Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Matching

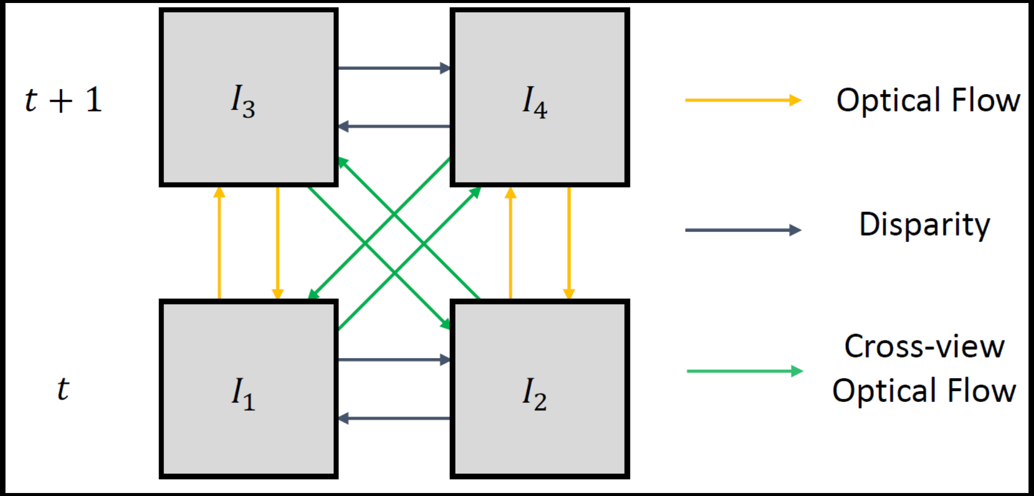
Pengpeng Liu, Irwin King, Michael Lyu, Jia Xu
The Chinese University of Hong Kong
Huya AI

We propose a unified method to jointly learn optical flow and stereo matching.

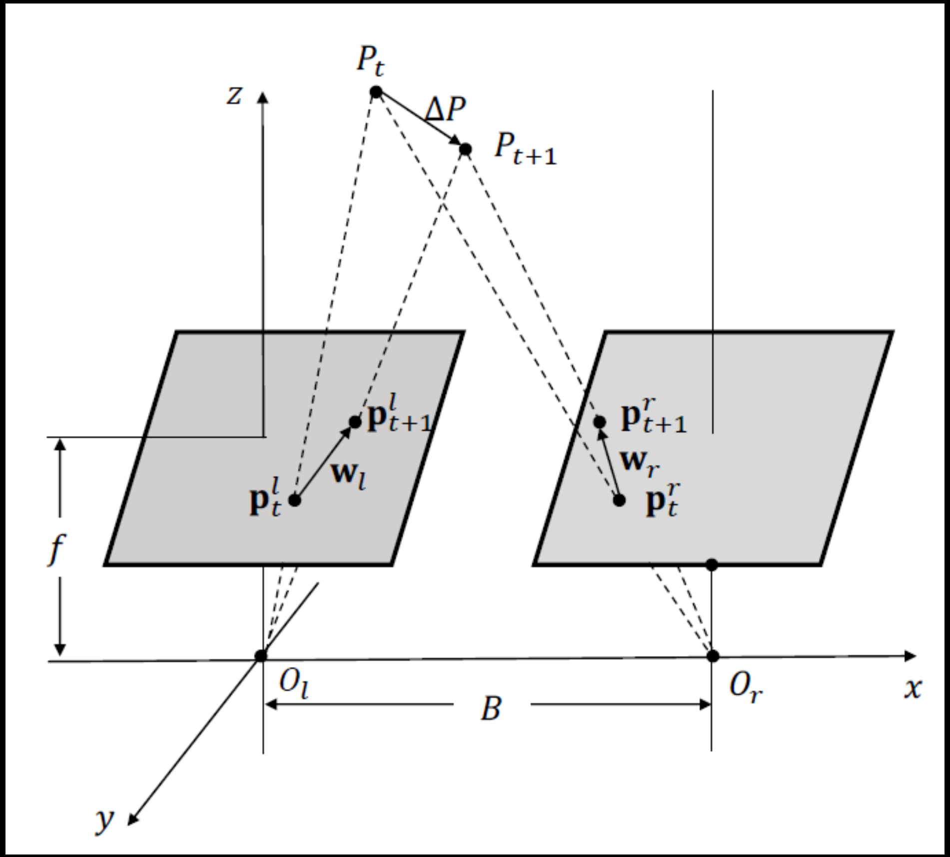
- Intuition 1: stereo matching can be modeled as a special case of optical flow, and we can leverage 3D geometric constraints behind stereoscopic videos to guide the learning of these two forms of correspondences.

