GMFlow: Learning Optical Flow via Global Matching

Haofei Xu[1] Jing Zhang[2] Jianfei Cai[1] Hamid Rezatofighi[1] Dacheng Tao[3,2]

[1]Department of Data Science and AI, Monash University, Australia

[2]The University of Sydney, Australia [3]JD Explore Academy, China

CVPR 2022, Oral

Haofei Xu

Video vs. Image

- Video: an additional *temporal* dimension

YouTube-8M

ImageNet

Optical Flow

- Apparent motion between two video frames

image 1 & 2

optical flow (image 1 to 2)

$$(\Delta x, \Delta y) : (x, y) \rightarrow (x + \Delta x, y + \Delta y)$$

Optical Flow

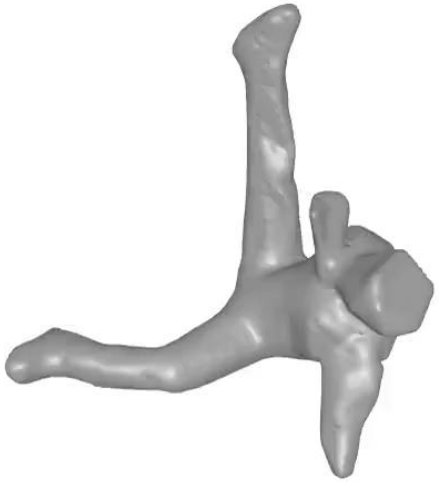
- Apparent motion between two video frames

image 1 & 2

optical flow (image 1 to 2)

What is Optical Flow for?

- 3D reconstruction from video

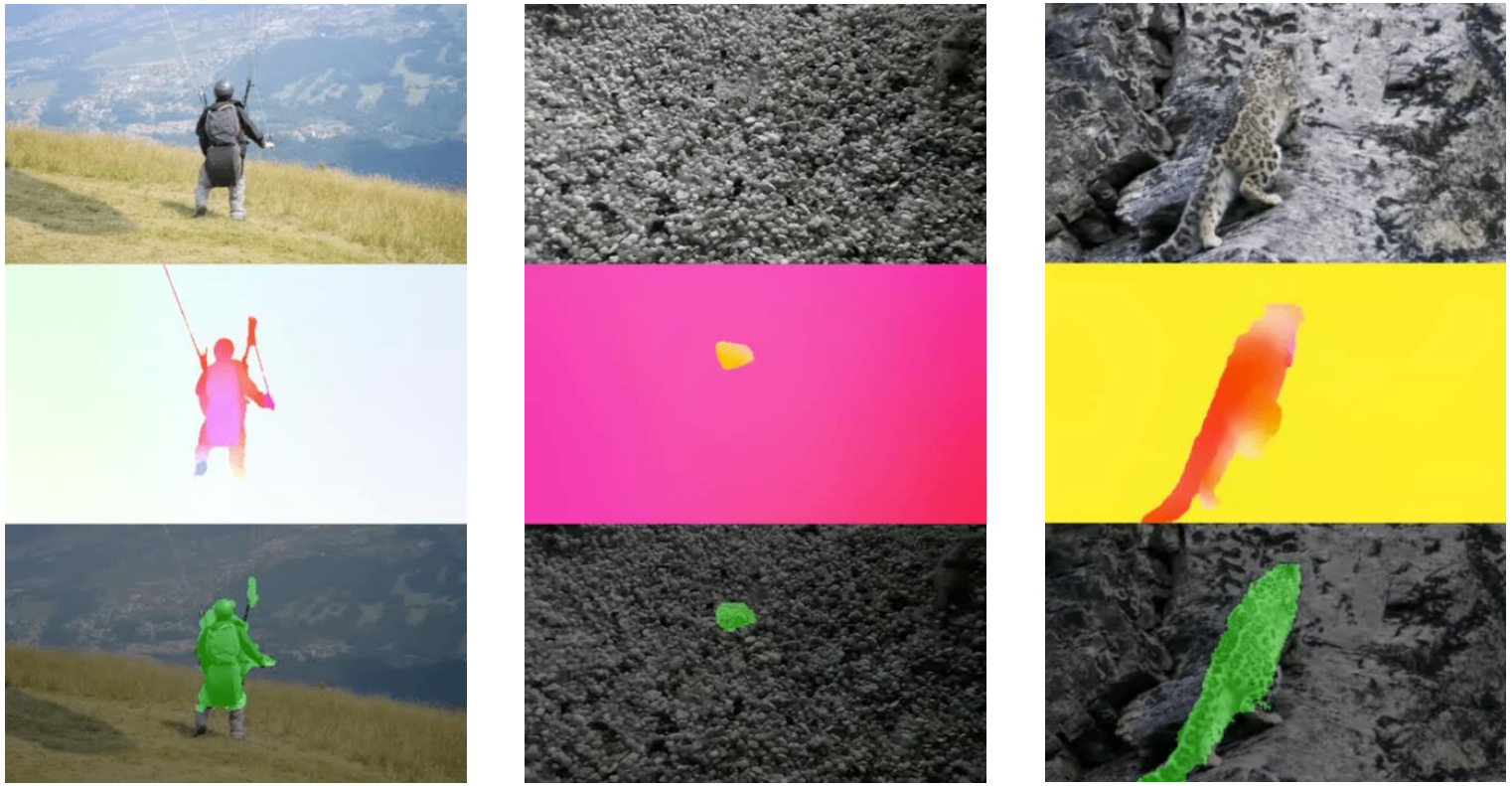


use optical flow as a constraint to solve the inverse problem

Yang et al. ViSER: Video-Specific Surface Embeddings for Articulated 3D Shape Reconstruction. NeurIPS 2021

What is Optical Flow for?

