SelFlow: Self-Supervised Learning of Optical Flow

Pengpeng Liu, Michael Lyu, Irwin King, Jia Xu

Presenter: Tiantian Li, Yuhan Huang, Sicheng Mu

Outline

- Introduction
- Review of related work
- Method
- Experiments and Main results
- Conclusions
- Pos and cons

What is optical flow?

- Track the apparent motion (correspondence) of object in a video

Visual World is Continuous

Object Permanence

1

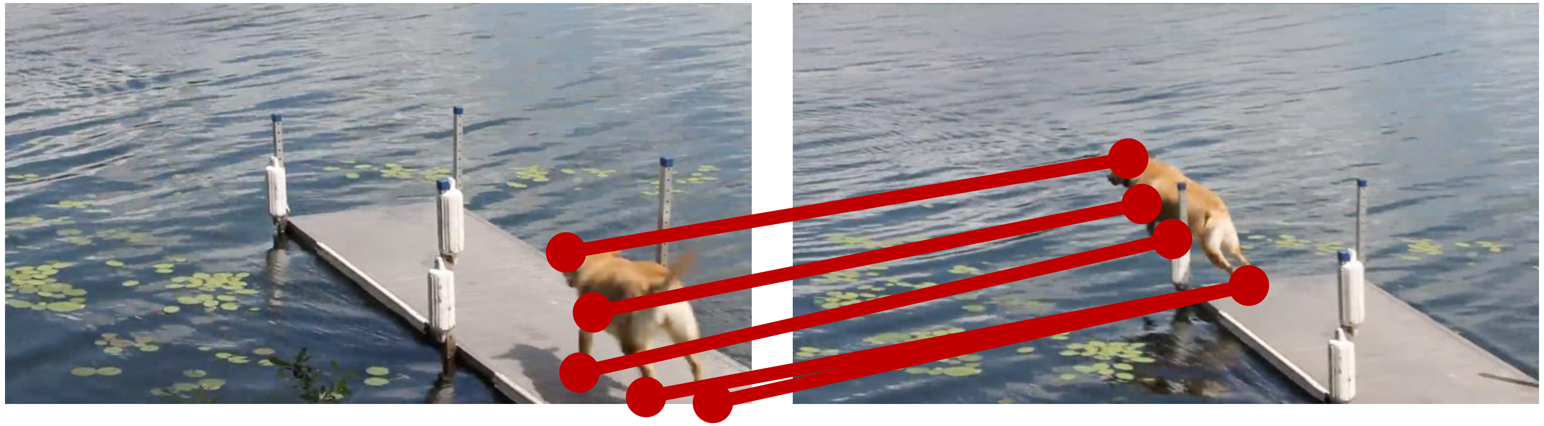
2

3

4

Wang et al. Learning Correspondence from the Cycle-consistency of Time, ICCV 2019.

Correspondence in Time



Learning correspondence without human supervision

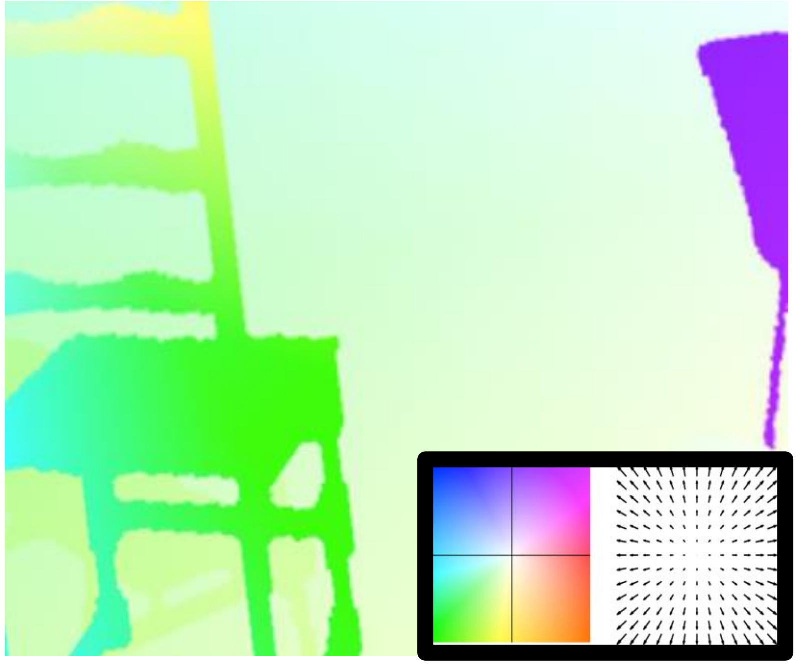
Labeling correspondence is very expensive!

Optical Flow Estimation

Frame 1

Frame 2

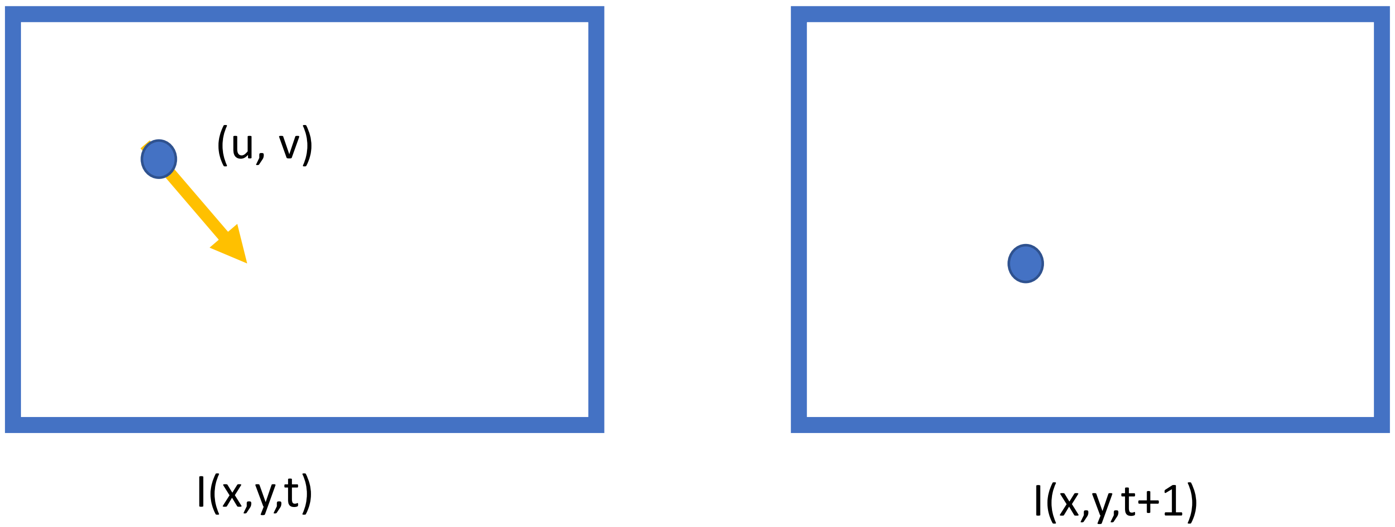
Optical Flow



Pixel-level correspondence Sensitive to local perturbation

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

Optical Flow Constraints



1) Brightness constancy constraint (equation)

