MemFlow: Optical Flow Estimation and Prediction with Memory

Supplementary Material

1. Details of Future Prediction

MemFlow-P. We present an overview of our Memory module for future Prediction of optical Flow (MemFlow-P) as in Fig. 1. Specifically, given current frame $\mathbf{I}_t$, we should first calculate the 2D motion feature f_m with previous frame $\mathbf{I}_{t-1}$. We are now able to update the memory buffer with f_m and the context feature $\mathcal{C}_\theta(\mathbf{I}_{t-1})$ from $\mathbf{I}_{t-1}$. Then we extract the context feature $\mathcal{C}_\theta(\mathbf{I}_t)$ from the current frame $\mathbf{I}_t$, which also serves as a query and reads out the aggregated motion feature f_{am} from the memory buffer. Besides, we also forward warp the previous flow $f_{t-1\to t}$ as a base f_p for flow prediction. Finally, we concatenate the aggregated motion feature from history, context feature from the current frame, and forward warped flow f_p for flow prediction with a simple CNN: $f = \mathrm{Convs}(f_c, f_{am}, f_p)$. Our CNN has similar convolutional layers as the original GRU. It consists of two SKBlocks as introduced by SKFlow [10]. Each SKBlock consists of two Feed Forward Networks (FFN), two depth-wise convolutional layers, and one point-wise convolutional layer. The total parameter of our MemFlow-P is 5.1 M. Our loss function is the l_1 distance between our predicted flow and the groundtruth:

$$\mathcal{L} = ||f_{gt} - f||_1. \tag{1}$$

MemFlow-P for Video Prediction. As shown in Fig. 2, we first predict the optical flow $f_{t\to t+1}$ for the last video frame $\mathbf{I}_t$. Besides, we also estimate the monocular depth from DPT [8] for the last video frame. We then utilize the Softmax Splatting [7] for forward warping the last video frame. As shown in the right part of Fig. 2, we get the splatted frame and a disocclusion mask indicating the blank regions. We finally inpaint the disocclusion region with image inpainting method ZITS [5] and get the synthesised frame $\hat{\mathbf{I}}_{t+1}$.

2. Implementation Details

Network Details. Our **MemFlow** shares the same network architecture with SKFlow [10]. Specifically, our feature encoder and context encoder consist of 6 residual blocks, 2 at 1/2 resolution, 1/4 resolution, and 1/8 resolution, respectively. Besides, our motion encoder and GRU are based on 6 and 2 SKBlocks as in SKFlow [10], respectively. And our **MemFlow-P** only replaces the GRU with a small CNN as illustrated in Sec. 1. As for our **MemFlow-T**, we utilize the first two stages of ImageNet-pretrained Twins-SVT [1] as our feature and context encoder.

Training Details. During training, we employ FlashAttention-2 [3, 4] for faster memory read-out.

Training Schedule. We first pre-train our networks with 2-frame in FlyingChair and FlyingThings3D for 120k (batch size 8) and 150k (batch size 6) iterations, respectively. Then, we train our networks with 3-frame and batch size 8 on the following datasets, for

- **MemFlow**, we train on FlyingThings3D for additional 600k iterations for generalization evaluation. Then, we finetune our model for 600k iterations on Sintel, KITTI, HD1K, and FlyingThings3D for Sintel submission. Finally, we finetune on KITTI for 40k and on Spring for 400k iterations, respectively.
- **MemFlow-T**, we train on FlyingThings3D for additional 600k iterations for generalization evaluation. Then, we finetune our model for 300k iterations on Sintel, KITTI, HD1K, and FlyingThings3D for Sintel submission. Finally, we finetune on KITTI for 40k iterations.
- **MemFlow-P**, we randomly initialized the newly added CNN. We then train MemFlow-P on FlyingThings3D for an additional 40k iterations for generalization evaluation. For the experiment of video prediction, we train our models on Sintel, KITTI, HD1K, and FlyingThings3D with 300k iterations.

Evaluation Protocol of Video Prediction. We evaluate the performance of video prediction on four sequences from the KITTI test set following previous works [6, 11]. The four sequences we employed are:

- "2011_09_26_drive_0060_sync",
- "2011_09_26_drive_0084_sync",
- "2011_09_26_drive_0093_sync", and
- "2011_09_26_drive_0096_sync".

Besides, as in prior works [6, 11], we use a context of T=4 past frames as input. All algorithms synthesize the next frame based on past frames.

