3D视觉和图形学
在直播内容生产中的应用和挑战

许佳

虎牙是一家技术驱动的内容公司

2012：YY直播诞生，国内首家游戏直播

2014：更名为虎牙直播，首次推出1080P服务

2016：虎牙公司独立

2017：A轮融资7500万美元

2018：3月获得腾讯领投的B轮4.62亿美元
5月纽交所上市

2019：海外 NIMO, YOME

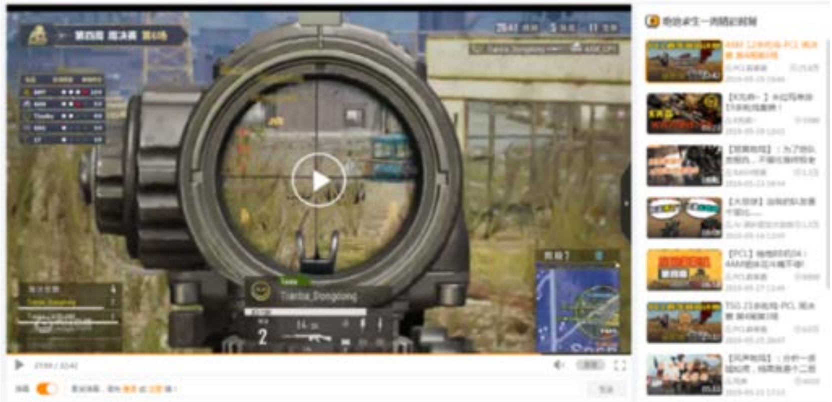
技术驱动娱乐 - 应用样例

美颜

内容识别

智能弹幕

虚拟主播



游戏视频集锦自动剪辑

广告商业化

直播内容生产

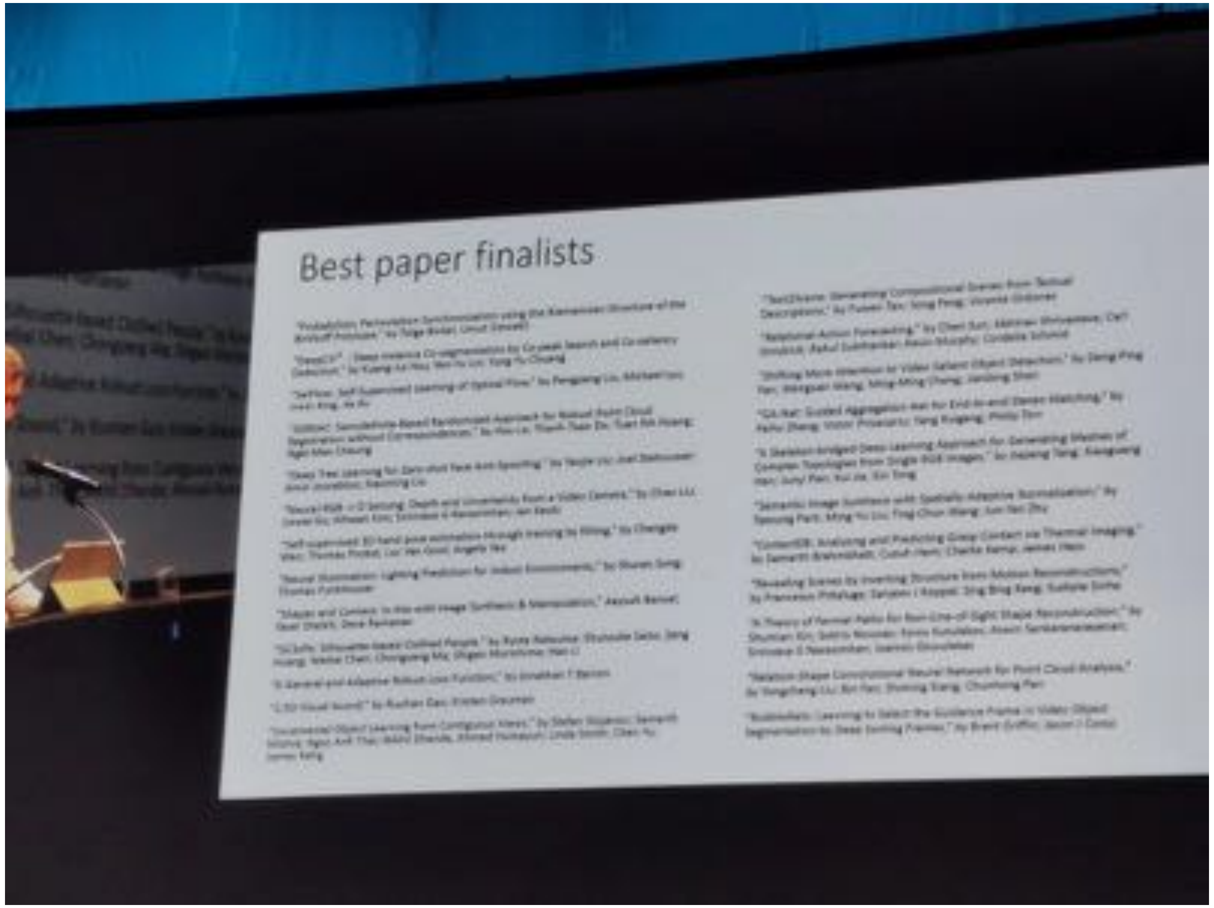
挑战：实时生产高质量内容

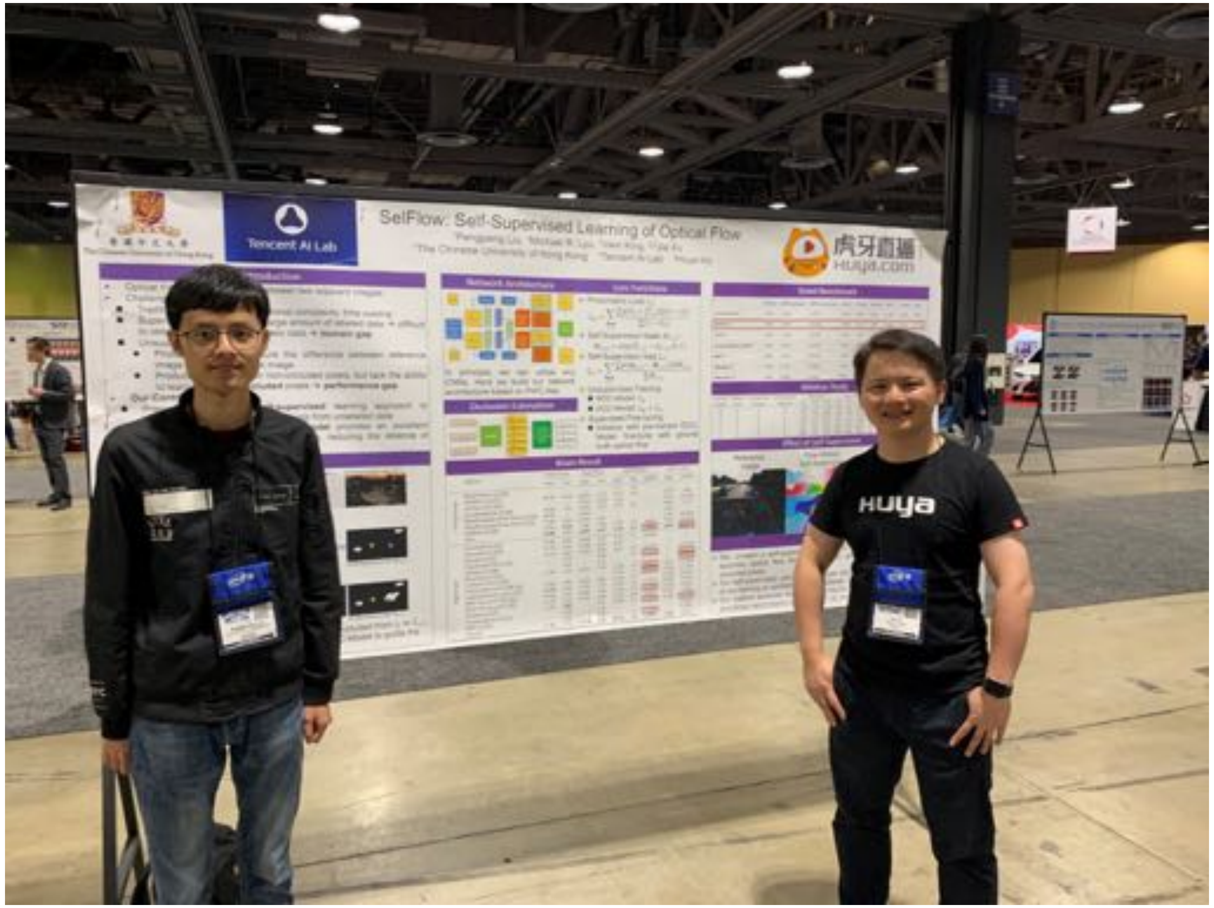
排名	片名	年份	票房	观影人次	产地	导演	备注
1	战狼2	2017	56.83亿	1.6亿	中国大陆	吴京	首部票房突破40亿及50亿
2	哪吒之魔童降世	2019	49亿		中国大陆	饺子	票房最高动画电影[注 2]
3	流浪地球	2019	46.54亿	1.05亿	中国大陆	郭帆	