- Moving object segmentation

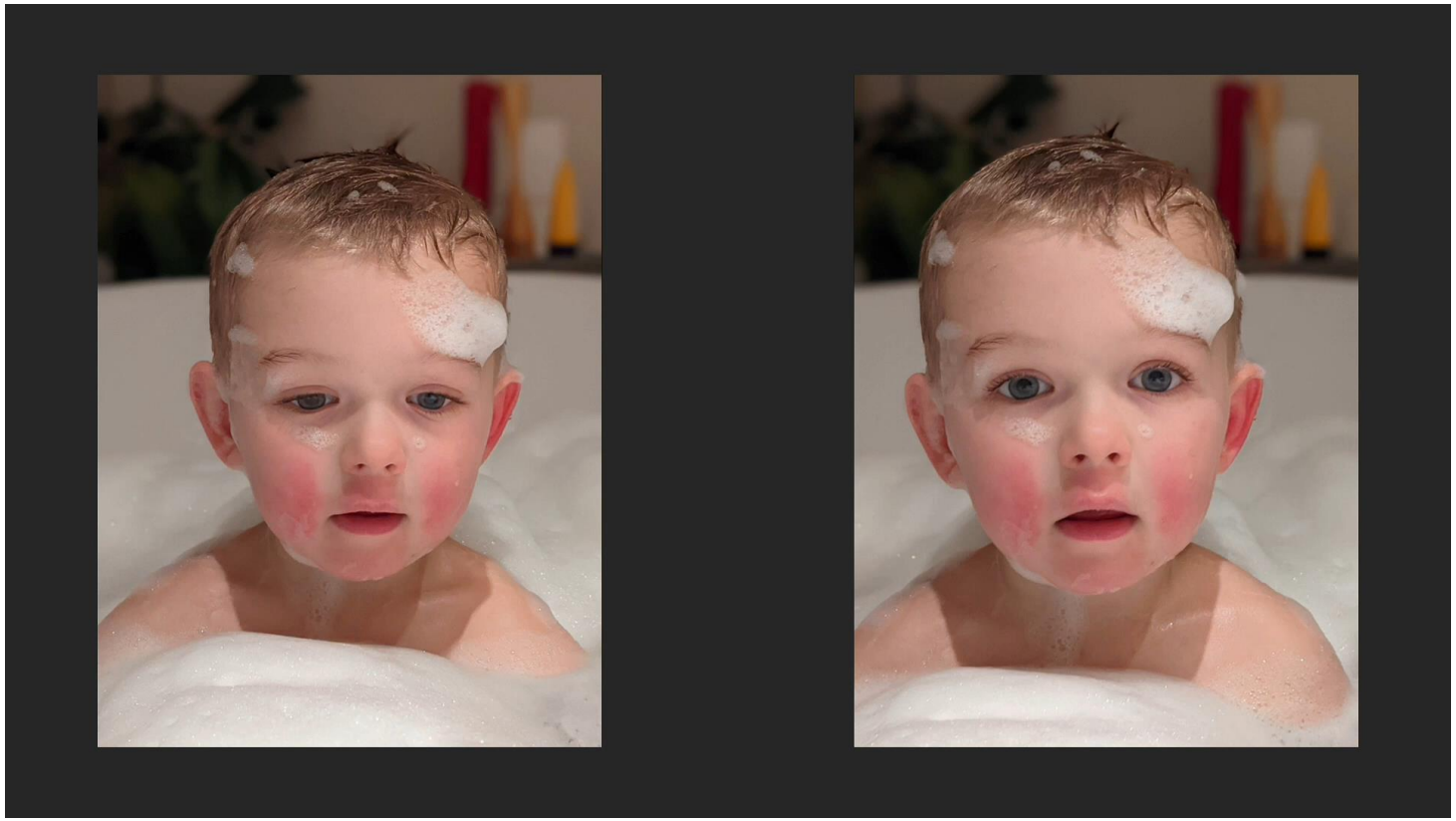


use optical flow as input

Yang et al. Self-supervised Video Object Segmentation by Motion Grouping. ICCV 2021

What is Optical Flow for?