   I(x,y,t) = I(x+u, y+v, t+1)

2) Small motion: (u and v are less than 1 pixel or smooth)

   Taylor series expansion of I:

$$I(x+u, y+v) = I(x,y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v \quad + \text{[higher order terms]}$$

$$I(x+u, y+v) \approx I(x,y) + \frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v$$

Challenging: occlusion

- Occlusion destroys the consistency constraint in optical flow estimation

cause an error estimation

Related Work

- Classical Optical Flow Estimation

  - Energy minimization problem based on brightness constancy and spatial smoothness

$$E(u,v) = \iint (I_2(p+w) - I_1)^2 + \alpha^2 \left(\left| |\nabla u| \right|^2 + \left| |\nabla v| \right|^2 \right) dxdy$$

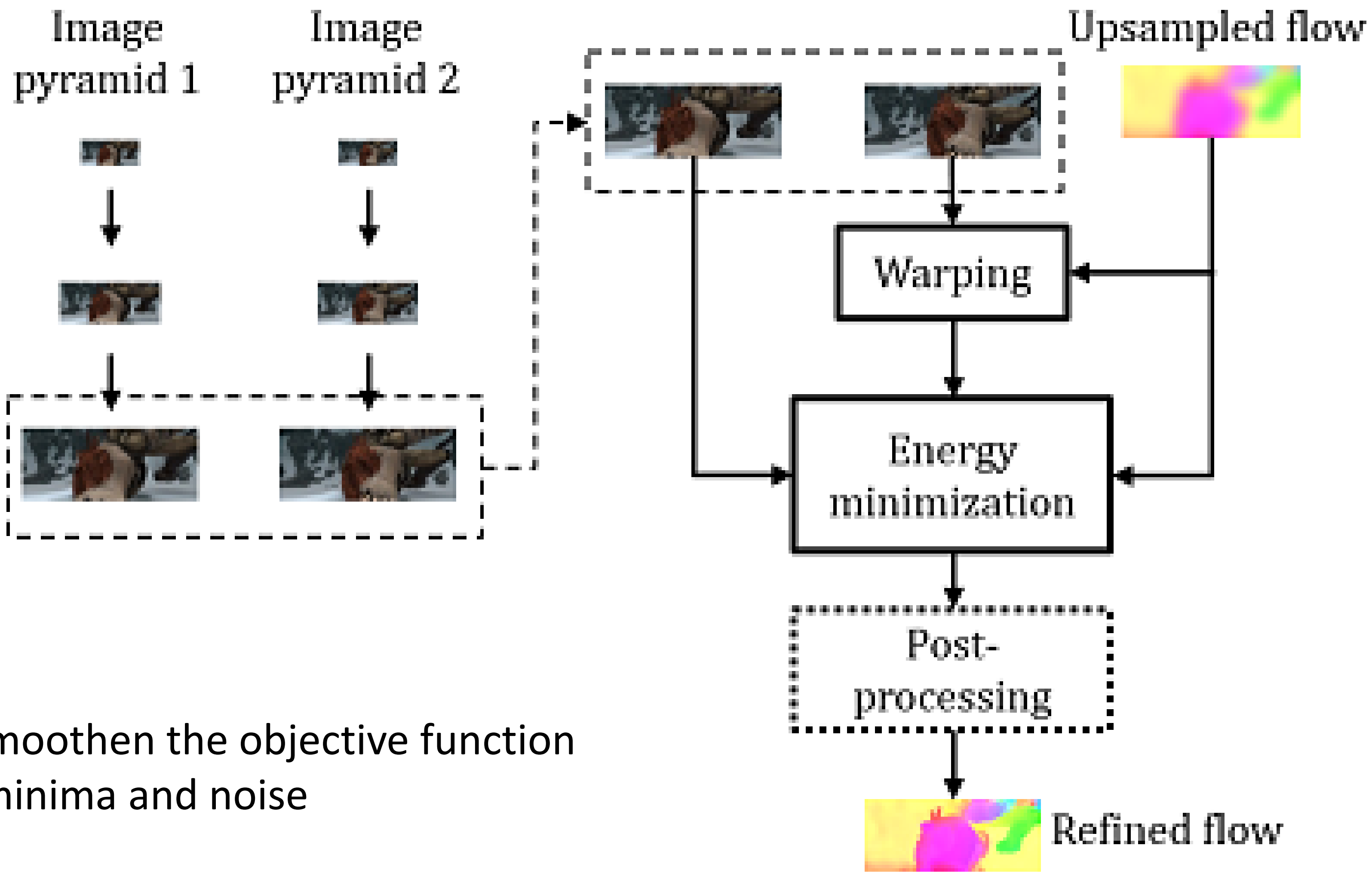
p = (x,y) ; w(p) = (u(p), v(p))