- Intuition 2: we unveil the bottlenecks in prior self-supervised learning approaches and propose to create a new set of challenging proxy tasks to boost performance.

Geometric relationship between flow and stereo

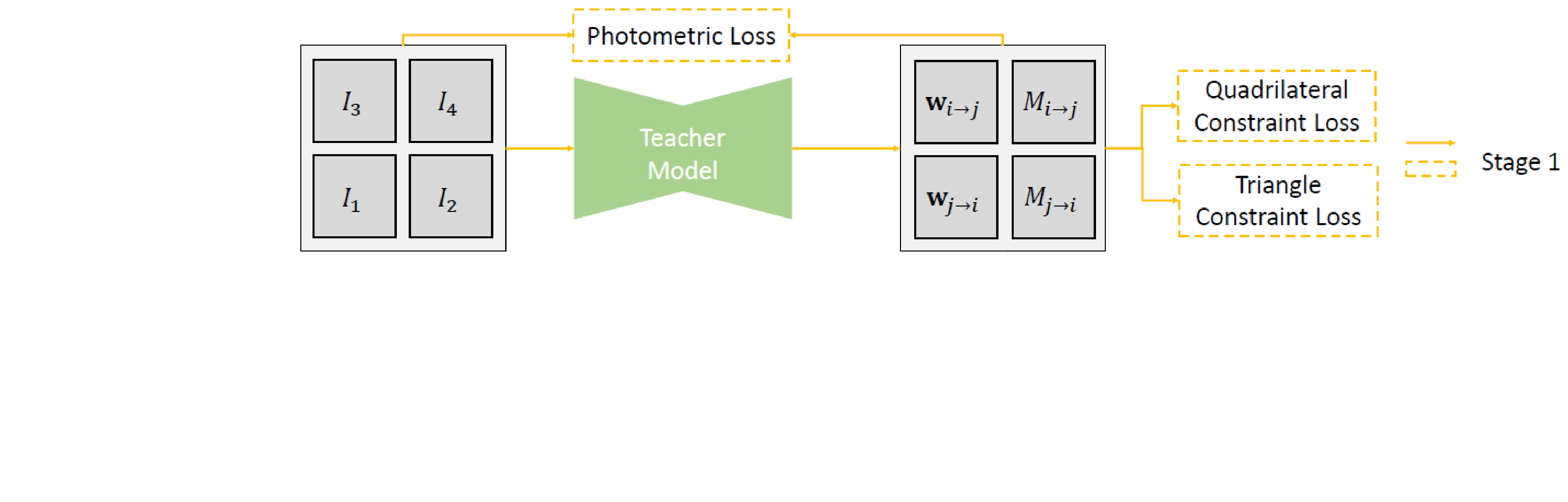


12 cross-view correspondnce maps among 4 stereoscopic frames.



