香港中文大學
The Chinese University of Hong Kong

Self-Supervised Learning of Dense Correspondence

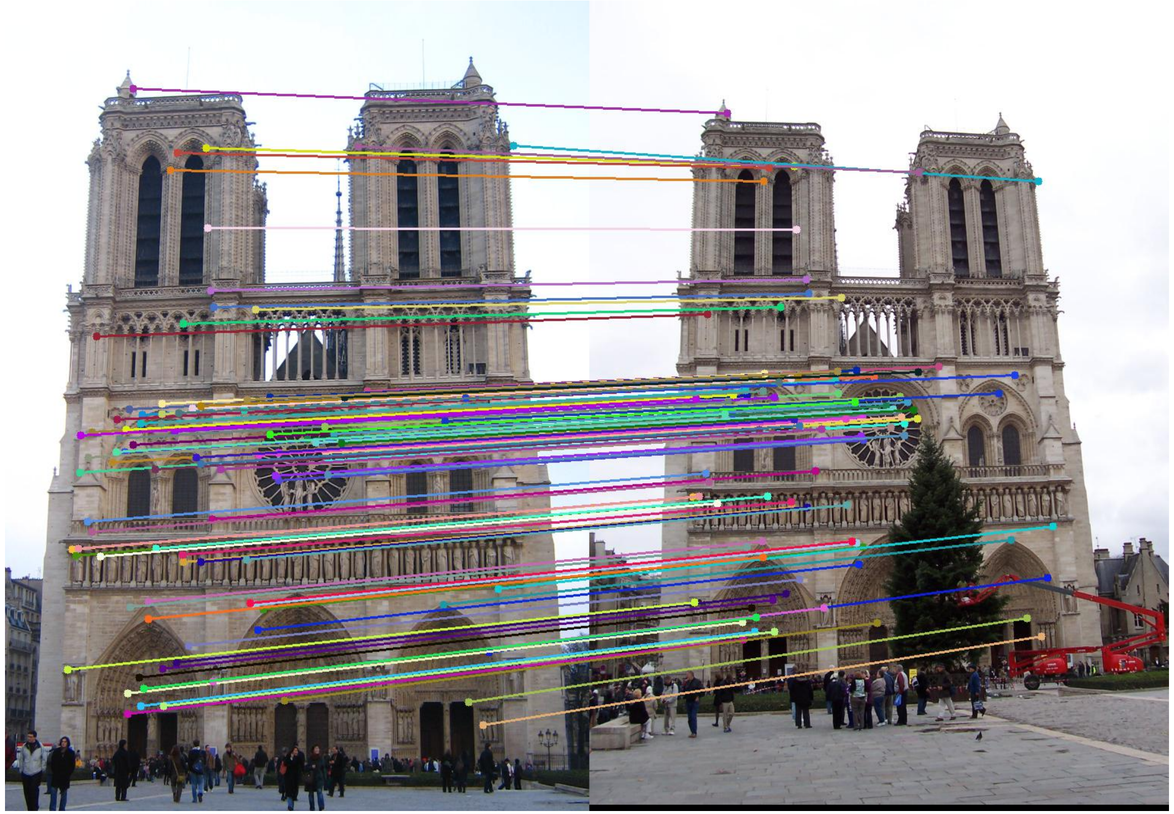
LIU, Pengpeng

Ph.D. Oral Defense

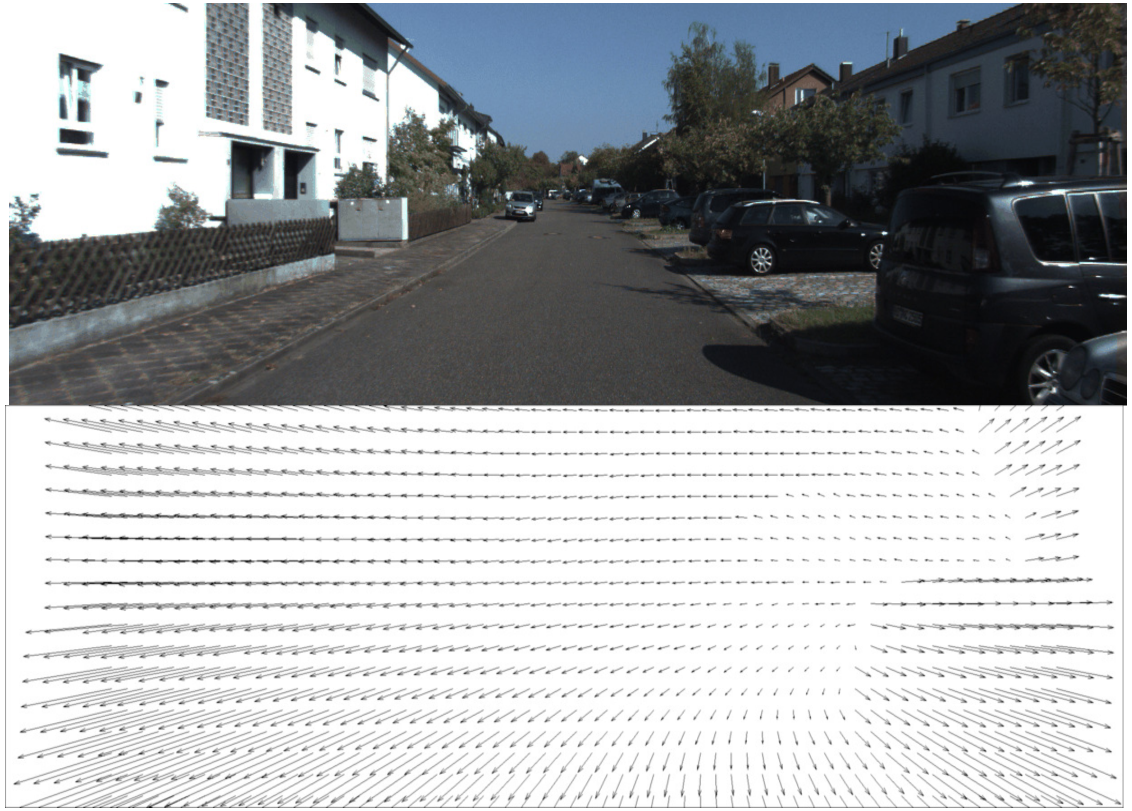
Supervisors: Prof. Michael R. Lyu and Prof. Irwin King

2020/11/19

Correspondence is a Matching Problem



Sparse Correspondence

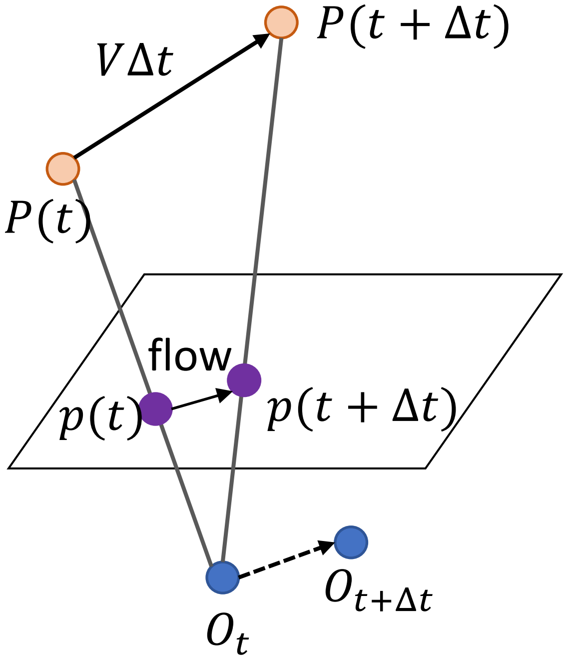


Dense Correspondence

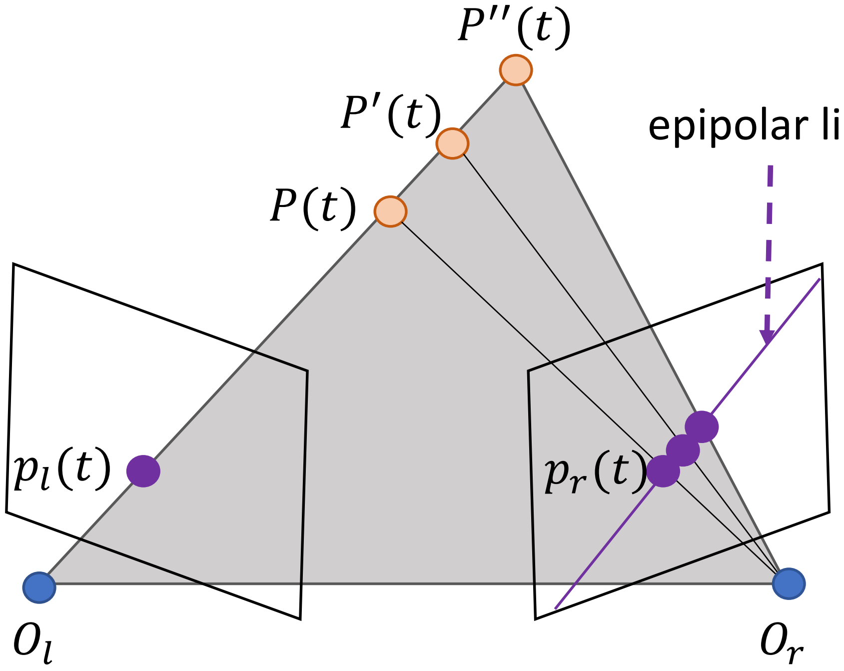
The three fundamental problems of computer vision are: "Correspondence, correspondence, and correspondence!" --- *Takeo Kanade*

Dense Correspondence Tasks

- Optical flow and stereo matching



Flow Geometry



Stereo Geometry

Relative locations and orientations of the cameras are **not fixed**: 2D matching

Relative locations and orientations of the cameras are **fixed**: 1D matching

Stereo matching can be regarded as a special case of optical flow.

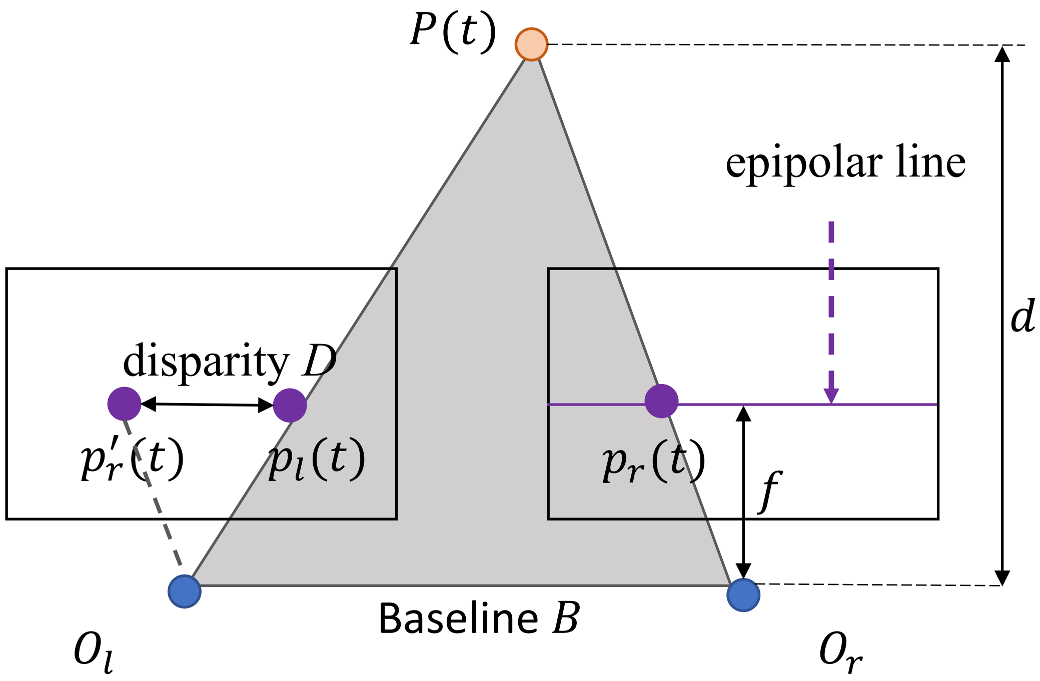
Correspondence is Crucial

- Optical flow: motion analysis

Image Sequences

Optical Flow

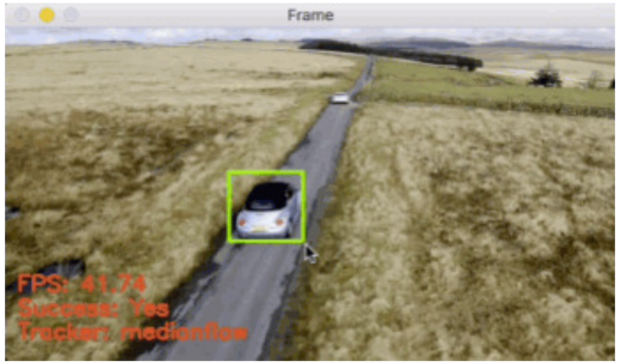
- Stereo matching: 3D understanding



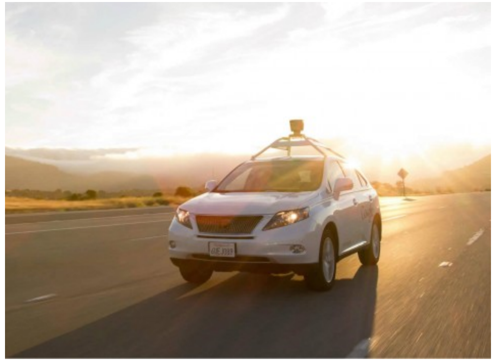
Depth $d = fB/D$.
Disparity is inversely proportional to depth!

Correspondence is Everywhere

Image Stitching

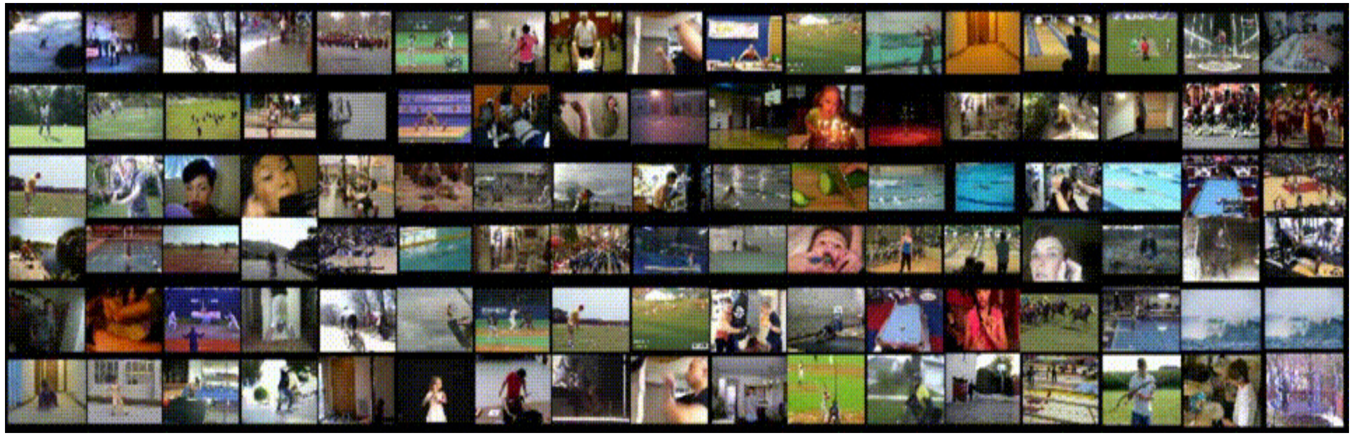


Object Tracking



Autonomous Driving

3D Reconstruction



Video Action Recognition

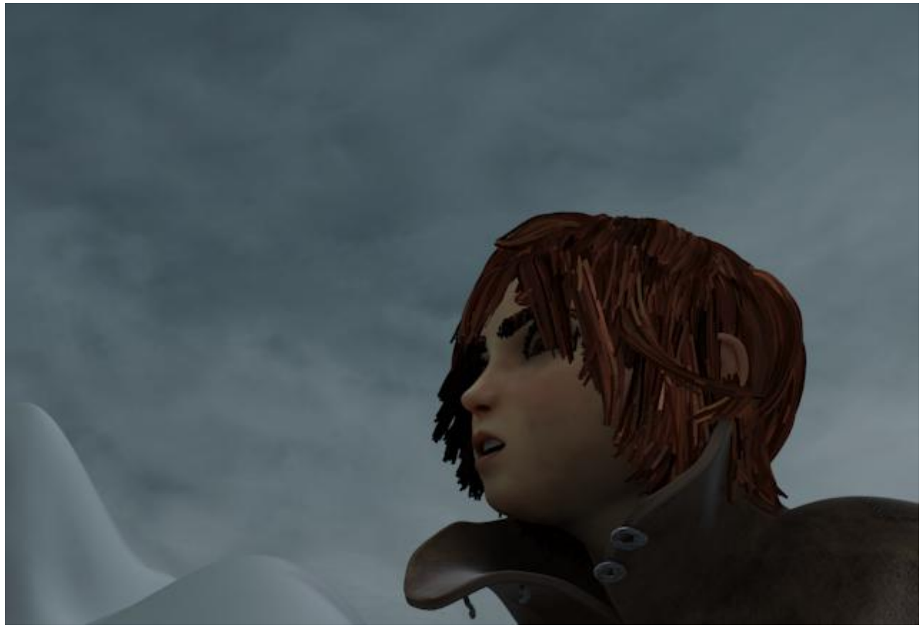
Correspondence Estimation is Challenging

- Occlusion

Where is the finger in the right image?

Correspondence Estimation is Challenging

- Illumination change



The right image is darker due to underexposure.

Correspondence Estimation is Challenging

- Motion blur and atmospheric effects

Object boundaries are blurry.

Correspondence Estimation is Challenging

- Hard to obtain ground truth

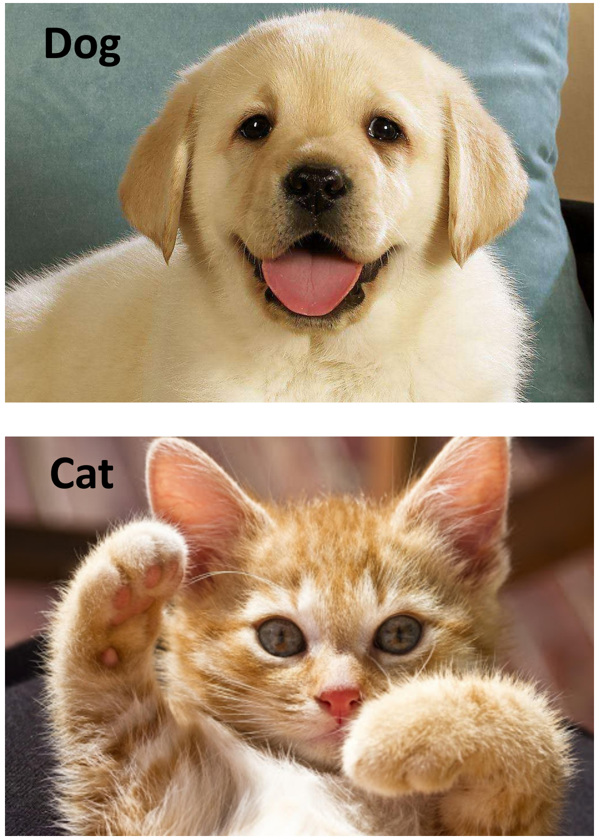


Image Classification

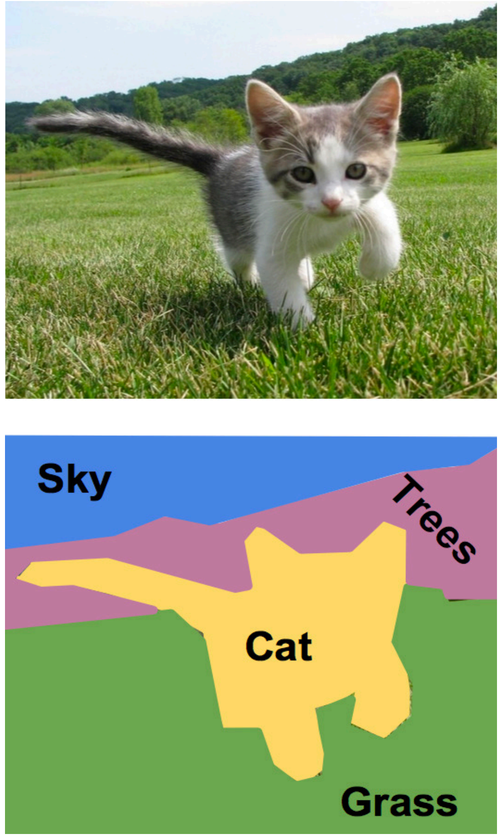
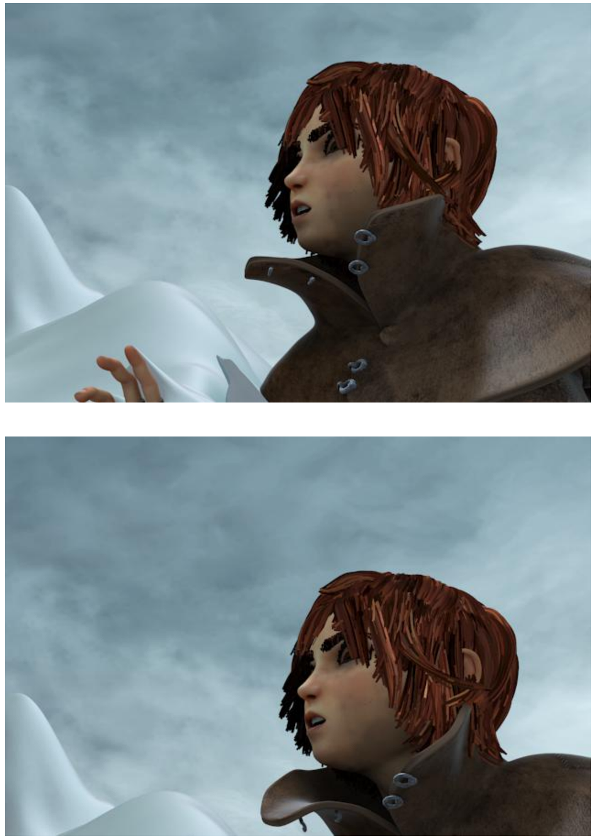


Image Segmentation



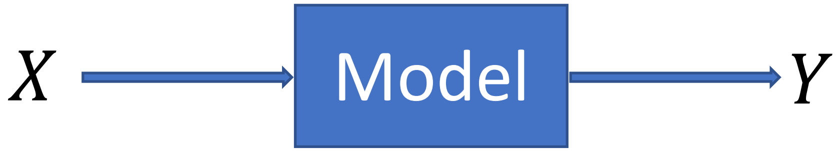
Optical Flow

Can you label the correspondence of each pixel between these two images?