4	复仇者联盟4：终局之战	2019	42.4亿	8641万	美国	罗素兄弟	
5	红海行动	2018	36.5亿	9292.1万	中国大陆 香港	林超贤	

技术人才的需求

- 3D Face/Body Reconstruction
- 3D Human Pose Estimation
- Scene Reconstruction
- Optical Flow/Depth
- Video Segmentation
- Motion Capture
- Character Rigging/ Animation
- Real-time Rendering
- AR/VR

SelFlow论文入围CVPR Best Paper Finalist





AI团队成员2019年在CVPR/ICCV/ICML/NeurIPS等顶会均有论文发表

SelFlow: Self-Supervised Learning of Optical Flow

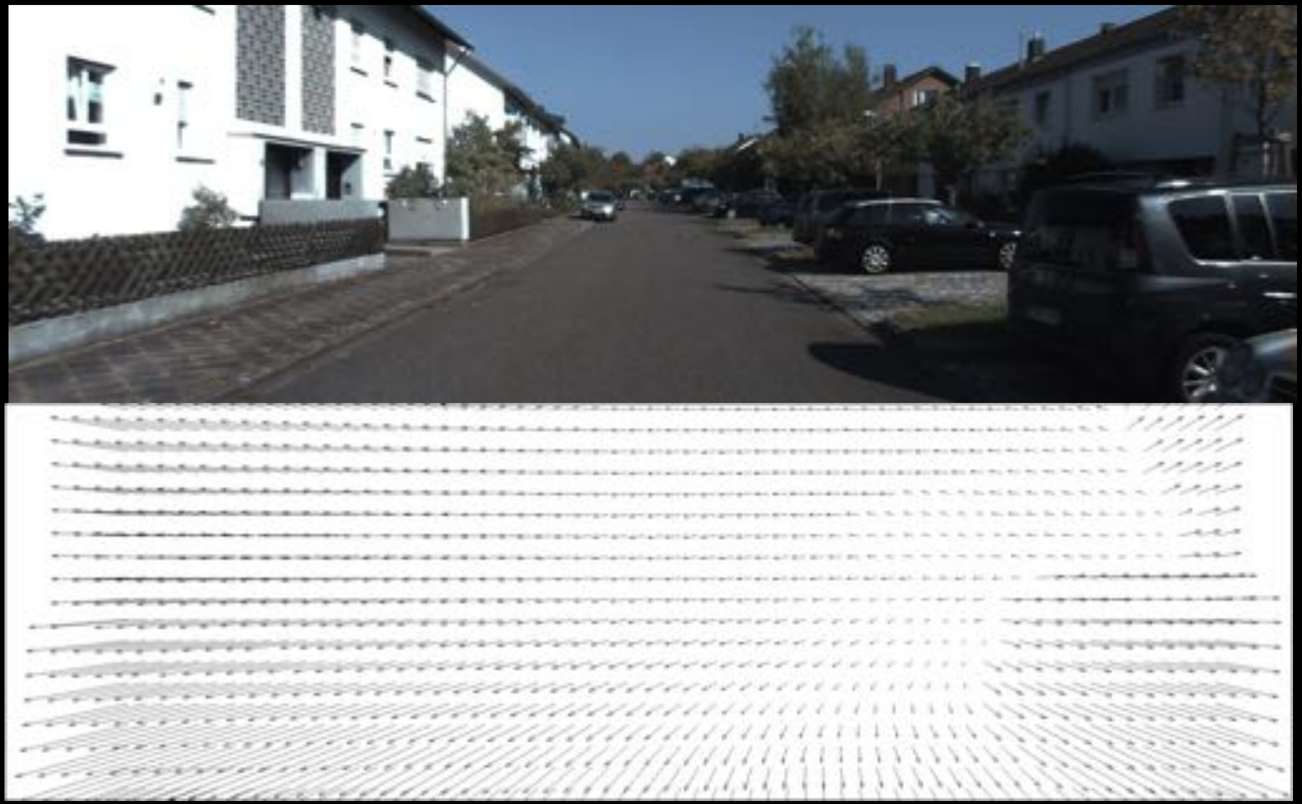
Pengpeng Liu, Michael Lyu, Irwin King, Jia Xu

