A Supplemental materials

A.1 Comparison of latency using memory bank

As discussed in previous ections, to better demonstrate that the temporal advantage of StreamFlow is not solely due to the memory bank, this section explores the efficiency comparison between StreamFlow and other methods when using a memory bank. Given that the model's runtime is closely related to the coding implementation, this comparison prioritizes officially open-sourced multi-frame optical flow methods. However, as of the writing of this paper, the choices for leading open-source multi-frame methods are quite limited, and thus VideoFlow [8] was selected for comparison. The experimental setup and the machine are consistent with those described in previous sections, and the measured time is the average of five tests. The input is resized to 432×1024, and the model is trained via (C+)T manner. As shown in Table Appendix A.1, it can be observed that StreamFlow still exhibits good efficiency in time. This is because, in addition to the memory bank, it further optimizes the average estimation time in the decoder.

Method	Sintel (clean)	Sintel (final)	Fl-EPE	Fl-all	Latency	Hardware
VideoFlow-BOF [8]	1.03	2.19	3.96	15.3	122.37ms	A100-40G
StreamFlow (Ours)	**0.87**	**2.11**	**3.85**	**12.6**	**85.53ms**	A100-40G

Table 1: Comparison of latency using memory bank.

A.2 Qualitative analysis on real-world scenes

In this section, we facilitate our visualizations and evaluations using two prominent real-world datasets, namely DAVIS [7]. The DAVIS dataset, short for Densely Annotated VIdeo Segmentation, is a widely recognized benchmark in the field of computer vision. It comprises high-quality video sequences captured in diverse scenarios, encompassing a broad range of challenging visual conditions such as occlusions, motion blur, and dynamic object interactions. The dataset provides pixel-level annotations for every frame, facilitating precise evaluation and comparison of various video segmentation methods. The visualizations on the DAVIS dataset are shown in Figure 1. Our model is pretrained using the "T" and "T+S+H+K" schedule and then fine-tuned on KITTI [6]. "T" denotes the FlyingThings [5] dataset and "T+S+H+K" refers to the combination of the FlyingThings, Sintel [2], HD1K [4], and KITTI datasets. Then we infer our models on the DAVIS dataset. The number of refinements is set to 12. The number of input frames for each non-overlapping group is 3. We could learn that StreamFlow demonstrates remarkable adaptability across real-world datasets, showing its robust performance in challenging scenes for optical flow estimation. This is particularly evident in scenarios such as the occlusion of the bear's hind legs in the first row, first column, and the small motion of the small tennis ball in the last column. Additionally, it can be observed that in the motion captured in the first row, second, and third columns, the hind legs of the camel and the leg movements of the dancer are also vividly delineated. These instances reaffirm its efficacy in diverse and demanding environments for optical flow estimation.

A.3 Qualitative analysis on occluded regions

In this section, we focus on the performance of the occluded regions. As discussed in previous works [3, 9], here we term occlusions as areas where pixels appear in the current frame while disappearing in the next frame. We visualize the flow-error map on occluded regions of the Sintel dataset with the official occlusion masks. All models are trained using the (C+)T schedule. As shown in Figure 2, significant occluded areas are highlighted using red boxes. A darker color in the flow-error map denotes a more significant error. We could learn that StreamFlow achieves better overall performance, and attains leading performance on the occluded regions.

A.4 Initialization of GTR

In this section, we investigate the impact of different GTR initialization methods. Previous works in spatio-temporal modeling such as [1] have suggested initializing the temporal modules with zero values. We employed two distinct initialization approaches, namely zero initialization and PyTorch's

