VideoFlow: Supplementary Material

Xiaoyu Shi[1,2] Zhaoyang Huang[1,2*] Weikang Bian[1] Dasong Li[1] Manyuan Zhang[1]
Ka Chun Cheung[2] Simon See[2] Hongwei Qin[3] Jifeng Dai[4] Hongsheng Li[1,5,6*]

[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]NVIDIA AI Technology Center [3]SenseTime Research [4]Tsinghua University
[5]Centre for Perceptual and Interactive Intelligence (CPII) [6]Shanghai AI Laboratory
{xiaoyushi@link, drinkingcoder@link, hsli@ee}.cuhk.edu.hk

1. More Implementation Details

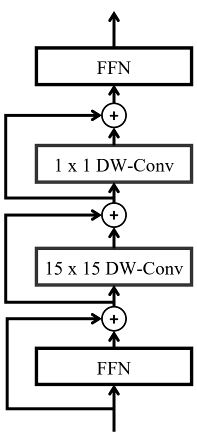


Figure 1. Network details of SKBlock.

Following SKFlow, we replace the vanilla convolutions with super-kernel blocks [5, 3] for all encoders. As shown in Figure 1, SKBlock utilizes large-kernel depth-wise convolutions to enlarge the receptive fields while maintaining low computational cost.

As shown in Table 1, we follow SKFlow to implement these modules. We simply concatenate the correlation feature and bi-directional flows along the channel as inputs to CorrEncoder and FlowEncoder, respectively. The MotionEncoder of TROF is the same as that of SKFlow. For MOP, it additionally takes the 144-dim concatenation of motion features as input and additionally outputs a 48-dim updated motion feature.

2. More Details About Training Data

Note that the KITTI benchmark provides 20-frame sequences and each sequence has one GT flow map between frames 10-11. Therefore, we use the same number of GT for training as other two-frame methods. Since only the FlyingThings dataset provides bi-directional ground-truth optical flows, for other datasets, we randomly flip the input sequence and ground-truth optical flows with a probability of 0.3 to supervise the predicted backward flows. We follow FlowFormer for other settings of data augmentation.

*Corresponding author: Zhaoyang Huang and Hongsheng Li

Block	CorrEncoder	FlowEncoder	MotionEncoder	Updater	FlowHead
Type	SKBlock	SKBlock	SKBlock	SKBlock	SKBlock
Input Dim.	648	4	256	512	128
Output Dim.	192	64	124	128	4
Hidden Dim.	256	96	256	512	128

Table 1. Implementation details of encoders.

3. Online Mode v.s. Offline Mode

We conducted an experiment where only past frames are used (online setting). Our 5-frame VideoFlow achieves 4.08 Fl-all on the KITTI test set, outperforming 3-frame VideoFlow (4.44) and all previous two-frame methods. Specifically, we take KITTI images with indices 7, 8, 9, 10, 11 as inputs. The KITTI benchmark only evaluates the accuracy between images 10-11. The improvement reflects the superiority of VideoFlow in online setting.

4. Additional Qualitative Results

We provide additional visualisations in Figure 4, comparing our VideoFlow with previous best model FlowFormer++ [4, 2, 1].

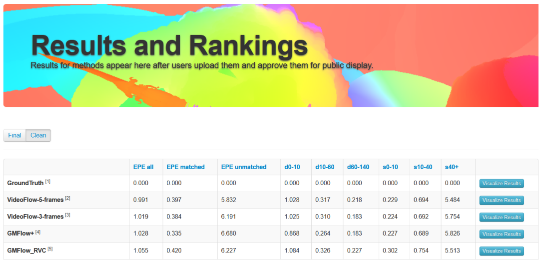
5. Screenshots of Sintel Leaderboard

Figures 2 and 3 are the screenshots of evaluation results on the Sintel benchmarks. Our five-frame and three-frame models take the first and second places.

References

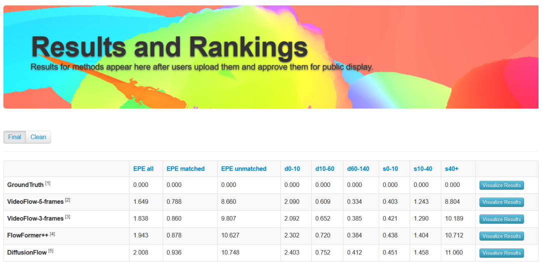
[1] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022. 1

[2] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Yijin Li, Hongwei Qin, Jifeng Dai, Xiaogang Wang, and Hongsheng Li. Flowformer: A transformer architecture and its masked cost volume autoencoding for optical flow. *arXiv preprint arXiv:2306.05442*, 2023. 1



	EPE all	EPE matched	EPE unmatched	d0-10	d10-60	d60-140	s0-10	s10-40	s40+	
GroundTruth [1]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Visualize Results
VideoFlow-5-frames [2]	0.991	0.397	5.832	1.028	0.317	0.218	0.229	0.694	5.484	Visualize Results
VideoFlow-3-frames [3]	1.019	0.384	6.191	1.025	0.310	0.183	0.224	0.692	5.754	Visualize Results
GMFlow+ [4]	1.028	0.335	6.680	0.868	0.264	0.183	0.227	0.689	5.826	Visualize Results
GMFlow_RVC [5]	1.055	0.420	6.227	1.084	0.326	0.227	0.302	0.754	5.513	Visualize Results