Hard to Collect Dense Correspondence Labels

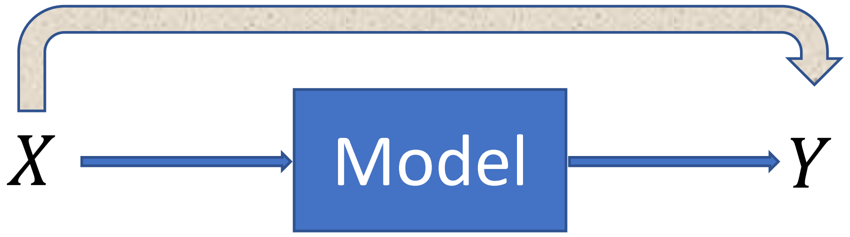
We aim to design self-supervised learning methods to learn dense correspondence from unlabeled data.

Self-Supervised Learning



Supervised Learning

Pretext task:
automatically generate



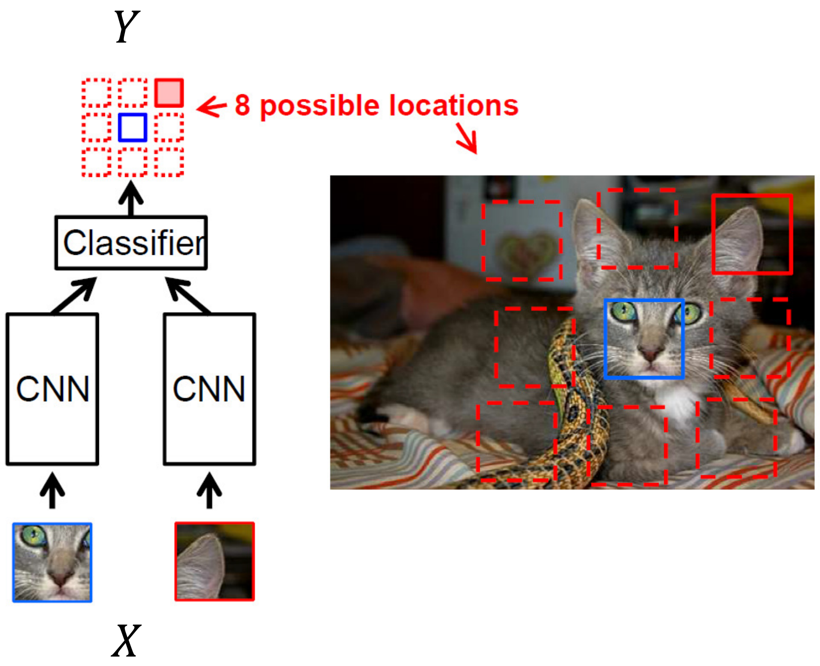
Self-Supervised Learning

Definition: a form of unsupervised learning where the supervision signal is purely generated from the data itself.

Self-Supervised Learning

- Pretext task: image inpainting, image colorization, image super-resolution, order prediction, video frame prediction, etc

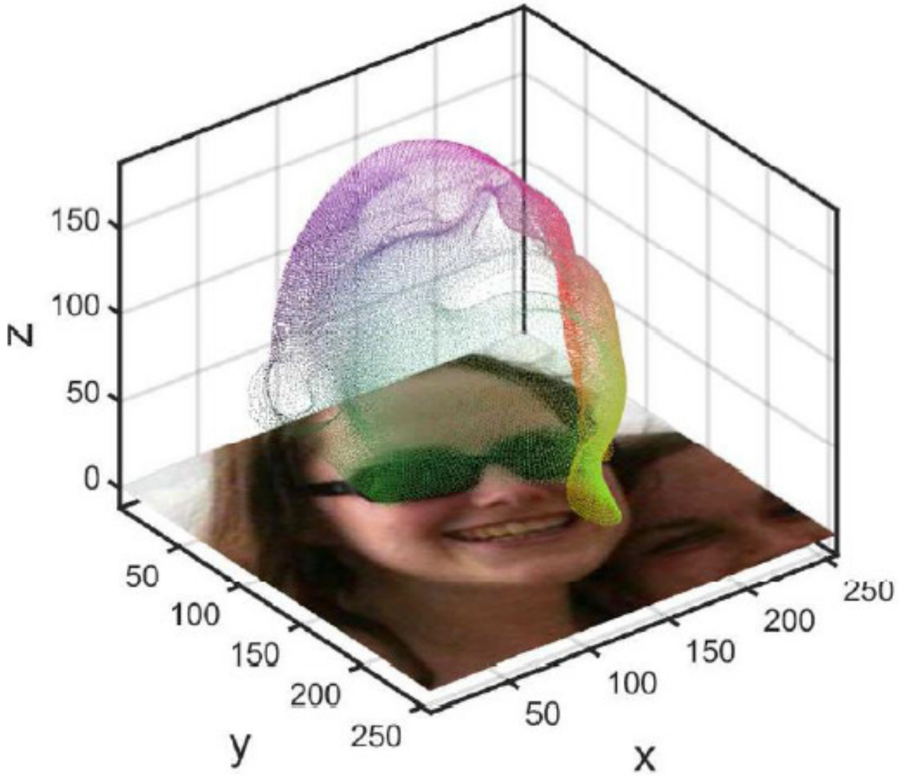
Image Inpainting



Relative Position Prediction

3D Face Reconstruction

- 3D face reconstruction: a special case of dense correspondence



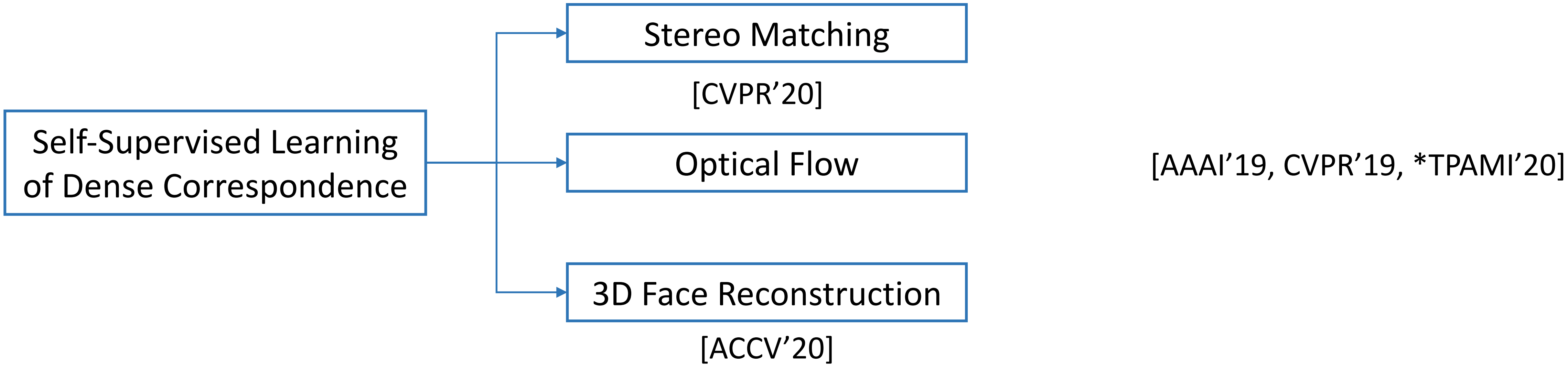
Dense correspondence between a 2D face image and a 3D face model

Learn 3D face reconstruction from videos and employ optical flow as a 2D constraint.

3D face reconstruction can be regarded as an application of optical flow.

3D Face Reconstruction can be regarded as an application of optical flow.

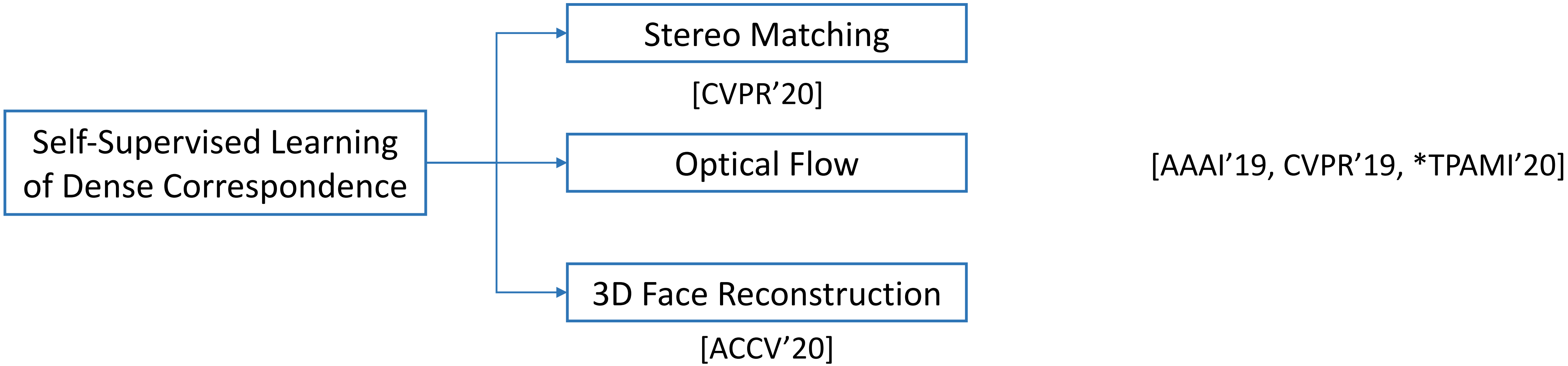
Thesis Contributions



- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels.
- Stereo Matching: explore the geometric relationship between flow and stereo.
- 3D Face Reconstruction: pose guidance network and multi-image consistency.

* In Submission

Thesis Contributions



- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D Face Reconstruction: pose guidance network and multi-image consistency

* In Submission

Thesis Contributions

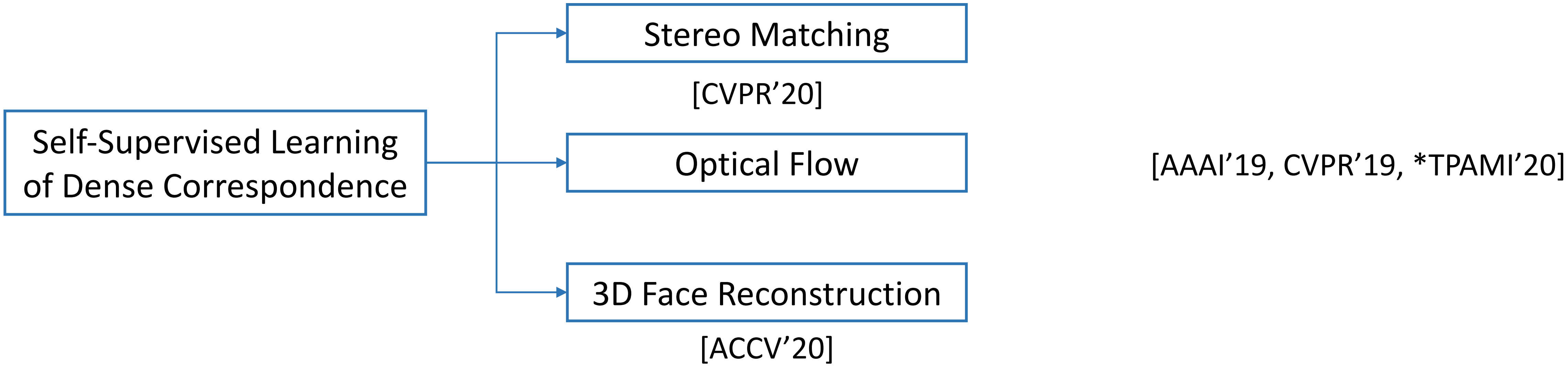
Self-Supervised Learning of Dense Correspondence

- Stereo Matching [CVPR’20]
- Optical Flow [AAAI’19, CVPR’19, *TPAMI’20]
- 3D Face Reconstruction [ACCV’20]

- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D Face Reconstruction: pose guidance network and multi-image consistency

* In Submission

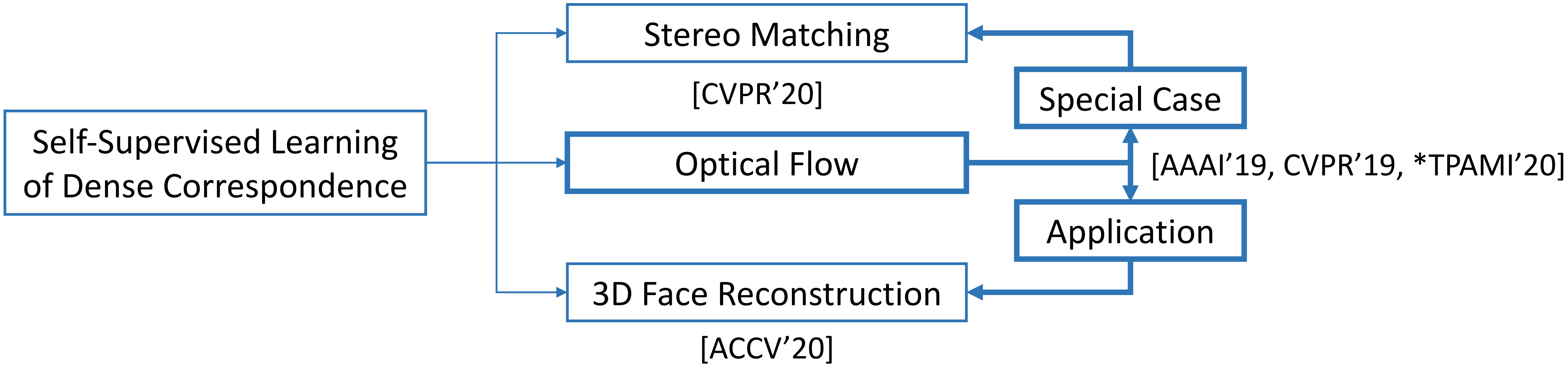
Thesis Contributions



- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D Face Reconstruction: pose guidance network and multi-image consistency

* In Submission

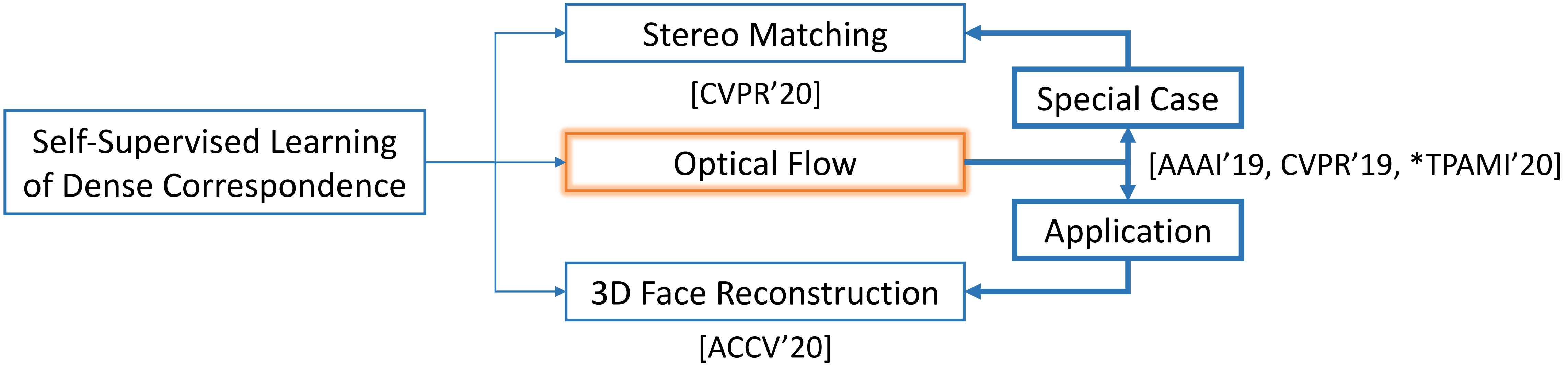
Thesis Contributions



- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D face reconstruction: pose guidance network and multi-image consistency

* In Submission

Thesis Contributions



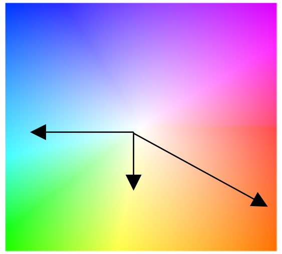
- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D face reconstruction: pose guidance network and multi-image consistency

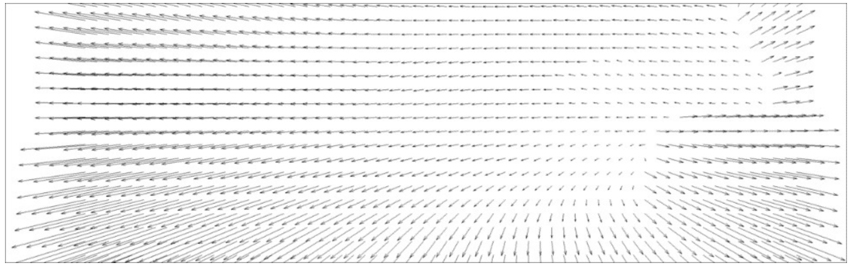
Optical flow and its applications

* In Submission

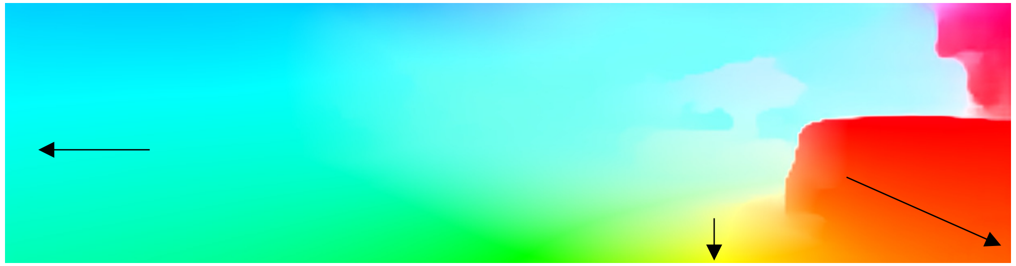
Optical Flow: Task Definition

Color Coding: hue denotes the direction of the motion, and saturation denotes the magnitude of the motion.