1
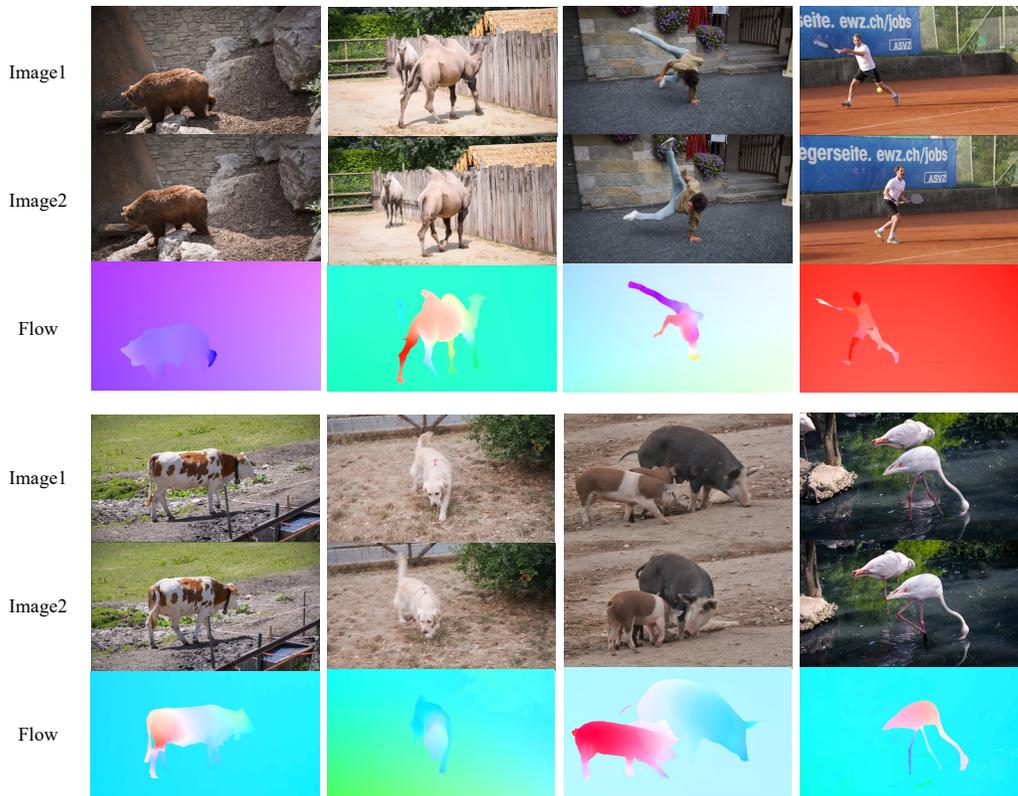
Figure 1: Visualizations of predicted flows on DAVIS [7]. StreamFlow demonstrates robust generalization to other real-world datasets, performing well in challenging scenarios for optical flow estimation, as evidenced by instances such as the occluded hind legs of the bear in the first column and the small tennis ball in the last column.

default initialization, and the corresponding results are presented in Table 2. Following training on the FlyingThings dataset, the model was tested on the Sintel and KITTI datasets. It is evident from the results that the zero initialization could contribute to a better overall performance.

Method	Sintel (Clean)	Sintel (Final)	KITTI (EPE)	KITTI (Fl-all)
Default	**0.91**	2.20	4.05	13.44
Zero-init	0.93	**2.15**	**3.92**	**12.36**

Table 2: Comparison of different ways of initialization. All models are trained under the FlyingThings.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[3] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021.

2
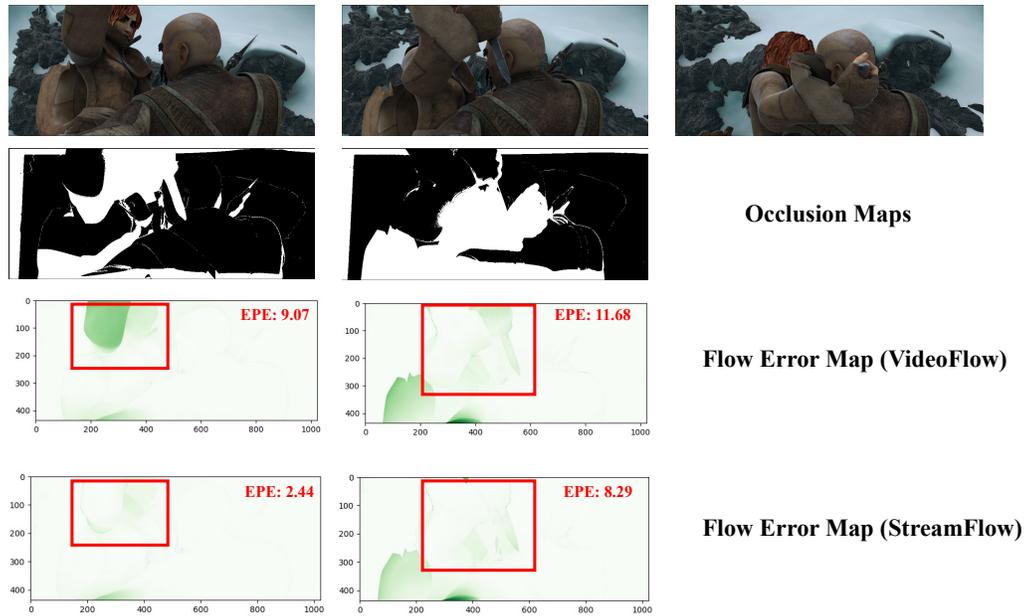
Figure 2: Visualizations of the performance on the occluded regions. StreamFlow achieves comparable performance even with advanced methods. All models are trained on the FlyingThings dataset. A darker color in the flow error map denotes a higher estimation error compared with ground truth.

[4] Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.

[5] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[6] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015.

[7] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[8] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

[9] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *Advances in Neural Information Processing Systems*, 35:11313–11326, 2022.

3