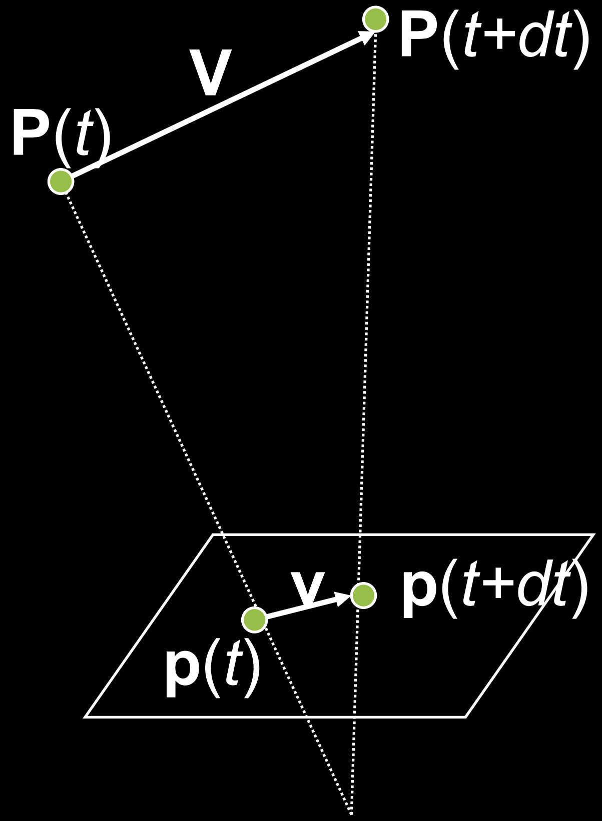
CVPR 2019 (Oral Presentation)

Optical flow describes the pixel motion between two adjacent images. The motion field is the projection of 3D scene motion into the image.

Images

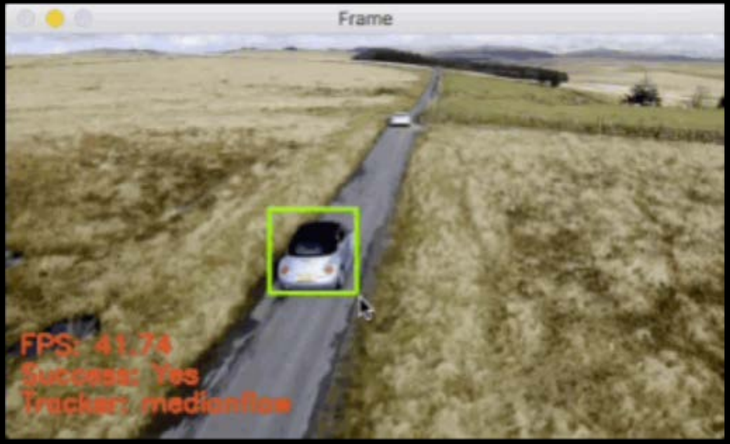
Optical Flow





Optical flow has a wide range of applications.