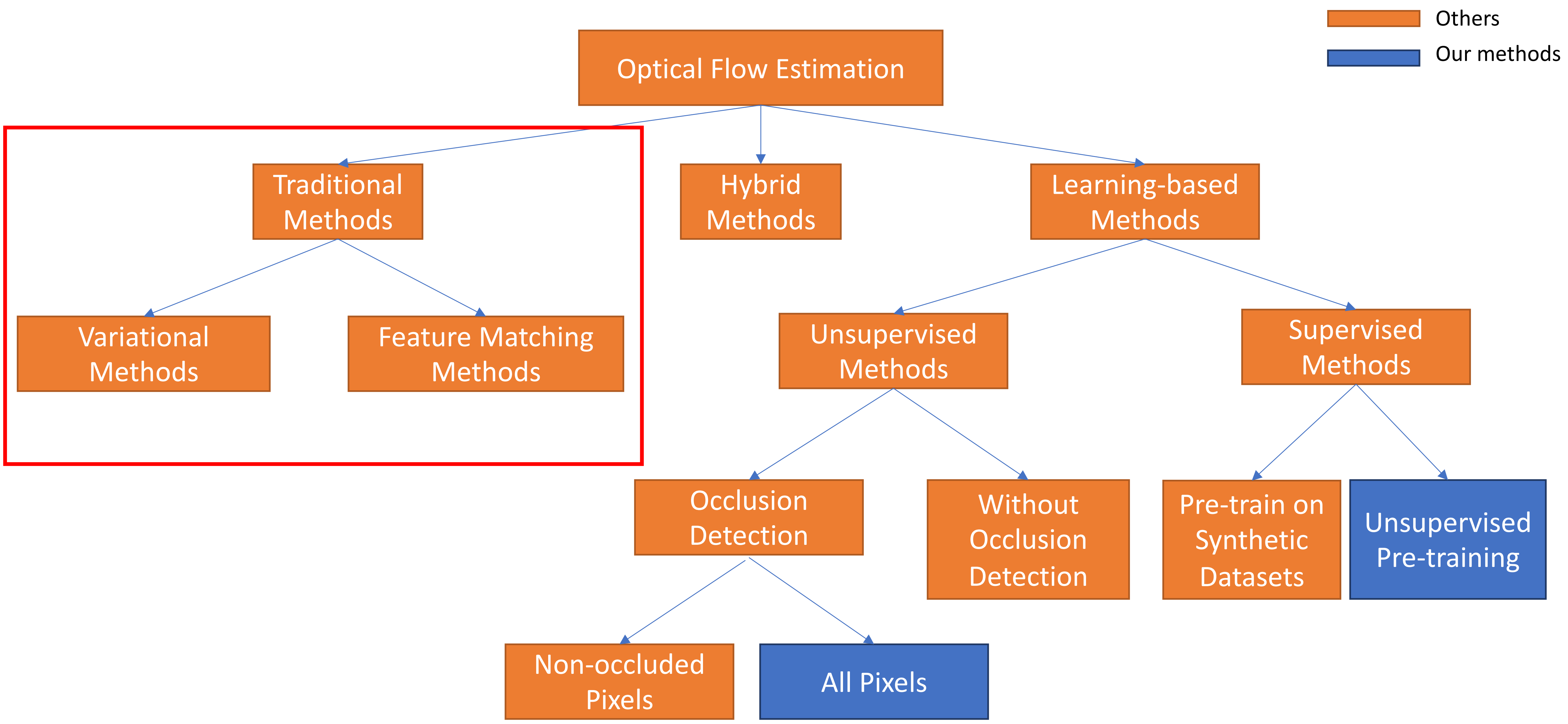


Optical flow represented with arrow



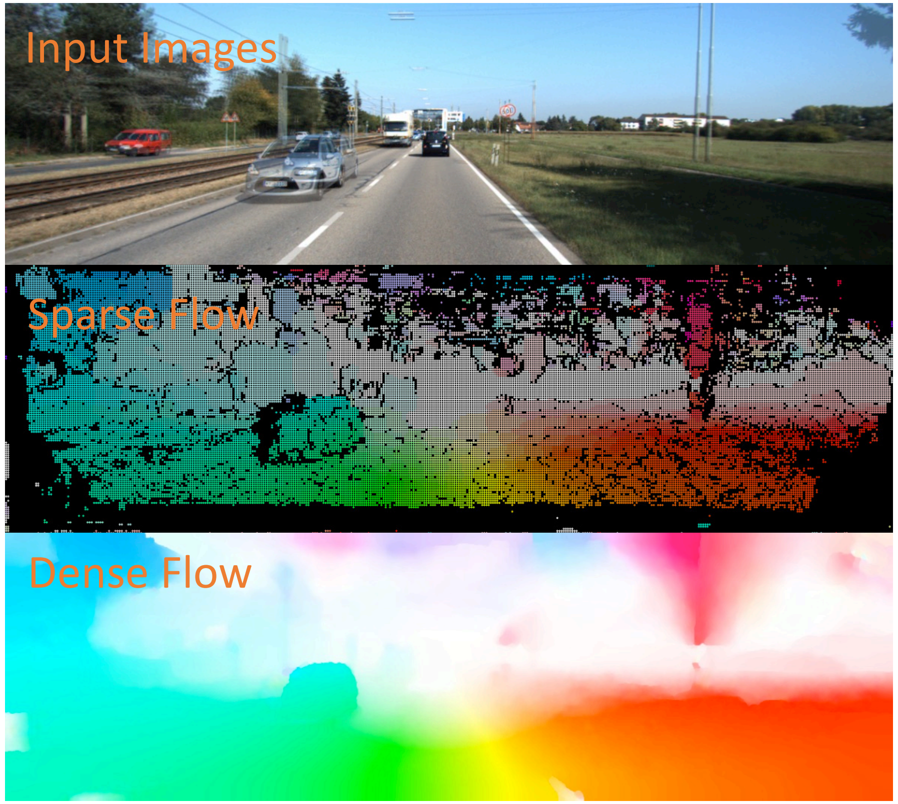
Optical flow represented with color coding

Background Review

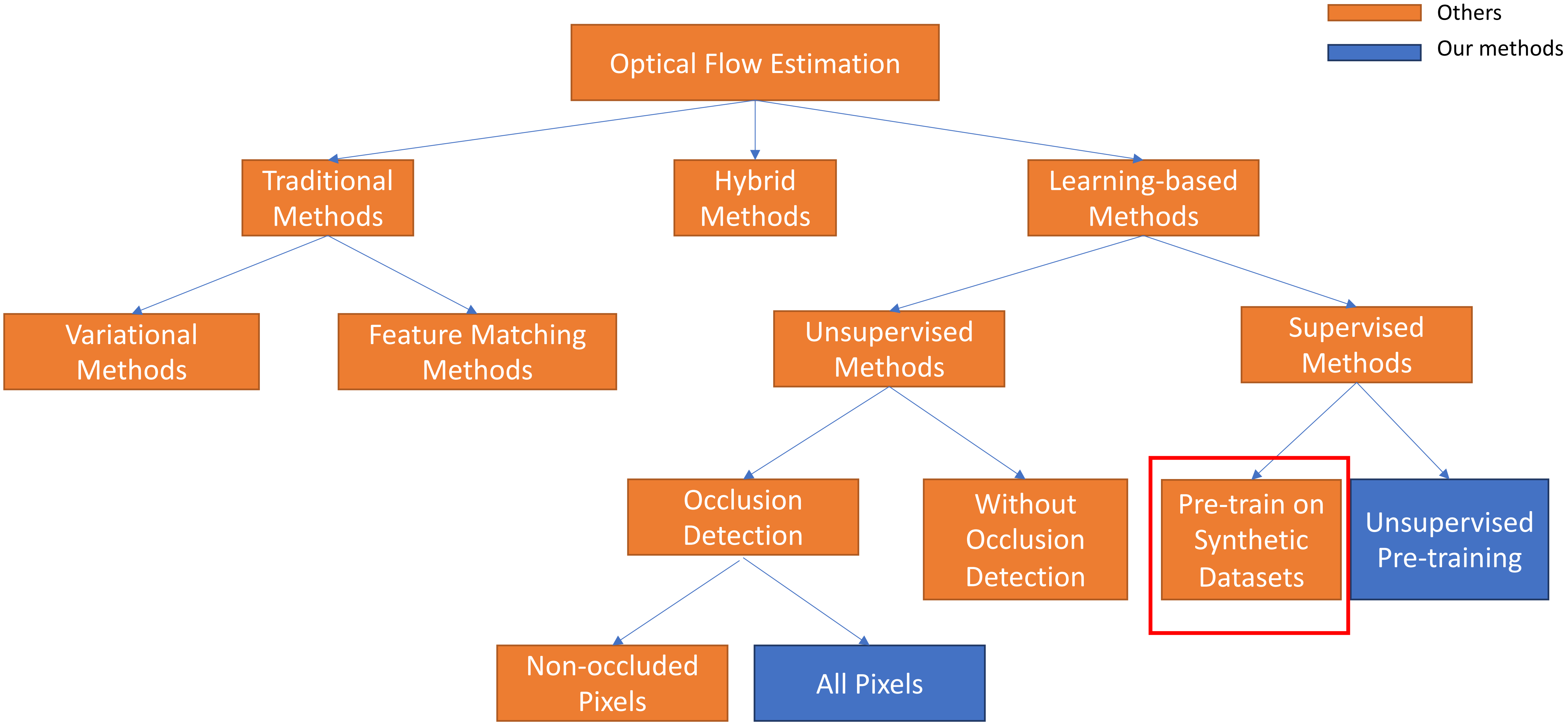


Traditional Methods

- Variational approaches: coarse-to-fine optical flow estimation
- Feature matching: sparse to dense
- Disadvantages: slow, not work well for large motion

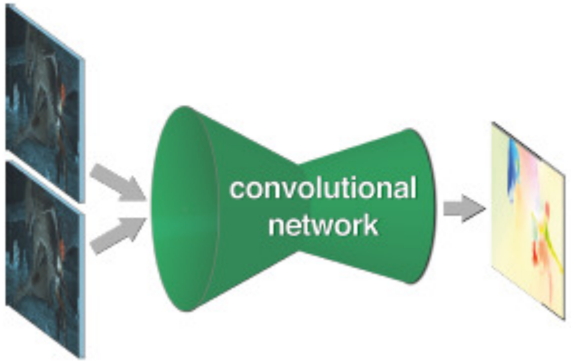


Background Review

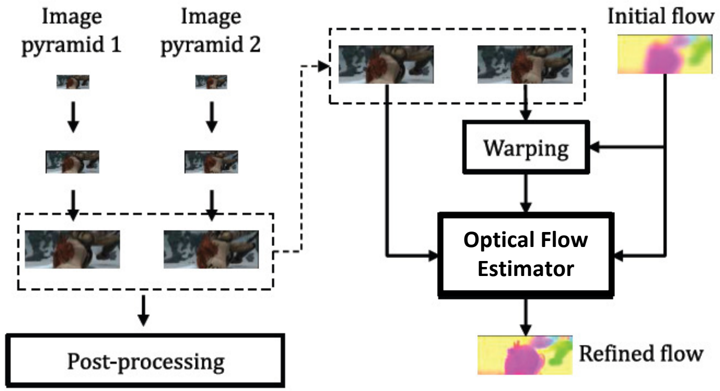


Supervised Learning Methods

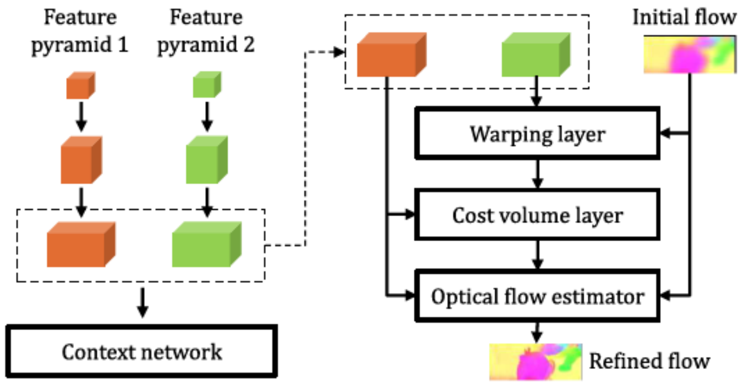
- Input two images, output a dense optical flow map with CNNs
  - FlowNet [Dosovitskiy et al. CVPR 2015]
  - FlowNet 2.0 [Ilg et al. CVPR 2017]
  - SpyNet [Ranjan et al. CVPR 2017]
  - PWC-Net [Sun et al. CVPR 2018]



FlowNet



SpyNet



PWC-Net

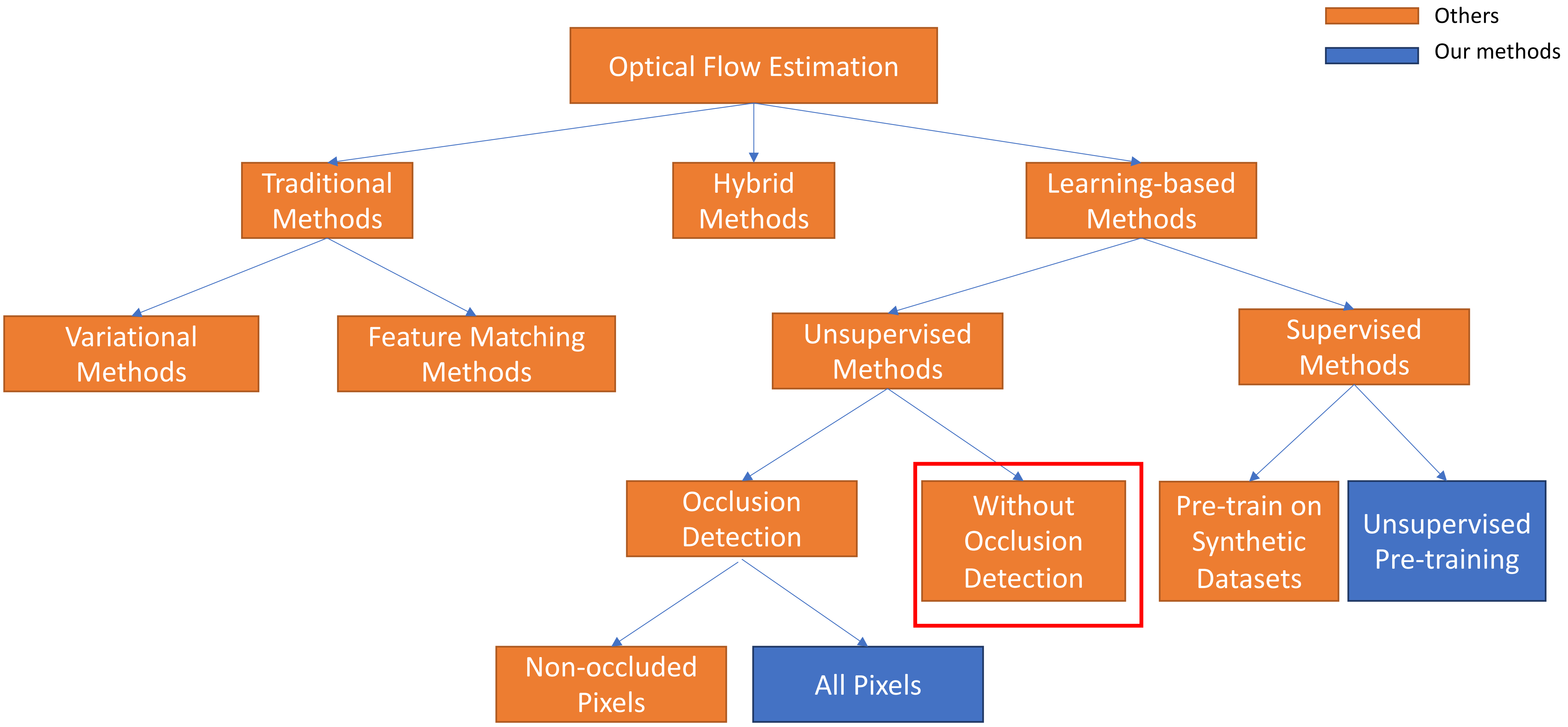
Supervised Learning Methods

- Advantages: high performance, high speed
- Disadvantages: need a large amount of labeled data → difficult to obtain
  ➔ pre-train on synthetic data ➔ domain gap