3. More Qualitative Comparison

More qualitative results on Sintel training set and KITTI training set after pre-training on FlyingChair and FlyingThings3D are given in Figs. 3 and 4. We highlight the areas where our MemFlow(-T) achieves substantial improvements with bounding boxes, compared to previous state-of-the-art VideoFlow-MOF [9] and our baseline SKFlow [10]. Please zoom in for more details.

We also provide more qualitative results on the 1080p Spring test set as shown in Fig. 5. The qualitative results show superior cross-resolution generalization performance of our MemFlow, which is trained with the image resolution of 368x768.

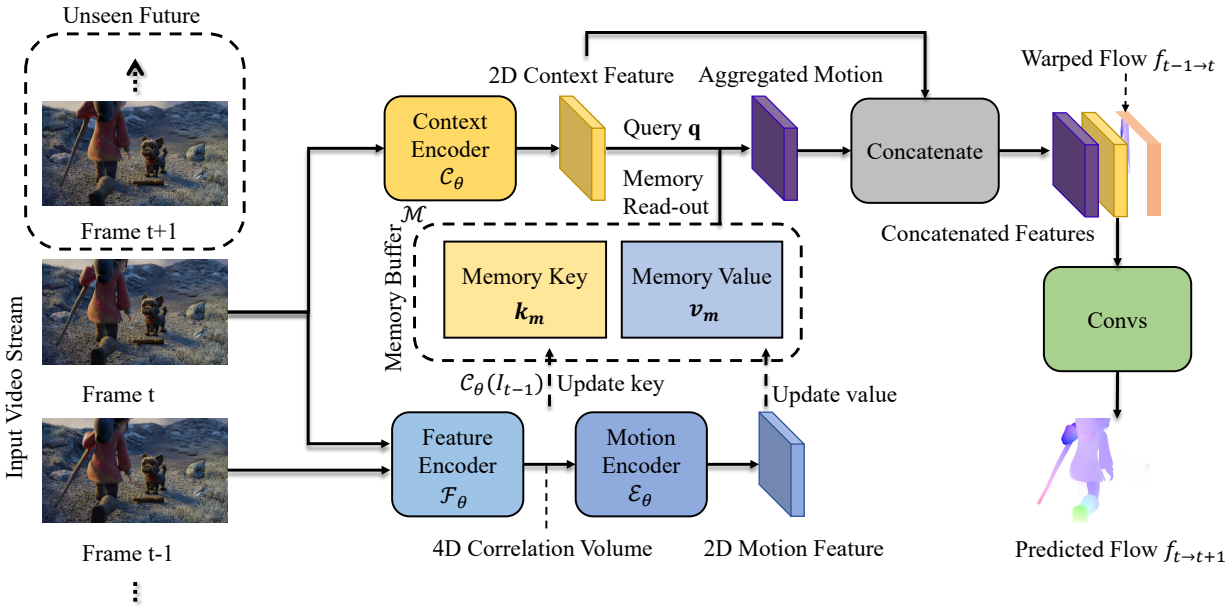

Figure 1. Overview of our MemFlow-P for future prediction of optical flow.

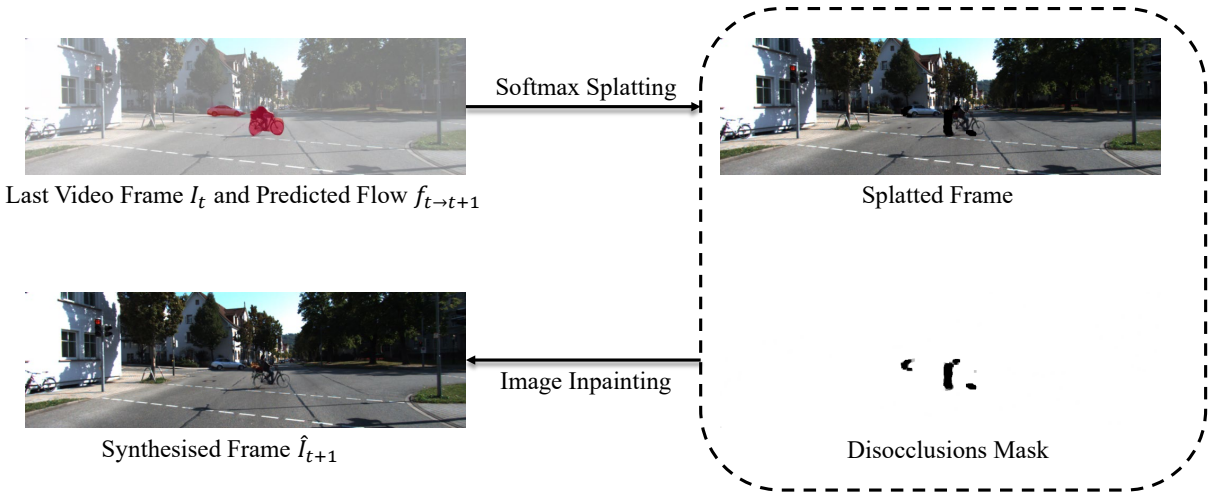

Figure 2. Overview of our MemFlow-P for video prediction.