Autonomous Driving



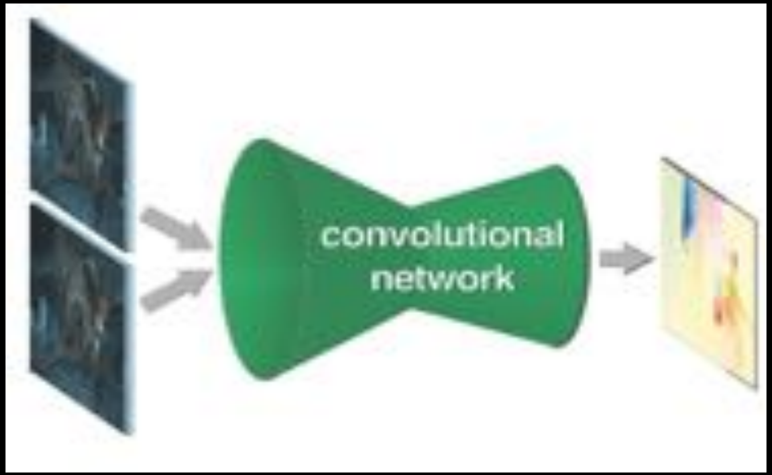
Object Tracking

3D Shape Reconstruction

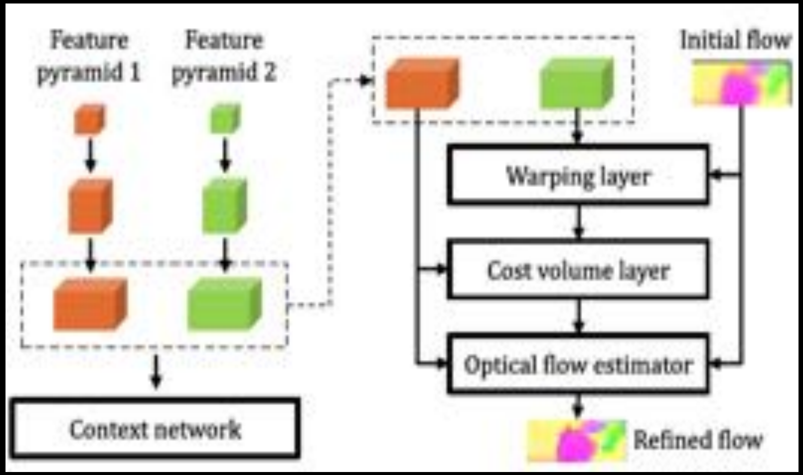
Video Action Recognition

CNNs for Optical Flow

- Advantage: high performance while run at real time.
- Disadvantage: need a large amount of labeled data → difficult to obtain.



FlowNet



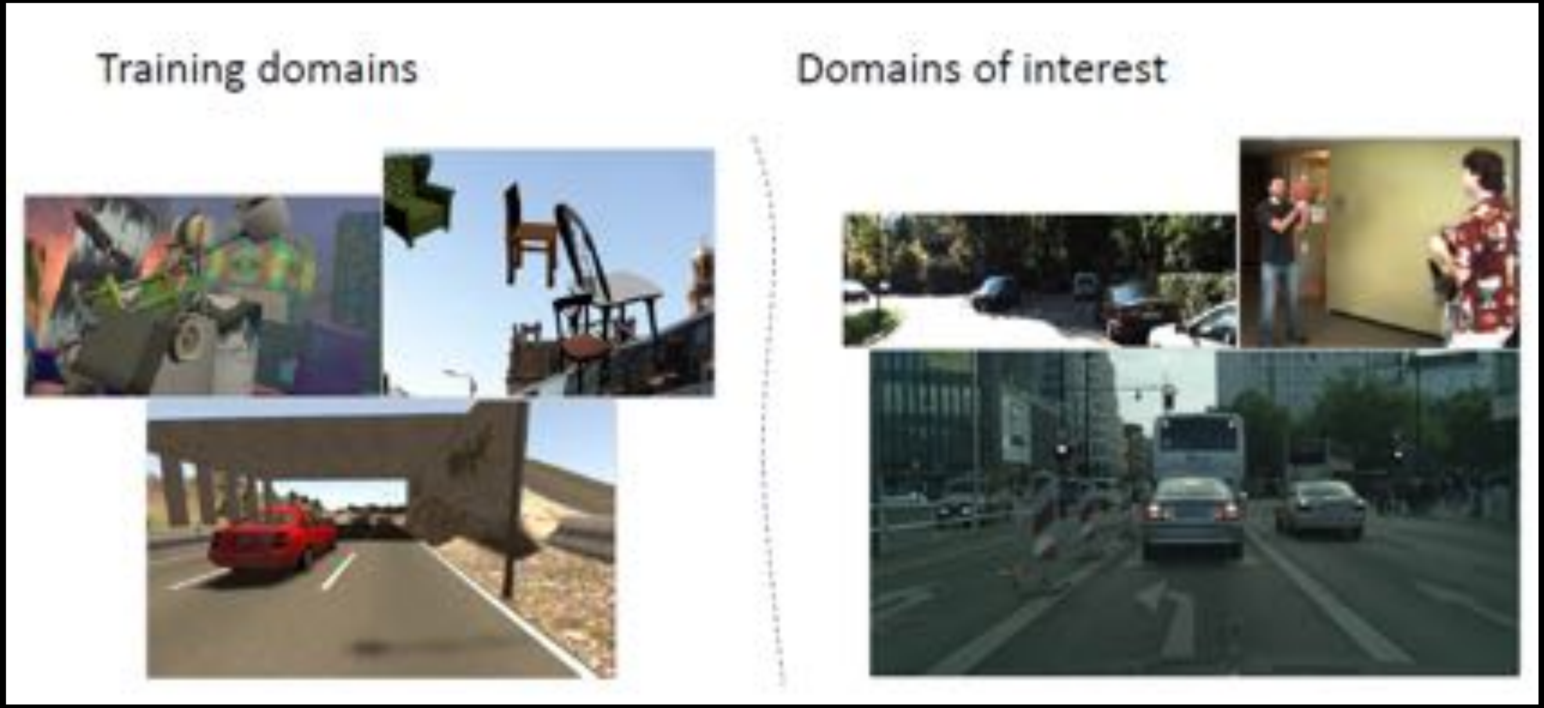
PWC-Net

Fischer et al. 2015, "FlowNet: Learning Optical Flow with Convolutional Networks"

Sun et al. 2018, "PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume"

CNNs for Optical Flow

- Advantage: high performance while running at real time.
- Disadvantage: need a large amount of labeled data → difficult to obtain.
  - Pre-train on synthetic datasets → domain gap.
- Unsupervised Learning: produce reliable flow for non-occluded pixels, but lack the ability to learn the flow of occluded pixels → performance gap.