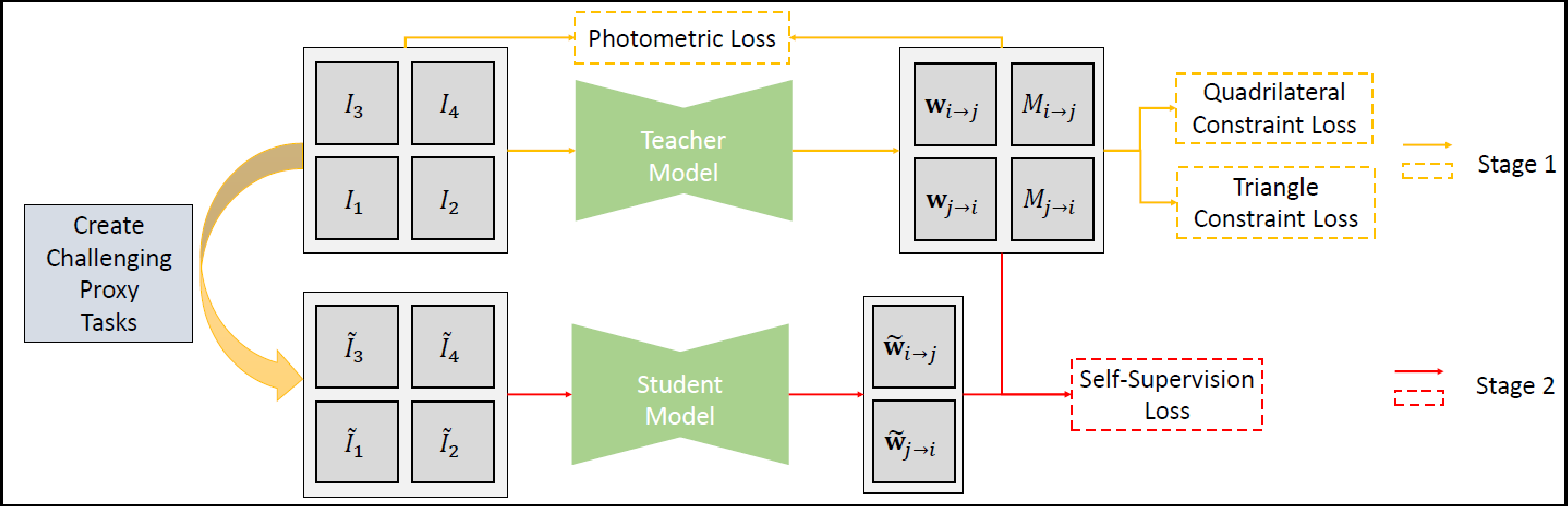
3D geometric constraints between optical flow $\mathbf{w}_l$ and $\mathbf{w}_r$) and stereo disparity from time t to $t+1$ in the 3D projection view.

Self-Supervised Learning: stage 1



Stage 1: we add geometric constraints between optical flow and stereo disparity to improve the quality of confident predictions.

Self-Supervised Learning: stage 2



Stage 2: we create challenging proxy tasks to guide the student model for effective self-supervised learning.

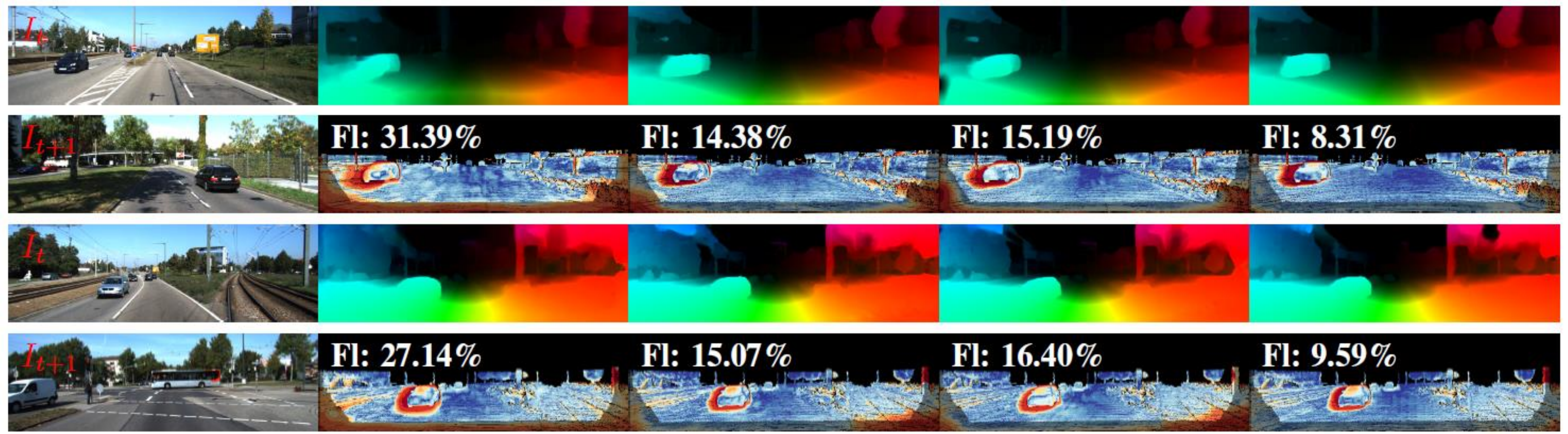
Our method outperforms all existing unsupervised optical flow methods on KITTI datasets. Our self-supervised method even outperforms several state-of-the-art fully supervised methods.