Training domains

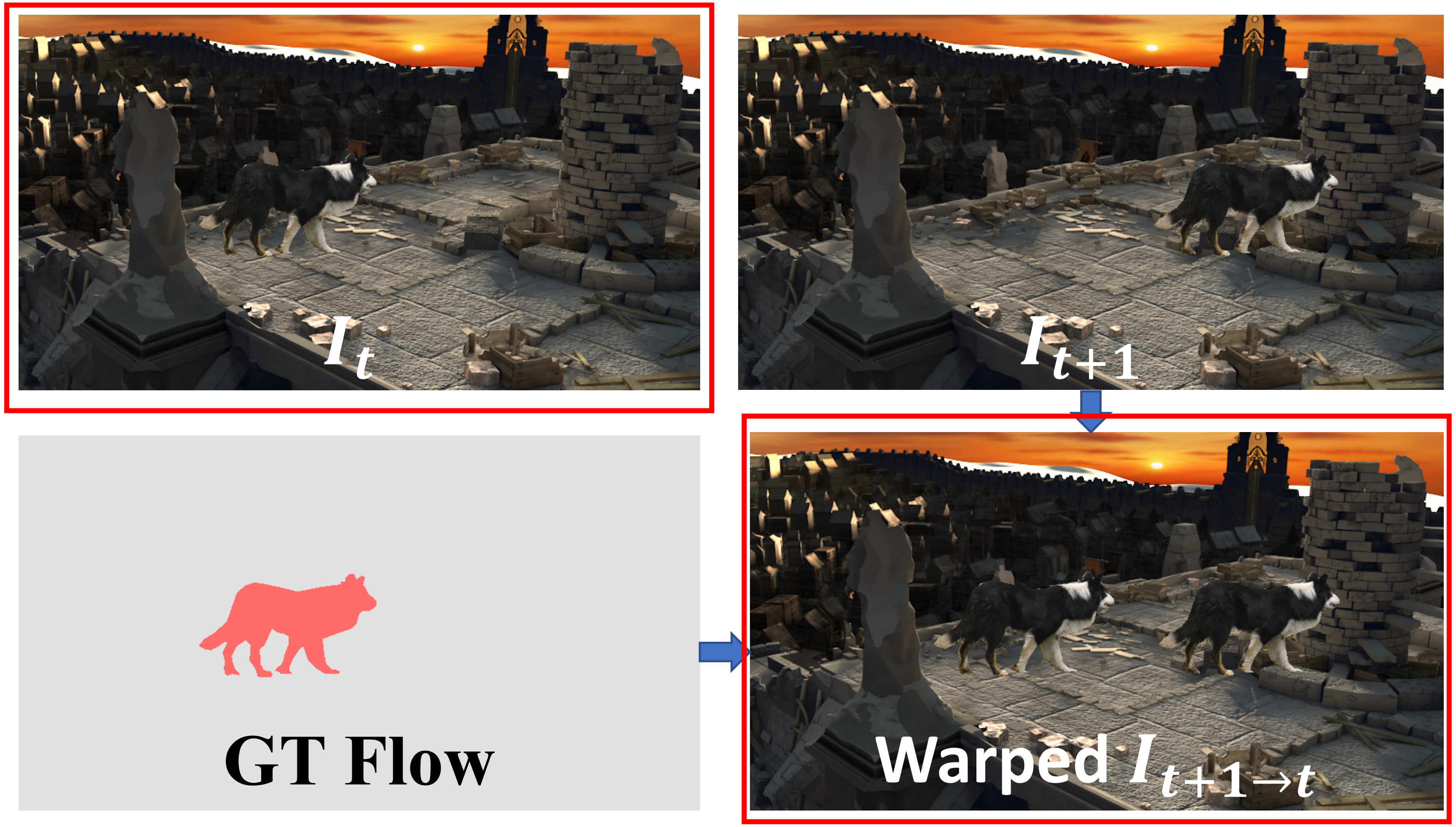
Domains of interest

Background Review



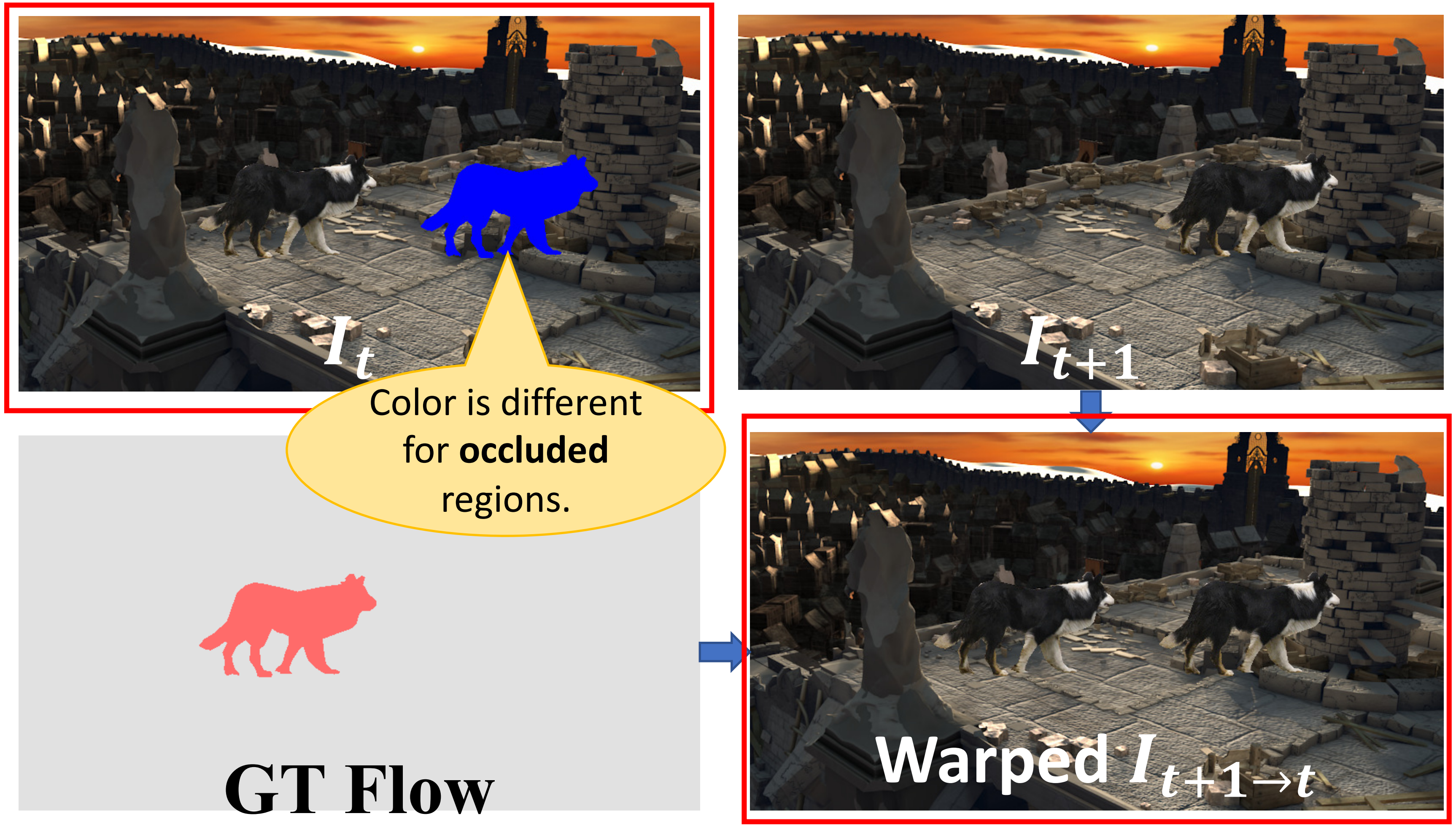
Unsupervised Learning Methods

- Advantage: infinite training data

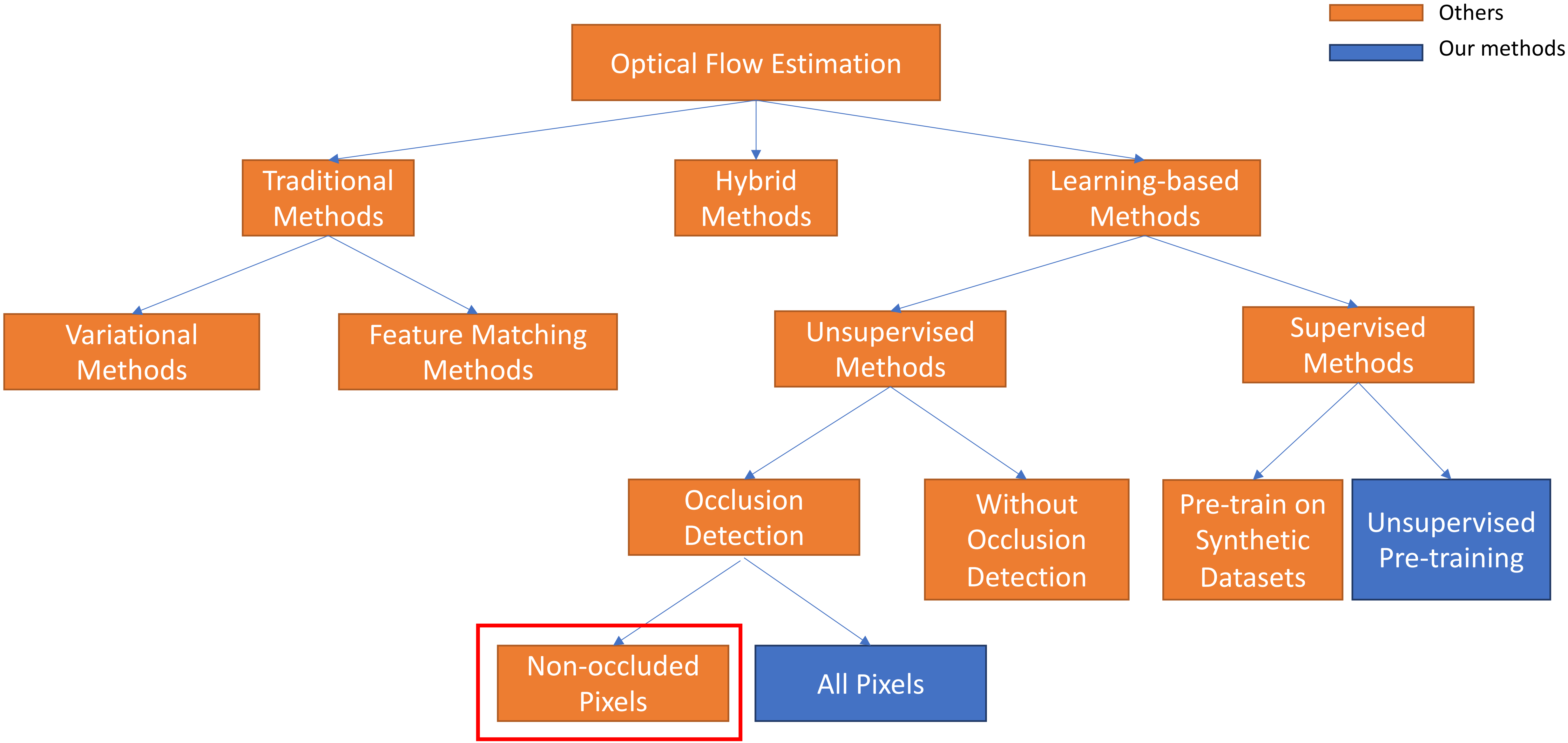


Unsupervised Learning Methods

- Problem: brightness consistency does not hold for **occluded** pixels

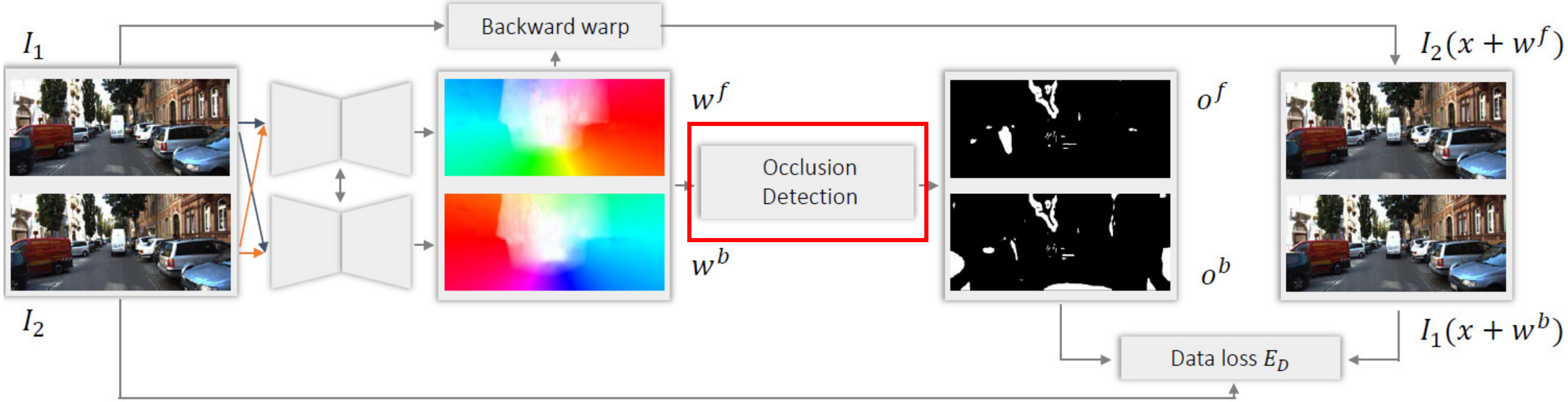


Background Review

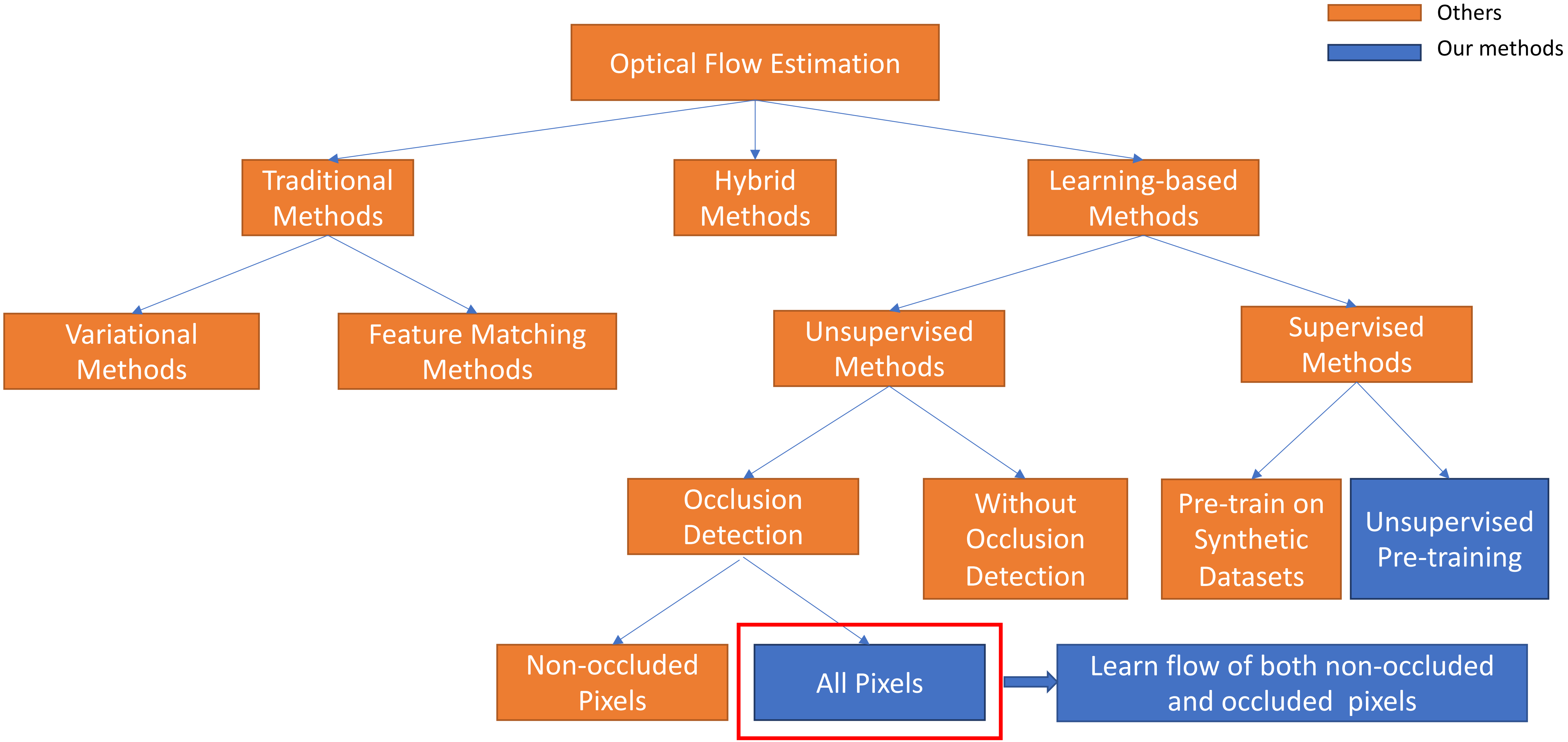


Unsupervised Learning Methods

- Advantage: infinite training data, learn flow of non-occluded pixels
- Disadvantage: lack the ability to predict flow of **occluded** pixels



Motivation


Others
Our methods
Optical Flow Estimation
Traditional Methods
Hybrid Methods
Learning-based Methods
Variational Methods
Feature Matching Methods
Unsupervised Methods
Supervised Methods
Occlusion Detection
Without Occlusion Detection
Pre-train on Synthetic Datasets
Unsupervised Pre-training
Non-occluded Pixels
All Pixels
Learn flow of both non-occluded and occluded pixels

Method

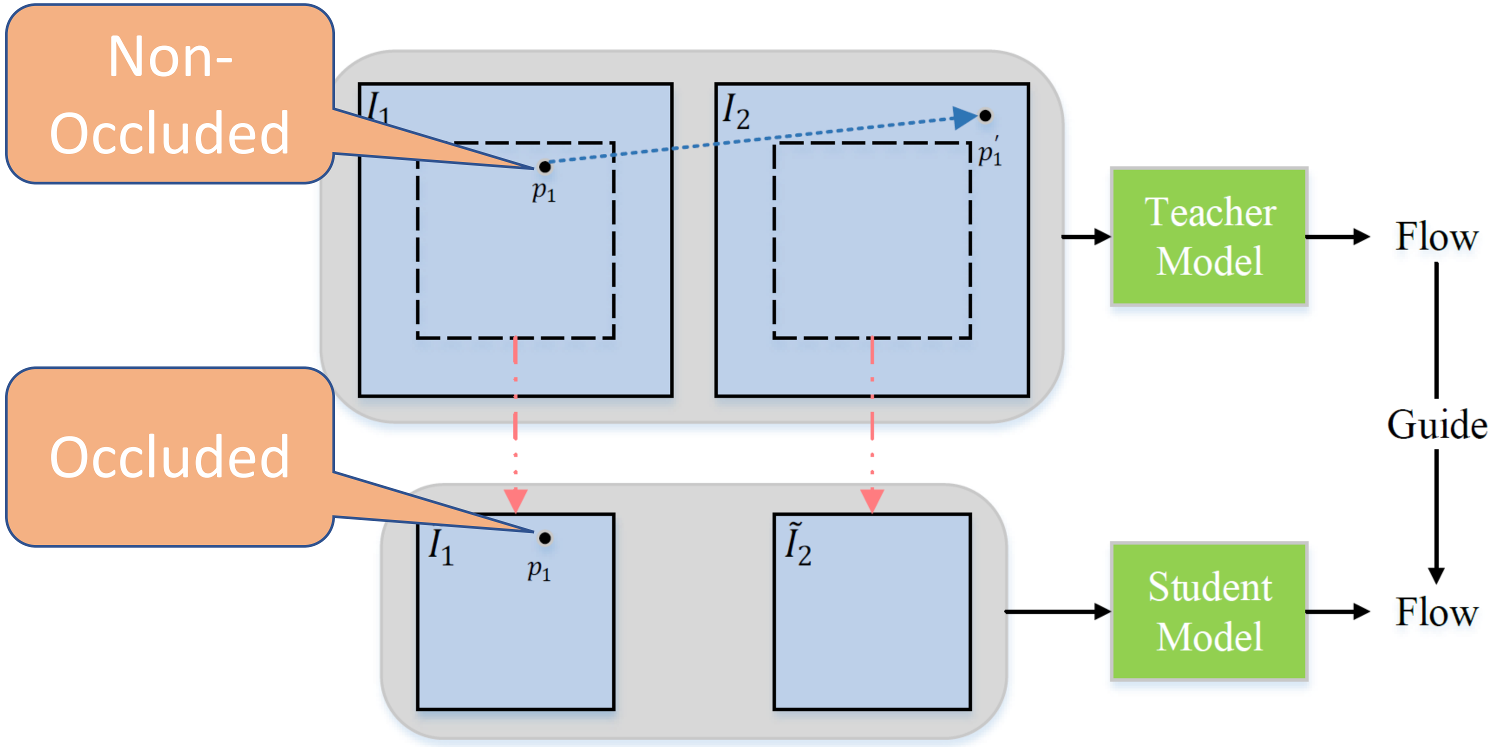
- We propose a series of self-supervised learning methods
  - DDFlow [AAAI'19]
  - SelFlow [CVPR'19]
  - Flow2Stereo [CVPR'20]
  - DistillFlow [*TPAMI'20]

- Advantages
  - Make use of infinite unlabeled data
  - Learn flow of both occluded and non-occluded pixels from unlabeled data
  - Reduce the performance gap compared with supervised methods
  - Reduce the reliance of synthetic data

* In Submission

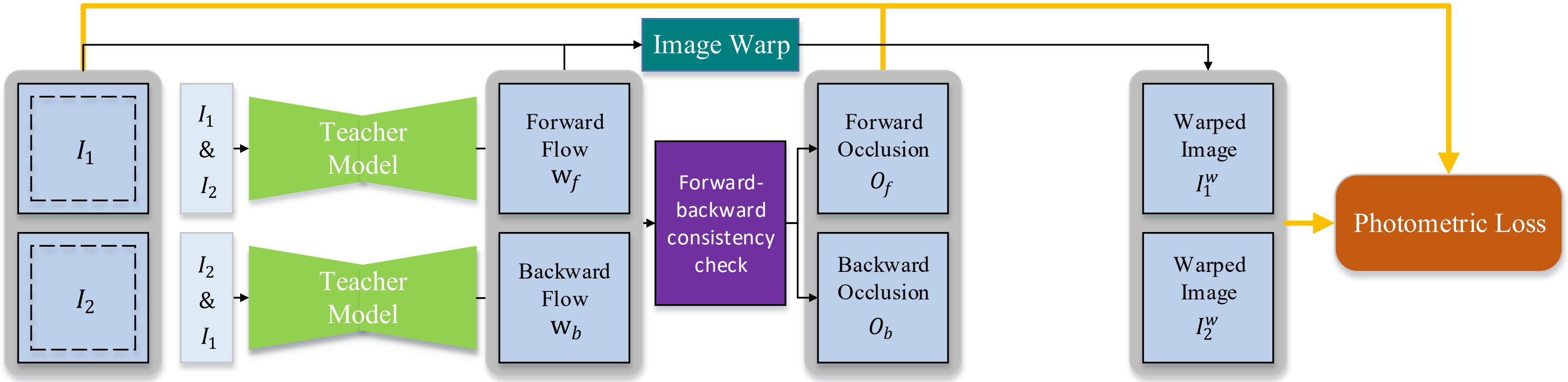
DDFlow: Observation

- The optical flow of **non-occluded** pixels can be **accurately** estimated.
- How do we fully utilize those reliable predictions?
- We can create artificial occlusions for self-supervision.