Meister et al. 2018, “UnFlow: Unsupervised Learning of Optical Flow with a Bidirectional Census Loss"

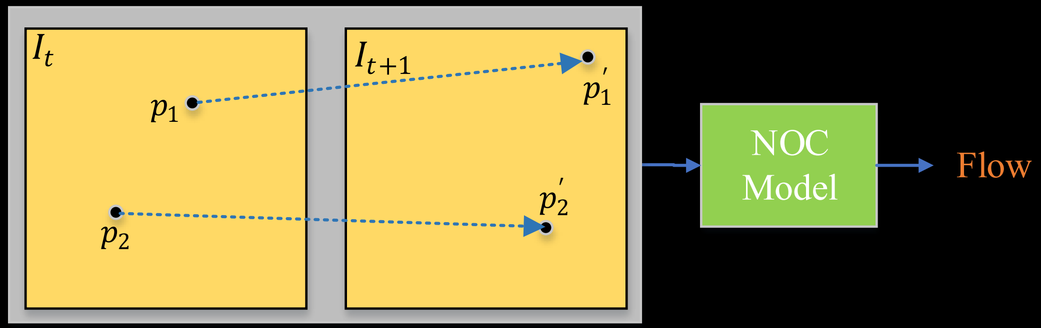
Our Approach

- We present a self-supervised learning approach to learning optical flow from unlabeled data. Our method distills reliable flow estimations from non-occluded pixels, and uses these predictions to learn optical flow for hallucinated occlusions.

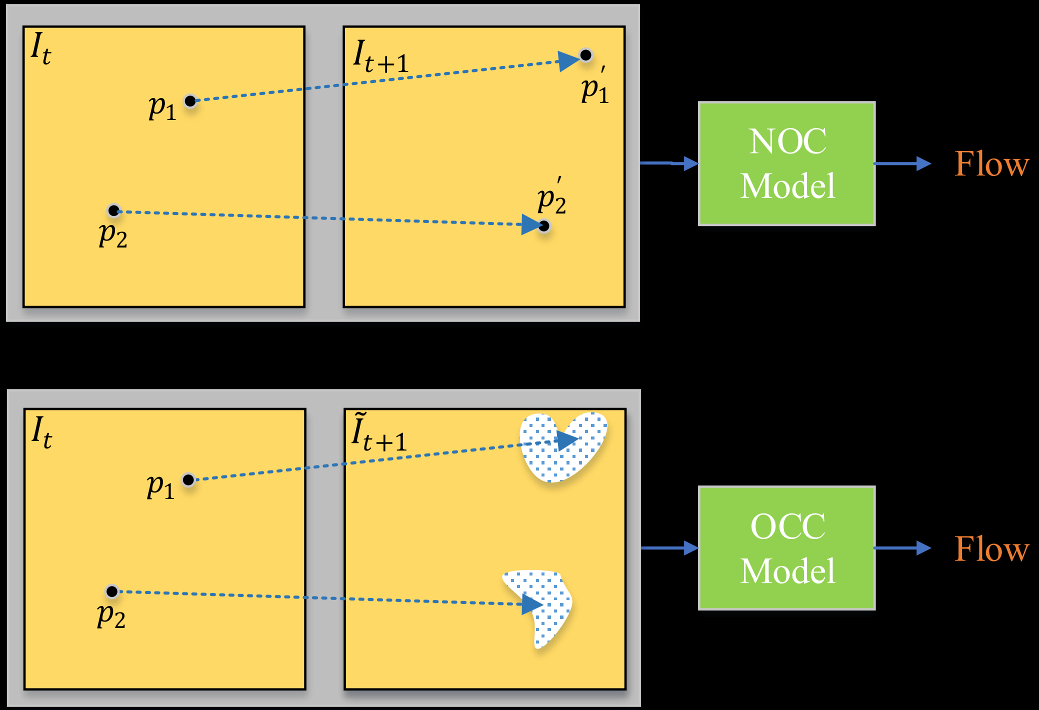
- After fine-tuning with the pre-trained self-supervised model, we achieve state-of-the-art supervised learning results, reducing the reliance of synthetic data.

Main Idea

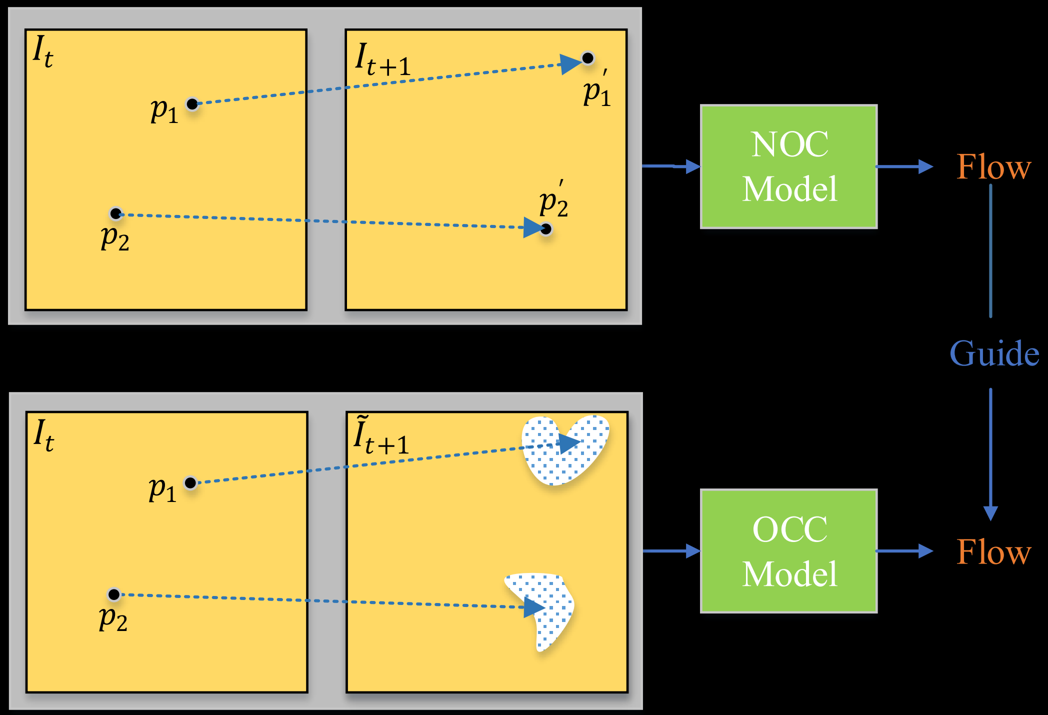
Initially, p_1 and p_2 are non-occluded from I_t to I_{t+1}, p_1' and p_2' are their corresponding pixels. NOC-Model can accurately estimate the flow of p_1 and p_2 using photometric loss.



