Spike Transformer: Monocular Depth Estimation for Spiking Camera *Appendix*

Jiyuan Zhang[1,*], Lulu Tang[2,3,*], Zhaofei Yu[1,†], Jiwen Lu[3], and Tiejun Huang[1,2]

[1] Department of Computer Science, Peking University
[2] Beijing Academy of Artificial Intelligence
[3] Department of Automation, Tsinghua University
jyzhang@stu.pku.edu.cn, {lulutang, lujiwen}@tsinghua.edu.cn, {yuzf12, tjhuang}@pku.edu.cn
*Joint First Authors, †Corresponding Author

1 Ablation study on different temporal window *T* and different time-scale partition *n*

Tab. 1 reports results with different T and n on the "DENSE-spike" dataset. $T = 128$ and $n = 4$ are finally adopted.

Table 1. Ablation study on different T **and** n**.**

Size of T	**Abs Rel** ↓	**Sq Rel** ↓	**RMS log** ↓	**SI log** ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
128	0.606	**18.388**	**0.706**	0.395	**0.682**	**0.762**	**0.813**
64	**0.595**	19.179	0.734	0.419	0.648	0.740	0.798
32	0.697	22.015	0.730	**0.393**	0.632	0.734	0.793
Number of n	**Abs Rel** ↓	**Sq Rel** ↓	**RMS log** ↓	**SI log** ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
4	0.606	**18.388**	**0.706**	**0.395**	**0.682**	**0.762**	**0.813**
2	**0.602**	19.290	0.714	0.412	0.670	0.753	0.804
8	0.695	23.681	0.748	0.446	0.668	0.748	0.800

2 Comparison to methods using images or events.

In this section, we give more comparative results with the depth estimation methods using images and events. We compare with an existing depth estimation method, RAM-Net [1], which uses the modalities of image and event from "DENSE" dataset. We train our Spike-T using spikes from the synthetic "DENSE-spike" dataset. As reported in Tab. 2, our Spike-T achieves comparable performance with RAM-Net [1] trained on event and image.

In addition, we compare Spike-T with the two-stage method, which first reconstructs images from spike streams [6] and then estimates depth based on the

Table 2. Comparisons with different methods and data input.

Method	**Modality**	**Abs Rel** ↓	**RMS log** ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Spike-T(Ours)	Spike	**0.606**	0.706	0.682	0.762	0.813
I baseline [1]	Image	0.752	0.786	0.655	0.745	0.802
E baseline [1]	Event	0.849	0.836	0.633	0.734	0.795
RAM Net [1]	Event+Image	0.717	**0.671**	**0.705**	**0.797**	**0.849**

recovered images [1]. On the "DENSE-spike" dataset, our method gets a more satisfactory depth map than the two-stage one (Abs Rel↓: 0.606 v.s. 0.759). Moreover, the inference speed for the two-stage strategy can hardly satisfy the real-time requirement due to the extra computational overhead in the reconstruction process.

3 High-speed depth recovery using Spike v.s. Image

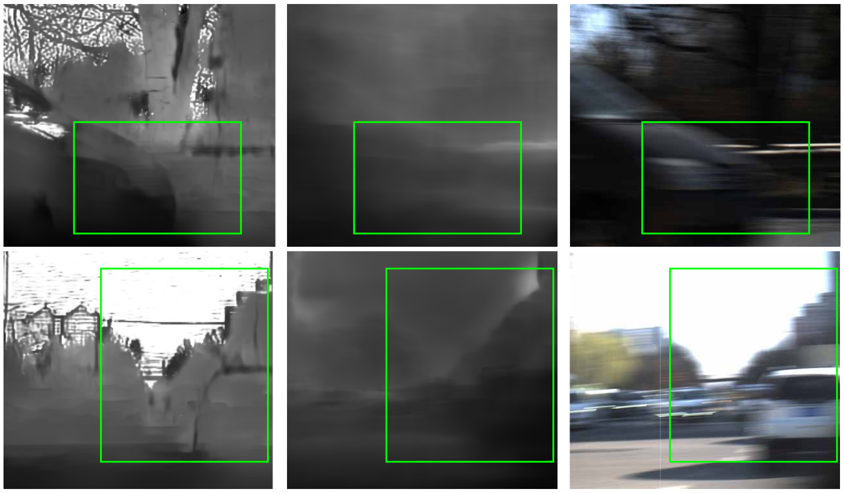
Fig. 1 illustrates examples of fast-moving and shaky scenes captured by a synchronized spike camera and a traditional one with a beam splitter. (a) Depth recovery from **real** spike streams by Spike-T trained on "DENSE-spike"; (b) Depth recovery from real images using DPT-Hybrid [4] pre-trained on large-scale datasets and fine-tuned on KITTI [2]; (c) Images captured by a traditional camera. Visualization results demonstrate the advantage of spike cameras over conventional ones under some challenging scenarios.

