MEMFOF: High-Resolution Training for Memory-Efficient Multi-Frame Optical Flow Estimation

Supplementary Material

7. Definitions

Here we will provide more detailed definitions used in the main text.

7.1. WAUC

In optical flow, weighted area under curve (WAUC), originally from VIPER [24], is formally defined as the integral

$$\frac{2}{5}\int_0^5 f(x)\cdot\frac{5-x}{5}\,dx, \tag{15}$$

where $f(x)$ is equal to the percentage of pixels where the flow error does not exceed x pixels. The metric ranges from 0 at worst to 100 at best.

7.2. Mixture-of-Laplace Loss

For a single flow vector coordinate, the Mixture-of-Laplace (MoL) in SEA-RAFT is defined as:

$$\begin{aligned}\text{MixLap}(\mu_{gt};\alpha,\beta,\mu) = -\log\Big[\frac{\alpha}{2}\cdot e^{-|\mu_{gt}-\mu|}+ \\ +\frac{1-\alpha}{2e^{\beta}}\cdot e^{-\frac{|\mu_{gt}-\mu|}{e^{\beta}}}\Big],\end{aligned} \tag{16}$$

where μ_{gt} is the target flow coordinate, μ is the predicted flow coordinate, α is the predicted mixing coefficient, and β is the predicted scale parameter. For a single optical flow frame prediction, the MoL loss is defined as:

$$\begin{aligned}\mathcal{L}_{MoL} = \frac{1}{2HW}\sum_{u,v}\sum_{d\in\{x,y\}}\text{MixLap}\big(\mu_{\text{gt}}(u,v)_d; \\ \alpha(u,v),\beta(u,v),\mu(u,v)_d\big).\end{aligned} \tag{17}$$

7.3. 2D Motion histogram

In order to visually demonstrate the discrepancy in motion magnitudes between common training datasets and Spring, we construct 2D histograms of motion vectors. Final results can be seen in Figure 4. The histograms are constructed in the following way:

$$\begin{aligned}H(u,v) = \sum_{n=1}^{N}\sum_{h=1}^{H}\sum_{w=1}^{W}[u \le f_n(h,w,0) \le u+1] \\ \cdot[v \le f_n(h,w,1) \le v+1],\end{aligned}$$

where $f_n \in \mathbb{R}^{H\times W\times 2}$ is the nth flow field from a dataset, (u,v) is the motion vector ($u \in [-H', H']$ and $v \in [-W', W']$) and $[\cdot]$ is the Iverson bracket. We set $H' = 1080$, $W' = 1920$, therefore our final histograms all have the same 2160×3840 resolution, for illustration purposes, we take the logarithm of bin counts. Maximum motion boundaries are derived as twice the size of images in the dataset, since the largest motion possible is to move diagonally from one corner of an image to the other one.

Table 7. Performance of our main model depending on the number of iterative refinements (N). Metrics are calculated on the Spring train dataset after the TSKH stage. Speed (runtime) was measured on an Nvidia RTX 3090 GPU (24 GB).

N	1px ↓	EPE ↓	WAUC ↑	Fl ↓	Speed, ms
0	6.170	0.893	90.898	2.625	71
1	3.752	0.397	94.731	1.212	172
2	3.300	0.350	95.322	0.979	215
4	3.133	**0.339**	95.565	0.863	299
6	3.081	$\underline{0.340}$	$\underline{95.603}$	0.835	385
8	3.061	0.341	**95.604**	0.823	472
10	$\underline{3.050}$	0.342	95.601	$\underline{0.820}$	557
12	**3.045**	0.342	95.598	**0.819**	642

Table 8. FullHD, method configurations taken from leaderboard sumbissions. Speed (runtime) was measured on an Nvidia RTX 3090 GPU (24 GB).

Method	Standard corr.		Alt. corr.	
	GB	ms	GB	ms
RAFT	7.97	557	1.32	1302
VideoFlow-BOF	17.74	1648	7.41	3275
MEMFOF	2.09	472	1.52	1235

8. Additional ablations

In this section, we provide ablations or ablation data not included in the main text.

8.1. Number of iterative refinements

We study our method's behavior depending on the number of iterative refinements. The results are provided in Table 7. For a balance between speed and accuracy, we choose to perform 8 iterative refinements.

Table 9. Full correlation volume and number of frames ablation table.