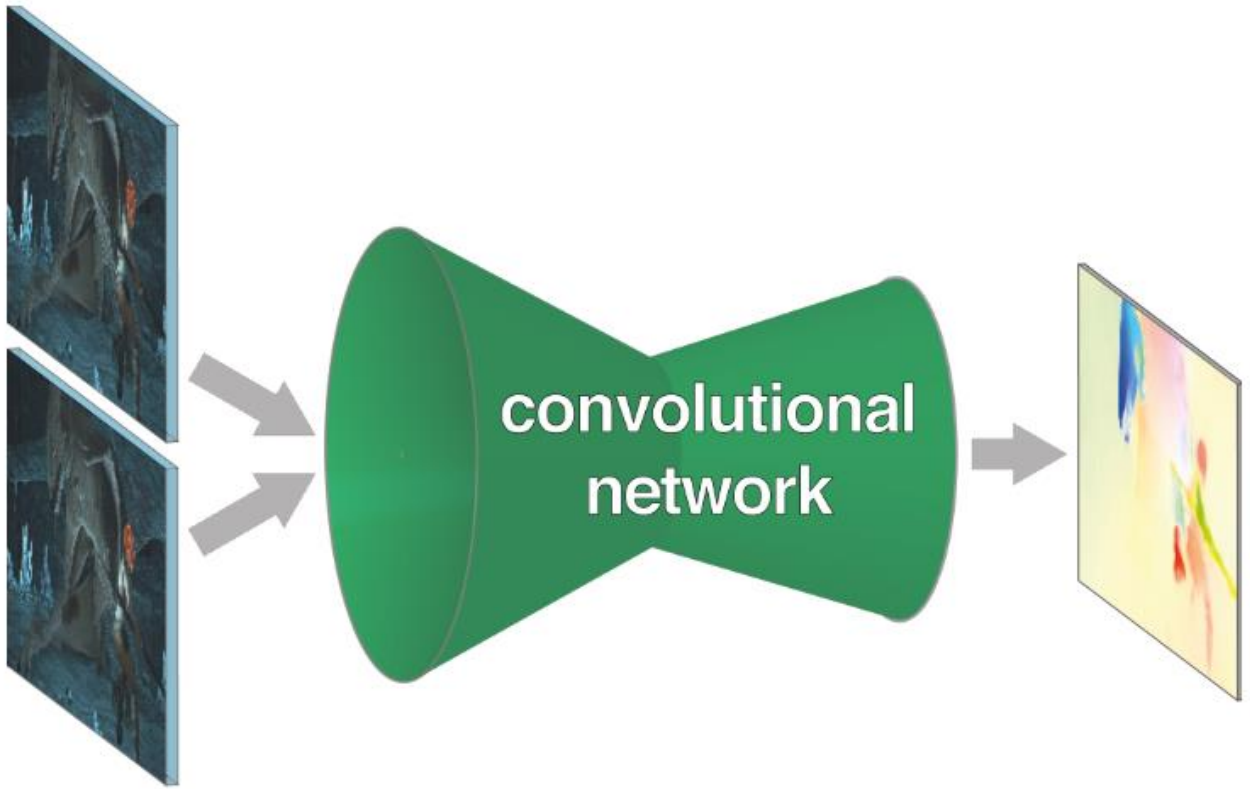
- Video frame interpolation



use optical flow as an intermediate component

Reda et al. FILM: Frame Interpolation for Large Motion. arXiv 2022

Learning Optical Flow

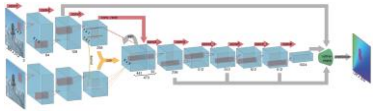


FlowNet

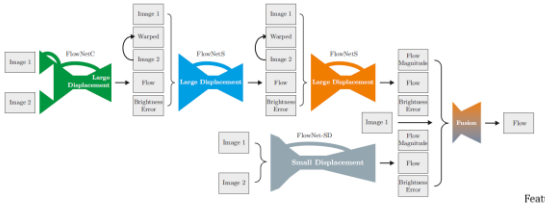
Dosovitskiy et al. FlowNet: Learning Optical Flow with Convolutional Networks. ICCV 2015

