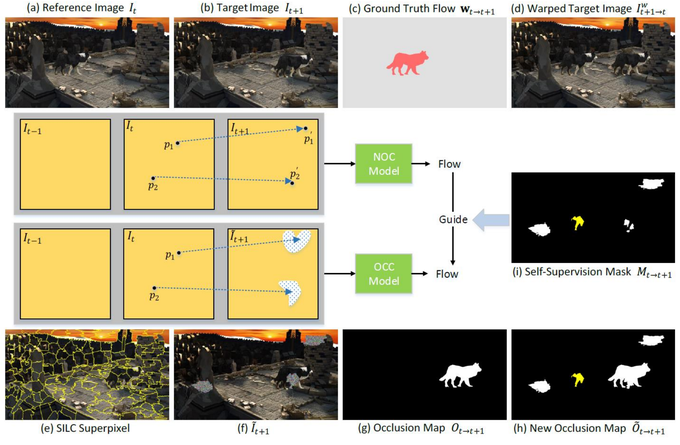
SelFlow: Self-Supervised Learning of Optical Flow

[1]Pengpeng Liu, [1]Michael R. Lyu, [1]Irwin King, [2,3]Jia Xu
[1]The Chinese University of Hong Kong [2]Tencent AI Lab [3]Huya Inc.

Introduction

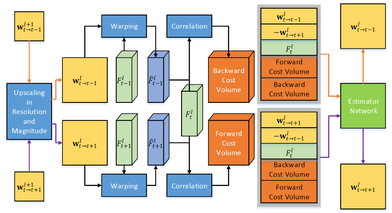
- Optical Flow: motion of pixels between two adjacent images
- Challenges
  - Traditional methods: computational complexity, time costing
  - Supervised learning: need a large amount of labeled data → difficult to obtain → pre-train on synthetic data → **domain gap**
  - Unsupervised learning
    - Photometric loss: measure the difference between reference image and warped target image
    - Produce reliable flow for non-occluded pixels, but lack the ability to learn the flow of **occluded** pixels → **performance gap**
- **Our Contribution**
  - Present a **two-stage self-supervised** learning approach to learning optical flow of occluded pixels from unlabeled data
  - Our **self-supervised pre-trained model** provides an excellent **initialization** for supervised fine-tuning, reducing the reliance of pre-training on synthetic datasets.

Method



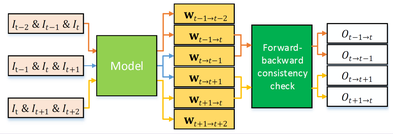
- p_1 is non-occluded from I_t to I_{t+1}, but becomes occluded from I_t to $\tilde{I}_{t+1}$
- We distill reliable flow estimations of p_1 from NOC-Model to guide the learning of OCC-Model

Network Architecture



In principle, we can utilize any CNNs. Here we build our network architecture based on PWC-Net.

Occlusion Estimation



Loss Functions

- Photometric Loss L_p:

$$L_p = \sum_{i,j} \frac{\sum \psi(I_i - I^w_{j\to i}) \odot (1 - O_i)}{\sum (1 - O_i)}$$

- Self-Supervision Mask $M_{i\to j}$:

$$M_{i\to j} = \text{clip}(\tilde{O}_{i\to j} - O_{i\to j}, 0, 1)$$

- Self-Supervision loss L_o:

$$L_o = \sum_{i,j} \frac{\sum \psi(\mathbf{w}_{i\to j} - \tilde{\mathbf{w}}_{i\to j}) \odot M_{i\to j}}{\sum M_{i\to j}}$$

- Unsupervised Training
  - NOC-Model: L_p
  - OCC-Model: $L_p + L_o$
- Supervised Fine-tuning
  - Initialize with pre-trained OCC-Model, fine-tune with ground truth optical flow