effective for small motion

fail when displacements are large

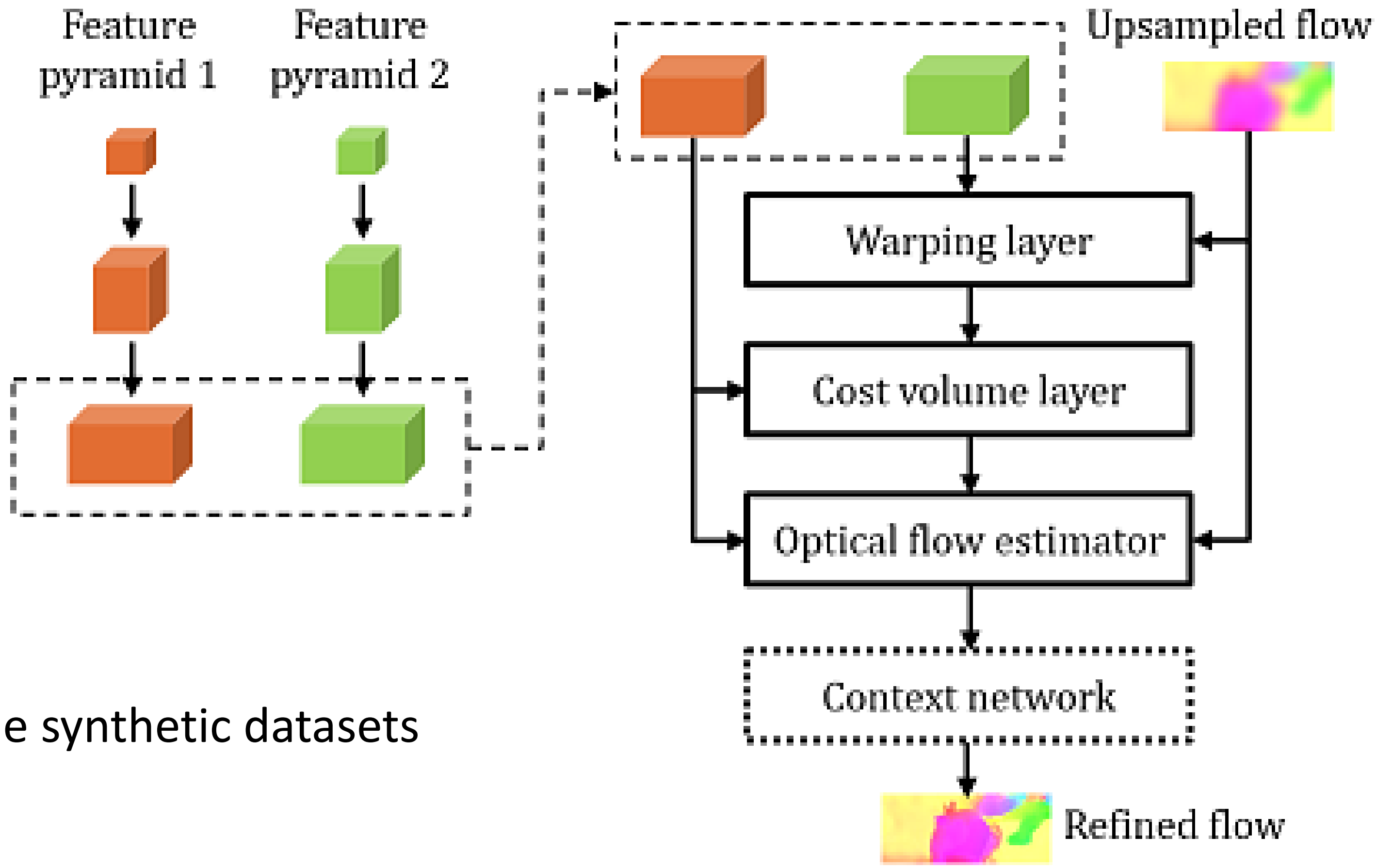
- Classical Optical Flow Estimation

  - Coarse to fine manner



Blurring can smoothen the objective function
Reduce local minima and noise

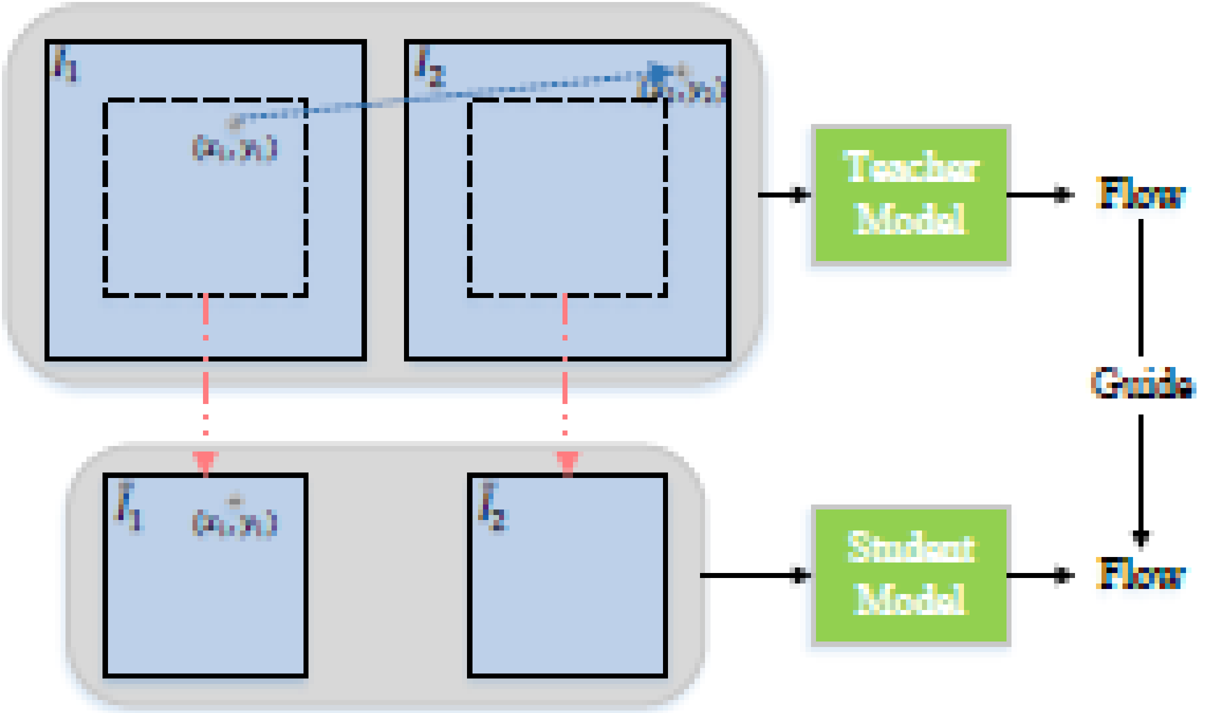
- Supervised Learning of Optical Flow
  - Warp features extracted from CNNs
- PWC-Net



pre-training on multiple synthetic datasets

PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume, CVPR 2018

- Unsupervised Learning of Optical Flow
  - Photometric loss (pixel-wised difference)
  - Does not hold for occluded pixels
- DDFlow
  - <u>Data distillation</u> approach to learning the optical flow of occluded pixels



DDFlow: Learning Optical Flow with Unlabeled Data Distillation

Methods

Problem

- Supervised methods requires a large amount of labeled training data, which is difficult to obtain for optical flow, especially when there are occlusions.