Milestones of Optical Flow

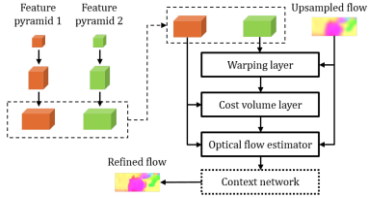
error



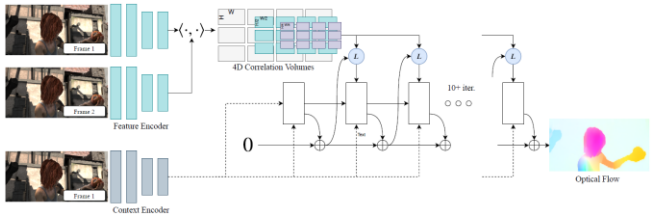
FlowNet



FlowNet 2



PWC-Net

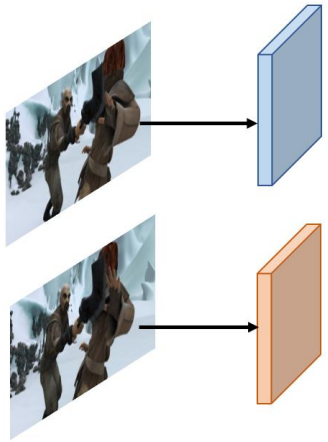


RAFT

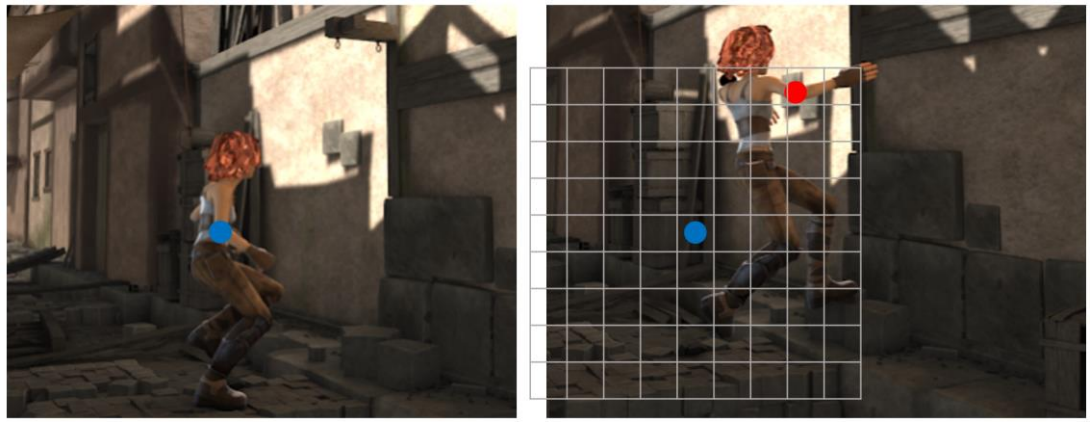
2015.04 2016.12 2017.09 2020.03

time

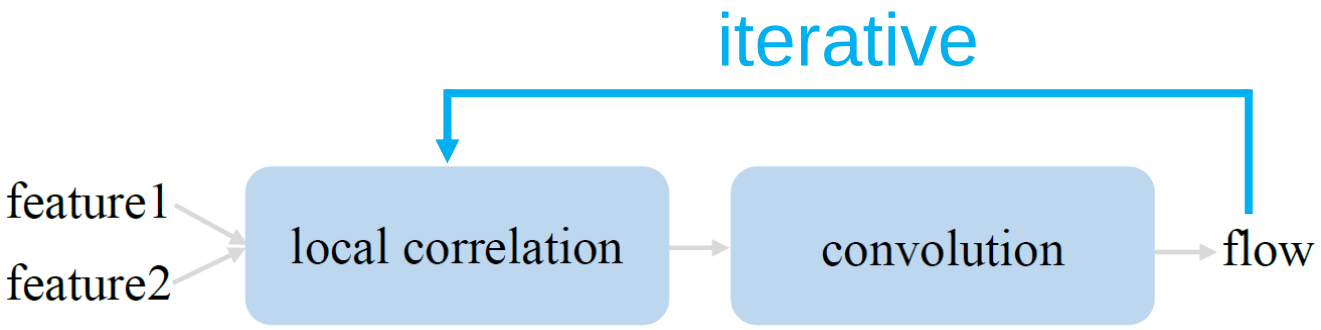
Typical Flow Estimation Approach

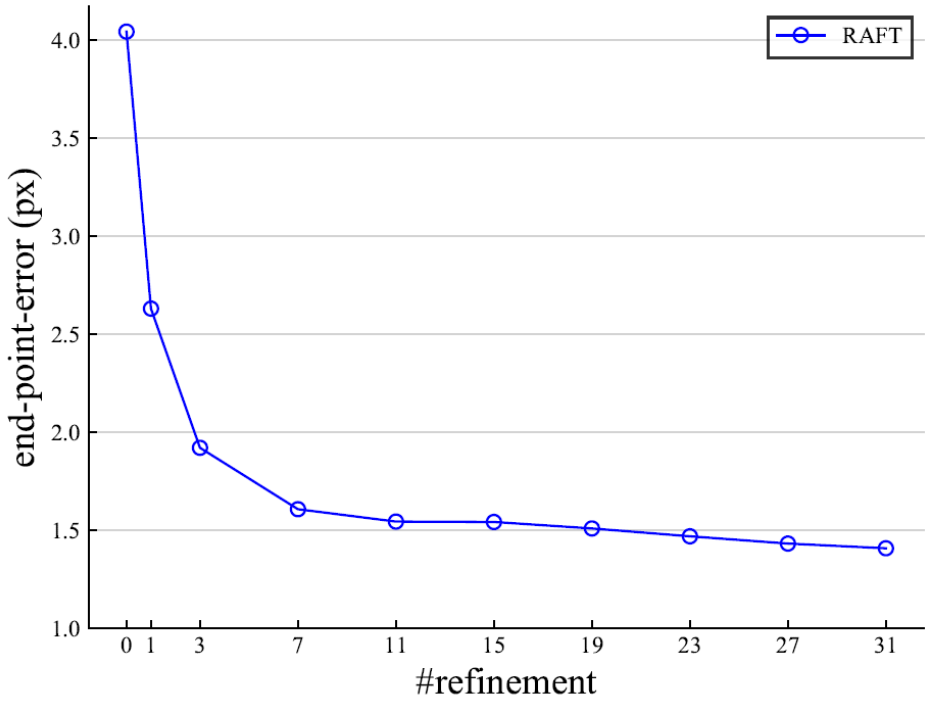


feature extraction



RAFT: Iterative Refinement





- Is it possible to achieve both high accuracy & efficiency *without* requiring such a large number of refinements?

Correspondence Search: Intuition

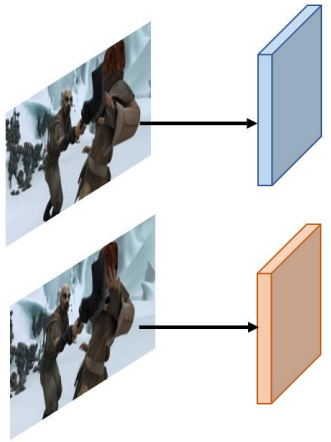
- Regression or matching?

Global Matching for Optical Flow

- Inputs: $\boldsymbol{I}_1$ $\boldsymbol{I}_2$
- Feature extraction: $\boldsymbol{F}_1, \boldsymbol{F}_2 \in \mathbb{R}^{H \times W \times D}$
- Global correlation: $\boldsymbol{C} = \frac{\boldsymbol{F}_1 \boldsymbol{F}_2^T}{\sqrt{D}} \in \mathbb{R}^{H \times W \times H \times W}$
- Softmax normalization: $\boldsymbol{M} = \mathrm{softmax}(\boldsymbol{C}) \in \mathbb{R}^{H \times W \times H \times W}$
- Correspondence: $\hat{\boldsymbol{G}} = \boldsymbol{M}\boldsymbol{G} \in \mathbb{R}^{H \times W \times 2}$
- Optical flow: $\boldsymbol{V} = \hat{\boldsymbol{G}} - \boldsymbol{G} \in \mathbb{R}^{H \times W \times 2}$

probability wrt $\boldsymbol{G} \in \mathbb{R}^{H \times W \times 2}$

Feature Enhancement



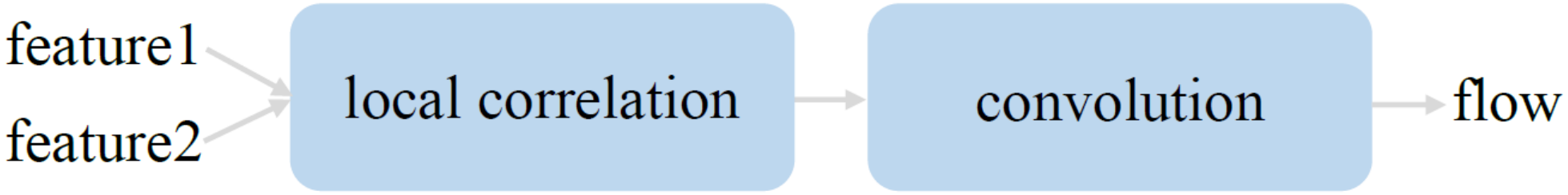
- Previous features are extracted *independently* from a CNN, without considering their *mutual dependencies* yet
- Goal: $f : (\boldsymbol{F}_1, \boldsymbol{F}_2) \rightarrow (\hat{\boldsymbol{F}}_1, \hat{\boldsymbol{F}}_2)$
- f : Transformer, stacked self-, cross-attentions and FFN

$$\hat{\boldsymbol{F}}_1 = \mathcal{T}(\boldsymbol{F}_1+\boldsymbol{P}, \boldsymbol{F}_2+\boldsymbol{P}), \quad \hat{\boldsymbol{F}}_2 = \mathcal{T}(\boldsymbol{F}_2+\boldsymbol{P}, \boldsymbol{F}_1+\boldsymbol{P})$$

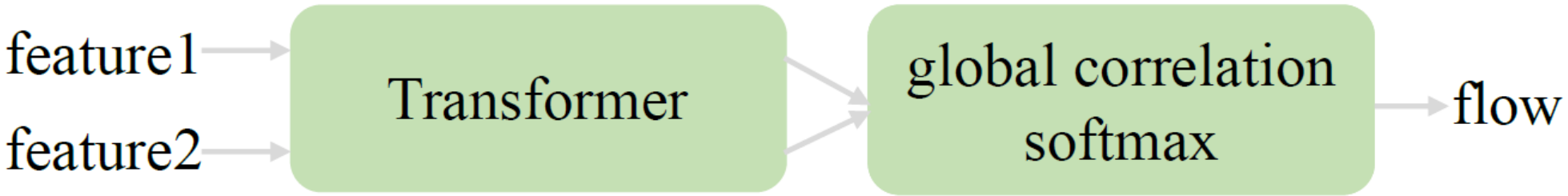
- Efficient implementation: shifted local window attention (Swin)

Methodology Comparison

- Local regression *vs.* Global matching



(a) previous flow estimation approach



(b) GMFlow

Methodology Comparison

average error flow magnitude

Method	#blocks	Things (val, clean)				Sintel (train, clean)				Sintel (train, final)				Param (M)
		EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}	
cost volume + conv	0													
	4													
	8													
	12													
	18													
Transformer + softmax	0													
	1													
	2													
	4													
	6													