4. More Results on Future Prediction of Optical Flow

Flow Prediction Results. We further show the full results of generalization performance evaluation for flow prediction in Tab. 1. MemFlow-P still outperforms other competitors in terms of the EPE on the clean pass of datasets and the Fl-all on KITTI-15 by a large margin, showing great dataset-specific and cross-dataset performance.

Ablation Studies. In this section, we report the ablation studies of flow prediction. First, we train a baseline model

Table 1. Generalization evaluation of flow prediction on FlyingThings3D, Sintel, and KITTI-15.

Method	Things		Sintel		KITTI-15	
	Clean	Final	Clean	Final	Fl-epe	Fl-all
Warped Oracle	14.76	14.76	5.76	5.76	-	-
MemFlow	15.55	15.70	5.92	6.23	12.95	54.48
OFNet [2]	13.73	13.76	5.78	6.03	12.43	59.17
MemFlow-P	**7.81**	**7.56**	**4.97**	**5.38**	**8.82**	**43.93**

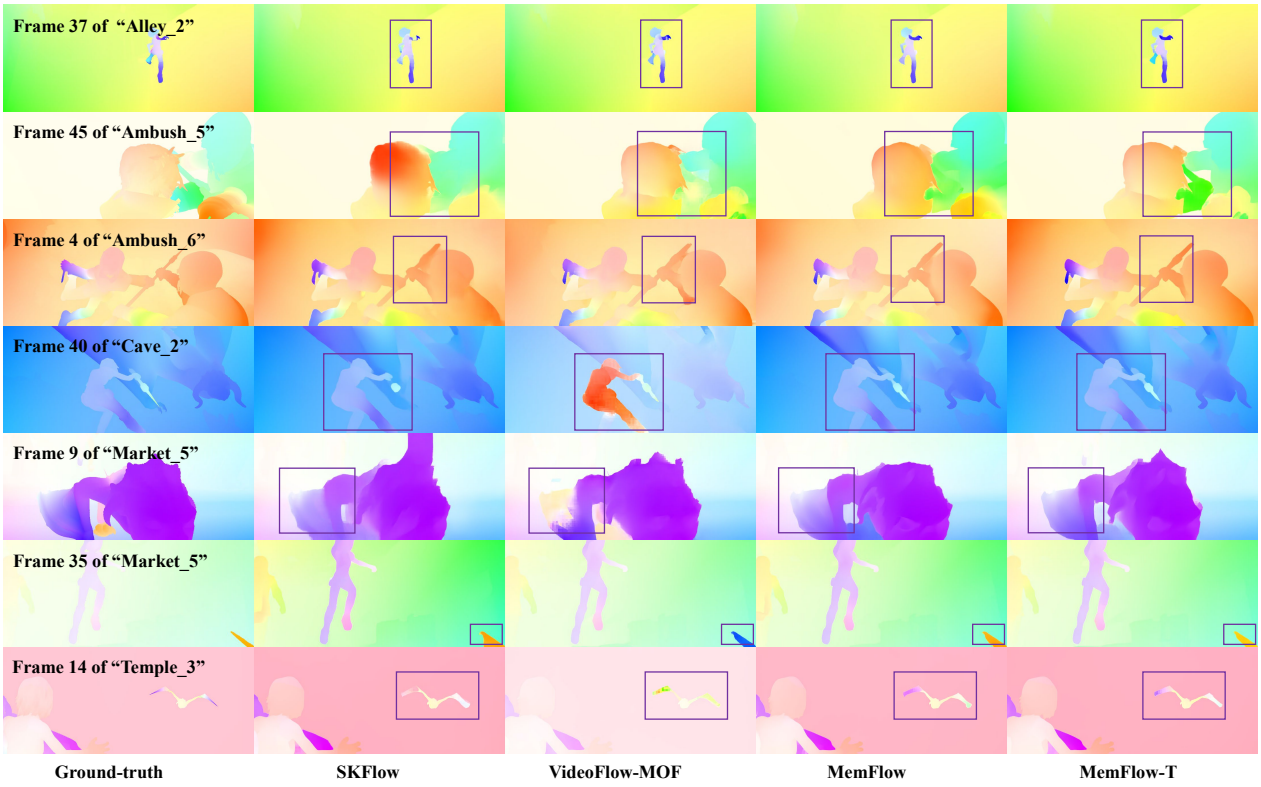

Figure 3. More qualitative results on Sintel training set final pass after pre-training on FlyingChair and FlyingThings3D. Bounding boxes mark the regions of substantial improvements. Please zoom in for details.