	Method	Train	KITTI 2012						KITTI 2015				
			train		test				train		test		
		Stereo	EPE-all	EPE-noc	EPE-all	EPE-noc	Fl-all	Fl-noc	EPE-all	EPE-noc	Fl-all	Fl-fg	Fl-bg
Supervised	SpyNet [32]	✗	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [1]	✗	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [45]	✗	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [15]	✗	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [30]	✗	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [14]	✗	(1.05)	–	1.6	**0.8**	7.27%	**3.27%**	(1.62)	–	9.38%	7.99%	9.66%
	PWC-Net [39]	✗	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	MFF [34]	✗	–	–	1.7	0.9	7.87%	4.19%	–	–	**7.17%**	**7.25%**	**7.15%**
	SelFlow [26]	✗	**(0.76)**	–	**1.5**	0.9	**6.19%**	3.32%	**(1.18)**	–	8.42%	7.61%	12.48%
Unsupervised	BackToBasic [17]	✗	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [35]	✗	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [30]	✗	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [44]	✗	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	MultiFrameOccFlow-None [16]	✗	–	–	–	–	–	–	6.65	3.24	–	–	–
	MultiFrameOccFlow-Soft [16]	✗	–	–	–	–	–	–	6.59	3.22	22.94%	–	–
	DDFlow [25]	✗	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [26]	✗	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Lai *et al.* [22]	✓	2.56	1.39	–	–	–	–	7.134	4.306	–	–	–
	UnOS [43]	✓	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	Our+L_p+L_q+L_t	✓	4.91	0.84	–	–	–	–	7.88	2.24	–	–	–
	Ours+L_p+L_q+L_t+Self-Supervision	✓	**1.45**	**0.82**	**1.7**	**0.9**	**7.63%**	**4.02%**	**3.54**	**2.12**	**11.10%**	**16.67%**	**9.99%**

We directly apply our optical flow model to estimate stereo disparity, it achieves state-of-the-art unsupervised stereo matching performance.

Method	KITTI 2012						KITTI 2015					
	EPE-all	EPE-noc	EPE-occ	D1-all	D1-noc	D1-all (test)	EPE-all	EPE-noc	EPE-occ	D1-all	D1-noc	D1-all (test)
Joung *et al.* [18]	–	–	–	–	–	13.88%	–	–	–	13.92%	–	–
Godard *et al.* [8] *	2.12	1.44	30.91	10.41%	8.33%	–	1.96	1.53	24.66	10.86%	9.22%	–
Zhou *et al.* [51]	–	–	–	–	–	–	–	–	–	9.41%	8.35%	–
OASM-Net [23]	–	–	–	8.79%	6.69%	8.60%	–	–	–	–	–	8.98%
SeqStereo *et al.* [46] *	2.37	1.63	33.62	9.64%	7.89%	–	1.84	1.46	26.07	8.79%	7.7%	–
Liu *et al.* [24] *	1.78	1.68	6.25	11.57%	10.61%	–	1.52	1.48	4.23	9.57%	9.10%	–
Guo *et al.* [9] *	1.16	1.09	4.14	6.45%	5.82%	–	1.71	1.67	4.06	7.06%	6.75%	–
UnOS [43]	–	–	–	–	–	5.93%	–	–	–	**5.94%**	–	6.67%
Ours+L_p	1.73	1.13	27.03	7.88%	5.87%	–	1.79	1.40	25.24	9.83%	7.74%	–
Ours+L_p+L_q+L_t	1.62	0.94	29.26	6.69%	4.69%	–	1.67	**1.31**	19.55	8.62%	7.15%	–
Ours+L_p+L_q+L_t+Self-Supervision	**1.01**	**0.93**	**4.52**	**5.14%**	**4.59%**	**5.11%**	**1.34**	**1.31**	**2.56**	6.13%	**5.93%**	**6.61%**