We inject random noise to I_{t+1} and let noise cover p_1' and p_2', then p_1 and p_2 become occluded from I_t to $\tilde{I}_{t+1}$. OCC-Model cannot accurately estimate flow of p_1 and p_2 using photometric loss.



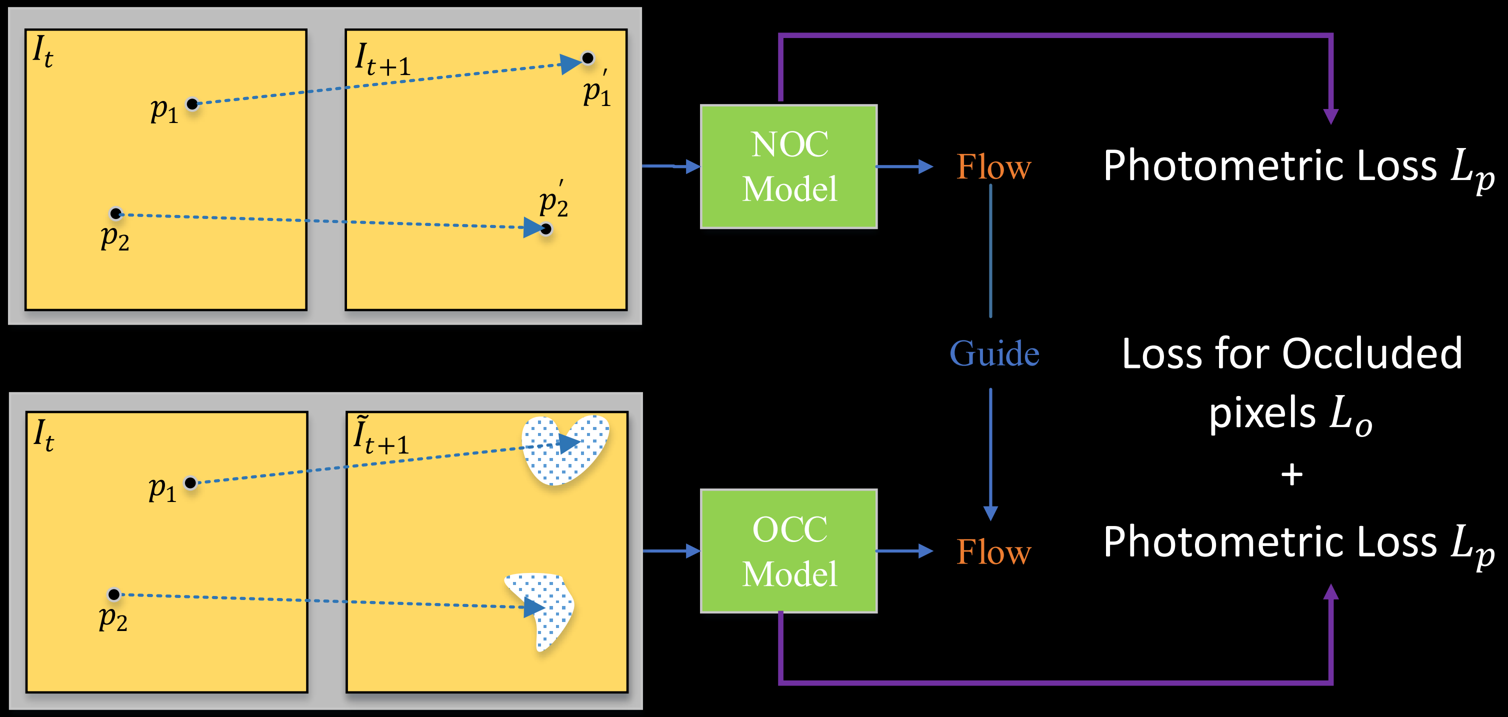
We distill reliable flow estimations of p_1 and p_2 from NOC-Model to guide the flow learning for OCC-Model. The guidance is only employed to pixels that are occluded from I_t to $\tilde{I}_{t+1}$ but non-occluded from I_t to I_{t+1}, such as p_1 and p_2.