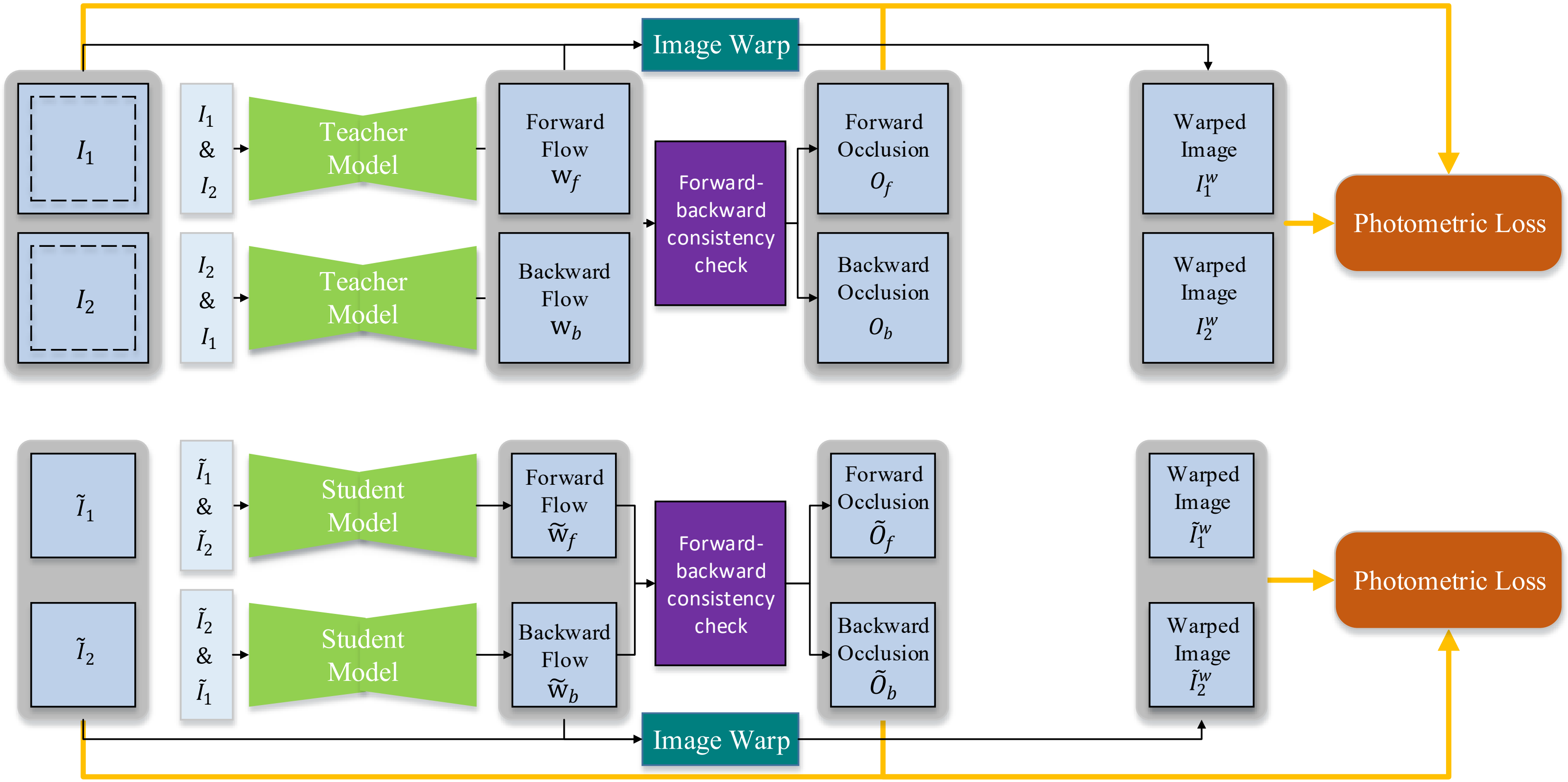
Self-Supervised Learning Framework

- The teacher model is trained with the photometric loss L_p for non-occluded pixels.



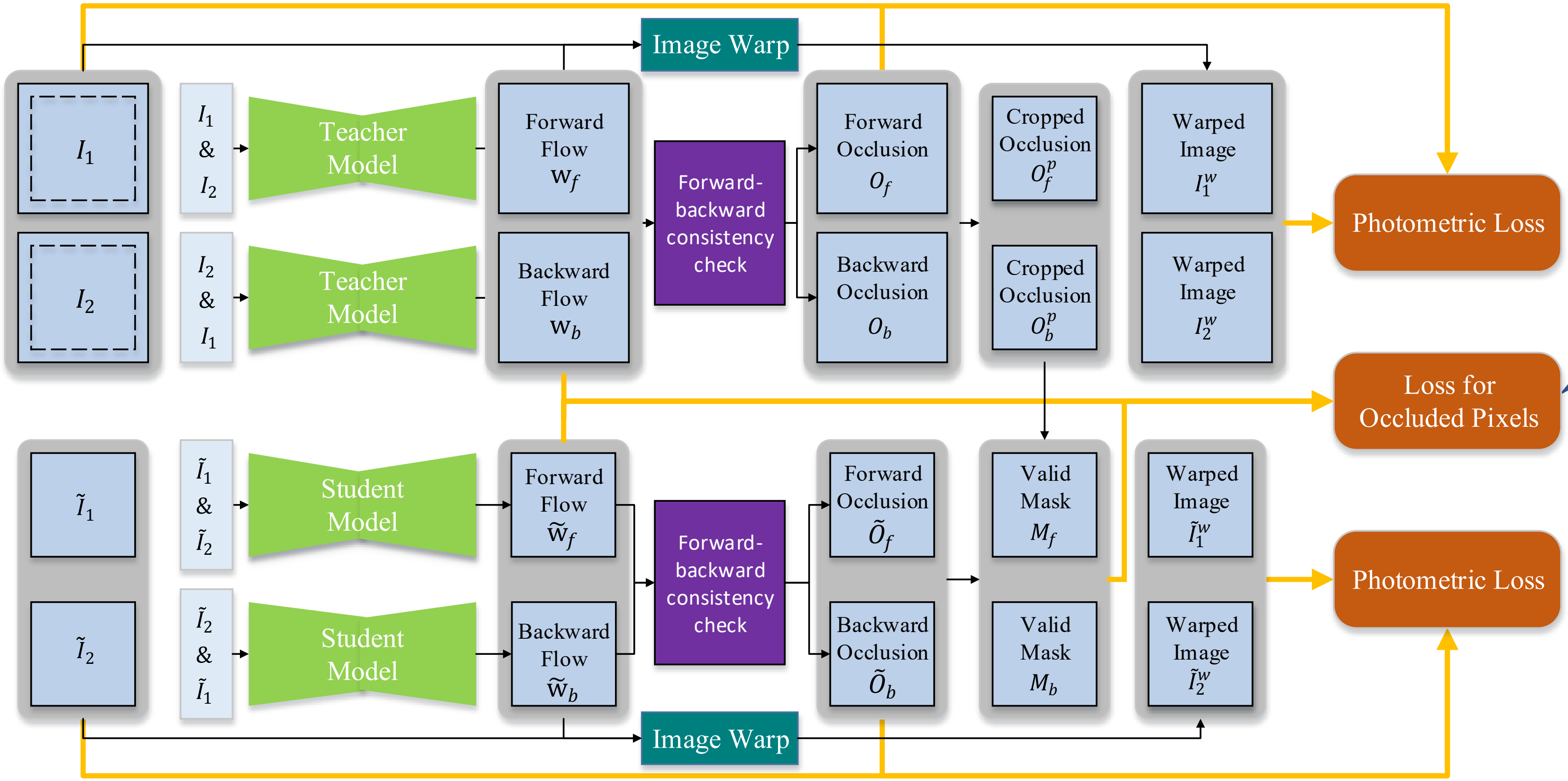
Self-Supervised Learning Framework

- The student model shares the same network structure with teacher model.



Self-Supervised Learning Framework

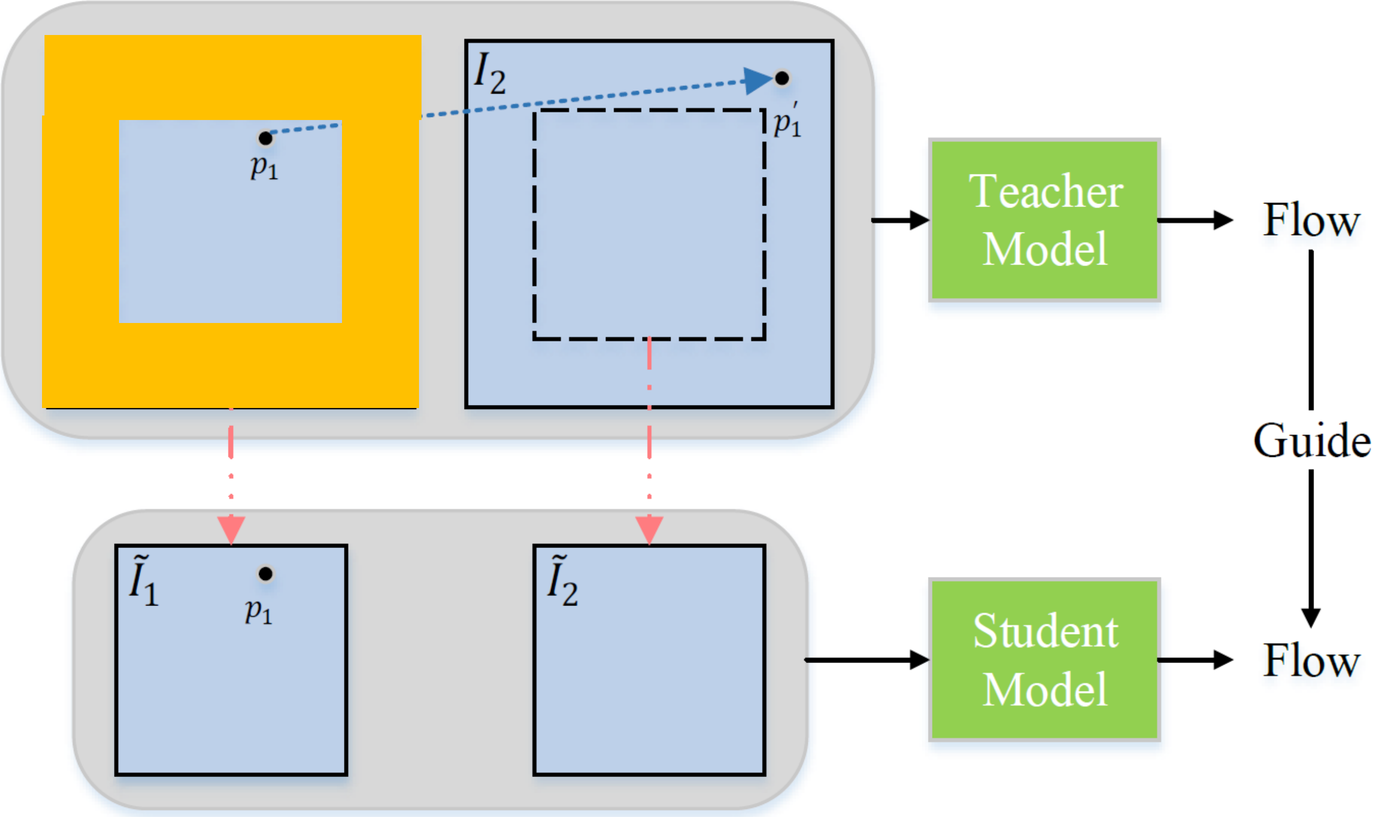
- The student model is trained with photometric loss L_p and self-supervised loss L_o for occluded pixels using predictions from the teacher model.



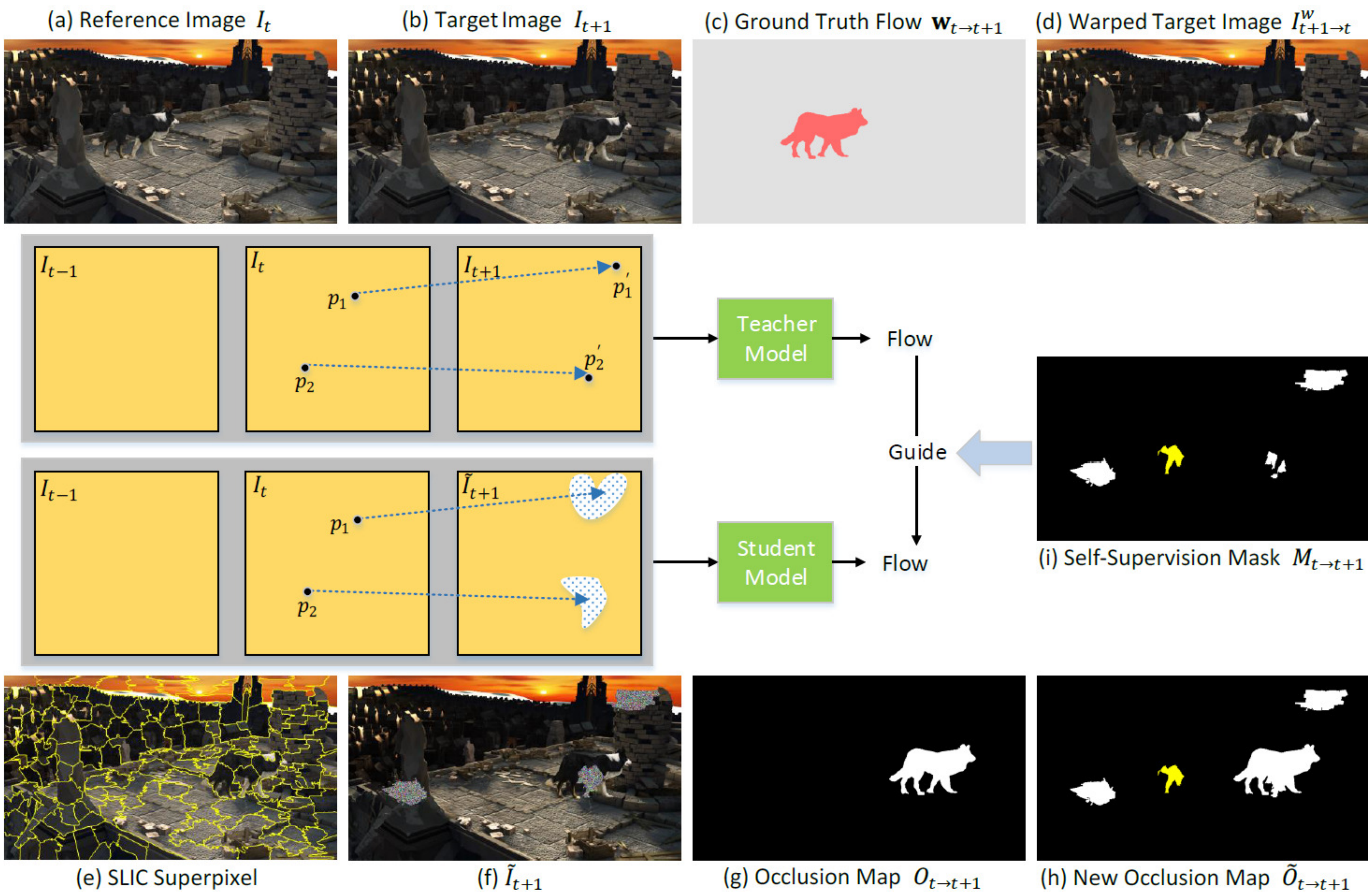
L_o only functions on pixels that are non-occluded in original images but occluded in cropped patches.

Rethink Occlusion

- **Cropping** strategy only works well for occlusions **near image boundary.**
- How to cope with occlusions **elsewhere**?

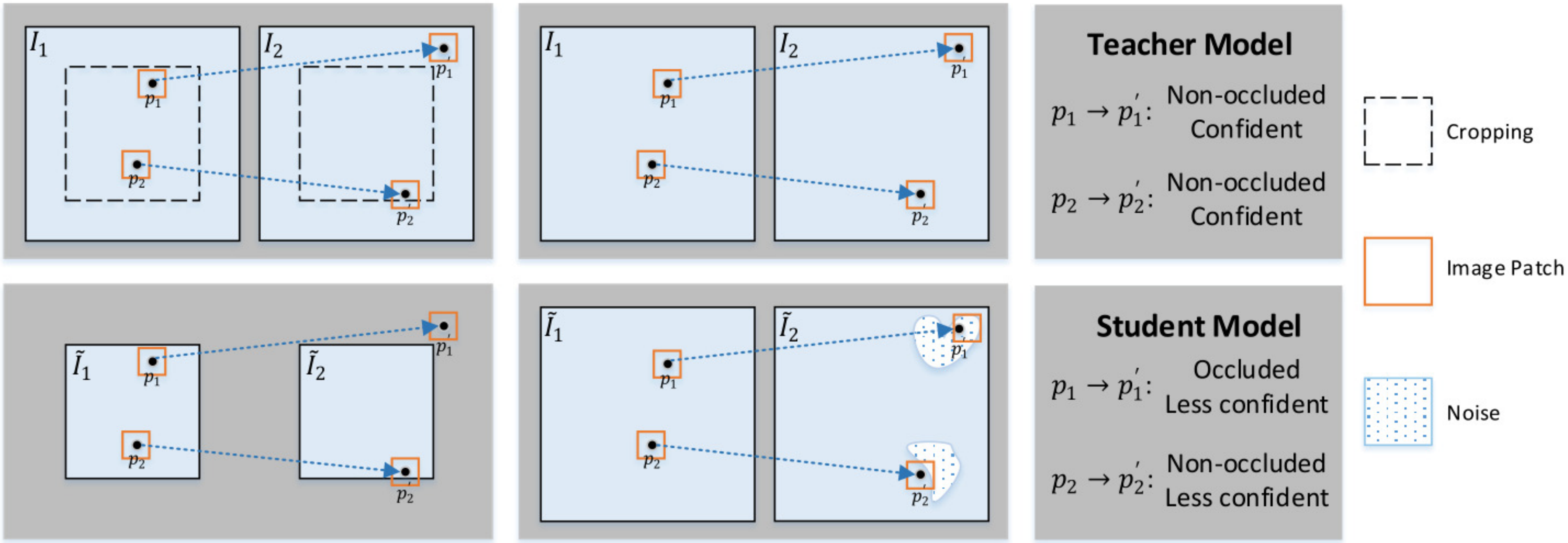


SelFlow: Superpixel-based Occlusion Hallucination



Key of Self-Supervision

- Observation: self-supervision also improves the flow learning of **non-occluded** pixels
- Key: create **challenging transformations** and let **confident** predictions supervise less **confident** predictions (Flow2Stereo)

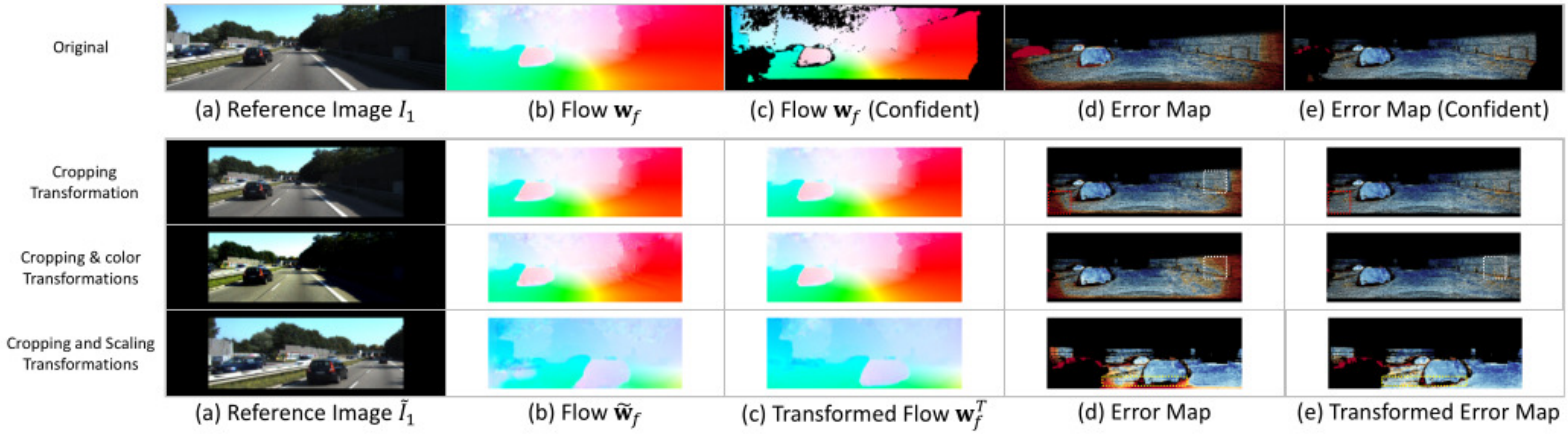


(a) Cropping occlusion hallucination

(b) Superpixel occlusion hallucination

Challenging Transformations

- Three kinds of challenging transformations (DistillFlow):
  - Occlusion hallucination-based transformations
  - Color transformations
  - Geometric transformations



Limitations

- The performance of the teacher model determines the upper bound of the student model

- We propose three improvements:

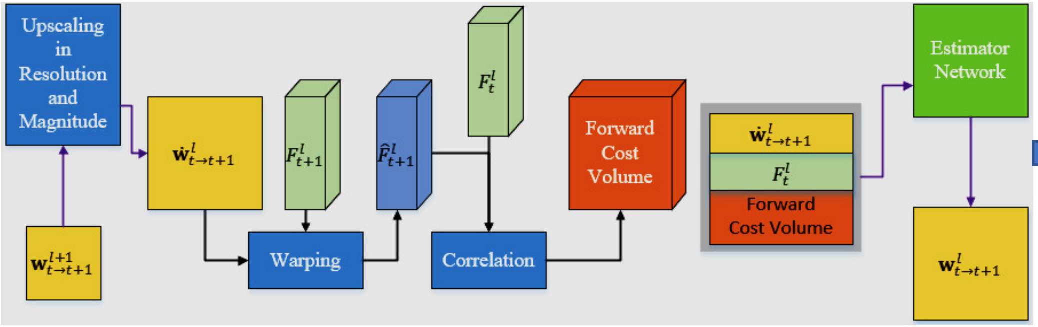
  - Utilize multiple frames: explore temporal consistency (SelFlow)

  - Use stereo videos: explore the geometric constraints between optical flow and stereo disparity (Flow2Stereo)

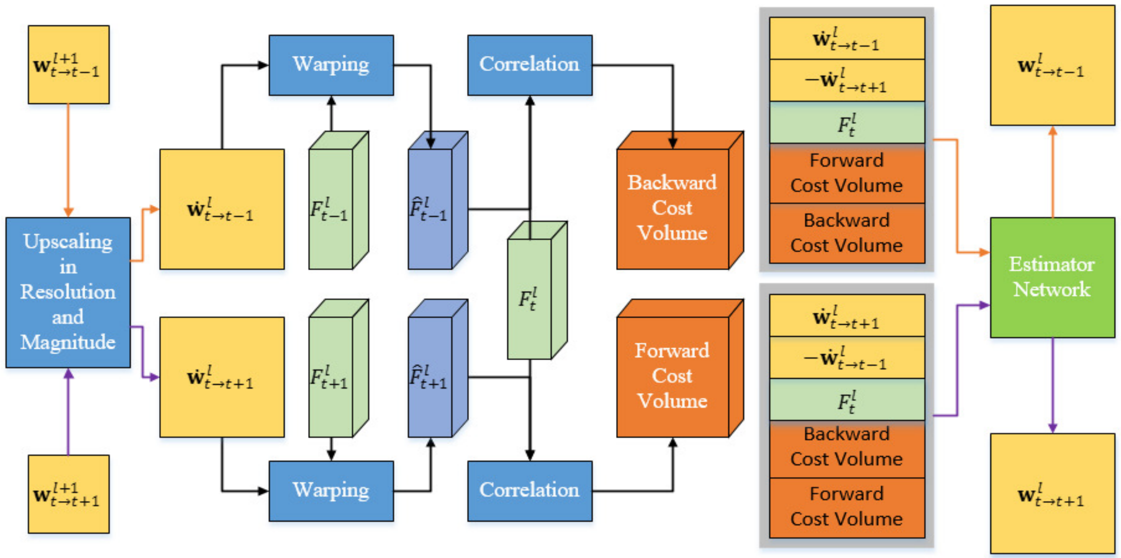
  - Model distillation: employ multiple teacher models and ensemble multiple predictions (DistillFlow)

Direction 1: Multi-frame Optical Flow Estimation

- Our three-frame flow estimation network:
  - Compute bidirectional flow and cost volume
  - Combine reversed backward flow and backward cost volume information
  - Swap initial flow and cost volume to estimate forward and backward flow concurrently



Two-frame PWC-Net network structure at each level

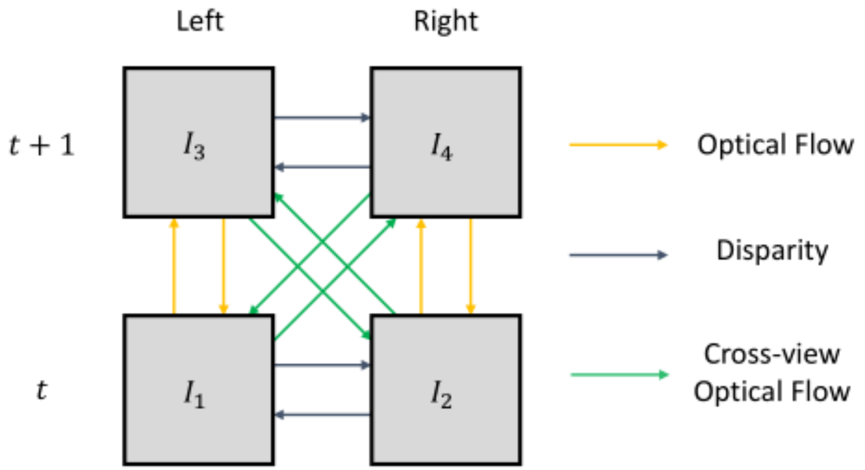


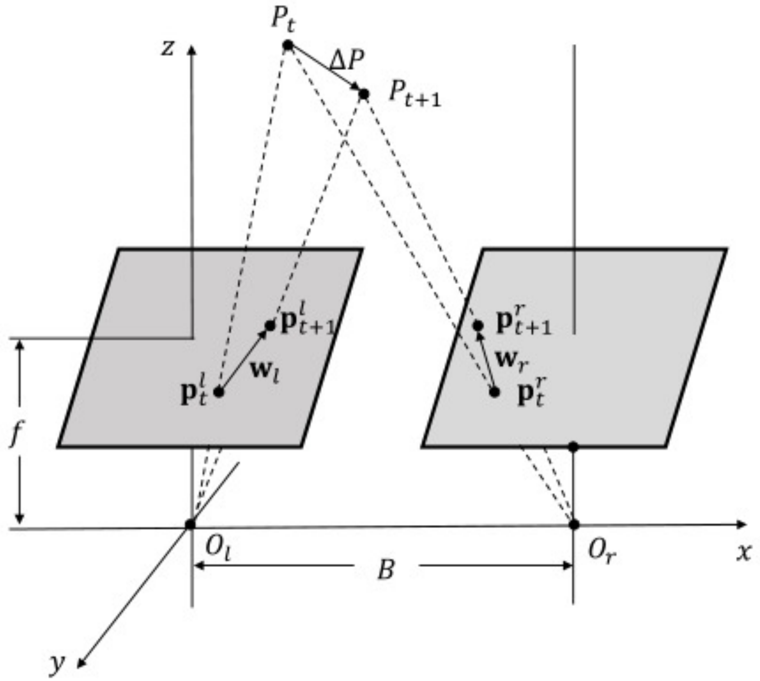
Three-frame network structure at each level

Direction 2: Use Stereo Data

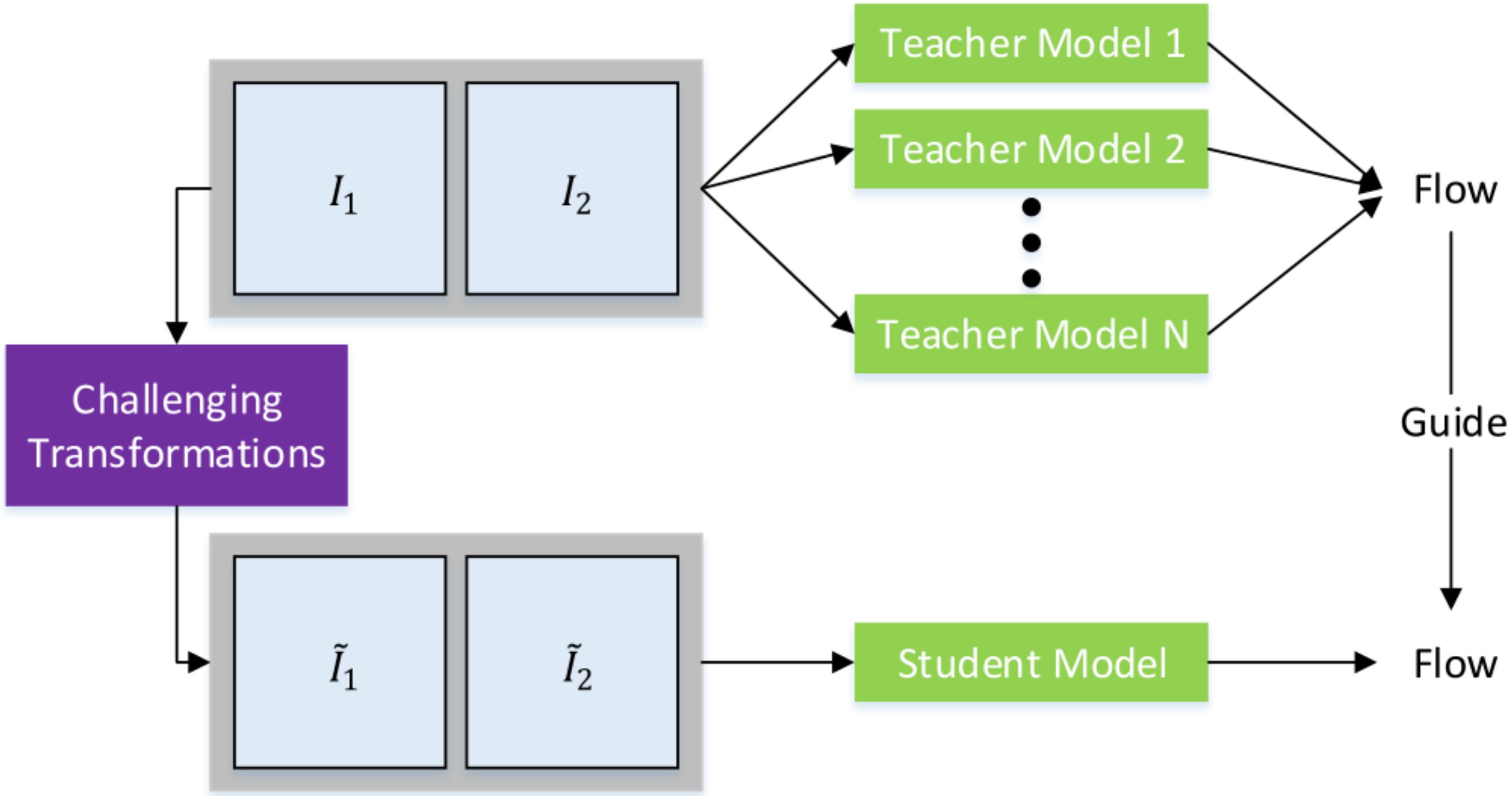
- We regard stereo matching as a special case of optical flow, and use one unified network to predict both optical flow and stereo disparity
- Geometric constrains

$$\begin{cases} u_r - u_l = (-d_{t+1}) - (-d_t) \\ v_r - v_l = 0 \end{cases}$$

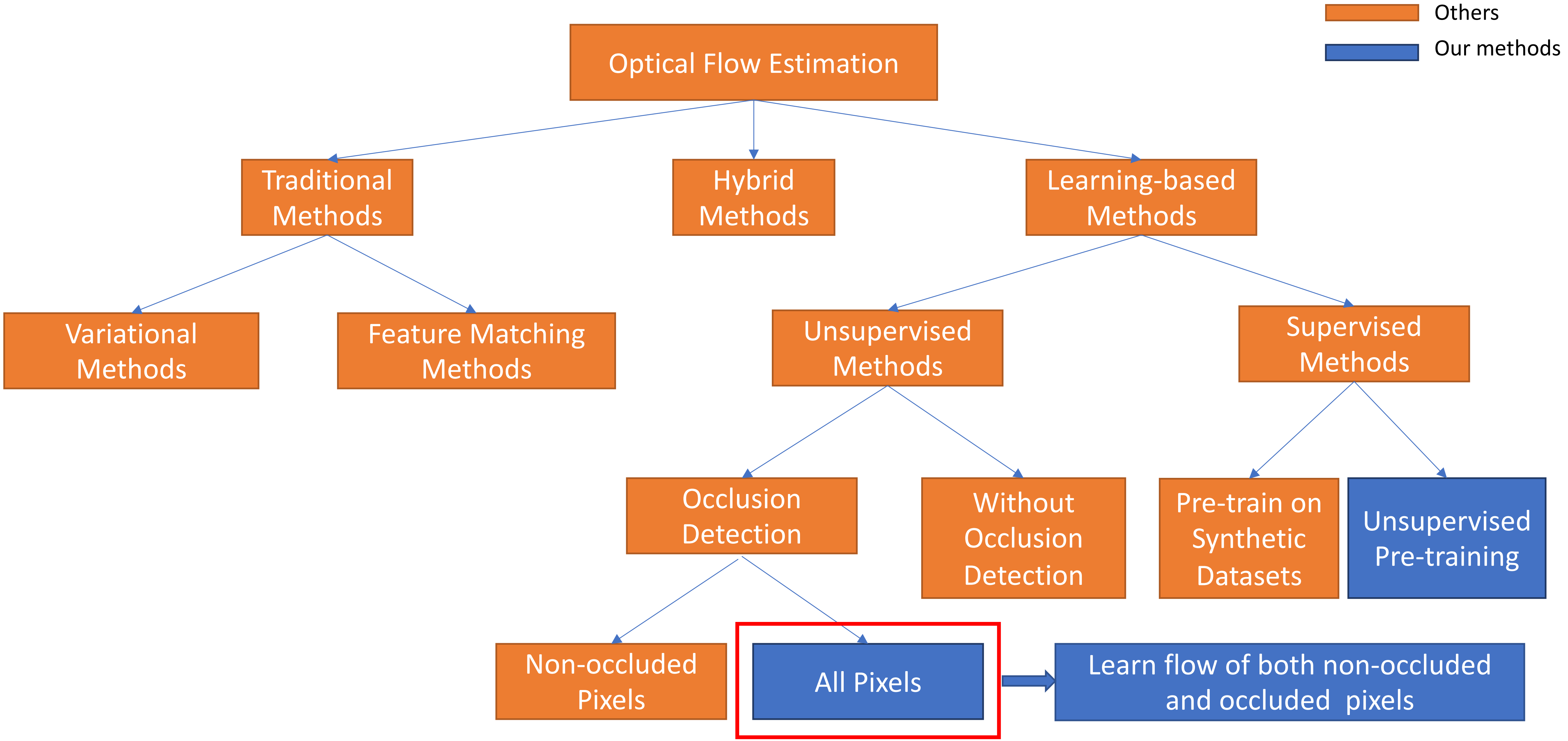




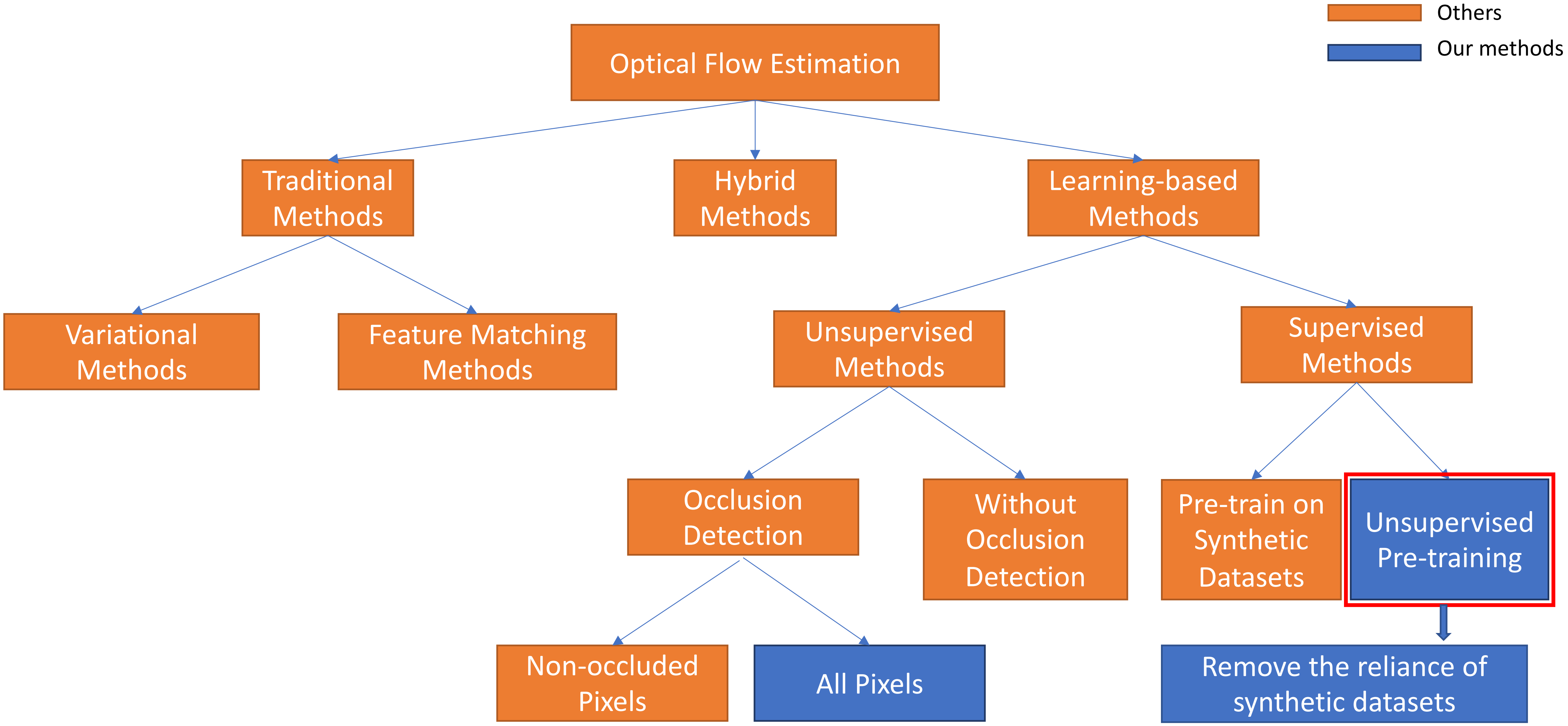
Direction 3: Model Distillation



Motivation


Others
Our methods
Optical Flow Estimation
Traditional Methods
Hybrid Methods
Learning-based Methods
Variational Methods
Feature Matching Methods
Unsupervised Methods
Supervised Methods
Occlusion Detection
Without Occlusion Detection
Pre-train on Synthetic Datasets
Unsupervised Pre-training
Non-occluded Pixels
All Pixels
Learn flow of both non-occluded and occluded pixels

Motivation


Others
Our methods
Optical Flow Estimation
Traditional Methods
Hybrid Methods
Learning-based Methods
Variational Methods
Feature Matching Methods
Unsupervised Methods
Supervised Methods
Occlusion Detection
Without Occlusion Detection
Pre-train on Synthetic Datasets
Unsupervised Pre-training
Non-occluded Pixels
All Pixels
Remove the reliance of synthetic datasets

Supervised Fine-tuning

- Self-supervised pre-training achieves excellent initializations for supervised fine-tuning: remove the reliance of synthetic data

- Previous methods: pre-train on synthetic data → fine-tune with limited labeled data

- Our method: pre-train with unlabeled data → fine-tune with limited labeled data

A new perspective in supervised learning of optical flow

Experiments: Datasets

- Labeled datasets

Dataset	Training	Test	Annotations
KITTI 2012	194 pairs	195 pairs	sparse
KITTI 2015	200 paris	200 pairs	sparse
Sintel Clean	23 videos	12 videos	Dense
Sintel Final			

- Unlabeled datasets
  - Both KITTI and Sintel contain large quantities of unlabeled raw data

Experiments: Evaluation Metrics

- Optical Flow
  - **EPE:** average endpoint error between the predicted flow and the ground truth flow.
  - **Fl:** percentage of erroneous pixels

- Occlusion Detection
  - F-score: the harmonic average of the precision and recall

Experiments: Quantitative Results

- We achieve the best unsupervised optical flow estimation performance on all datasets