Figure 2. **Screenshot of Sintel (Clean) leaderboard.**



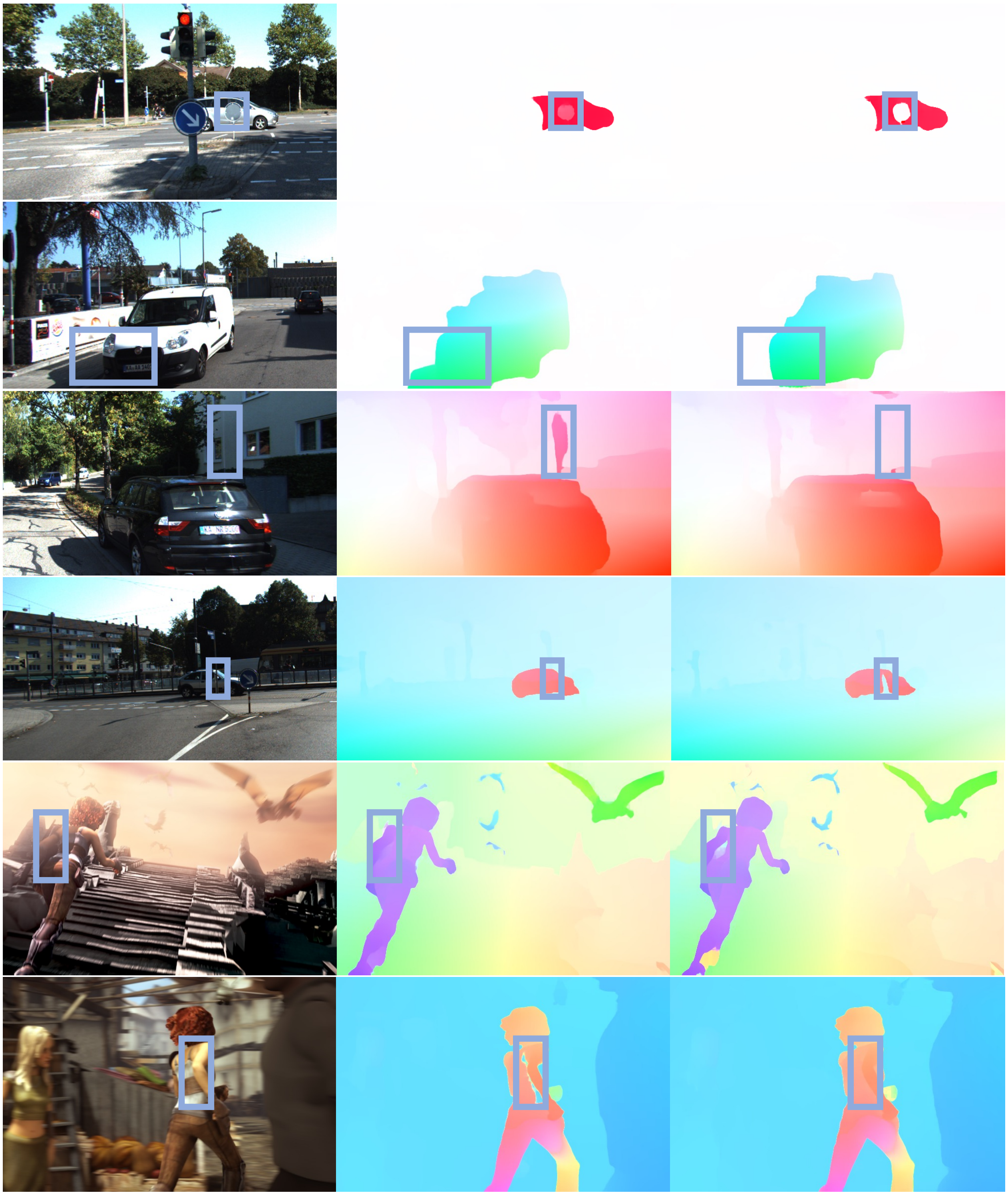
	EPE all	EPE matched	EPE unmatched	d0-10	d10-60	d60-140	s0-10	s10-40	s40+	
GroundTruth [1]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Visualize Results
VideoFlow-5-frames [2]	1.649	0.788	8.660	2.090	0.609	0.334	0.403	1.243	8.804	Visualize Results
VideoFlow-3-frames [3]	1.838	0.860	9.807	2.092	0.652	0.385	0.421	1.290	10.189	Visualize Results
FlowFormer++ [4]	1.943	0.878	10.627	2.302	0.720	0.384	0.438	1.404	10.712	Visualize Results
DiffusionFlow [5]	2.008	0.936	10.748	2.403	0.752	0.412	0.451	1.458	11.060	Visualize Results

Figure 3. **Screenshot of Sintel (Final) leaderboard.**

[3] Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. A simple baseline for video restoration with grouped spatial-temporal shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9822–9832, June 2023. 1

[4] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. *arXiv preprint arXiv:2303.01237*, 2023. 1

[5] Shangkun Sun, Yuanqi Chen, Yu Zhu, Guodong Guo, and Ge Li. Skflow: Learning optical flow with super kernels. *arXiv preprint arXiv:2205.14623*, 2022. 1



Input FlowFormer++ VideoFlow (ours)

Figure 4. **VideoFlow can better distinguish foreground and background objects and is more robust to noise such as light reflection.**