Main Result

	Method	Sintel Clean train	Sintel Clean test	Sintel Final train	Sintel Final test	KITTI 2012 train	KITTI 2012 test	KITTI 2012 test(Fl)	KITTI 2015 train	KITTI 2015 test(Fl)
Unsupervised	BackToBasic+ft [20]	–	–	–	–	11.3	9.9	–	–	–
	DSTFlow+ft [37]	(6.16)	10.41	(6.81)	11.27	10.43	12.4	–	16.79	39%
	UnFlow-CSS [29]	–	–	(7.91)	10.22	3.29	–	–	8.10	23.30%
	OccAwareFlow+ft [46]	(4.03)	7.95	(5.95)	9.15	3.55	4.2	–	8.88	31.2%
	MultiFrameOccFlow-None+ft [18]	(6.05)	–	(7.09)	–	–	–	–	6.65	–
	MultiFrameOccFlow-Soft+ft [18]	(3.89)	7.23	(5.52)	8.81	–	–	–	6.59	22.94%
	DDFlow+ft [26]	(2.92)	**6.18**	3.98	7.40	2.35	3.0	8.86%	5.72	14.29%
	Ours	**(2.88)**	6.56	**(3.87)**	**6.57**	**1.69**	**2.2**	**7.68%**	**4.84**	**14.19%**
Supervised	FlowNetS+ft [10]	(3.66)	6.96	(4.44)	7.76	7.52	9.1	44.49%	–	–
	FlowNetC+ft [10]	(3.78)	6.85	(5.28)	8.51	8.79	–	–	–	–
	SpyNet+ft [35]	(3.17)	6.64	(4.32)	8.36	8.25	10.1	20.97%	–	35.07%
	FlowFieldsCNN+ft [2]	–	3.78	–	5.36	–	3.0	13.01%	–	18.68 %
	DCFlow+ft [49]	–	3.54	–	5.12	–	–	–	–	14.83%
	FlowNet2+ft [15]	(1.45)	4.16	(2.01)	5.74	(1.28)	1.8	8.8%	(2.3)	11.48%
	UnFlow-CSS+ft [29]	–	–	–	–	(1.14)	1.7	8.42%	(1.86)	11.11%
	LiteFlowNet+ft-CVPR [14]	(1.64)	4.86	(2.23)	6.09	(1.26)	1.7	–	(2.16)	10.24%
	LiteFlowNet+ft-axXiv [14]	**(1.35)**	4.54	(1.78)	5.38	(1.05)	1.6	7.27%	(1.62)	9.38%
	PWC-Net+ft-CVPR [43]	(2.02)	4.39	(2.08)	5.04	(1.45)	1.7	8.10%	(2.16)	9.60%
	PWC-Net+ft-axXiv [42]	(1.71)	3.45	(2.34)	4.60	(1.08)	**1.5**	6.82%	(1.45)	7.90%
	ProFlow+ft [27]	(1.78)	**2.82**	–	5.02	(1.89)	2.1	7.88%	(5.22)	15.04%
	ContinualFlow+ft [31]	–	3.34	–	4.52	–	–	–	–	10.03%
	MFF+ft [36]	–	3.42	–	4.57	–	1.7	7.87%	–	**7.17%**
	Ours+ft	(1.68)	3.74	**(1.77)**	**4.26**	**(0.76)**	**1.5**	**6.19%**	**(1.18)**	8.42%

Sintel Benchmark

	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
SelFlow [2]	4.262	2.040	22.369	4.083	1.715	1.287	0.582	2.343	27.154
VCN [3]	4.520	2.195	23.478	4.423	1.802	1.357	0.934	2.816	26.434
ContinualFlow_ROB [4]	4.528	2.723	19.248	5.050	2.573	1.713	0.872	3.114	26.063
MFF [5]	4.566	2.216	23.732	4.664	2.017	1.222	0.893	2.902	26.810
IRR-PWC [6]	4.579	2.154	24.355	4.165	1.843	1.292	0.709	2.423	28.998
PWC-Net+ [7]	4.596	2.254	23.696	4.781	2.045	1.234	0.945	2.978	26.620

Ablation Study

Occlusion Handling	Multiple Frame	Self-Supervision Rectangle	Self-Supervision Superpixel	Sintel Clean ALL	Sintel Clean NOC	Sintel Clean OCC	Sintel Final ALL	Sintel Final NOC	Sintel Final OCC	KITTI 2012 ALL	KITTI 2012 NOC	KITTI 2012 OCC	KITTI 2015 ALL	KITTI 2015 NOC	KITTI 2015 OCC
✗	✗	✗	✗	(3.85)	(1.53)	(33.48)	(5.28)	(2.81)	(36.83)	7.05	1.31	45.03	13.51	3.71	75.51
✗	✓	✗	✗	(3.67)	(1.54)	(30.80)	(4.98)	(2.68)	(34.42)	6.52	1.11	42.44	12.13	3.47	66.91
✓	✗	✗	✗	(3.35)	(1.37)	(28.70)	(4.50)	(2.37)	(31.81)	4.96	0.99	31.29	8.99	3.20	45.68
✓	✓	✗	✗	(3.20)	(1.35)	(26.63)	(4.33)	(2.32)	(29.80)	3.32	0.94	19.11	7.66	2.47	40.99
✓	✗	✗	✓	(2.96)	(1.33)	(23.78)	(4.06)	(2.25)	(27.19)	1.97	0.92	8.96	5.85	2.96	24.17
✓	✓	✓	✗	(2.91)	(1.37)	(22.58)	(3.99)	(2.27)	(26.01)	1.78	0.96	7.47	5.01	2.55	21.86
✓	✓	✗	✓	**(2.88)**	**(1.30)**	**(22.06)**	**(3.87)**	**(2.24)**	**(25.42)**	**1.69**	**0.91**	**6.95**	**4.84**	**2.40**	**19.68**

Effect of Self-Supervision

Conclusion

- We present a self-supervised approach to learning accurate optical flow for both occluded and non-occluded pixels
- Our self-supervised pre-training reduces the reliance of pre-training on synthetic labeled datasets
- Our method achieves state-of-the-art results on KITTI and Sintel benchmarks (currently No.1 on Sintel)

Code Link