4 Instructions on More Visualization Results

More visualization results can be found at **THIS LINK**. We evaluate our method by training networks on the synthetic dataset 'DENSE-spike' and testing on both the synthetic dataset and the real dataset 'Outdoor-Spike'. There are three videos, two of them are full validating sequences on the synthetic dataset, named as **town06.mp4** and **town07.mp4**, while the other is the full testing sequence of real dataset, dubbed as **outdoor.mp4**. In addition, we present demos of sequential results for U-Net[5], E2Depth[3] and our Spike-T.

5 Details of Metrics

For a thorough evaluation of the proposed model, we introduce several important metrics, including absolute relative error **(Abs Rel.)**, square relative error (**Sq Rel.**), mean absolute depth error (**MAE**), root mean square logarithmic error (**RMSE log**) and the accuracy metric (**Acc.** δ). Detailed formulations are as follows.



(a) Spike-T (b) DPT-Hybrid (c) Real Image

Fig. 1. Depth recovery from scenes at a relative high speed of 120 km/h (Row 1), and from a shaky scenario (Row 2).

- **Absolute Relative Error (Abs Rel.)** computes average errors on the normalized depth map for every pixel, formulated as $\frac{1}{N}\sum_p \frac{|\mathcal{D}_p - \hat{\mathcal{D}}_p|}{|\mathcal{D}_p|}$, which normalizes the value of depth to the range [0,1].
- **Square Relative Error (Sq Rel.)**, formulated as $\frac{1}{N}\sum_p \frac{|\mathcal{D}_p - \hat{\mathcal{D}}_p|^2}{|\mathcal{D}_p|}$, which focuses on large depth errors due to its square numerator.
- **Mean Absolute Error (MAE)** can be formulated as $\frac{1}{N}\sum_p |\mathcal{D}_p - \hat{\mathcal{D}}_p|$.
- **Root Mean Square Error (RMSE)** is a classic metric for per-pixel prediction error and the logarithm version (**RMSE log**) can be denoted as $\sqrt{\frac{1}{N}\sum_p |\log \mathcal{D}_p - \log \hat{\mathcal{D}}_p|^2}$.
- **The Accuracy (Acc.) as** δ denotes the percentage of all pixels $\mathcal{D}_p$ that satisfy $\max(\frac{\hat{\mathcal{D}}_p}{\mathcal{D}_p}, \frac{\mathcal{D}_p}{\hat{\mathcal{D}}_p}) < thr$, where $thr = 1.25, 1.25^2, 1.25^3$.

Where N is the number of all valid pixels $\mathbf{p}$, $\mathcal{D}$ and $\hat{\mathcal{D}}$ are the ground truth depth and the predicted depth respectively.)

6 Measurements on the Model

We test our model on one NVIDIA A100-SXM4-80GB GPU and the inference speed is about 22.6 FPS. In addition, The total number of trainable parameters of the model is 20.55 MB.

References

1. Gehrig, D., Rüegg, M., Gehrig, M., Hidalgo-Carrió, J., Scaramuzza, D.: Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction. IEEE Robotics and Automation Letters **6**(2), 2822–2829 (2021)
2. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
3. Hidalgo-Carrió, J., Gehrig, D., Scaramuzza, D.: Learning monocular dense depth from events. In: 2020 International Conference on 3D Vision (3DV). pp. 534–542. IEEE (2020)
4. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (2021)
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI. pp. 234–241. Springer (2015)
6. Zheng, Y., Zheng, L., Yu, Z., Shi, B., Tian, Y., Huang, T.: High-speed image reconstruction through short-term plasticity for spiking cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6358–6367 (2021)