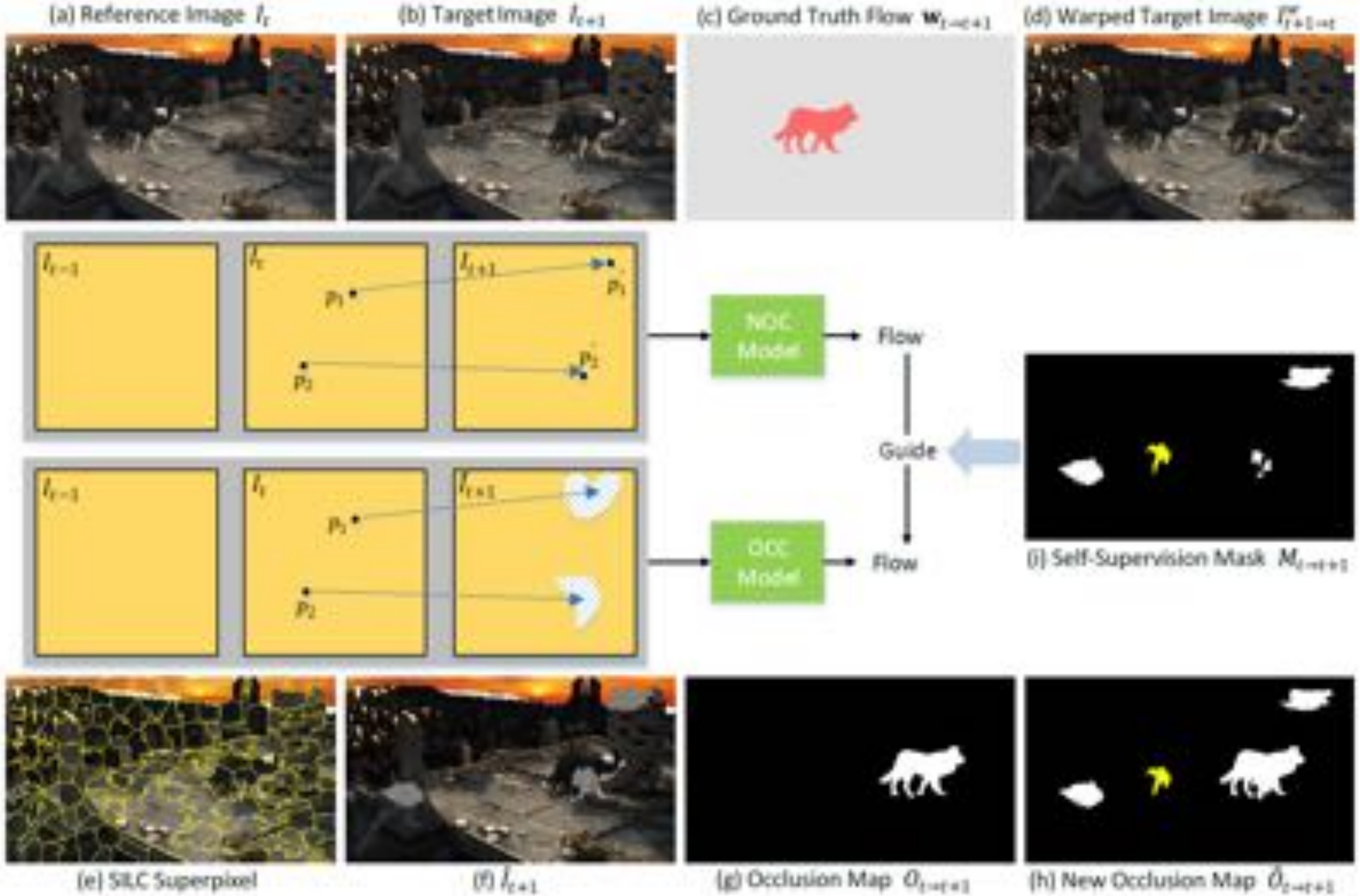


Self-supervision mask

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

Our method removes the reliance of pre-training on synthetic data.





(a) Reference Image I_t
(b) Target Image I_{t+1}
(c) Ground Truth Flow $\mathbf{w}_{t\to t+1}$
(d) Warped Target Image $I^w_{t+1\to t}$
I_{t-1}
I_t
I_{t+1}
p_1
p_2
NOC Model
Flow
Guide
OCC Model
Flow
(i) Self-Supervision Mask $M_{t\to t+1}$
(e) SILC Superpixel
(f) $\tilde{I}_{t+1}$
(g) Occlusion Map $O_{t\to t+1}$
(h) New Occlusion Map $\tilde{O}_{t\to t+1}$

Quantitative Results

Our unsupervised method outperforms all existing unsupervised methods on all datasets except Sintel Clean.

	Method	Sintel Clean		Sintel Final		KITTI 2012			KITTI 2015	
		train	test	train	test	train	test	test(Fl)	train	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%

Our unsupervised method even outperforms several famous fully-supervised methods.

	Method	Sintel Clean		Sintel Final		KITTI 2012			KITTI 2015	
		train	test	train	test	train	test	test(Fl)	train	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%

Our fine-tuned models achieve state-of-the-art results without using any external labeled data.

	Method	Sintel Clean		Sintel Final		KITTI 2012			KITTI 2015	
		train	test	train	test	train	test	test(Fl)	train	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%

Our fine-tuned result on Sintel achieves EPE=4.26, outperforming all submitted methods (Rank 1 until now).

	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
CompactFlow [8]	4.626	2.099	25.253	4.192	1.825	1.233	0.845	2.677	28.120
HD3-Flow [9]	4.666	2.174	24.994	3.786	1.719	1.647	0.657	2.182	30.579
LiteFlowNet2-MD+ [10]	4.728	2.249	24.939	4.010	1.925	1.504	0.783	2.634	29.369
VCN-small [11]	4.733	2.436	23.453	4.781	2.038	1.524	1.108	3.069	26.726
CompactFlow-woscv [12]	4.858	2.213	26.439	4.220	1.867	1.453	0.906	2.701	29.709
SF_Net [13]	4.860	2.301	25.732	4.121	1.991	1.493	0.812	2.606	30.402
Semantic_Lattice [14]	4.886	2.456	24.701	4.596	2.082	1.529	0.803	3.024	29.649

