Depth from Asymmetric Frame-Event Stereo: A Divide-and-Conquer Approach Supplementary Material

Xihao Chen Wenming Weng Yueyi Zhang Zhiwei Xiong
University of Science and Technology of China
{xhchen10, wmweng}@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

This supplementary document is organized as follows:

Sec. 1 provides more implementation details of different methods.

Sec. 2 provides the definitions of different evaluation metrics.

Sec. 3 provides the efficiency comparison of different methods.

1. More Implementation Details.

All comparison methods and ours are implemented with Pytorch on an NVIDIA GeForce RTX 3090 GPU. The training setups of different methods are consistent, including event representation, optimizer, learning rate schedule, and batch size.

PSMNet [1] & AANet [11] & RAFT-Stereo [4]. For these three representative stereo matching networks, we use their official implementation except for the identical yet independent feature extractors, given the asymmetry of the input frame and event images. We adopt the smooth L_1 loss [1] as recommended.

E2VID [5]+PSMNet/AANet/RAFT-Stereo. For this category, we use the same implementation as aforementioned, except that the network inputs are frame images and the intensity images reconstructed from event streams by E2VID [5]. The weights of E2VID we adopt are retrained by [7] which present better reconstruction results. We do not finetune E2VID on the DSEC dataset, since it requires aligned event and intensity data which is not available.

DCNet [10] is a three-step depth estimation method from SAFE systems: (i) estimate *sparse* disparity maps from the binary edge images of event and frame images by minimizing a matching cost; (ii) estimate *dense* disparity maps directly from event and frame images with a stereo matching network, similar to the methods in the first category; (iii) fuse the sparse and dense disparity maps with a U-Net [6]. We use the implementation provided by the authors in the first step. For the second step, we use PSMNet [1] as an embodiment given its superior performance on DSEC over other networks.

HDES [13] is an end-to-end depth estimation network based on U-Net architecture. A pyramid attention module is adopted to help focus on the important areas for different modalities. Rather than the event queue [8], the same event representation [12] as our method is adopted to achieve significantly faster convergence speed and comparable performance as recommended in [3].

Our Method. We adopt the feature extractor, matching module, and regularization module recommended by PDS [9], considering their distinct trade-off between performance and efficiency. We decrease the downsample scale of the regularization module from 16 to 8, because we empirically found that it decreases the computational consumption and the number of network parameters without sacrificing the performance of our method on the DSEC dataset. In 3D ConvLSTM cell, we replace the $\tanh$ activation with ELU to achieve faster convergence and better performance following [2].

2. Evaluation Metrics

The adopted disparity or depth metrics are calculated as

i. $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|d_i - \hat{d}_i\right|$

ii. $\text{MAE}_{rel} = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|d_i - \hat{d}_i\right|}{d_i}$

iii. $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - \hat{d}_i)^2}$

iv. $\text{Inlier Ratio}(\Delta < 1.05^j) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[\frac{d_i}{1.05^j} < \hat{d}_i < 1.05^j \times d_i] \times 100\%$

v. $\text{NPE} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[\left|d_i - \hat{d}_i\right| < N] \times 100\%$

where $\hat{d}_i$ and d_i denote the predicted and ground-truth depth (or disparity) value for a given pixel i with valid ground-truth value, n is the number of valid pixels, and $\mathbb{1}$ is the indicator function.

3. Efficiency Comparison

In Tab. 1, we compare the computational consumption (*i.e.*, FLOPs) and the numbers of network parameters (*i.e.*, Params) for different methods. FLOPs is computed with stereo inputs of size 256×256 and maximum disparity value 96. The FLOPs of our method with temporal fusion is computed for each time step.

Table 1. Efficiency of different comparison methods. 'FLOPs' is the number of floating point operations while 'Params' denotes the number of network parameters. DCNet [10] computes sparse disparity maps using an optimization method, which is not counted in FLOPs.

Method	FLOPs (G)	Params (M)
RAFT-Stereo [4]	190.31	12.22
AANet [11]	**41.44**	11.19
PSMNet [1]	78.88	8.57
E2VID [5] + RAFT-Stereo [4]	249.84	22.93
E2VID [5] + AANet [11]	100.96	21.90
E2VID [5] + PSMNet [1]	138.41	19.28
DCNet [10] *	141.86	21.22
HDES [13]	254.41	88.32
Ours	79.80	**6.47**

References

[1] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, 2018. 1, 2

[2] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In *CVPR*, pages 15324–15333, 2021. 1

[3] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 1

[4] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *3DV*, pages 218–227, 2021. 1, 2

[5] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE Transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019. 1, 2

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[7] Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *ECCV*, pages 534–549. Springer, 2020. 1

[8] Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, and Michael Hirsch. Learning an event sequence embedding for dense event-based deep stereo. In *ICCV*, pages 1527–1537, 2019. 1

[9] Stepan Tulyakov, Anton Ivanov, and François Fleuret. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. In *NeurIPS*, 2018. 1

[10] Ziwei Wang, Liyuan Pan, Yonhon Ng, Zheyu Zhuang, and Robert Mahony. Stereo hybrid event-frame (shef) cameras for 3d perception. In *IROS*, pages 9758–9764, 2021. 1, 2

[11] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *CVPR*, pages 1959–1968, 2020. 1, 2

[12] Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, and Kostas Daniilidis. Unsupervised event-based learning of optical flow, depth, and egomotion. In *CVPR*, pages 989–997, 2019. 1

[13] Yi-Fan Zuo, Li Cui, Xin Peng, Yanyu Xu, Shenghua Gao, Xia Wang, and Laurent Kneip. Accurate depth estimation from a hybrid event-rgb stereo setup. In *IROS*, pages 6833–6840, 2021. 1, 2