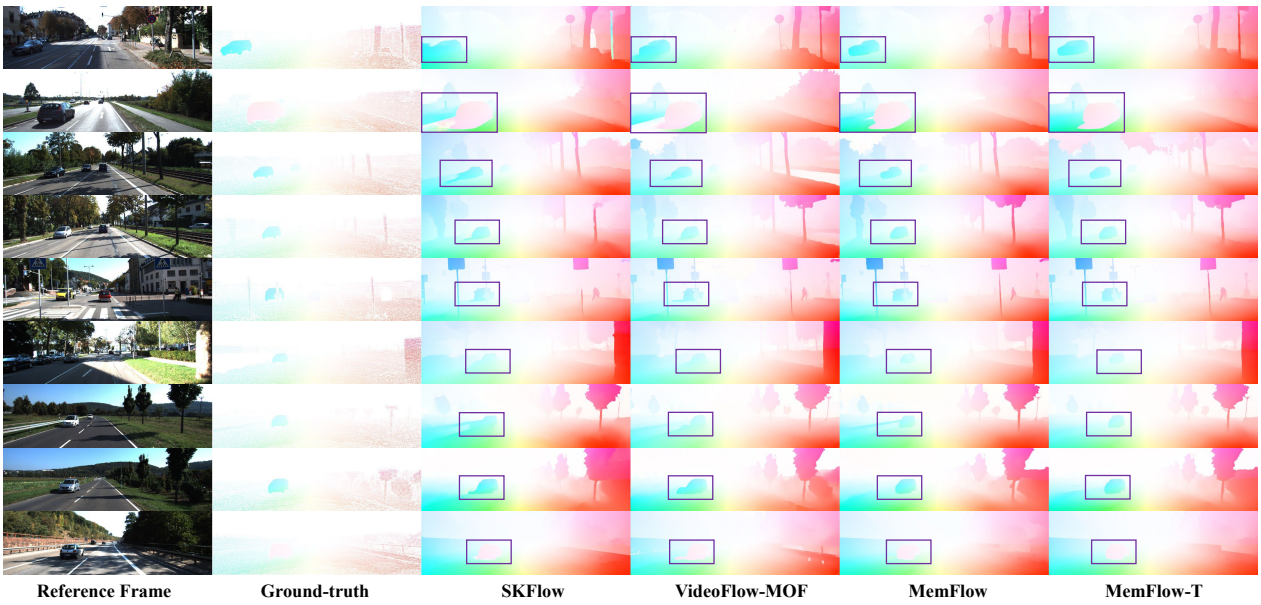

Figure 4. Qualitative results on KITTI training set after pre-training on FlyingChair and FlyingThings3D. Bounding boxes mark the regions of substantial improvements. Please zoom in for details.

for flow prediction without the forward warped past flow as input of CNN. The model is trained with 6-frame videos sampled from FlyingThings3D. As shown in Tab. 2, concatenating the forward warped flow can improve the cross-