	Method	Sintel Clean EPE-train	Sintel Clean EPE-test	Sintel Final EPE-train	Sintel Final EPE-test
Unsupervised	DSTFlow [110]	(6.16)	10.41	(6.81)	11.27
	UnFlow-CSS [92]	–	–	(7.91)	10.22
	OccAwareFlow [136]	(4.03)	7.95	(5.95)	9.15
	Back2FutureFlow-None [53]*	(6.05)	–	(7.09)	–
	Back2FutureFlow-Soft [53]*	(3.89)	7.23	(5.52)	8.81
	EpipolarFlow [159]	(3.54)	7.00	(4.99)	8.51
	DDFlow [79]	(2.92)	6.18	(3.98)	7.40
	SelFlow [80]*	(2.88)	6.56	(3.87)	6.57
	DistillFlow (trained on KITTI)	4.21	–	5.06	–
	DistillFlow	**(2.61)**	**4.23**	**(3.70)**	**5.81**
Supervised	FlowNetS [26]	(3.66)	6.96	(4.44)	7.76
	FlowNetC [26]	(3.78)	6.85	(5.28)	8.51
	SpyNet [106]	(3.17)	6.64	(4.32)	8.36
	FlowFieldsCNN [4]	–	3.78	–	5.36
	DCFlow [140]	–	3.54	–	5.12
	FlowNet2 [50]	(1.45)	4.16	(2.01)	5.74
	LiteFlowNet [48]	**(1.35)**	4.54	(1.78)	5.38
	LiteFlowNet2 [49]	(1.41)	3.48	(1.83)	4.69
	PWC-Net [121]	(2.02)	4.39	(2.08)	5.04
	PWC-Net+ [122]	(1.71)	3.45	(2.34)	4.60
	ContinualFlow [97]	–	3.34	–	4.52
	HD^{3}Flow [146]	(1.70)	4.79	**(1.17)**	4.67
	IRR-PWC [1]	(1.92)	3.84	(2.51)	4.58
	MFF [109]*	–	3.42	–	4.57
	VCN [143]	(1.66)	2.81	(2.24)	4.40
	SENSE [56]	(1.54)	3.60	(2.05)	4.86
	ScopeFlow [6]	–	3.59	–	**4.10**
	MaskFlowNet-S [158]	–	2.77	–	4.38
	MaskFlowNet [158]	–	**2.52**	–	4.17
	SelFlow [80]*	(1.68)	3.74	(1.77)	4.26
	DistillFlow	(1.63)	3.49	(1.76)	**4.10**

	Method	KITTI 2012 train EPE-all	KITTI 2012 train EPE-noc	KITTI 2012 test EPE-all	KITTI 2012 test EPE-noc	KITTI 2012 test Fl-all	KITTI 2012 test Fl-noc	KITTI 2015 train EPE-all	KITTI 2015 train EPE-noc	KITTI 2015 test Fl-all	KITTI 2015 test Fl-fg	KITTI 2015 test Fl-bg
Unsupervised	BackToBasic [55]	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [110]	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [92]	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [136]	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	Back2FutureFlow-None [53]*	–	–	–	–	–	–	6.65	3.24	–	–	–
	Back2FutureFlow-Soft [53]*	–	–	–	–	–	–	6.59	3.22	22.94%	24.27%	22.67%
	EpipolarFlow [159]	(2.51)	(0.99)	3.4	1.3	–	–	(5.55)	(2.46)	16.95%	–	–
	Lai *et al.* [70] (+Stereo)	2.56	1.39	–	–	–	–	7.13	4.31	–	–	–
	UnOS [135] (+Stereo)	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	DDFlow [79]	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [80]*	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Flow2Stereo [81] (+Stereo)	1.45	**0.82**	1.7	**0.9**	7.63%	4.02%	3.54	2.12	11.10%	**16.67%**	9.99%
	DistillFlow (trained on Sintel)	2.33	1.08					8.16	4.20			
	DistillFlow	**1.38**	0.83	**1.6**	**0.9**	**7.18%**	**3.91%**	**2.93**	**1.96**	**10.54%**	16.98%	**9.26%**
Supervised	FlowNetS [26]	7.52	–	9.1	–	44.49%	–	–	–	–	–	–
	SpyNet [106]	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [4]	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [140]	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [50]	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [92]	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [48]	(1.05)	–	1.6	0.8	7.27%	3.27%	(1.62)	–	9.38%	7.99%	9.66%
	LiteFlowNet2 [49]	(0.95)	–	1.4	0.7	6.16%	2.63%	(1.33)	–	7.62%	7.64%	7.62%
	PWC-Net [121]	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	PWC-Net+ [122]	(1.08)	–	1.4	0.8	6.72%	3.36%	(1.45)	–	7.72%	7.88%	7.69%
	ContinualFlow [97]	–	–	–	–	–	–	–	–	10.03%	17.48%	8.54%
	HD^{3}Flow [146]	(0.81)	–	1.4	0.7	5.41%	2.26%	(1.31)	–	6.55%	9.02%	6.05%
	IRR-PWC [1]	–	–	1.6	0.9	6.70%	3.21%	(1.45)	–	7.65%	7.52%	7.68%
	MFF [109]*	–	–	1.7	0.9	7.87%	4.19%	–	–	7.17%	**7.25%**	7.15%
	VCN [143]	–	–	–	–	–	–	(1.16)	–	6.30%	8.66%	5.83%
	SENSE [56]	(1.18)	–	1.5	–	–	3.03%	(2.05)	–	8.16%	–	–
	ScopeFlow [6]	–	–	1.3	0.7	5.66%	2.68%	–	–	6.82%	7.36%	6.72%
	MaskFlowNet-S [158]	–	–	**1.1**	**0.6**	5.24%	2.29%	–	–	6.81%	8.21%	6.53%
	MaskFlowNet [158]	–	–	**1.1**	**0.6**	**4.82%**	**2.07%**	–	–	6.11%	7.70%	5.79%
	SelFlow [80]*	**(0.76)**	(0.47)	1.5	0.9	6.19%	3.32%	(1.18)	(0.82)	8.42%	7.61%	12.48%
	DistillFlow	(0.79)	**(0.45)**	1.2	**0.6**	5.23%	2.33%	**(1.14)**	**(0.74)**	**5.94%**	7.96%	**5.53%**

Experiments: Quantitative Results

- Our unsupervised results even outperform several famous fully-supervised methods

	Method	Sintel Clean EPE-train	Sintel Clean EPE-test	Sintel Final EPE-train	Sintel Final EPE-test
Unsupervised	DSTFlow [110]	(6.16)	10.41	(6.81)	11.27
	UnFlow-CSS [92]	–	–	(7.91)	10.22
	OccAwareFlow [136]	(4.03)	7.95	(5.95)	9.15
	Back2FutureFlow-None [53]*	(6.05)	–	(7.09)	–
	Back2FutureFlow-Soft [53]*	(3.89)	7.23	(5.52)	8.81
	EpipolarFlow [159]	(3.54)	7.00	(4.99)	8.51
	DDFlow [79]	(2.92)	6.18	(3.98)	7.40
	SelFlow [80]*	(2.88)	6.56	(3.87)	6.57
	DistillFlow (trained on KITTI)	4.21	–	5.06	–
	DistillFlow	**(2.61)**	**4.23**	**(3.70)**	**5.81**
Supervised	FlowNetS [26]	(3.66)	6.96	(4.44)	7.76
	FlowNetC [26]	(3.78)	6.85	(5.28)	8.51
	SpyNet [106]	(3.17)	6.64	(4.32)	8.36
	FlowFieldsCNN [4]	–	3.78	–	5.36
	DCFlow [140]	–	3.54	–	5.12
	FlowNet2 [50]	(1.45)	4.16	(2.01)	5.74
	LiteFlowNet [48]	**(1.35)**	4.54	(1.78)	5.38
	LiteFlowNet2 [49]	(1.41)	3.48	(1.83)	4.69
	PWC-Net [121]	(2.02)	4.39	(2.08)	5.04
	PWC-Net+ [122]	(1.71)	3.45	(2.34)	4.60
	ContinualFlow [97]	–	3.34	–	4.52
	HD^{3}Flow [146]	(1.70)	4.79	**(1.17)**	4.67
	IRR-PWC [1]	(1.92)	3.84	(2.51)	4.58
	MFF [109]*	–	3.42	–	4.57
	VCN [143]	(1.66)	2.81	(2.24)	4.40
	SENSE [56]	(1.54)	3.60	(2.05)	4.86
	ScopeFlow [6]	–	3.59	–	**4.10**
	MaskFlowNet-S [158]	–	2.77	–	4.38
	MaskFlowNet [158]	–	**2.52**	–	4.17
	SelFlow [80]*	(1.68)	3.74	(1.77)	4.26
	DistillFlow	(1.63)	3.49	(1.76)	**4.10**

	Method	KITTI 2012 train EPE-all	KITTI 2012 train EPE-noc	KITTI 2012 test EPE-all	KITTI 2012 test EPE-noc	KITTI 2012 test Fl-all	KITTI 2012 test Fl-noc	KITTI 2015 train EPE-all	KITTI 2015 train EPE-noc	KITTI 2015 test Fl-all	KITTI 2015 test Fl-fg	KITTI 2015 test Fl-bg
Unsupervised	BackToBasic [55]	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [110]	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [92]	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [136]	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	Back2FutureFlow-None [53]*	–	–	–	–	–	–	6.65	3.24	–	–	–
	Back2FutureFlow-Soft [53]*	–	–	–	–	–	–	6.59	3.22	22.94%	24.27%	22.67%
	EpipolarFlow [159]	(2.51)	(0.99)	3.4	1.3	–	–	(5.55)	(2.46)	16.95%	–	–
	Lai *et al.* [70] (+Stereo)	2.56	1.39	–	–	–	–	7.13	4.31	–	–	–
	UnOS [135] (+Stereo)	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	DDFlow [79]	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [80]*	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Flow2Stereo [81] (+Stereo)	1.45	**0.82**	1.7	**0.9**	7.63%	4.02%	3.54	2.12	11.10%	**16.67%**	9.99%
	DistillFlow (trained on Sintel)	2.33	1.08	–	–	–	–	8.16	4.20	–	–	–
	DistillFlow	**1.38**	0.83	**1.6**	**0.9**	**7.18%**	**3.91%**	**2.93**	**1.96**	**10.54%**	16.98%	**9.26%**
Supervised	FlowNetS [26]	7.52	–	9.1	–	44.49%	–	–	–	–	–	–
	SpyNet [106]	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [4]	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [140]	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [50]	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [92]	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [48]	(1.05)	–	1.6	0.8	7.27%	3.27%	(1.62)	–	9.38%	7.99%	9.66%
	LiteFlowNet2 [49]	(0.95)	–	1.4	0.7	6.16%	2.63%	(1.33)	–	7.62%	7.64%	7.62%
	PWC-Net [121]	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	PWC-Net+ [122]	(1.08)	–	1.4	0.8	6.72%	3.36%	(1.45)	–	7.72%	7.88%	7.69%
	ContinualFlow [97]	–	–	–	–	–	–	–	–	10.03%	17.48%	8.54%
	HD^{3}Flow [146]	(0.81)	–	1.4	0.7	5.41%	2.26%	(1.31)	–	6.55%	9.02%	6.05%
	IRR-PWC [1]	–	–	1.6	0.9	6.70%	3.21%	(1.45)	–	7.65%	7.52%	7.68%
	MFF [109]*	–	–	1.7	0.9	7.87%	4.19%	–	–	7.17%	**7.25%**	7.15%
	VCN [143]	–	–	–	–	–	–	(1.16)	–	6.30%	8.66%	5.83%
	SENSE [56]	(1.18)	–	1.5	–	–	3.03%	(2.05)	–	8.16%	–	–
	ScopeFlow [6]	–	–	1.3	0.7	5.66%	2.68%	–	–	6.82%	7.36%	6.72%
	MaskFlowNet-S [158]	–	–	**1.1**	**0.6**	5.24%	2.29%	–	–	6.81%	8.21%	6.53%
	MaskFlowNet [158]	–	–	**1.1**	**0.6**	**4.82%**	**2.07%**	–	–	6.11%	7.70%	5.79%
	SelFlow [80]*	**(0.76)**	(0.47)	1.5	0.9	6.19%	3.32%	(1.18)	(0.82)	8.42%	7.61%	12.48%
	DistillFlow	(0.79)	**(0.45)**	1.2	**0.6**	5.23%	2.33%	**(1.14)**	**(0.74)**	**5.94%**	7.96%	**5.53%**

Experiments: Quantitative Results

- With more challenging transformations, DistillFlow achieves great performance improvement over SelFlow

	Method	Sintel Clean		Sintel Final	
		EPE-train	EPE-test	EPE-train	EPE-test
Unsupervised	DSTFlow [110]	(6.16)	10.41	(6.81)	11.27
	UnFlow-CSS [92]	–	–	(7.91)	10.22
	OccAwareFlow [136]	(4.03)	7.95	(5.95)	9.15
	Back2FutureFlow-None [53]*	(6.05)	–	(7.09)	–
	Back2FutureFlow-Soft [53]*	(3.89)	7.23	(5.52)	8.81
	EpipolarFlow [159]	(3.54)	7.00	(4.99)	8.51
	DDFlow [79]	(2.92)	6.18	(3.98)	7.40
	SelFlow [80]*	(2.88)	6.56	(3.87)	6.57
	DistillFlow (trained on KITTI)	4.21	–	5.06	–
	DistillFlow	**(2.61)**	**4.23**	**(3.70)**	**5.81**
Supervised	FlowNetS [26]	(3.66)	6.96	(4.44)	7.76
	FlowNetC [26]	(3.78)	6.85	(5.28)	8.51
	SpyNet [106]	(3.17)	6.64	(4.32)	8.36
	FlowFieldsCNN [4]	–	3.78	–	5.36
	DCFlow [140]	–	3.54	–	5.12
	FlowNet2 [50]	(1.45)	4.16	(2.01)	5.74
	LiteFlowNet [48]	**(1.35)**	4.54	(1.78)	5.38
	LiteFlowNet2 [49]	(1.41)	3.48	(1.83)	4.69
	PWC-Net [121]	(2.02)	4.39	(2.08)	5.04
	PWC-Net+ [122]	(1.71)	3.45	(2.34)	4.60
	ContinualFlow [97]	–	3.34	–	4.52
	HD^{3}Flow [146]	(1.70)	4.79	**(1.17)**	4.67
	IRR-PWC [1]	(1.92)	3.84	(2.51)	4.58
	MFF [109]*	–	3.42	–	4.57
	VCN [143]	(1.66)	2.81	(2.24)	4.40
	SENSE [56]	(1.54)	3.60	(2.05)	4.86
	ScopeFlow [6]	–	3.59	–	**4.10**
	MaskFlowNet-S [158]	–	2.77	–	4.38
	MaskFlowNet [158]	–	**2.52**	–	4.17
	SelFlow [80]*	(1.68)	3.74	(1.77)	4.26
	DistillFlow	(1.63)	3.49	(1.76)	**4.10**

	Method	KITTI 2012						KITTI 2015				
		train		test				train		test		
		EPE-all	EPE-noc	EPE-all	EPE-noc	Fl-all	Fl-noc	EPE-all	EPE-noc	Fl-all	Fl-fg	Fl-bg
Unsupervised	BackToBasic [55]	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [110]	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [92]	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [136]	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	Back2FutureFlow-None [53]*	–	–	–	–	–	–	6.65	3.24	–	–	–
	Back2FutureFlow-Soft [53]*	–	–	–	–	–	–	6.59	3.22	22.94%	24.27%	22.67%
	EpipolarFlow [159]	(2.51)	(0.99)	3.4	1.3	–	–	(5.55)	(2.46)	16.95%	–	–
	Lai *et al.* [70] (+Stereo)	2.56	1.39	–	–	–	–	7.13	4.31	–	–	–
	UnOS [135] (+Stereo)	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	DDFlow [79]	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [80]*	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Flow2Stereo [81] (+Stereo)	1.45	**0.82**	1.7	**0.9**	7.63%	4.02%	3.54	2.12	11.10%	**16.67%**	9.99%
	DistillFlow (trained on Sintel)	2.33	1.08					8.16	4.20			
	DistillFlow	**1.38**	0.83	**1.6**	**0.9**	**7.18%**	**3.91%**	**2.93**	**1.96**	**10.54%**	16.98%	**9.26%**
Supervised	FlowNetS [26]	7.52	–	9.1	–	44.49%	–	–	–	–	–	–
	SpyNet [106]	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [4]	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [140]	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [50]	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [92]	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [48]	(1.05)	–	1.6	0.8	7.27%	3.27%	(1.62)	–	9.38%	7.99%	9.66%
	LiteFlowNet2 [49]	(0.95)	–	1.4	0.7	6.16%	2.63%	(1.33)	–	7.62%	7.64%	7.62%
	PWC-Net [121]	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	PWC-Net+ [122]	(1.08)	–	1.4	0.8	6.72%	3.36%	(1.45)	–	7.72%	7.88%	7.69%
	ContinualFlow [97]	–	–	–	–	–	–	–	–	10.03%	17.48%	8.54%
	HD^{3}Flow [146]	(0.81)	–	1.4	0.7	5.41%	2.26%	(1.31)	–	6.55%	9.02%	6.05%
	IRR-PWC [1]	–	–	1.6	0.9	6.70%	3.21%	(1.45)	–	7.65%	7.52%	7.68%
	MFF [109]*	–	–	1.7	0.9	7.87%	4.19%	–	–	7.17%	**7.25%**	7.15%
	VCN [143]	–	–	–	–	–	–	(1.16)	–	6.30%	8.66%	5.83%
	SENSE [56]	(1.18)	–	1.5	–	–	3.03%	(2.05)	–	8.16%	–	–
	ScopeFlow [6]	–	–	1.3	0.7	5.66%	2.68%	–	–	6.82%	7.36%	6.72%
	MaskFlowNet-S [158]	–	–	**1.1**	**0.6**	5.24%	2.29%	–	–	6.81%	8.21%	6.53%
	MaskFlowNet [158]	–	–	**1.1**	**0.6**	**4.82%**	**2.07%**	–	–	6.11%	7.70%	5.79%
	SelFlow [80]*	**(0.76)**	(0.47)	1.5	0.9	6.19%	3.32%	(1.18)	(0.82)	8.42%	7.61%	12.48%
	DistillFlow	(0.79)	**(0.45)**	1.2	**0.6**	5.23%	2.33%	**(1.14)**	**(0.74)**	**5.94%**	7.96%	**5.53%**

Experiments: Quantitative Results

- In Flow2Stereo, 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%**

Experiments: Quantitative Results

- We achieve the state-of-the-art occlusion estimation results on Sintel and KITTI datasets

Method	KITTI 2012	KITTI 2015	Sintel Clean	Sintel Final
MODOF [141]	–	–	–	0.48
OccAwareFlow [136]	0.95	0.88	(0.54)	(0.48)
Back2Future [53]*	–	**0.91**	(0.49)	(0.44)
DDFlow [79]	0.94	0.86	**(0.59)**	(0.52)
SelFlow [80]*	0.95	0.88	**(0.59)**	(0.52)
DistillFlow	**0.96**	0.89	**(0.59)**	**(0.53)**

Experiments: Quantitative Results

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

	Method	Sintel Clean EPE-train	Sintel Clean EPE-test	Sintel Final EPE-train	Sintel Final EPE-test
Unsupervised	DSTFlow [110]	(6.16)	10.41	(6.81)	11.27
	UnFlow-CSS [92]	–	–	(7.91)	10.22
	OccAwareFlow [136]	(4.03)	7.95	(5.95)	9.15
	Back2FutureFlow-None [53]*	(6.05)	–	(7.09)	–
	Back2FutureFlow-Soft [53]*	(3.89)	7.23	(5.52)	8.81
	EpipolarFlow [159]	(3.54)	7.00	(4.99)	8.51
	DDFlow [79]	(2.92)	6.18	(3.98)	7.40
	SelFlow [80]*	(2.88)	6.56	(3.87)	6.57
	DistillFlow (trained on KITTI)	4.21	–	5.06	–
	DistillFlow	**(2.61)**	**4.23**	**(3.70)**	**5.81**
Supervised	FlowNetS [26]	(3.66)	6.96	(4.44)	7.76
	FlowNetC [26]	(3.78)	6.85	(5.28)	8.51
	SpyNet [106]	(3.17)	6.64	(4.32)	8.36
	FlowFieldsCNN [4]	–	3.78	–	5.36
	DCFlow [140]	–	3.54	–	5.12
	FlowNet2 [50]	(1.45)	4.16	(2.01)	5.74
	LiteFlowNet [48]	**(1.35)**	4.54	(1.78)	5.38
	LiteFlowNet2 [49]	(1.41)	3.48	(1.83)	4.69
	PWC-Net [121]	(2.02)	4.39	(2.08)	5.04
	PWC-Net+ [122]	(1.71)	3.45	(2.34)	4.60
	ContinualFlow [97]	–	3.34	–	4.52
	HD^{3}Flow [146]	(1.70)	4.79	**(1.17)**	4.67
	IRR-PWC [1]	(1.92)	3.84	(2.51)	4.58
	MFF [109]*	–	3.42	–	4.57
	VCN [143]	(1.66)	2.81	(2.24)	4.40
	SENSE [56]	(1.54)	3.60	(2.05)	4.86
	ScopeFlow [6]	–	3.59	–	**4.10**
	MaskFlowNet-S [158]	–	2.77	–	4.38
	MaskFlowNet [158]	–	**2.52**	–	4.17
	SelFlow [80]*	(1.68)	3.74	(1.77)	4.26
	DistillFlow	(1.63)	3.49	(1.76)	**4.10**