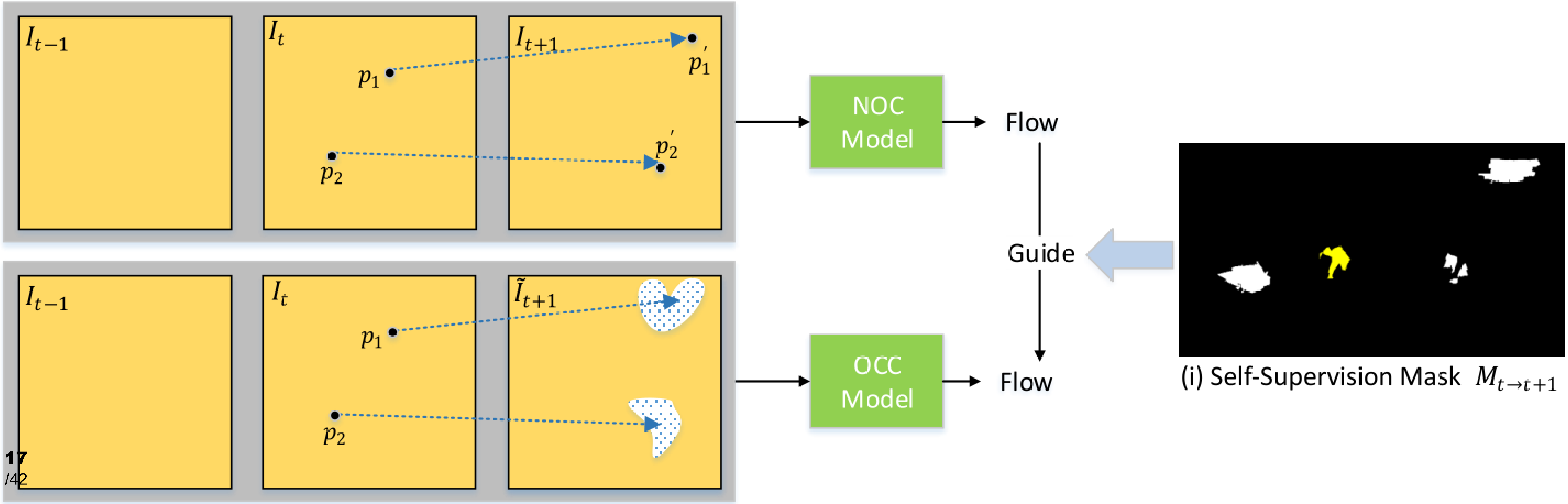
- Previous unsupervised learning methods only handle specific cases of occluded pixels. They lack the ability to reason about the optical flow of all possible occluded pixels.

Method

- Self-supervised learning
- Superpixel-based Occlusion Hallucination
- Multi-frame input

Self-Supervision

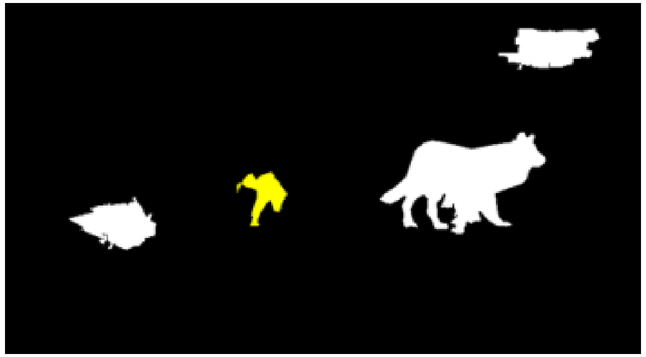
- Use the flow estimation of NOC-Model as annotations to guide OCC-Model



(i) Self-Supervision Mask $M_{t\to t+1}$

Notation

- I_t: image of frame t
- $w_{i\to j}$: flow from I_i to I_j
- $I^w_{j\to i}$: warping I_j to I_i with flow $w_{i\to j}$
- $O_{i\to j}$: occlusion map from $I_i$to I_j
- $\tilde{I}_t$: image with random noise
- $\widetilde{w}, \tilde{O}, \widetilde{I^w}$