Methodology Comparison

Method	#blocks	Things (val, clean)				Sintel (train, clean)				Sintel (train, final)				Param (M)
		EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}	
cost volume + conv	0	18.83	3.42	6.49	49.65	6.45	1.75	7.17	38.19	7.75	2.10	8.88	45.29	1.8
	4	10.99	1.70	3.41	29.78	3.32	0.73	3.84	20.58	4.93	0.99	5.71	31.16	4.6
	8	9.59	1.44	2.96	26.04	2.89	0.65	3.36	17.75	4.32	0.88	4.95	27.33	8.0
	12	9.04	1.37	2.84	24.46	2.78	0.65	3.32	16.69	4.07	0.84	4.76	25.44	11.5
	18	8.67	1.33	2.74	23.43	2.61	0.59	3.07	15.91	3.94	0.82	4.62	24.58	15.7
Transformer + softmax	0	22.93	8.57	11.13	52.07	8.44	2.71	11.60	42.10	10.28	3.11	13.83	53.34	1.0
	1	11.45	2.98	4.68	28.35	4.12	1.27	5.08	22.25	6.11	1.70	7.89	33.52	1.6
	2	8.59	1.80	3.28	21.99	3.09	0.90	3.66	17.37	4.54	1.24	5.44	26.00	2.1
	4	7.19	1.40	2.62	18.66	2.43	0.67	2.73	14.23	3.78	1.01	4.27	22.37	3.1
	6	**6.67**	**1.26**	**2.40**	**17.37**	**2.28**	**0.58**	**2.49**	**13.89**	**3.44**	**0.80**	**3.97**	**21.02**	4.2

- Our formulation outperforms previous method by a large margin, especially for large motion

Ablation: Transformer Components

setup	Things (val)	Sintel (train)		Param (M)
	clean	clean	final	
full	**6.67**	**2.28**	**3.44**	4.2
w/o cross	10.84	4.48	6.32	3.8
w/o pos	8.38	2.85	4.28	4.2
w/o FFN	8.71	3.10	4.43	1.8
w/o self	7.04	2.49	3.69	3.8

- Cross-attention contributes most

Ablation: Global *vs.* Local Matching

matching space	Things (val, clean)			
	EPE	s_{0-10}	s_{10-40}	s_{40+}
global	**6.67**	1.26	**2.40**	**17.37**
local 3×3	31.78	1.19	12.40	85.39
local 5×5	26.51	**0.89**	6.67	76.76
local 9×9	19.88	1.01	2.44	61.06

- Global matching is significantly better for large motion

When Matching Fails?

