Supplementary Material: Data Association between Event Streams and Intensity Frames under Diverse Baselines

Dehao Zhang[1,2] Qiankun Ding[3] Peiqi Duan[1] Chu Zhou[1] Boxin Shi[1,2,4,5] ✉

[1] NERCVT, School of Computer Science, Peking University
[2] AI Innovation Center, School of Computer Science, Peking University
[3] Yuanpei College, Peking University
[4] Institute for Artificial Intelligence, Peking University
[5] Beijing Academy of Artificial Intelligence

6 Appendix

6.1 Positional Encoding

In the Transformer module of LSparse-Net and SDense-Net, we apply the 2D extension of positional encoding following [1] as:

$$\mathrm{PE}(x,y)^i := \begin{cases} \sin(\omega_k \cdot x), i = 4k \\ \cos(\omega_k \cdot x), i = 4k+1 \\ \sin(\omega_k \cdot y), i = 4k+2 \\ \cos(\omega_k \cdot y), i = 4k+3 \end{cases}, \tag{7}$$

where $\omega_k = \frac{1}{10000^{(2k/d)}}$, d is the number of channels which are applied with positional encoding, and i is the index for feature channels.

6.2 Mutual Nearest Neighbor Filtering

Scores on two directions form matches are calculate through Softmax:

$$s_p = \frac{\exp(M^l)}{\sum_{i,j}\exp(M^l(i,j,m,n))} \quad \text{and} \quad s_e = \frac{\exp(M^l)}{\sum_{m,n}\exp(M^l(i,j,m,n))}, \tag{8}$$

We believe that $I_p^l(\hat{i},\hat{j})$ matches $I_e^l(\hat{m},\hat{n})$ when the following equation (Eq. (9)) is satisfied. We denote the final match matrix as $\bar{M}^l$.

$$(\hat{m},\hat{n}) = \arg\max_{m,n} s_p(i,j,m,n) \quad \text{and} \quad (\hat{i},\hat{j}) = \arg\max_{i,j} s_e(i,j,m,n). \tag{9}$$

✉ Corresponding author: shiboxin@pku.edu.cn

6.3 Dataset Preparation

We sample a part of the synthetic ScanNet [2] dataset for training to synthesize event streams by V2E [3]. Indexes for the sub-set range from 0 to 699. When generating the event stream corresponding to each intensity frame, we include the latest 20,000 events earlier than the timestamp of the intensity frame.

6.4 Additional Results

Pose Estimation. To better demonstrate the performances of our framework on pose estimation, we show additional results for data association on the synthetic data of ScanNet [2] dataset in Fig. 5. It is shown that, in scenes with large baselines and sparse textures, our framework can establish correct matches between event streams and intensity frames.