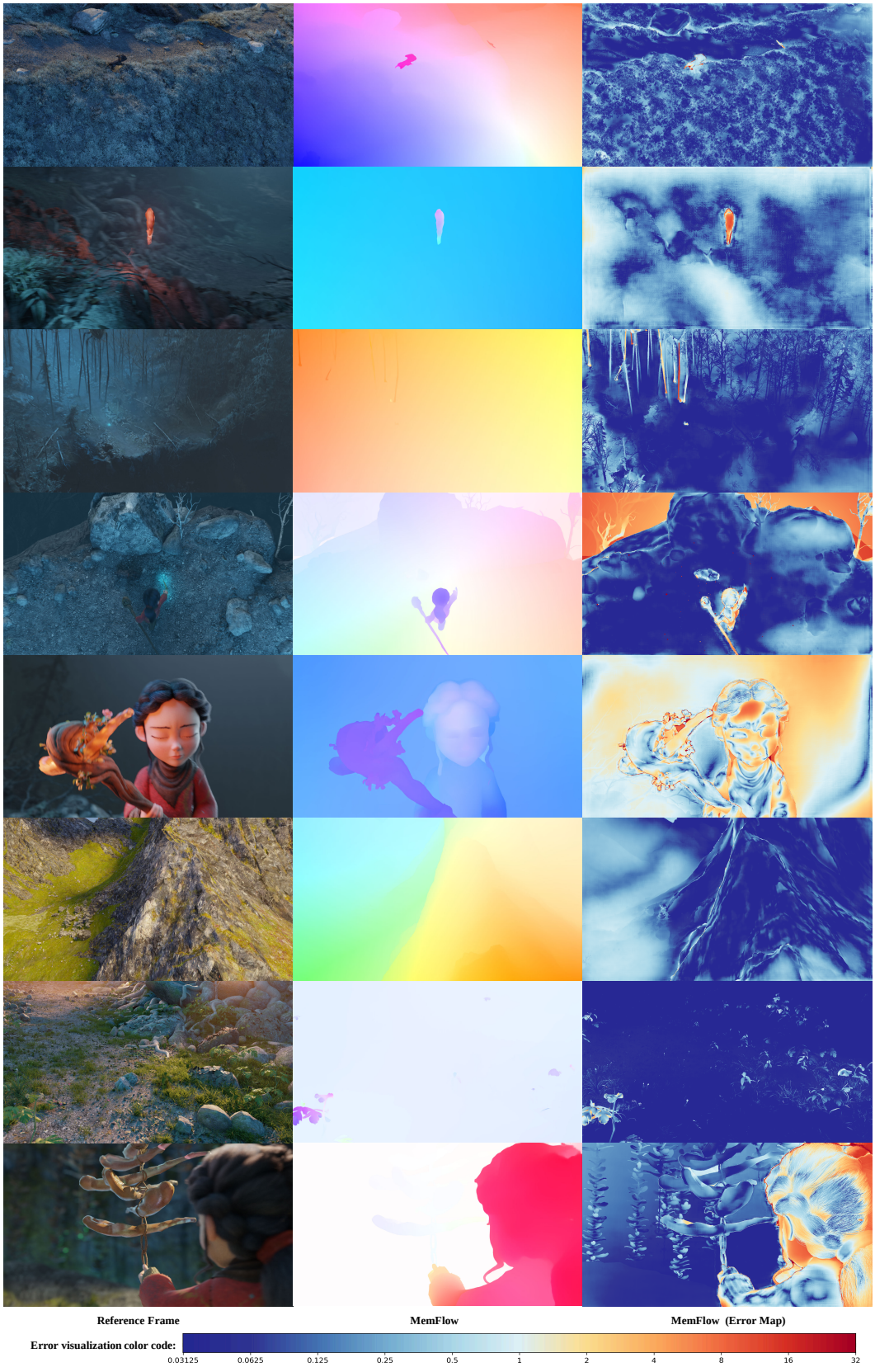

Figure 5. Qualitative results on 1080p Spring test set after finetuning on Spring. Error maps are downloaded from the official website. Please zoom in for details.

Table 2. Ablation studies on optical flow prediction.

Experiment	Things		Sintel		KITTI-15	
	Clean	Final	Clean	Final	Fl-epe	Fl-all
Baseline	**7.62**	**6.58**	5.25	5.79	8.84	56.63
+Forward Warped Flow	7.76	7.57	**4.96**	5.47	**8.57**	53.15
+Training with 3-frame	7.81	7.56	4.97	**5.38**	8.82	**43.93**

dataset generalization performance a lot, though with little worse results on the FlyingThings3D test split. Moreover, we find that training MemFlow-P with 3-frame videos can achieve similar results as the one trained with 6-frame. Therefore, we choose to train our MemFlow-P with 3-frame videos and forward warped flow.

Qualitative Results of Future Prediction by Optical Flow. We further provide several qualitative results of future prediction by optical flow as shown in Fig. 6. Our MemFlow-P can predict credible flow for the last video frame, and successfully synthesize the next frame.

Limitations of Long-term Future Prediction by Optical Flow. Our approach can generate nice results for short-term (one time step) future prediction as shown in Fig. 6. However, in the long term, as the predicted frame deviates from the distribution of training images, performance will drop quickly due to error accumulation like other video prediction methods. We further provide the quantitative and qualitative results of long-term future prediction in Figs. 7 and 8.

