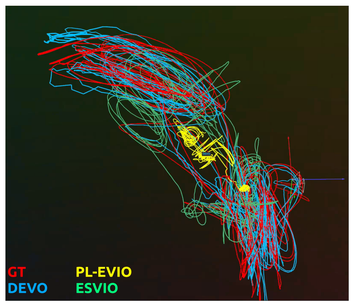
VECtor units-scooter



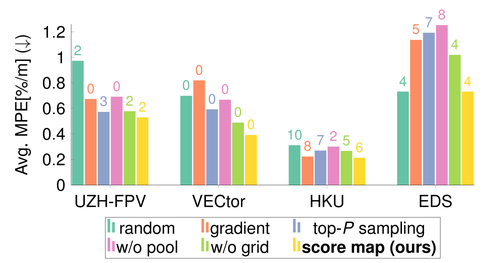
HKU aggressive_rotation



HKU hdr_dark_normal



Ablation Study



Ablation Study on Patch Selection and Sampling. DEVO trained with various patch selection methods and evaluated on four real-world benchmarks. Numbers near the bars indicate the total number of failures.

Conclusion

- DEVO demonstrates that **supervised learning** on simulated events of TartanAir dataset enables strong generalization to real-world event VO benchmark.
- We introduce a novel **patch selection** mechanism specifically tailored towards event data, which increases the accuracy and robustness of DEVO.

More Info

Paper, Code, Model, . . .