Corr. scale	#Frames	D_c	GMA	1px ↓				EPE ↓	WAUC ↑	Fl ↓	Memory, GB
				avg	s0-10	s10-40	s40+				
1/24	2	128	×	4.235	2.556	15.213	35.309	0.438	93.166	1.150	0.78
1/16	2	128	×	3.644	2.232	12.171	32.141	0.396	94.574	1.167	1.11
1/16	2	128	✓	3.547	2.132	12.101	32.025	0.408	94.617	1.035	1.29
1/16	2	256	×	3.420	2.072	11.440	30.941	0.372	94.761	1.018	1.12
1/16	2	512	×	3.375	2.047	11.201	30.614	0.350	95.130	0.888	1.30
1/24	3	512	✓	3.480	1.940	13.539	32.104	0.362	94.858	0.970	1.03
1/16	3	128	✓	3.560	2.154	12.176	31.543	0.380	94.859	1.094	1.78
1/16	3	256	✓	3.144	1.789	11.365	30.390	0.346	95.493	0.886	1.86
1/16	3	512	✓	**3.061**	**1.739**	11.156	**29.423**	0.341	95.604	**0.823**	2.09
1/16	3	512	×	3.151	1.833	**10.988**	30.165	**0.332**	**95.623**	0.896	1.82
1/24	5	512	✓	3.809	2.164	14.389	34.620	0.408	94.546	1.117	1.84

Table 10. Generalization performance of optical flow estimation on Sintel and KITTI-15 after the ”Things” stage. By default, all methods are trained on (FlyingChairs +) FlyingThings3D, additional datasets are listed in the ”Extra data” column.

Extra data	Method	Sintel (train)		KITTI-15 (train)	
		Clean ↓	Final ↓	Fl-epe ↓	Fl-all ↓
TartanAir	PWC-Net	2.55	3.93	10.4	33.7
	Flow1D	1.98	3.27	6.69	22.95
	MeFlow	1.49	2.75	5.31	16.65
	RAFT	1.43	2.71	5.04	17.40
	SEA-RAFT (S)	1.27	3.74	4.43	15.1
	SEA-RAFT (M)	1.21	4.04	4.29	14.2
	SEA-RAFT (L)	1.19	4.11	3.62	12.9
	MemFlow	0.93	2.08	3.88	13.7
	MemFlow-T	**0.85**	**2.06**	3.38	12.8
	VideoFlow-BOF	1.03	2.19	3.96	15.3
	VideoFlow-MOF	1.18	2.56	3.89	14.2
	StreamFlow	0.87	2.11	3.85	12.6
	MEMFOF (ours)	1.10	2.70	3.31	10.08
TartanAir	MEMFOF (ours)	1.20	3.91	**2.93**	**9.93**

8.2. Alternative correlation implementation

We additionally provide memory consumption and speed measurements for RAFT, VideoFlow and our method in Tab. 8 when using alternative correlation volume implementation that trades compute time for memory efficiency.

8.3. Corr. volume resolution and number of frames

We provide the full version of Table 5 with additional metrics as Table 9.

9. Additional results

In this section, we provide some other results that are not included in the main text.

9.1. Additional zero-shot evaluation

Following previous works, we evaluate the zero-shot performance of our method after the ”Things” training stage on Sintel (train) and KITTI (train). The results are provided in Table 10. Our method has the best zero-shot evaluation on KITTI and outperforms SEA-RAFT (L) on Sintel when trained on the same datasets.