(h) New Occlusion Map $\tilde{O}_{t\to t+1}$

(f) $\tilde{I}_{t+1}$

(a) Reference Image I_t

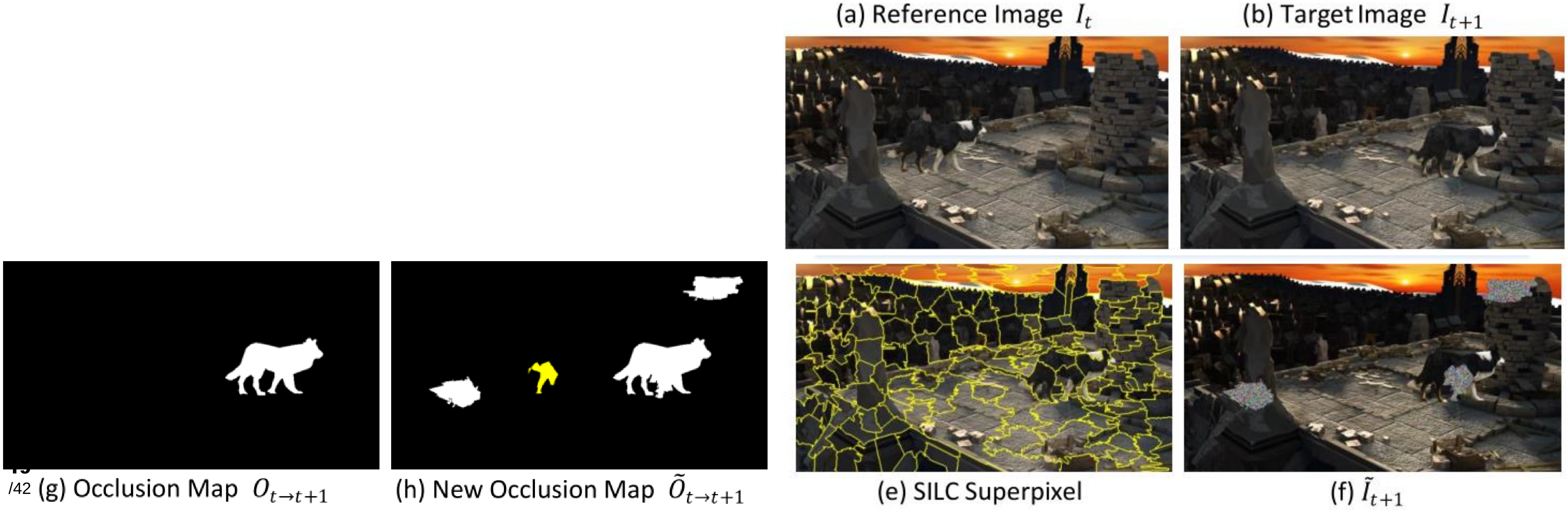
(b) Target Image I_{t+1}

(c) Ground Truth Flow $\mathbf{w}_{t\to t+1}$

(d) Warped Target Image $I^w_{t+1\to t}$

Occlusion Hallucination

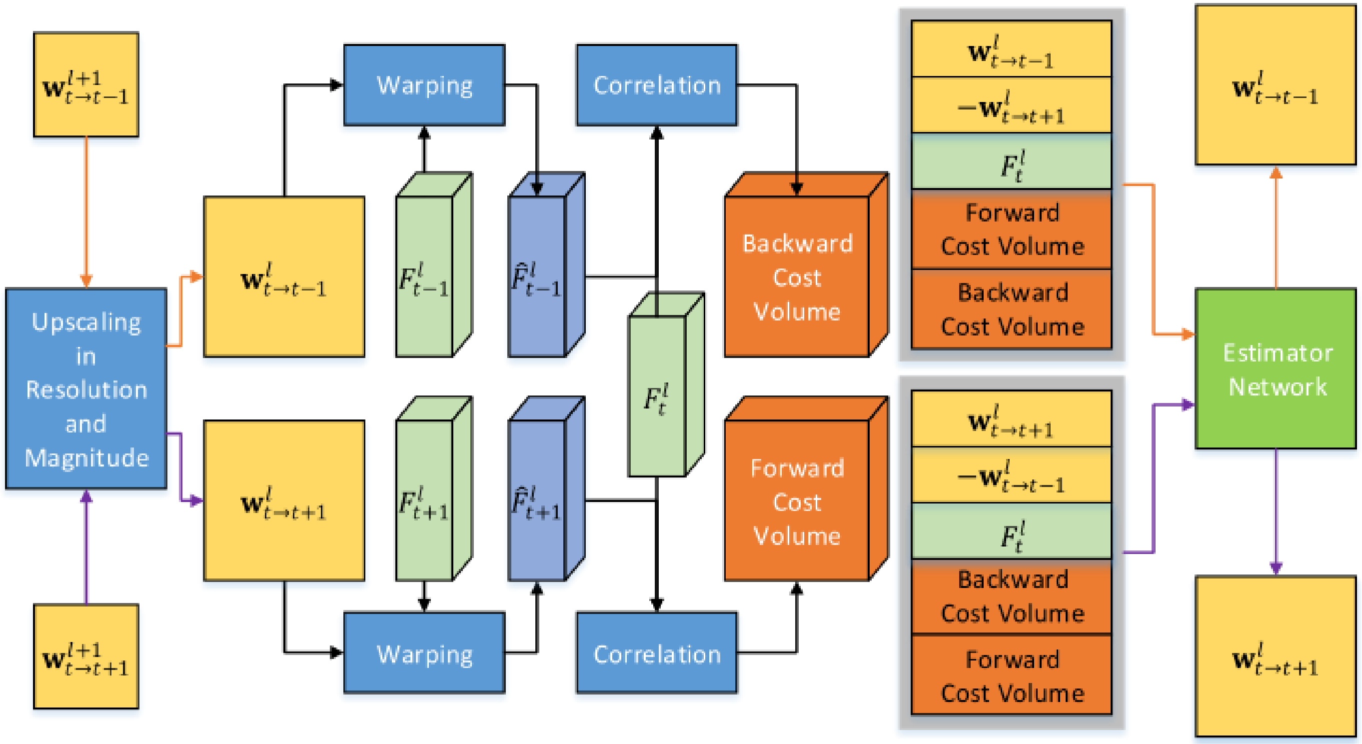
- 1. Generate superpixels;
- 2. Randomly select several superpixels and fill them with noise.



Network

- Based on PWC-Net
  - Pyramid network
  - Warping
  - Cost Volume

- Modifications
  - Three-frame input
  - Forward and backward flow

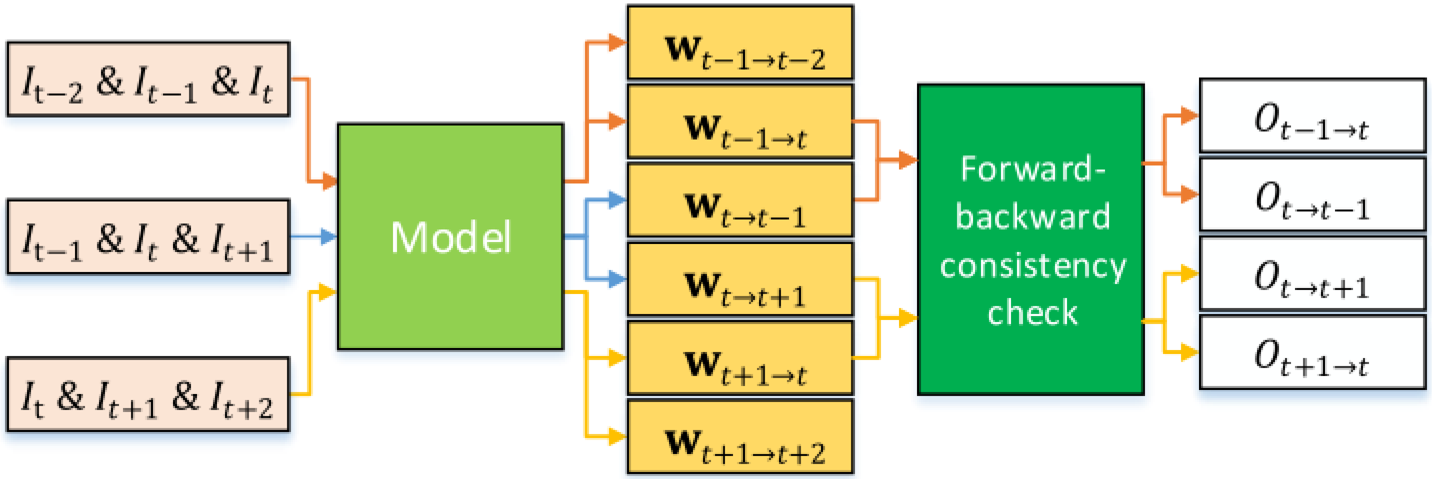


Occlusion Estimation

- Forward-backward consistency

- $\widehat{\boldsymbol{w}}_{t\to t+1} = \boldsymbol{w}_{t+1\to t}(\boldsymbol{p} + w_{t\to t+1}(\boldsymbol{p}))$

- $|\widehat{\boldsymbol{w}}_{t\to t+1} + \boldsymbol{w}_{t\to t+1}|^2 < \alpha_1(|\widehat{\boldsymbol{w}}_{t\to t+1}|^2 + |\boldsymbol{w}_{t\to t+1}|^2) + \alpha_2$



Loss Functions

- NOC: photometric loss
- $L_P = \sum_{i,j} \frac{\sum \psi\left(I_i - I_{j\to i}^{w}\right) \odot (1 - O_i)}{\sum (1 - O_i)}$
- Where $\psi(x) = (|x| + \epsilon)^q$

Loss Functions

- OOC: $L_O + L_P$
- $L_O = \sum_{i,j} \frac{\sum \psi(\boldsymbol{w}_{i\to j} - \widetilde{\boldsymbol{w}}_{i\to j}) \odot M_{i\to j}}{\sum M_{i\to j}}$
- $M_{i\to j} = clip(\tilde{O}_{i\to j} - O_{i\to j}, 0\,, 1)$

(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 $\tilde{O}_{t\to t+1}$

Supervised Fine-tuning

- Initialize with the pre-trained OCC-Model
- $L_s = \sum(\psi\left(w_{t\to t+1}^{gt} - w_{t\to t+1}\right)\odot V)/\sum V$
  - $w_{t\to t+1}^{gt}$ is ground truth flow
  - V denotes whether the pixel has a label

Experiments and Main Results

Datasets

- ~10,000 frames from Sintel
- Multi-view extensions from KITTI 2012 & 2015
- Rescale pixel values to [0,1] and normalize each channel to the standard normal distribution
- Census Transform
- Data Augmentation