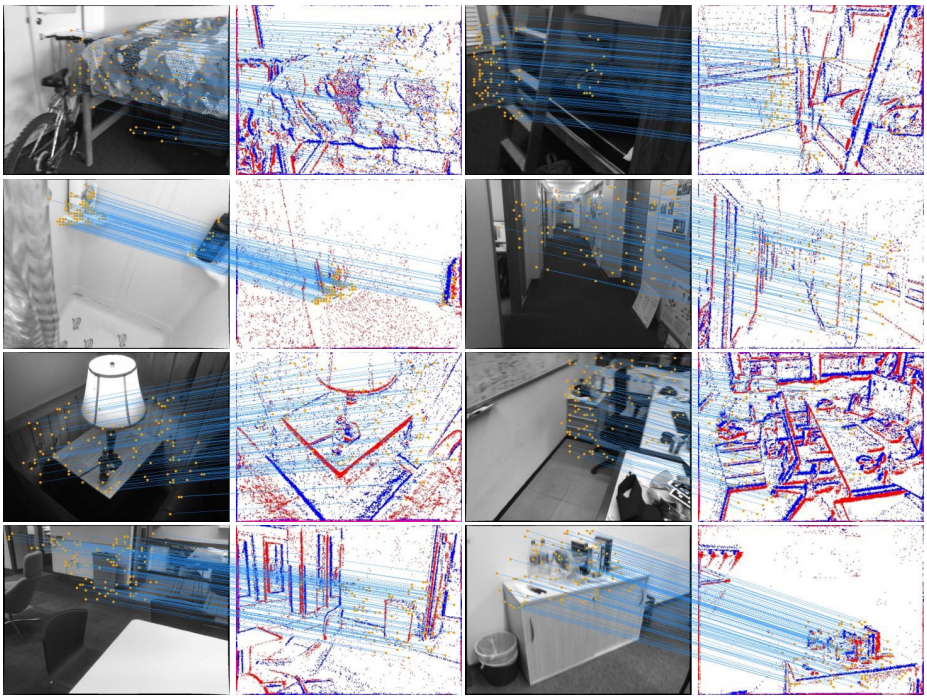

Fig. 5. Additional examples for pose estimation on the synthetic data. Our model can establish sound data association even when the views of the event streams and the intensity frames differ largely.

Stereo Depth Estimation. We further illustrate the capability of our framework to establish data association under small baselines by showing more results

in Fig. 6 on the Indoor Flying dataset from MVSEC [4]. Our framework is adaptive to the task of stereo depth estimation, as it outputs results close to the ground truths.

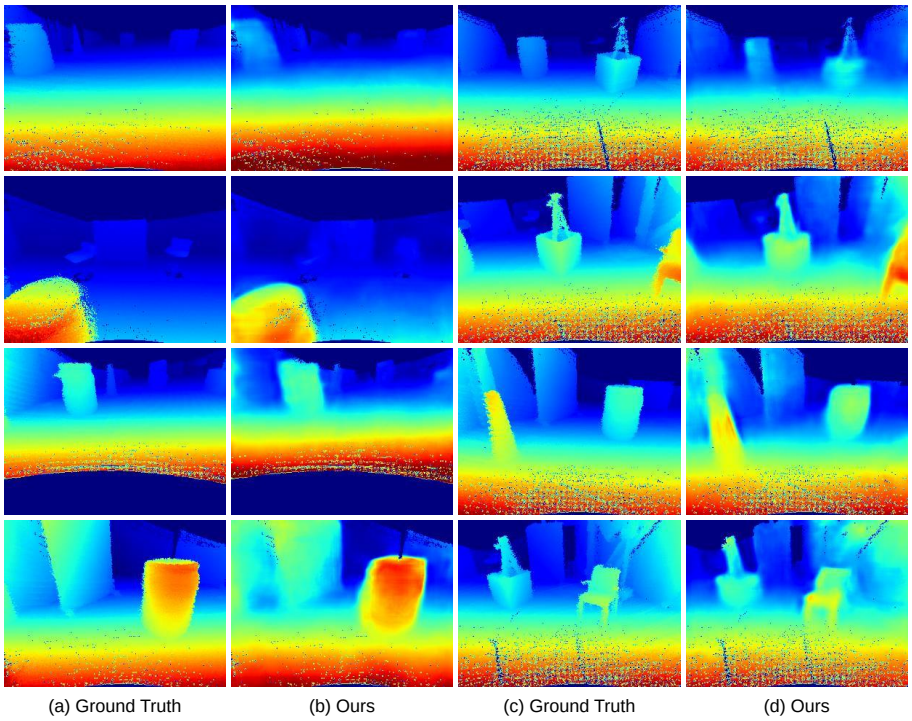

Fig. 6. Additional qualitative comparison on the Indoor Flying dataset of MVSEC [4]. The first and third columns show the ground truth, whereas the second and fourth columns show the outputs of our framework. We only select frames from sequence 1. In the first and second columns, from top to bottom, we select frame 150, 400, 700, and 925. In the third and fourth columns, from top to bottom, we select frame 250, 550, 850, and 1185.

References

1. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229 (2020)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. of Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5828–5839 (2017)

3. Hu, Y., Liu, S.C., Delbruck, T.: V2E: From video frames to realistic dvs events. In: Proc. of Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1312–1321 (2021)
4. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)