img0

img1

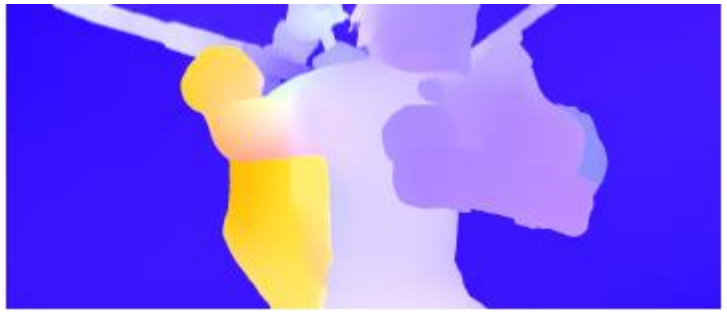
occlusion

out-of-boundary

Flow Propagation

img0

img1

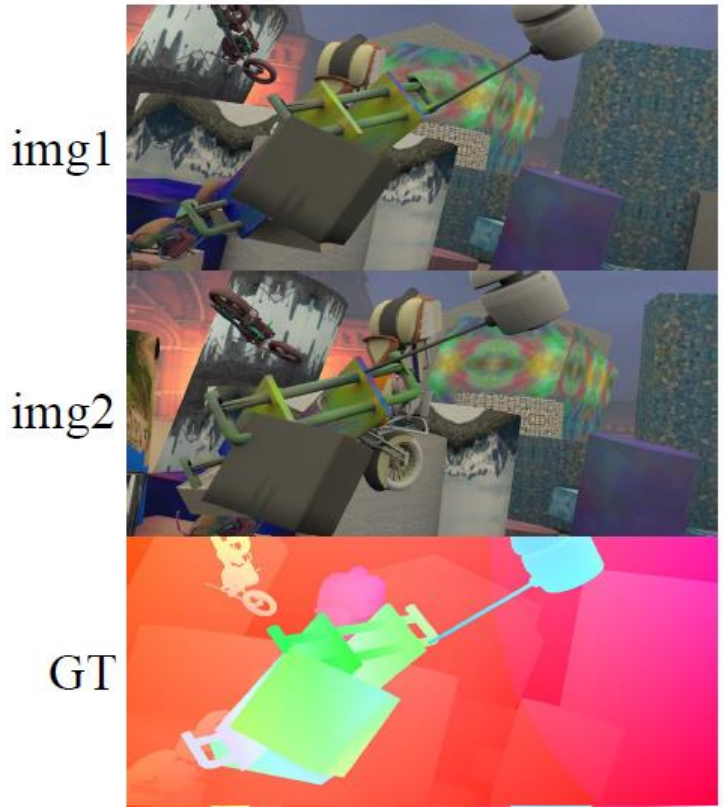


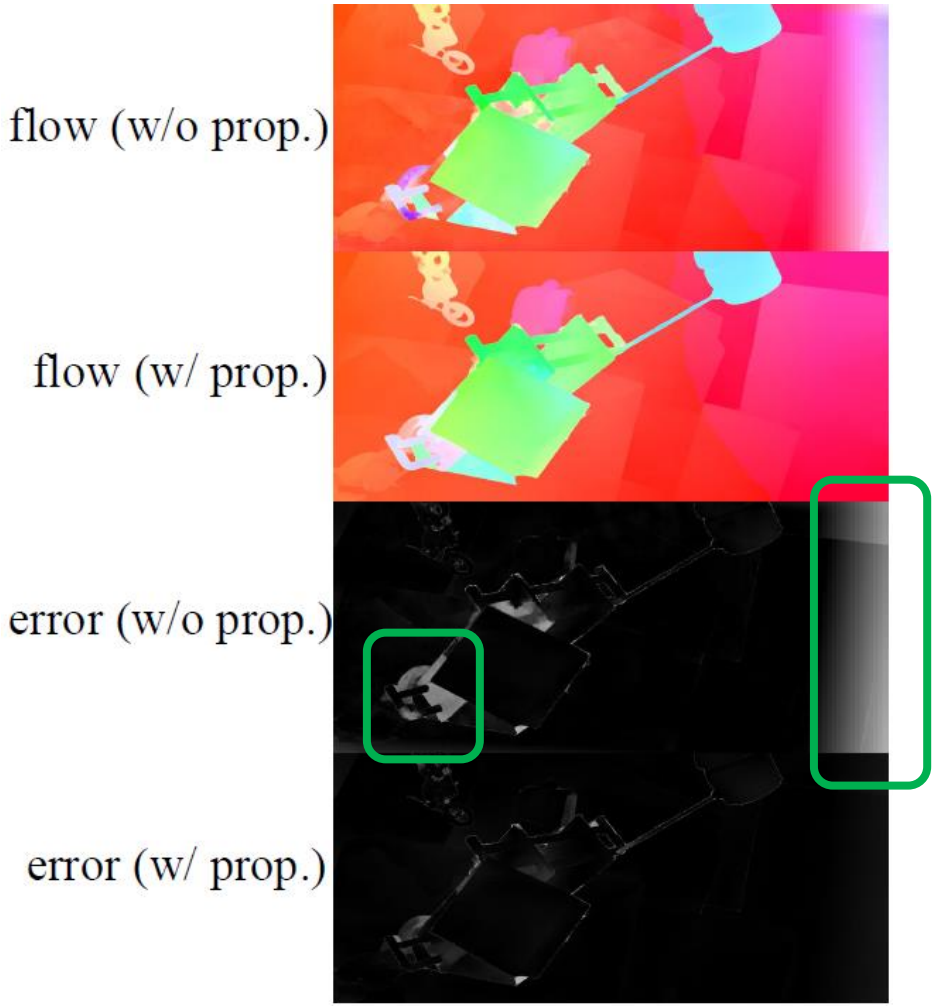
flow

- Observation: image and flow share structure similarity
- Use self-attention to propagate flow:

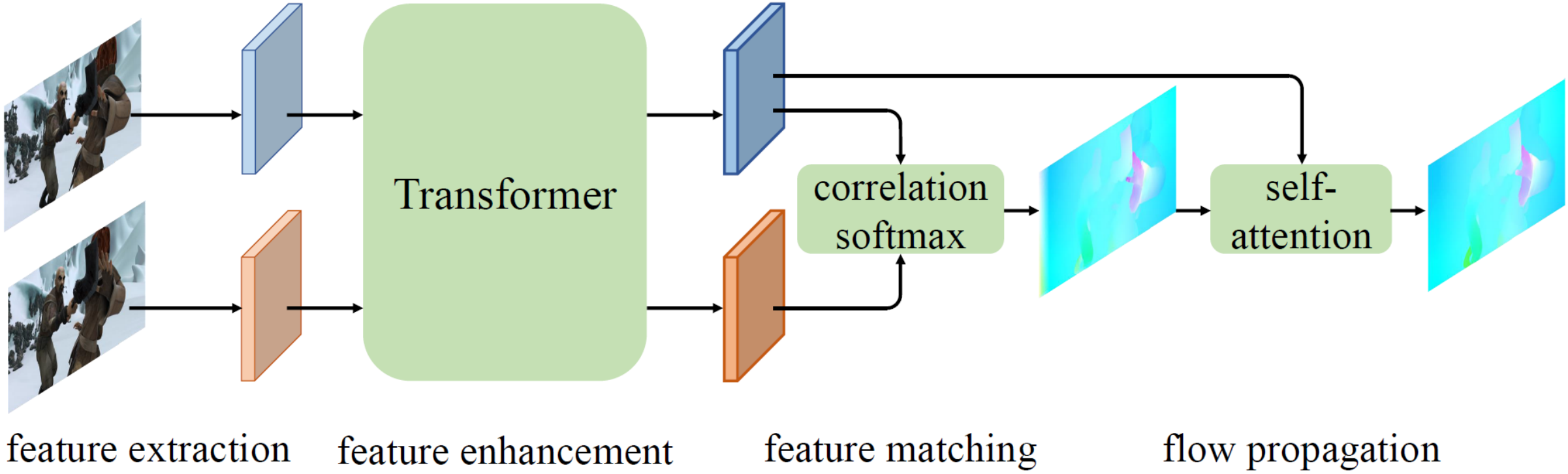
$$\tilde{\boldsymbol{V}} = \mathrm{softmax}\left(\frac{\hat{\boldsymbol{F}}_1 \hat{\boldsymbol{F}}_1^{\,T}}{\sqrt{D}}\right)\hat{\boldsymbol{V}} \in \mathbb{R}^{H\times W\times 2}$$