Clean Path **Final Path**

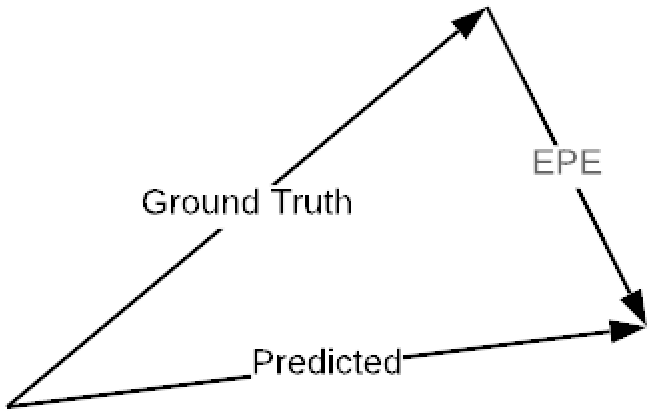
Sintel

KITTI

(Wulff *et al.*, CVPR 2012)

Evaluation Metrics

- Average EndPoint Error (EPE)



$$\|V_{est} - V_{gt}\|$$

- Percentage of Erroneous Pixels (Fl)

Outliers with the flow end-point error ≥ **3px or** ≥ **5%** (Uhrig *et al.*, 2017)

Quantitative Results

Method	Sintel Clean	Sintel Final	KITTI 2012	KITTI 2015
MODOF	–	0.48	–	–
OccAwareFlow	(0.54)	(0.48)	**0.95***	0.88*
MultiFrameOccFlow-Soft	(0.49)	(0.44)	–	**0.91***
DDFlow	**(0.59)**	**(0.52)**	0.94 *	0.86 *
Ours	**(0.59)**	**(0.52)**	**0.95** *	0.88*

$$\mathbf{F\ Measurement} = 2 \cdot \frac{\mathbf{Precision} \cdot \mathbf{Recall}}{\mathbf{Precision} + \mathbf{Recall}}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}$$

Asterisk denotes sparse occlusion annotation

Quantitative Results

	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%

The unsupervised results outperform several famous fully-supervised methods

Quantitative Results

	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%

For the first time, the supervised method achieved high performance without any external data

Quantitative Results

	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

Quantitative Results

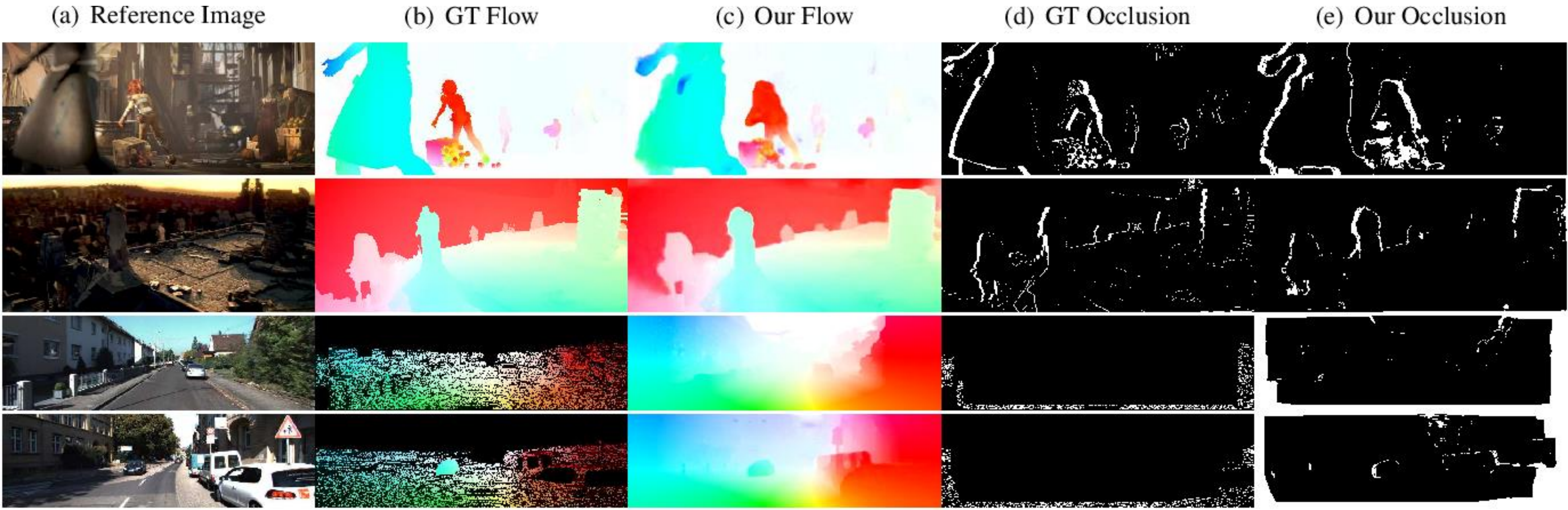
Occlusion Handling	Multiple Frame	Self-Supervision Rectangle	Self-Supervision Superpixel	Sintel Clean			Sintel Final			KITTI 2012			KITTI 2015		
				ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC	ALL	NOC	OCC
✗	✗	✗	✗	(3.85)	(1.53)	(33.48)	(5.28)	(2.81)	(36.83)	7.05	1.31	45.03	13.51	3.71	75.51
✗	✓	✗	✗	(3.67)	(1.54)	(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
✓	✓	✗	✓	**(2.88)**	**(1.30)**	**(22.06)**	**(3.87)**	**(2.24)**	**(25.42)**	**1.69**	**0.91**	**6.95**	**4.84**	**2.40**	**19.68**

Table 2. Ablation study. We report EPE of our unsupervised results under different settings over all pixels (ALL), non-occluded pixels (NOC) and occluded pixels (OCC).

Unsupervised Pre-training	Sintel Clean	Sintel Final	KITTI 2012	KITTI 2015
Without	1.97	2.68	3.93	3.10
With	**1.50**	**2.41**	**1.55**	**1.86**

Table 3. Ablation study. We report EPE of supervised fine-tuning results on our validation datasets with and without unsupervised pre-training.