5. Screenshots of 1080p Spring, Sintel, and KITTI Results

We further provide anonymous screenshots of Spring, Sintel, and KITTI results on the test server as in Figs. 9 to 11. Our MemFlow ranks first on the 1080p Spring benchmark. The one without finetuning on Spring also performs well in terms of cross-dataset generalization performance. On Sintel, our MemFlow-T and MemFlow take the third and fourth places on the final pass, which improves the performance of SKFlow a lot. We also achieved great improvement on the KITTI benchmark compared to the baseline SKFlow.

References

[1] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *Advances in Neural Information Processing Systems*, pages 9355–9366. Curran Associates, Inc., 2021. 1

[2] Andrea Ciamarra, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Forecasting future instance segmentation with learned optical flow and warping. In *International Conference on Image Analysis and Processing*, pages 349–361. Springer, 2022. 2

[3] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023. 1

[4] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. 1

[5] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 1

[6] Daniel Geng, Max Hamilton, and Andrew Owens. Comparing correspondences: Video prediction with correspondence-wise losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3365–3376, 2022. 1

[7] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[8] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 1

[9] 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. 1

[10] 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

[11] Yue Wu, Qiang Wen, and Qifeng Chen. Optimizing video prediction via video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17814–17823, 2022. 1

(a) (b) (c)

Figure 6. Qualitative Results of Future Prediction by Optical Flow. (a) Predicted optical flow superimposed on the last video frame. (b) Synthesized video frame based on our predicted flow. (c) Groundtruth next frame.

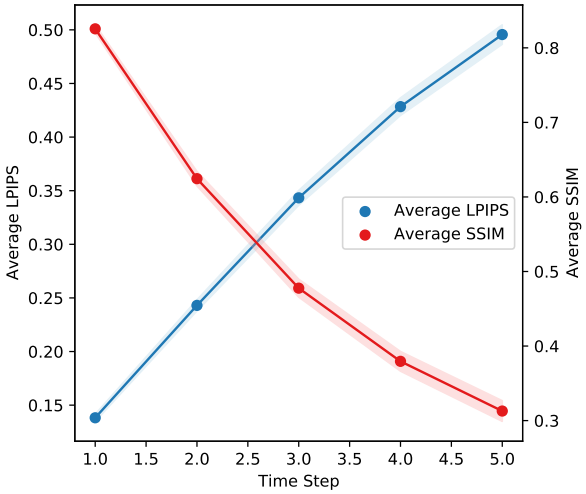

Figure 7. Quantitative results of long-term future prediction by optical flow. The plot shows the average LPIPS and SSIM-time step chart over KITTI test videos (256x832) and shadow is the 95% confidence interval. We calculate the metric with predicted frames for up to time step T + 5 from a context of T=4 past frames.