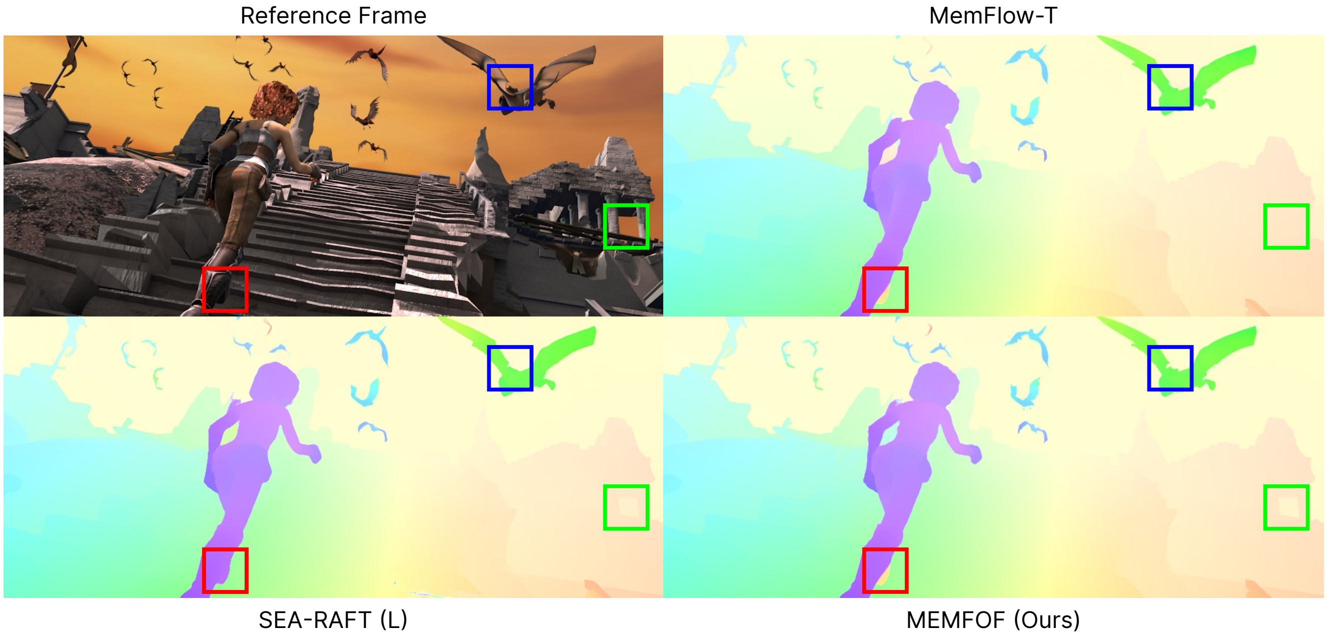

Figure 5. Qualitative comparison of MemFlow-T, SEA-RAFT (L), and our method on the Sintel benchmark. Sourced from official leaderboard submissions.

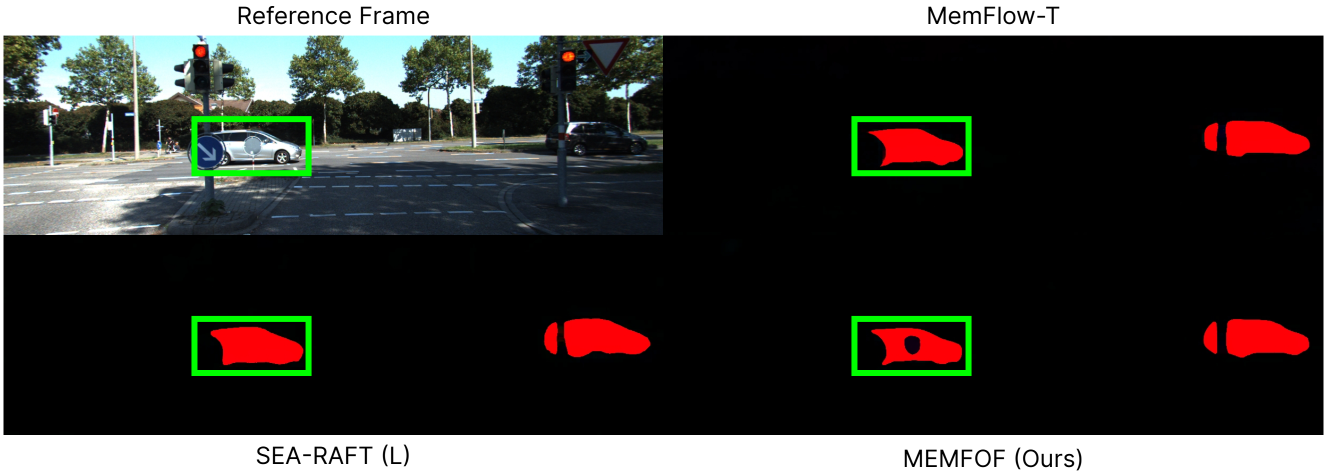

Figure 6. Qualitative comparison of MemFlow-T, SEA-RAFT (L), and our method on the KITTI-2015 benchmark. Sourced from official leaderboard submissions.

9.2. Qualitative comparison on Sintel and KITTI

We provide qualitative comparisons of our method on the Sintel and KITTI public benchmarks. As Figure 5 and Figure 6 show, our method has higher motion detail and coherence than our baseline or competitor.