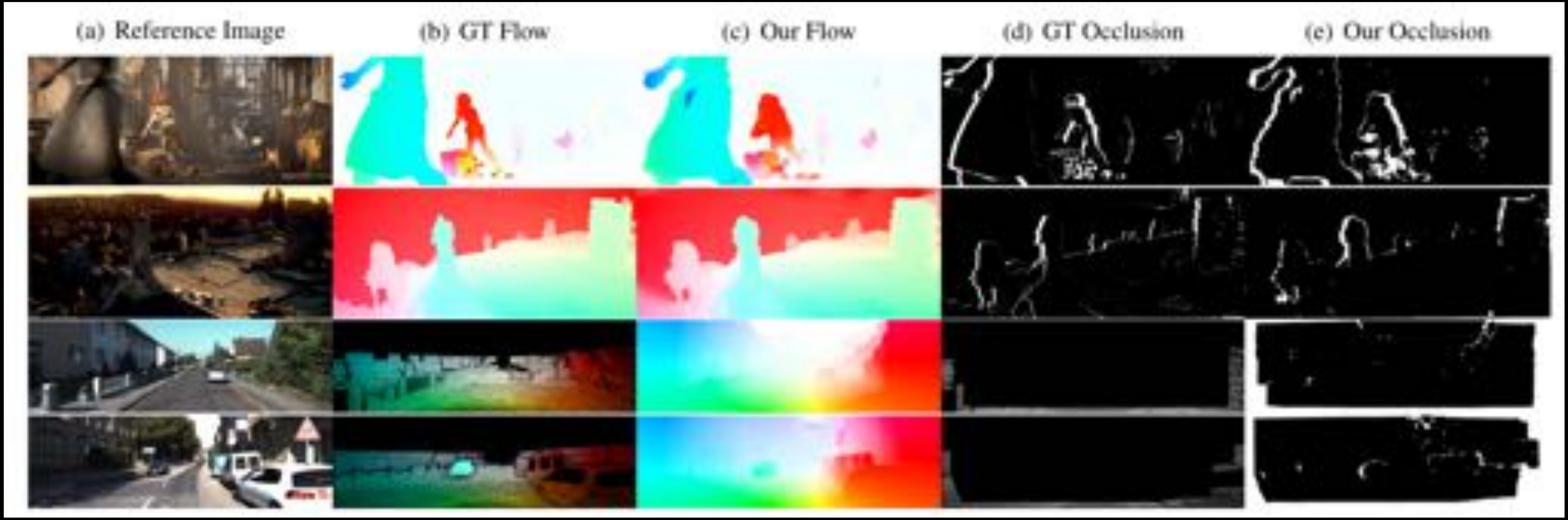
Ablation study

- Occlusion Handling, multiple-frame formulation and self-supervision consistently improve the flow estimation performance.
- We employ two extensions for noise injection: rectangle and superpixel. **Superpixel** setting works a litter better.
  - The shape of superpixels are usually **random**, which is consistent with the real-world cases.
  - The pixels within each superpixel usually **belong to the same object or have similar flow fields**. Low-level segmentation can be helpful for the optical flow estimation.

Occlusion Handling	Multiple Frame	Self-supervision Rectangle	Self-supervision Superpixel	Sintel Clean			Sintel Final			KITTI 2012			KITTI 2015		
				all	noc	occ	all	noc	occ	all	noc	occ	all	noc	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
✓	✓	✗	✓	**(2.82)**	**(1.30)**	**(22.06)**	**(3.87)**	**(2.24)**	**(25.42)**	**1.69**	**0.91**	**6.95**	**4.84**	**2.40**	**19.68**

Qualitative Results

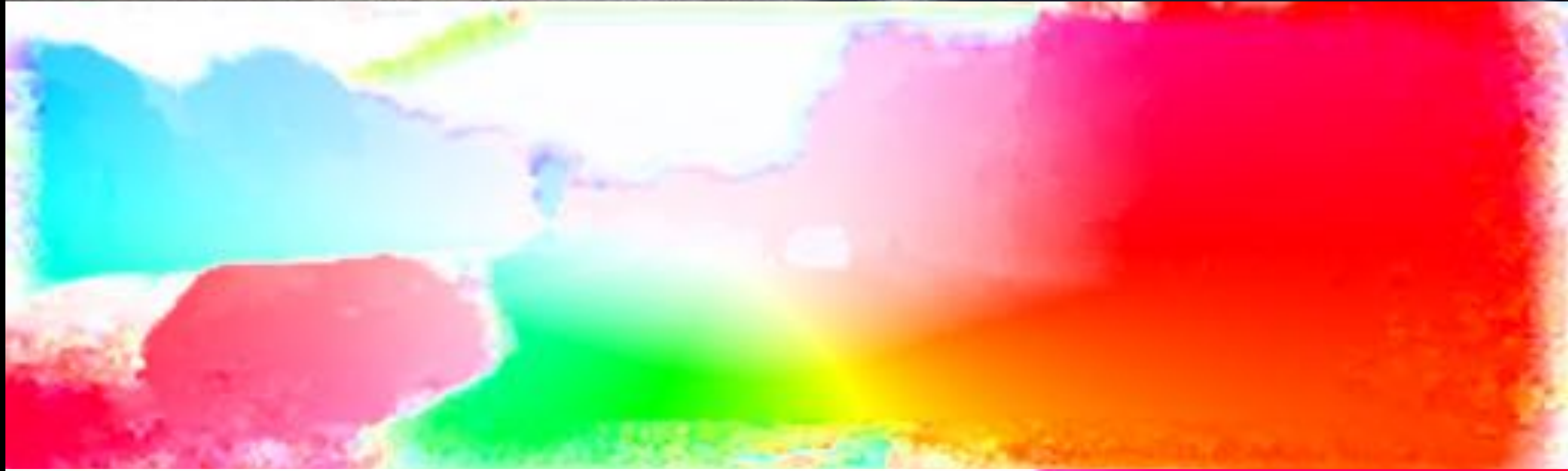
Sample unsupervised results on Sintel and KITTI dataset. From top to bottom, we show samples from Sintel Final, KITTI 2012 and KITTI 2015.



Effect of Self-Supervision

Reference Image

Flow Estimation
without Self-supervision



Flow Estimation
with Self-supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Comparing with PWC-Net, our fine-tuned model estimates optical flow with more accurate details.

Reference Image

Flow Estimation using PWC-Net

Flow Estimation using Our Fine-tuned Model

Reference Image

Flow Estimation using PWC-Net

Flow Estimation using Our Fine-tuned Model

Reference Image

Flow Estimation using PWC-Net

Flow Estimation using Our Fine-tuned Model

Reference Image

Flow Estimation using PWC-Net

Flow Estimation using Our Fine-tuned Model

To demonstrate the generalization ability of our model, we further show our flow estimation on real-word videos (from the DAVIS dataset).

Reference Image

Flow from Our Unsupervised Model

Flow from Our Fine-tuned Model

Reference Image

Flow from Our Unsupervised Model

Flow from Our Fine-tuned Model

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 available: https://github.com/ppliuboy/SelFlow

Q & A

Hiring Interns/Engineers/Researchers
in Vision and Graphics.

http://hr.huya.com

Email：xujia@huya.com