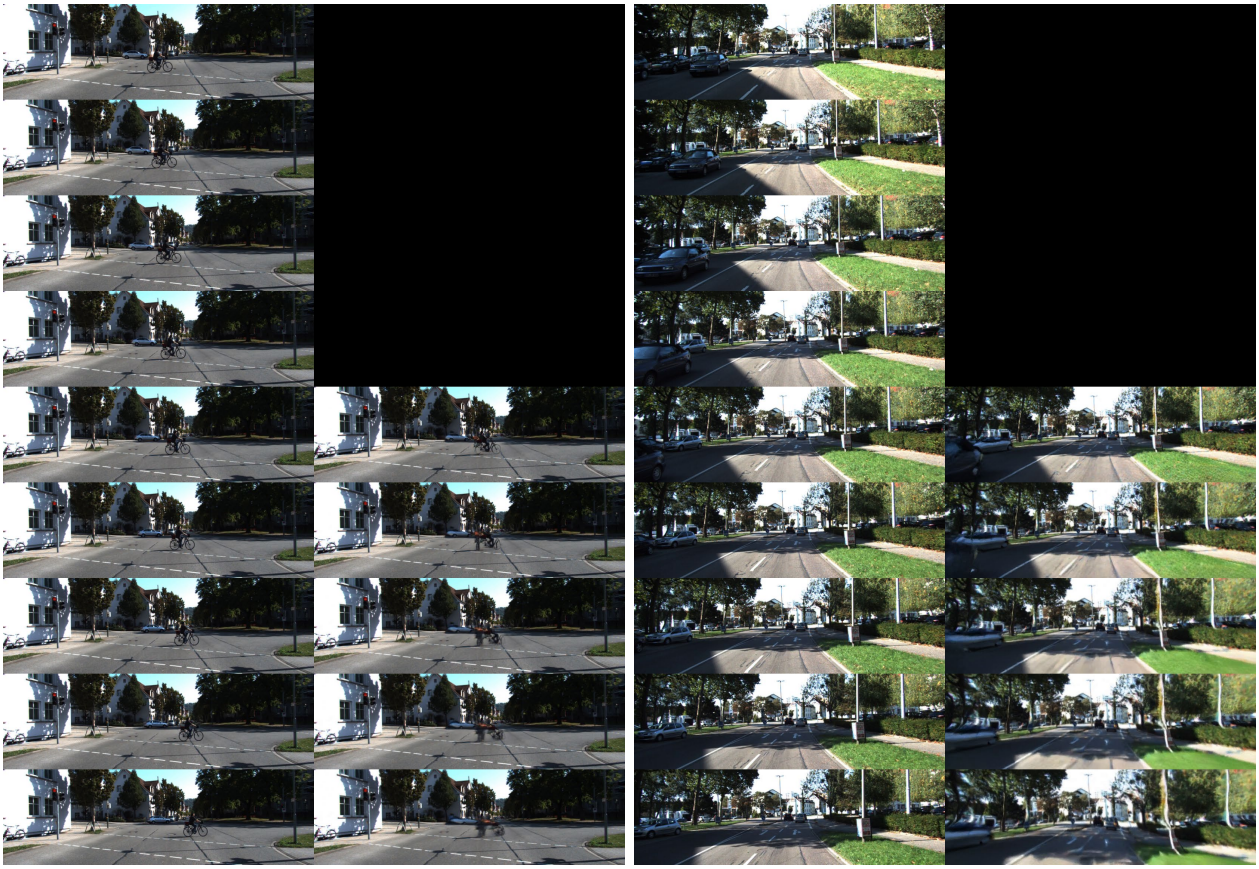

Figure 8. Our approach may fail to generate high-quality frames many steps into the future autoregressively due to error accumulation: Given 4 conditioning frames (top left), we show 5 predicted future frames in column 2 (bottom right) of two videos. Groundtruth frames are shown in the bottom left.

SPRING
Dataset & Benchmark

L. Mehl, J. Schmalfuss, A. Jahedi, Y. Nalivayko, A. Bruhn — University of Stuttgart

Download | Stereo | Optical Flow | Scene Flow | Submit | FAQ

Not logged in | Login | Create Account

	Name	1px ▲ total	1px low-det.	1px high-det.	1px matched	1px unmat.	1px rigid	1px non-rig.	1px not sky	1px sky	1px s0-10	1px s10-40	1px s40+	EPE	Fl	WAUC
1	MemFlow Anonymous.	4.482	4.119	61.703	3.742	35.115	2.391	20.306	3.934	12.809	1.305	4.437	31.184	0.471	1.416	93.855
2	XCAFlow Anonymous.	4.493	4.145	59.236	3.853	30.966	2.105	22.559	4.291	7.570	1.727	4.605	27.333	0.506	1.566	93.163
3	CroCo-Flow code CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. Weinzaepfel et al. ICCV 2023.	4.565	4.209	60.594	3.848	34.200	2.194	22.501	4.479	5.868	1.225	4.332	33.134	0.498	1.508	93.660
4	RPKNet Anonymous.	4.809	4.460	59.716	4.171	31.198	2.298	23.802	4.478	9.834	1.665	4.757	31.249	0.657	1.756	92.638
5	Win-Win Anonymous.	5.371	5.003	63.211	4.624	36.274	2.706	25.531	4.965	11.535	1.318	4.854	40.679	0.475	1.621	92.720
6	MS-RAFT+ code submitted by spring team | A. Jahedi, M. Luz, L. Mehl, M. Rivinius, and A. Bruhn. "High Resolution Multi-Scale RAFT ." In Robust Vision Challenge, 2022.	5.724	5.370	61.497	5.041	33.954	3.047	25.973	4.840	19.150	2.055	5.022	38.315	0.643	2.189	92.888
7	MemFlow(w/o ft) Anonymous.	5.759	5.394	63.348	5.107	32.755	3.293	24.422	4.494	24.990	2.918	4.820	32.071	0.627	2.114	92.253
8	FlowFormer code submitted by spring team | Z. Huang, X. Shi, C. Zhang, Q. Wang, K. C. Cheung, H. Qin, J. Dai, and H. Li. "FlowFormer: A Transformer Architecture for Optical Flow." In European Conference on Computer Vision (ECCV), 2022.	6.510	6.144	64.219	5.766	37.294	3.527	29.084	5.500	21.858	3.381	5.530	35.344	0.723	2.384	91.679
9	FlowNet2 code submitted by spring team | E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks." In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.	6.710	6.346	64.061	5.691	48.892	3.711	29.404	6.039	16.908	1.862	5.816	49.693	1.040	2.823	90.907
10	RAFT code submitted by spring team | Z. Teed, and J. Deng. "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow." In European Conference on Computer Vision (ECCV), 2020.	6.790	6.426	64.087	5.999	39.481	4.107	27.088	5.250	30.183	3.134	5.301	41.403	1.476	3.198	90.920