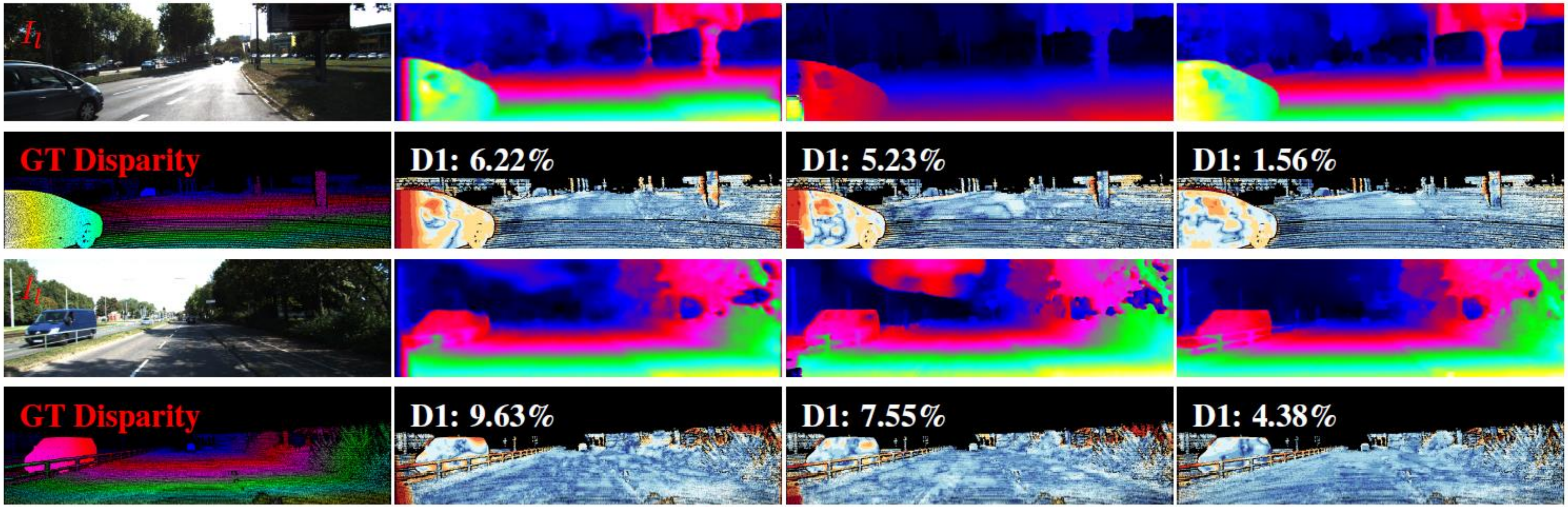
Optical flow qualitative evaluation: our model achieves much better results both quantitatively and qualitatively (e.g., shaded boundary regions).



(a) Input Images (b) MFO-Flow [16] (c) DDFlow [25] (d) SelFlow [26] (e) Ours

For each case, the top row is optical flow and the bottom row is error map.
Lower Fl is better.

Stereo matching qualitative evaluation: Our models estimate more accurate disparity maps (e.g., image boundary regions and moving-object boundary regions)



(a) I_l and GT Disparity (b) Godard *et al.* [8] (c) SeqStereo [46] (d) Ours

For each case, the top row is stereo disparity and the bottom row is error map. Lower D1 is better.

Conclusion

- We have presented a method to jointly learning optical flow and stereo matching with a unified model.
- We show that geometric constraints can improve the quality of those confident predictions.
- we unveil the bottlenecks in prior self-supervised learning approaches and propose to create a new set of challenging proxy tasks to boost performance.

- Code available: https://github.com/ppliuboy/Flow2Stereo