Flow Propagation



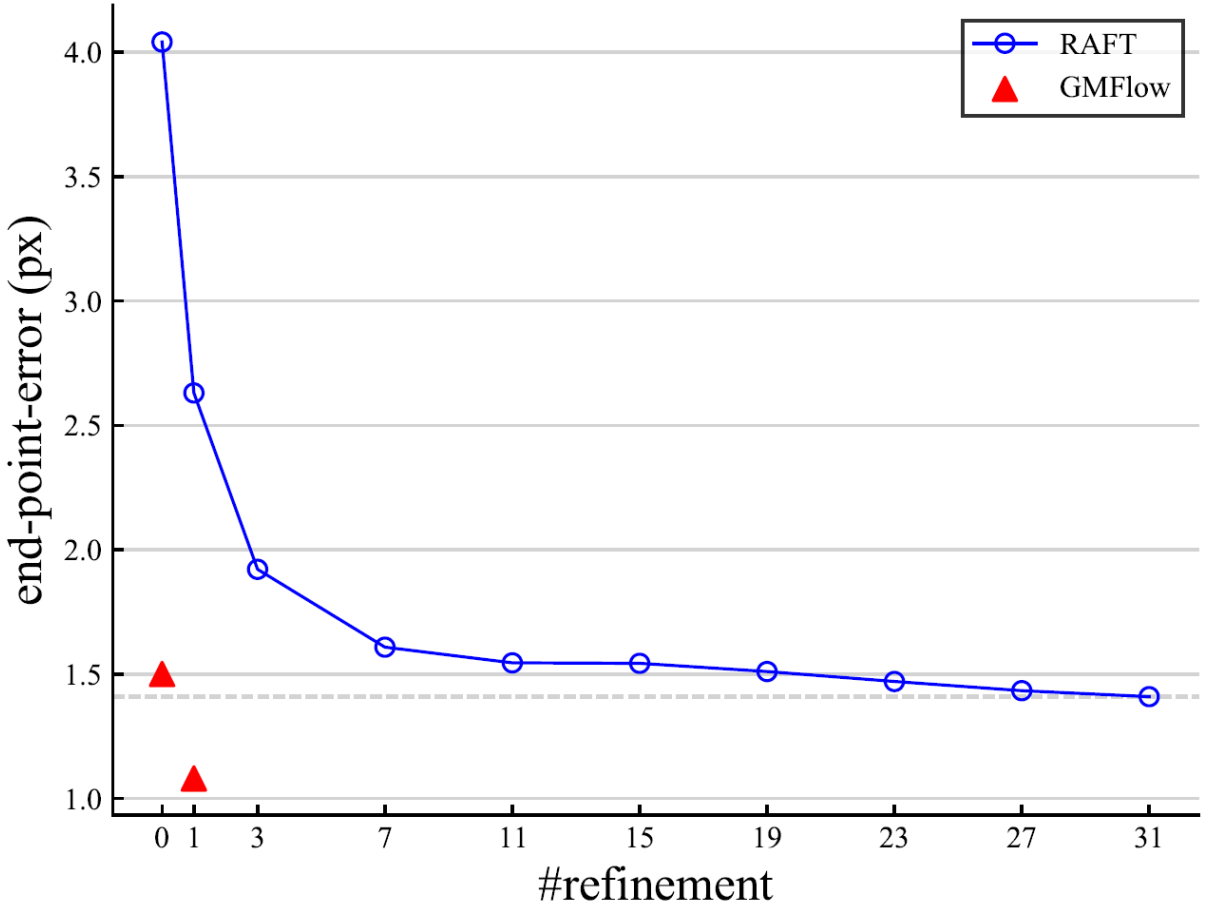


Framework



- The same framework can be used for an optional refinement at 1/4 resolution for residual flow prediction

Comparison with RAFT



- With only 1 refinement, GMFlow outperforms 31-refinements RAFT

Comparison with RAFT

Method	#refine.	Things (val, clean)				Sintel (train, clean)				Sintel (train, final)				Param (M)	Time (ms)
		EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}	EPE	s_{0-10}	s_{10-40}	s_{40+}		
RAFT [39]	0	14.28	1.47	3.62	40.48	4.04	0.77	4.30	26.66	5.45	0.99	6.30	35.19	5.3	25 (14)
	3	6.27	0.69	1.67	17.63	1.92	0.47	2.32	11.37	3.25	0.65	4.00	20.04		39 (21)
	7	4.66	0.55	1.38	12.87	1.61	0.39	1.90	9.61	2.80	0.53	3.30	17.76		58 (31)
	11	4.31	0.53	1.33	11.79	1.55	0.41	1.73	9.19	2.72	0.52	3.12	17.43		78 (41)
	23	4.22	0.53	1.32	11.52	1.47	0.36	1.63	9.00	2.69	0.52	3.05	17.28		133 (71)
	31	4.25	**0.53**	1.31	11.63	1.41	0.32	1.55	8.83	2.69	0.52	3.00	17.45		170 (91)
GMFlow	0	3.48	0.67	1.31	8.97	1.50	0.46	1.77	8.26	2.96	0.72	3.45	17.70	4.7	57 (26)
	1	**2.80**	**0.53**	**1.01**	**7.31**	**1.08**	**0.30**	**1.25**	**6.26**	**2.48**	**0.51**	**2.81**	**15.67**	4.7	151 (66)

V100(A100)

- GMFlow gains more speedup than RAFT (2.29x vs. 1.87x) on A100 since GMFlow doesn't require a large number of sequential computation

Results on Sintel

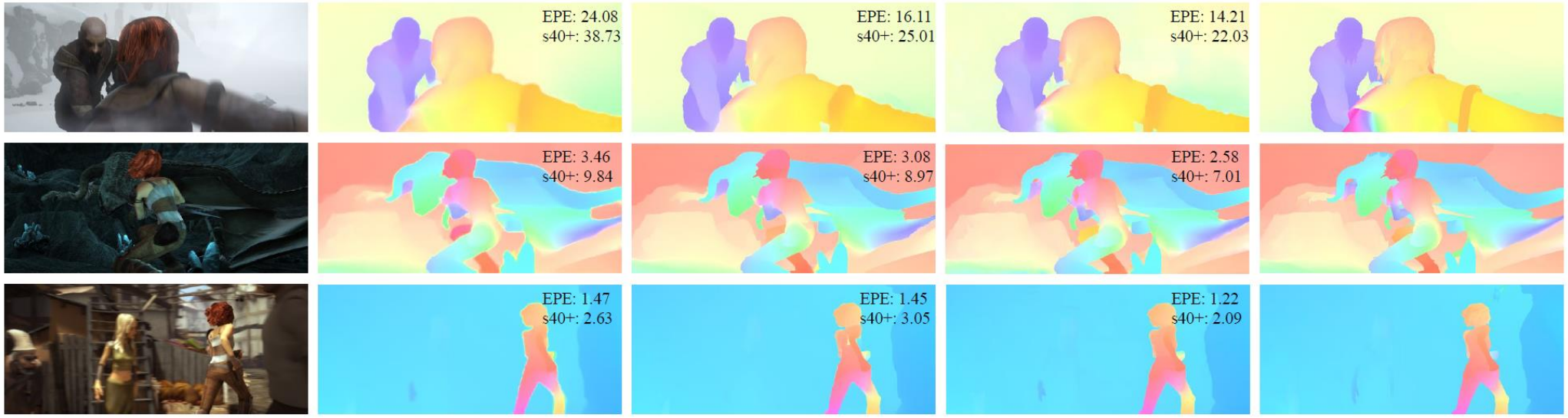


image DICL RAFT GMFlow GT

Method	Sintel (clean)			Sintel (final)		
	all	matched	unmatched	all	matched	unmatched
FlowNet2 [16]	4.16	1.56	25.40	5.74	2.75	30.11
PWC-Net+ [37]	3.45	1.41	20.12	4.60	2.25	23.70
HD^3 [50]	4.79	1.62	30.63	4.67	2.17	24.99
VCN [49]	2.81	1.11	16.68	4.40	2.22	22.24
DICL [42]	2.63	0.97	16.24	3.60	1.66	19.44
RAFT [39]	1.94	-	-	3.18	-	-
GMFlow	**1.74**	**0.65**	**10.56**	**2.90**	**1.32**	**15.80**