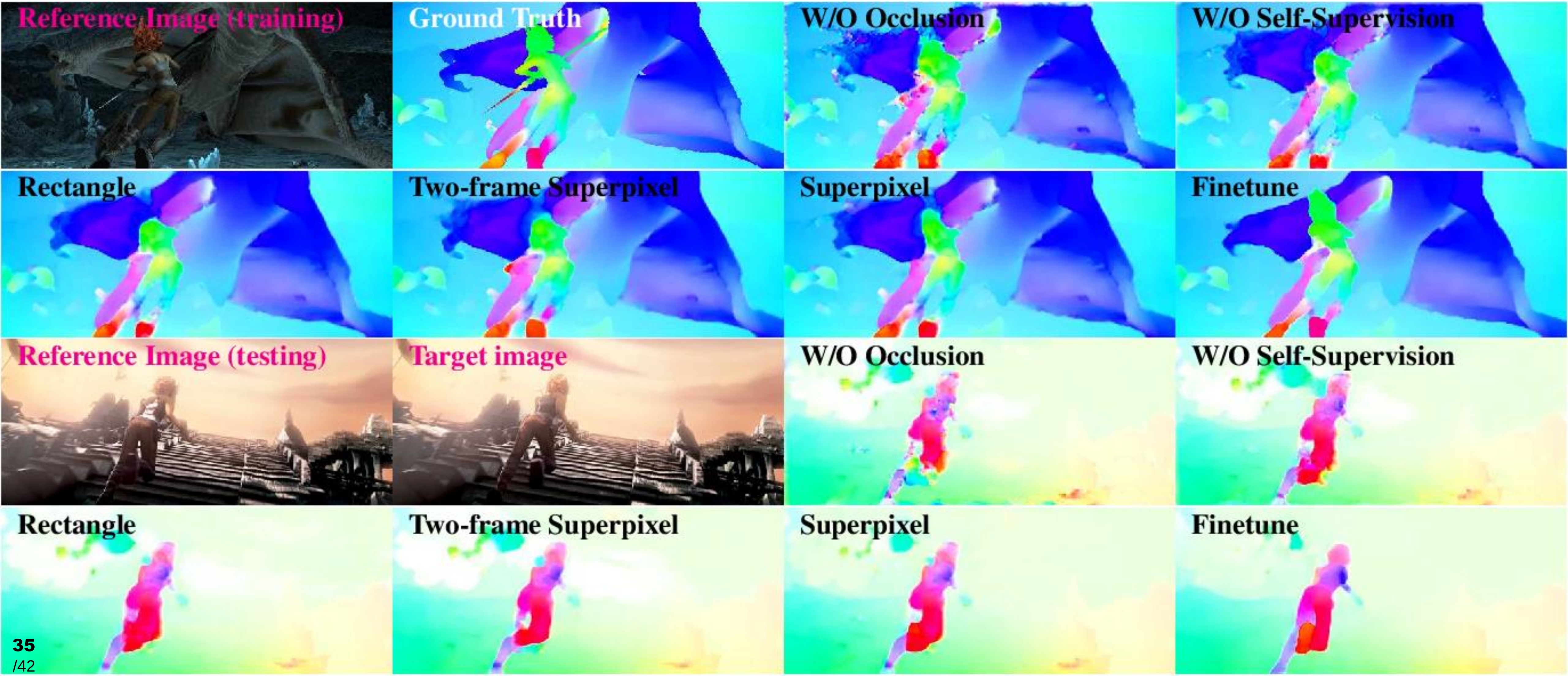
Qualitative Results



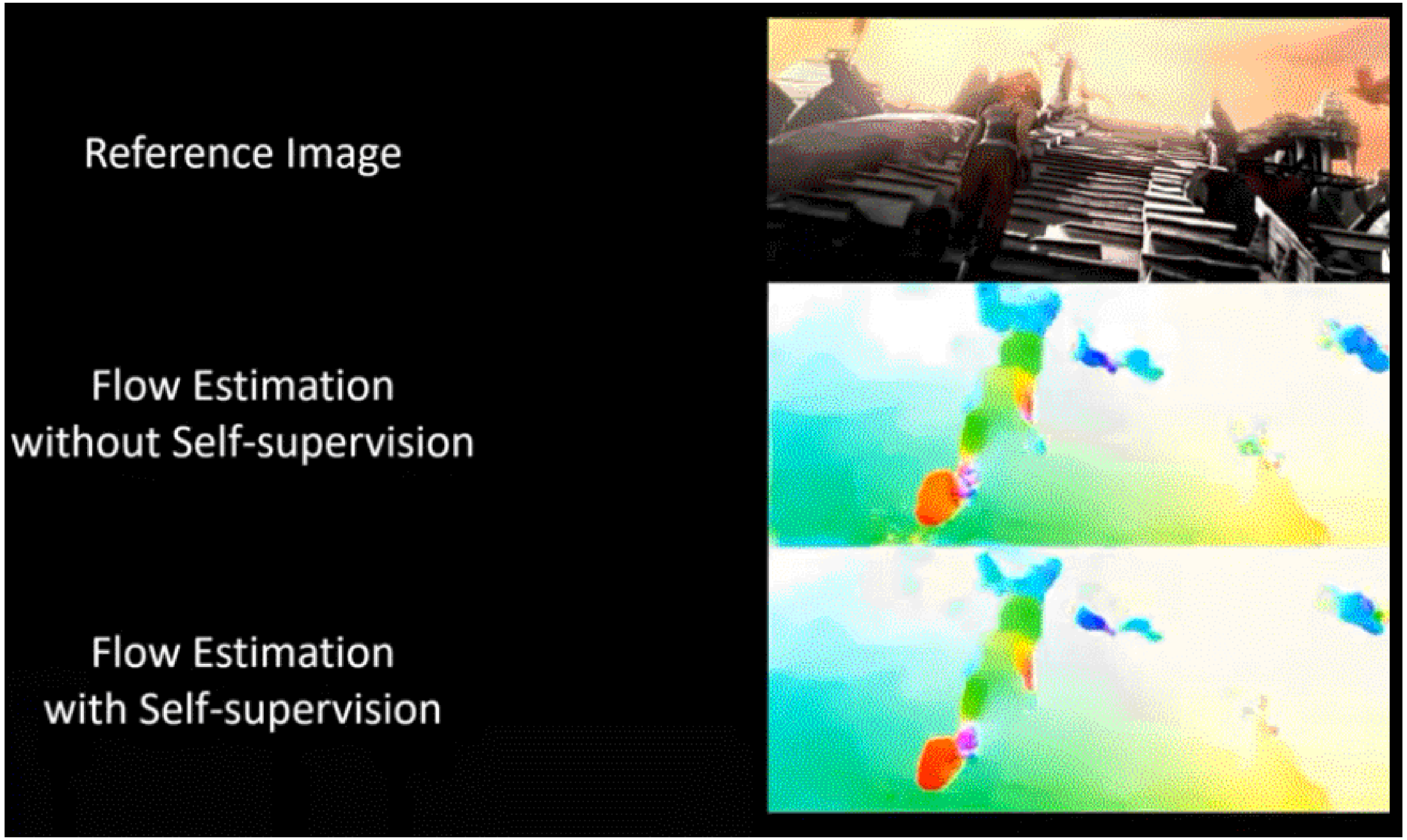
Qualitative Results

DAVIS dataset

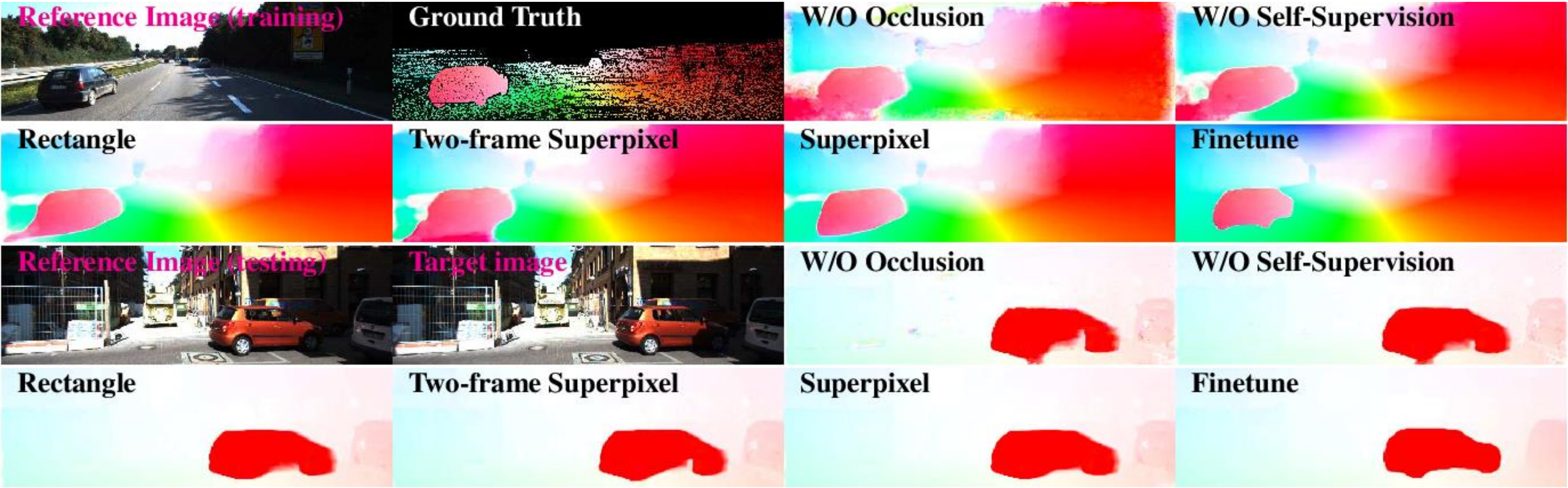
Qualitative Results (Sintel Datasets)



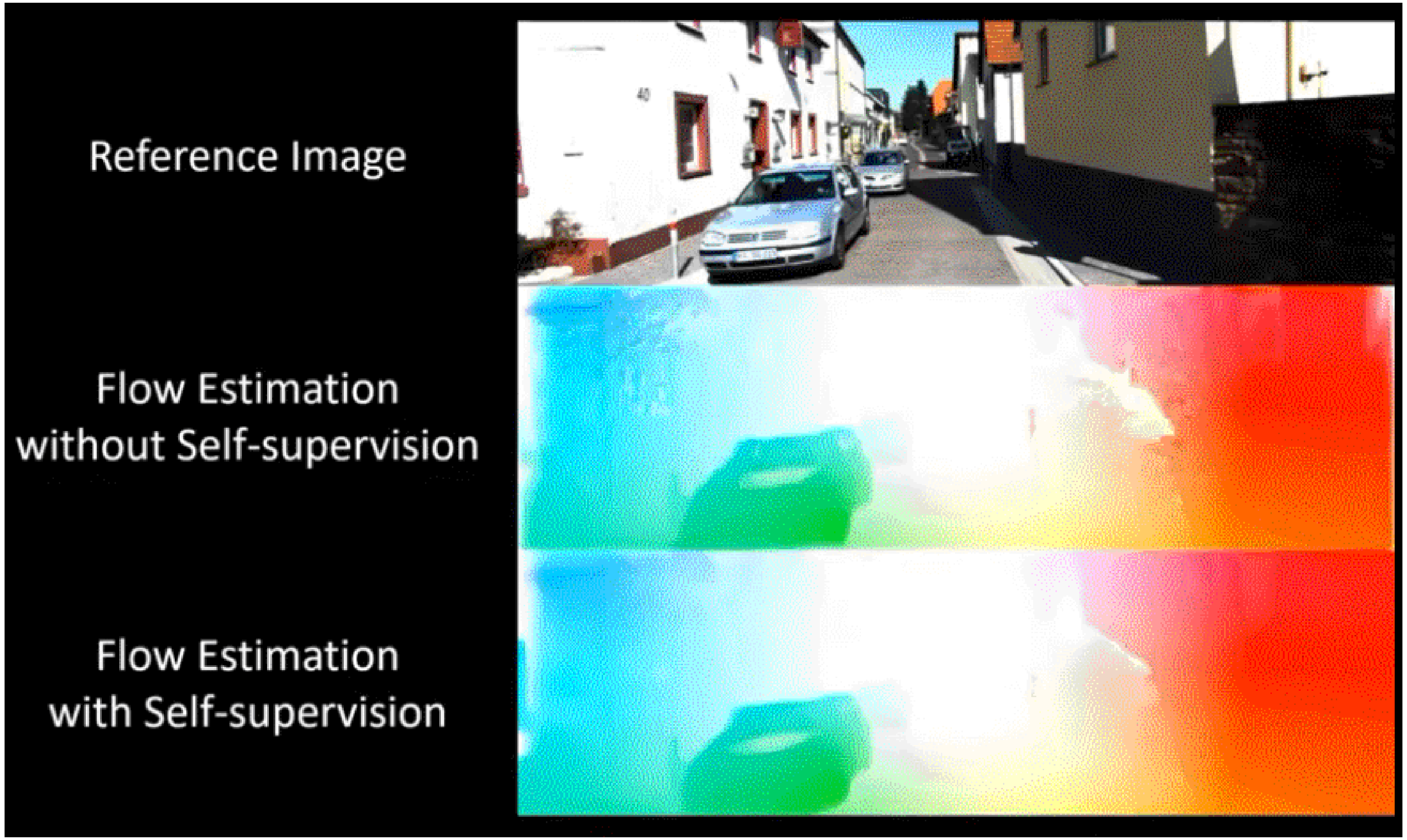
Qualitative Results (Sintel Datasets)



Qualitative Results (KITTI Datasets)



Qualitative Results (KITTI Datasets)



Comparison with PWC-Net

(Pengpeng Liu, CVPR 2019 Oral Presentation)

Conclusions

- A self-supervised approach to learning accurate optical flow for both occluded and non-occuluded pixels
- The method achieves state-of-art results on KITTI and Sintel benchmarks

(Pengpeng Liu, CVPR 2019 Oral Presentation)

Strengths

- Effectively aggregates temporal information from multiple frames to improve flow prediction.
- Significantly outperforms all existing unsupervised optical flow learning methods.
- Presents the potential of completely reduce the reliance of pre-training on synthetic labeled data

Weakness

- In terms of occlusion estimation only, the noise injection method does not seem to make a difference when compared with DDFlow.

Thank You