Figure 9. Screenshots for 1080p Spring optical flow evaluation on the official website.

Final Clean

	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-MOF [2]	1.649	0.788	8.660	2.090	0.609	0.334	0.403	1.243	8.804	Visualize Results
VideoFlow-BOF [3]	1.713	0.812	9.054	2.056	0.636	0.387	0.387	1.242	9.422	Visualize Results
MemFlow-T [4]	1.840	0.874	9.710	2.233	0.671	0.370	0.467	1.351	9.828	Visualize Results
MemFlow [5]	1.914	0.931	9.928	2.332	0.736	0.419	0.430	1.382	10.556	Visualize Results
FlowFormer++ [6]	1.943	0.878	10.627	2.302	0.720	0.384	0.438	1.404	10.712	Visualize Results

(a) Screenshot of Sintel Final results

Final Clean

	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-MOF [2]	0.991	0.397	5.832	1.028	0.317	0.218	0.229	0.694	5.484	Visualize Results
SAMFlow [3]	0.995	0.384	5.966	1.012	0.293	0.191	0.252	0.760	5.245	Visualize Results
VideoFlow-BOF [4]	1.005	0.389	6.023	1.029	0.310	0.189	0.229	0.695	5.605	Visualize Results
GMFlow+ [5]	1.028	0.335	6.680	0.868	0.264	0.183	0.227	0.689	5.826	Visualize Results
MemFlow [6]	1.046	0.426	6.091	1.169	0.308	0.206	0.253	0.778	5.623	Visualize Results
GMFlow_RVC [7]	1.055	0.420	6.227	1.084	0.326	0.227	0.302	0.754	5.513	Visualize Results
XCAFlow [8]	1.057	0.340	6.908	0.904	0.248	0.194	0.186	0.623	6.432	Visualize Results
TransFlow [9]	1.058	0.357	6.770	0.876	0.285	0.194	0.246	0.706	5.943	Visualize Results
CCMR+ [10]	1.067	0.311	7.235	0.832	0.262	0.143	0.148	0.560	6.864	Visualize Results
FlowFormer++ [11]	1.073	0.390	6.635	1.099	0.296	0.179	0.252	0.796	5.810	Visualize Results
NA [12]	1.077	0.398	6.610	1.142	0.297	0.179	0.250	0.792	5.865	Visualize Results
MemFlow-T [13]	1.081	0.430	6.384	1.171	0.351	0.184	0.246	0.750	6.024	Visualize Results

(b) Screenshot of Sintel Clean results

Figure 10. Screenshots for Sintel optical flow evaluation on the official website.

15	MemFlow-T			3.44 %	6.09 %	3.88 %	100.00 %			
16	RAFT-it+_RVC		code	3.62 %	5.33 %	3.90 %	100.00 %	0.14 s	1 core @ 2.5 Ghz (Python)	
D. Sun, C. Herrmann, F. Reda, M. Rubinstein, D. Fleet and W. Freeman: Disentangling Architecture and Training for Optical Flow. ECCV 2022.										
17	RRTC			3.77 %	4.70 %	3.93 %	100.00 %	0.3 s	1 core @ 2.5 Ghz (Python)	
18	RAFT-OCTC			3.72 %	5.39 %	4.00 %	100.00 %	0.2 s	GPU @ 2.5 Ghz (Python)	
J. Jeong, J. Lin, F. Porikli and N. Kwak: Imposing Consistency for Optical Flow Estimation (Qualcomm AI Research). CVPR 2022.										
19	RCA-Flow			3.67 %	6.25 %	4.10 %	100.00 %	0.16 s	1 core @ 2.5 Ghz (Python)	
20	MemFlow			3.67 %	6.27 %	4.10 %	100.00 %			

Figure 11. Screenshots for KITTI-15 optical flow evaluation on the official website.