Ranking on Sintel

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
GMFlow+ [2]	1.152	0.377	7.472	0.933	0.296	0.193	0.247	0.720	6.689	Visualize Results
FlowFormer [3]	1.178	0.460	7.020	1.216	0.340	0.218	0.305	0.867	6.254	Visualize Results
ETA [4]	1.258	0.527	7.209	1.491	0.405	0.229	0.284	0.915	6.914	Visualize Results
MS_RAFT [5]	1.374	0.479	8.678	1.340	0.379	0.224	0.221	0.767	8.572	Visualize Results
GMA [6]	1.388	0.582	7.963	1.537	0.461	0.278	0.331	0.963	7.662	Visualize Results
GMFlowNet [7]	1.390	0.520	8.486	1.275	0.395	0.293	0.314	0.991	7.698	Visualize Results
RFPM [8]	1.411	0.494	8.884	1.335	0.400	0.221	0.273	0.879	8.345	Visualize Results
RAFT-OCTC [9]	1.419	0.541	8.574	1.455	0.442	0.242	0.301	0.940	8.118	Visualize Results
AGFlow [10]	1.431	0.559	8.541	1.501	0.452	0.261	0.319	0.963	8.075	Visualize Results
DIP [11]	1.435	0.519	8.919	1.102	0.407	0.312	0.336	0.754	8.546	Visualize Results
CRAFT [12]	1.441	0.611	8.204	1.574	0.552	0.249	0.311	0.991	8.131	Visualize Results
SeparableFlow [13]	1.496	0.567	9.075	1.474	0.481	0.257	0.309	0.958	8.691	Visualize Results
C1 [14]	1.520	0.593	9.084	1.545	0.498	0.285	0.307	1.008	8.781	Visualize Results
FCTR-m [15]	1.524	0.575	9.264	1.512	0.468	0.250	0.325	0.979	8.791	Visualize Results
RAFTwarm+AOIR [16]	1.544	0.551	9.656	1.515	0.412	0.280	0.279	0.941	9.290	Visualize Results
MFR [17]	1.545	0.593	9.295	1.536	0.477	0.299	0.348	1.023	8.736	Visualize Results
RAFT-it [18]	1.554	0.612	9.242	1.664	0.514	0.273	0.287	0.971	9.261	Visualize Results
SKFlow [19]	1.558	0.627	9.155	1.594	0.514	0.342	0.326	0.990	9.050	Visualize Results
SCAR [20]	1.579	0.608	9.498	1.613	0.499	0.285	0.314	1.018	9.210	Visualize Results
RAFTwarm+OBS [21]	1.593	0.600	9.692	1.532	0.507	0.309	0.300	0.989	9.470	Visualize Results
RAFTv2-OER-warm-start [22]	1.594	0.625	9.487	1.567	0.512	0.339	0.328	1.014	9.271	Visualize Results
submission5367 [23]	1.601	0.636	9.471	1.613	0.545	0.312	0.326	0.971	9.456	Visualize Results
RAFT [24]	1.609	0.623	9.647	1.621	0.518	0.301	0.341	1.036	9.288	Visualize 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
GMFlow+ [2]	2.311	1.015	12.874	2.034	0.771	0.627	0.409	1.329	14.123	Visualize Results
FlowFormer [3]	2.364	1.113	12.567	2.604	0.933	0.599	0.613	1.597	12.910	Visualize Results
ETA [4]	2.374	1.235	11.666	3.059	1.065	0.598	0.546	1.653	13.190	Visualize Results
CRAFT [5]	2.417	1.163	12.637	2.837	1.012	0.547	0.538	1.623	13.656	Visualize Results
AGFlow [6]	2.469	1.221	12.643	2.892	0.991	0.698	0.560	1.692	13.816	Visualize Results
GMA [7]	2.470	1.241	12.501	2.863	1.057	0.653	0.566	1.817	13.492	Visualize Results
SKFlow [8]	2.535	1.221	13.255	2.968	1.050	0.617	0.633	1.829	13.679	Visualize Results
RAFT-OCTC [9]	2.574	1.243	13.435	2.880	1.045	0.667	0.578	1.701	14.594	Visualize Results
GMFlowNet [10]	2.648	1.271	13.882	2.818	1.050	0.776	0.699	1.784	14.417	Visualize Results
AGF-Flow [11]	2.651	1.275	13.853	2.605	0.877	0.828	0.612	1.520	15.489	Visualize Results
MS_RAFT [12]	2.667	1.190	14.706	2.635	0.941	0.749	0.468	1.511	16.377	Visualize Results
SeparableFlow [13]	2.667	1.275	14.013	2.937	1.056	0.620	0.580	1.738	15.269	Visualize Results
ALNF [14]	2.679	1.304	13.880	2.636	0.903	0.857	0.645	1.551	15.486	Visualize Results
FCTR-m [15]	2.687	1.261	14.310	2.897	1.032	0.709	0.578	1.761	15.386	Visualize Results
RAFT+NCUP [16]	2.692	1.323	13.854	3.139	1.086	0.636	0.635	1.844	14.949	Visualize Results
submission5367 [17]	2.742	1.282	14.656	3.027	1.110	0.644	0.562	1.743	15.980	Visualize Results
L2L-Flow-ext-warm [18]	2.780	1.319	14.697	3.098	1.145	0.637	0.656	1.879	15.502	Visualize Results
LCT-Flow2 [19]	2.781	1.349	14.465	2.720	0.989	0.895	0.620	1.582	16.405	Visualize Results
MFR [20]	2.801	1.380	14.385	3.075	1.112	0.772	0.674	1.829	15.703	Visualize Results
RAFTwarm+AOIR [21]	2.813	1.371	14.565	3.088	1.099	0.727	0.603	1.781	16.271	Visualize Results
NASFlow [22]	2.822	1.403	14.389	2.998	1.146	0.910	0.655	1.757	16.143	Visualize Results
RAFTwarm+OBS [23]	2.826	1.356	14.809	3.134	1.116	0.735	0.631	1.832	16.117	Visualize Results

- A variant GMFlow+ ranks **first** on both Sintel (clean) and Sintel (final)

More Visual Results

Conclusion

- A new *global matching* formulation for optical flow
- A new GMFlow framework

code available!

github.com/haofeixu/gmflow

- Strong performance without requiring a large number of refinements
- Can be served as a simple & strong baseline for further development

Thank you!