	Method	KITTI 2012 train EPE-all	KITTI 2012 train EPE-noc	KITTI 2012 test EPE-all	KITTI 2012 test EPE-noc	KITTI 2012 test Fl-all	KITTI 2012 test Fl-noc	KITTI 2015 train EPE-all	KITTI 2015 train EPE-noc	KITTI 2015 test Fl-all	KITTI 2015 test Fl-fg	KITTI 2015 test Fl-bg
Unsupervised	BackToBasic [55]	11.3	4.3	9.9	4.6	43.15%	34.85%	–	–	–	–	–
	DSTFlow [110]	10.43	3.29	12.4	4.0	–	–	16.79	6.96	39%	–	–
	UnFlow-CSS [92]	3.29	1.26	–	–	–	–	8.10	–	23.30%	–	–
	OccAwareFlow [136]	3.55	–	4.2	–	–	–	8.88	–	31.2%	–	–
	Back2FutureFlow-None [53]*	–	–	–	–	–	–	6.65	3.24	–	–	–
	Back2FutureFlow-Soft [53]*	–	–	–	–	–	–	6.59	3.22	22.94%	24.27%	22.67%
	EpipolarFlow [159]	(2.51)	(0.99)	3.4	1.3	–	–	(5.55)	(2.46)	16.95%	–	–
	Lai *et al.* [70] (+Stereo)	2.56	1.39	–	–	–	–	7.13	4.31	–	–	–
	UnOS [135] (+Stereo)	1.64	1.04	1.8	–	–	–	5.58	–	18.00%	–	–
	DDFlow [79]	2.35	1.02	3.0	1.1	8.86%	4.57%	5.72	2.73	14.29%	20.40%	13.08%
	SelFlow [80]*	1.69	0.91	2.2	1.0	7.68%	4.31%	4.84	2.40	14.19%	21.74%	12.68%
	Flow2Stereo [81] (+Stereo)	1.45	**0.82**	1.7	**0.9**	7.63%	4.02%	3.54	2.12	11.10%	**16.67%**	9.99%
	DistillFlow (trained on Sintel)	2.33	1.08	–	–	–	–	8.16	4.20	–	–	–
	DistillFlow	**1.38**	0.83	**1.6**	**0.9**	**7.18%**	**3.91%**	**2.93**	**1.96**	**10.54%**	16.98%	**9.26%**
Supervised	FlowNetS [26]	7.52	–	9.1	–	44.49%	–	–	–	–	–	–
	SpyNet [106]	3.36	–	4.1	2.0	20.97%	12.31%	–	–	35.07%	43.62%	33.36%
	FlowFieldsCNN [4]	–	–	3.0	1.2	13.01%	4.89%	–	–	18.68%	20.42%	18.33%
	DCFlow [140]	–	–	–	–	–	–	–	–	14.86%	23.70%	13.10%
	FlowNet2 [50]	(1.28)	–	1.8	1.0	8.80%	4.82%	(2.3)	–	10.41%	8.75%	10.75%
	UnFlow-CSS [92]	(1.14)	(0.66)	1.7	0.9	8.42%	4.28%	(1.86)	–	11.11%	15.93%	10.15%
	LiteFlowNet [48]	(1.05)	–	1.6	0.8	7.27%	3.27%	(1.62)	–	9.38%	7.99%	9.66%
	LiteFlowNet2 [49]	(0.95)	–	1.4	0.7	6.16%	2.63%	(1.33)	–	7.62%	7.64%	7.62%
	PWC-Net [121]	(1.45)	–	1.7	0.9	8.10%	4.22%	(2.16)	–	9.60%	9.31%	9.66%
	PWC-Net+ [122]	(1.08)	–	1.4	0.8	6.72%	3.36%	(1.45)	–	7.72%	7.88%	7.69%
	ContinualFlow [97]	–	–	–	–	–	–	–	–	10.03%	17.48%	8.54%
	HD^{3}Flow [146]	(0.81)	–	1.4	0.7	5.41%	2.26%	(1.31)	–	6.55%	9.02%	6.05%
	IRR-PWC [1]	–	–	1.6	0.9	6.70%	3.21%	(1.45)	–	7.65%	7.52%	7.68%
	MFF [109]*	–	–	1.7	0.9	7.87%	4.19%	–	–	7.17%	**7.25%**	7.15%
	VCN [143]	–	–	–	–	–	–	(1.16)	–	6.30%	8.66%	5.83%
	SENSE [56]	(1.18)	–	1.5	–	–	3.03%	(2.05)	–	8.16%	–	–
	ScopeFlow [6]	–	–	1.3	0.7	5.66%	2.68%	–	–	6.82%	7.36%	6.72%
	MaskFlowNet-S [158]	–	–	**1.1**	**0.6**	5.24%	2.29%	–	–	6.81%	8.21%	6.53%
	MaskFlowNet [158]	–	–	**1.1**	**0.6**	**4.82%**	**2.07%**	–	–	6.11%	7.70%	5.79%
	SelFlow [80]*	**(0.76)**	(0.47)	1.5	0.9	6.19%	3.32%	(1.18)	(0.82)	8.42%	7.61%	12.48%
	DistillFlow	(0.79)	**(0.45)**	1.2	**0.6**	5.23%	2.33%	**(1.14)**	**(0.74)**	**5.94%**	7.96%	**5.53%**

Experiments: Quantitative Results

- Our fine-tuned SelFlow model ranks first on Sintel dataset from November 2018 to November 2019

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
SelFlow [2]	4.262	2.040	22.369	4.083	1.715	1.287	0.582	2.343	27.154	Visualize Results
VCN [3]	4.520	2.195	23.478	4.423	1.802	1.357	0.934	2.816	26.434	Visualize Results
ContinualFlow_ROB [4]	4.528	2.723	19.248	5.050	2.573	1.713	0.872	3.114	26.063	Visualize Results
MFF [5]	4.566	2.216	23.732	4.664	2.017	1.222	0.893	2.902	26.810	Visualize Results
IRR-PWC [6]	4.579	2.154	24.355	4.165	1.843	1.292	0.709	2.423	28.998	Visualize Results
PWC-Net+ [7]	4.596	2.254	23.696	4.781	2.045	1.234	0.945	2.978	26.620	Visualize Results
CompactFlow [8]	4.626	2.099	25.253	4.192	1.825	1.233	0.845	2.677	28.120	Visualize Results
HD3-Flow [9]	4.666	2.174	24.994	3.786	1.719	1.647	0.657	2.182	30.579	Visualize Results
LiteFlowNet2-MD+ [10]	4.728	2.249	24.939	4.010	1.925	1.504	0.783	2.634	29.369	Visualize Results

Experiments: Quantitative Results

- Our fine-tuned DistillFlow model achieves Fl-all = 5.94%, rank 1st among all monocular methods on KITTI 2015 benchmark

Additional information used by the methods

- Stereo: Method uses left and right (stereo) images
- Multiview: Method uses more than 2 temporally adjacent images
- Motion stereo: Method uses epipolar geometry for computing optical flow
- Additional training data: Use of additional data sources for training (see details)

Evaluation ground truth: All pixels | Evaluation area: All pixels

	Method	Setting	Code	Fl-bg	Fl-fg	Fl-all	Density	Runtime	Environment	Compare
1	StereoExp-v2	Stereo		2.86 %	9.05 %	3.89 %	100.00 %	2 s	GPU @ 2.5 Ghz (Python)	
2	UberATG-DRISF	Stereo		3.59 %	10.40 %	4.73 %	100.00 %	0.75 s	CPU+GPU @ 2.5 Ghz (Python)	
W. Ma, S. Wang, R. Hu, Y. Xiong and R. Urtasun: Deep Rigid Instance Scene Flow. CVPR 2019.										
3	ACOSF	Stereo		4.56 %	12.00 %	5.79 %	100.00 %	5 min	1 core @ 3.0 Ghz (Matlab + C/C++)	
C. Li, H. Ma and Q. Liao: Two-Stage Adaptive Object Scene Flow Using Hybrid CNN-CRF Model. International Conference on Pattern Recognition (ICPR) 2020.										
4	DistillFlow+ft			5.53 %	7.96 %	5.94 %	100.00 %	0.03 s	1 core @ 2.5 Ghz (Python)	
5	VCN+MSDRNet			5.57 %	7.78 %	5.94 %	100.00 %	0.5 s	1 core @ 2.5 Ghz (C/C++)	
6	PCF-F			6.05 %	5.99 %	6.04 %	100.00 %	0.08 s	GPU @ 2.5 Ghz (Python)	
7	PPAC-HD3		code	5.78 %	7.48 %	6.06 %	100.00 %	0.19 s	NVIDIA GTX 1080 Ti	
A. Wannenwetsch and S. Roth: Probabilistic Pixel-Adaptive Refinement Networks. CVPR 2020.										
8	MaskFlownet		code	5.79 %	7.70 %	6.11 %	100.00 %	0.06 s	NVIDIA TITAN Xp	
S. Zhao, Y. Sheng, Y. Dong, E. Chang and Y. Xu: MaskFlownet: Asymmetric Feature Matching with Learnable Occlusion Mask. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.										
9	ISF	Stereo		5.40 %	10.29 %	6.22 %	100.00 %	10 min	1 core @ 3 Ghz (C/C++)	
A. Behl, O. Jafari, S. Mustikovela, H. Alhaija, C. Rother and A. Geiger: Bounding Boxes, Segmentations and Object Coordinates: How Important is Recognition for 3D Scene Flow Estimation in Autonomous Driving Scenarios?. International Conference on Computer Vision (ICCV) 2017.										

Experiments: Ablation Study

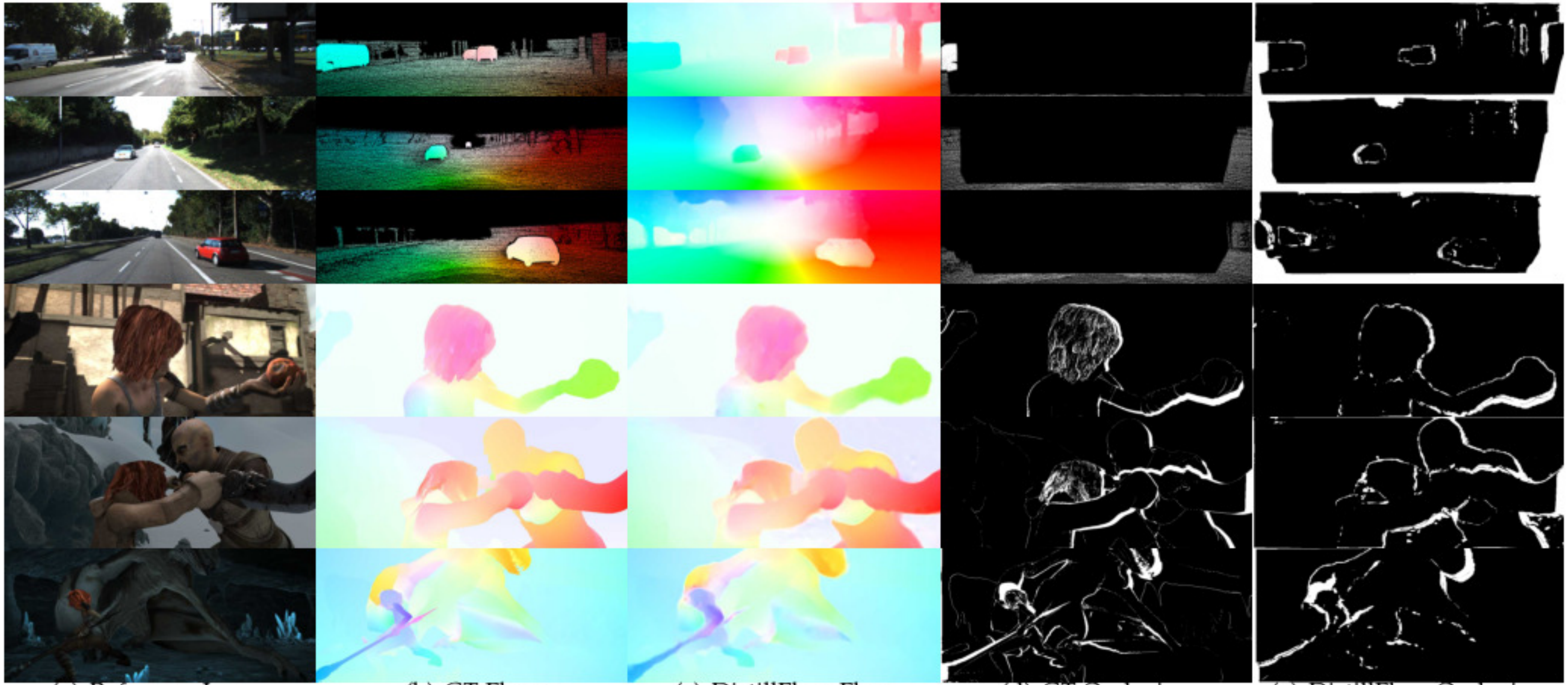
- Self-supervision greatly improves the optical flow estimation performance, especially for occluded pixels: more than 50% on KITTI
- Self-supervision is agnostic to network structures

Network Backbone	Occlusion Handling	Edge-Aware Smoothness	Data Distillation	Model Distillation	KITTI 2012					KITTI 2015				
					EPE-all	EPE-noc	EPE-occ	Fl-all	Fl-noc	EPE-all	EPE-noc	EPE-occ	Fl-all	Fl-noc
PWC-Net	✗	✗	✗	✗	7.73	1.41	49.63	18.08%	6.90%	14.02	4.57	73.74	25.34%	14.37%
	✓	✗	✗	✗	4.67	1.05	28.61	14.93%	5.32%	9.21	3.26	46.85	21.20%	11.07%
	✓	✓	✗	✗	3.36	0.97	19.18	13.31%	4.30%	7.83	3.28	36.55	19.91%	10.12%
	✓	✓	✓	✗	1.68	0.87	7.10	5.73%	3.56%	4.61	2.53	17.77	11.71%	8.66%
	✓	✓	✓	✓	**1.64**	**0.85**	**6.84**	**5.67%**	**3.53%**	**4.32**	**2.40**	**16.43**	**11.61%**	**8.64%**
PWC-Net†	✗	✗	✗	✗	7.33	1.30	47.26	16.27%	5.97%	12.49	3.59	68.82	23.07%	12.40%
	✓	✗	✗	✗	3.22	0.98	18.07	13.57%	4.40%	6.57	2.88	29.87	19.90%	10.01%
	✓	✓	✗	✗	2.92	0.93	16.06	12.44%	3.94%	6.45	2.59	30.90	19.08%	9.48%
	✓	✓	✓	✗	1.46	0.85	5.44	5.17%	3.38%	3.20	2.08	10.28	10.05%	8.03%
	✓	✓	✓	✓	**1.38**	**0.83**	**4.98**	**4.99%**	**3.25%**	**2.93**	**1.96**	**9.04**	**9.79%**	**7.81%**

Network Backbone	Knowledge Distillation	KITTI 2012			KITTI 2015			Sintel Clean			Sintel Final		
		EPE-all	EPE-noc	EPE-occ	EPE-all	EPE-noc	EPE-occ	EPE-all	EPE-noc	EPE-occ	EPE-all	EPE-noc	EPE-occ
FlowNetS	✗	4.26	1.53	22.34	8.85	3.82	40.63	(5.05)	(3.09)	(30.01)	(5.38)	(3.38)	(31.00)
	✓	**2.70**	**1.38**	**11.44**	**6.33**	**3.44**	**24.59**	**(4.20)**	**(2.36)**	**(27.66)**	**(4.83)**	**(2.90)**	**(29.49)**
FlowNetC	✗	3.63	1.26	19.31	8.11	3.45	37.61	(4.20)	(2.36)	(27.66)	(4.83)	(2.90)	(29.49)
	✓	**2.18**	**1.16**	**8.97**	**5.47**	**2.95**	**21.38**	**(3.45)**	**(1.90)**	**(23.27)**	**(4.17)**	**(2.52)**	**(25.36)**

Experiments: Qualitative Results

- Sample unsupervised results on KITTI and Sintel dataset. From top to bottom, samples are from KITTI 2015 and Sintel Final

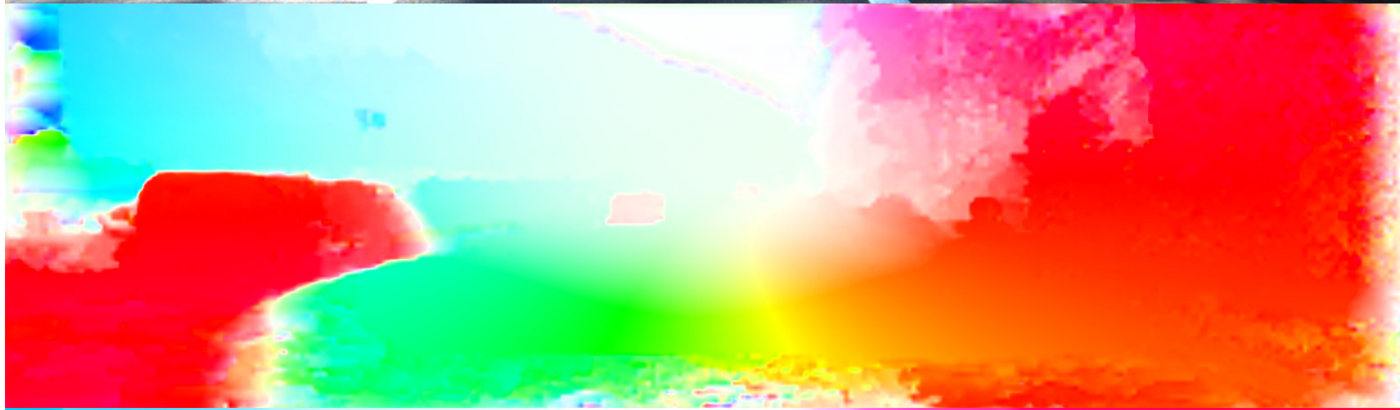


(a) Reference Image (b) GT Flow (c) DistillFlow Flow (d) GT Occlusion (e) DistillFlow Occlusion

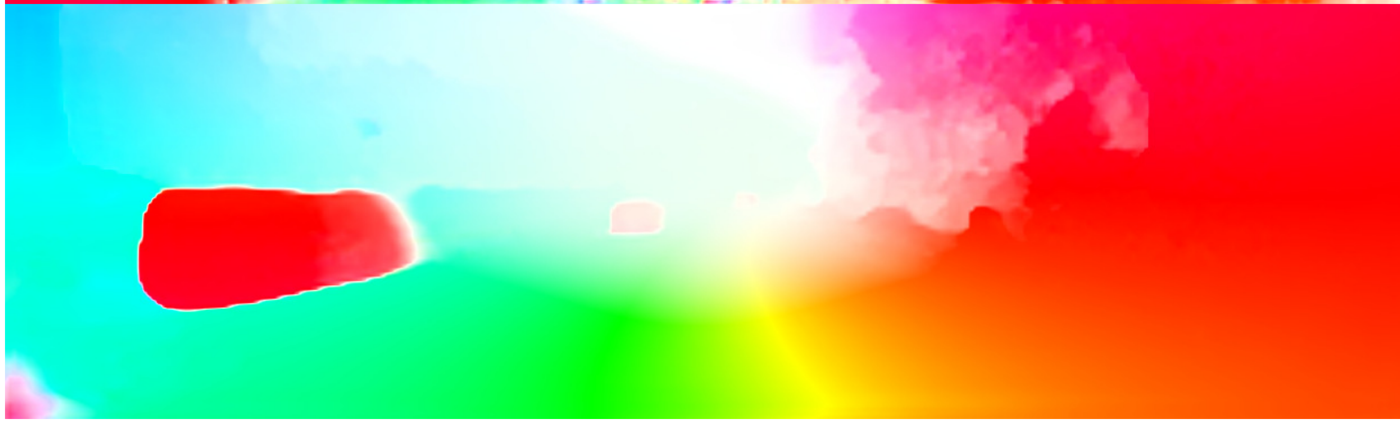
Experiments: Effect of Self-Supervision

Reference Image

Flow Estimation
without Self-supervision



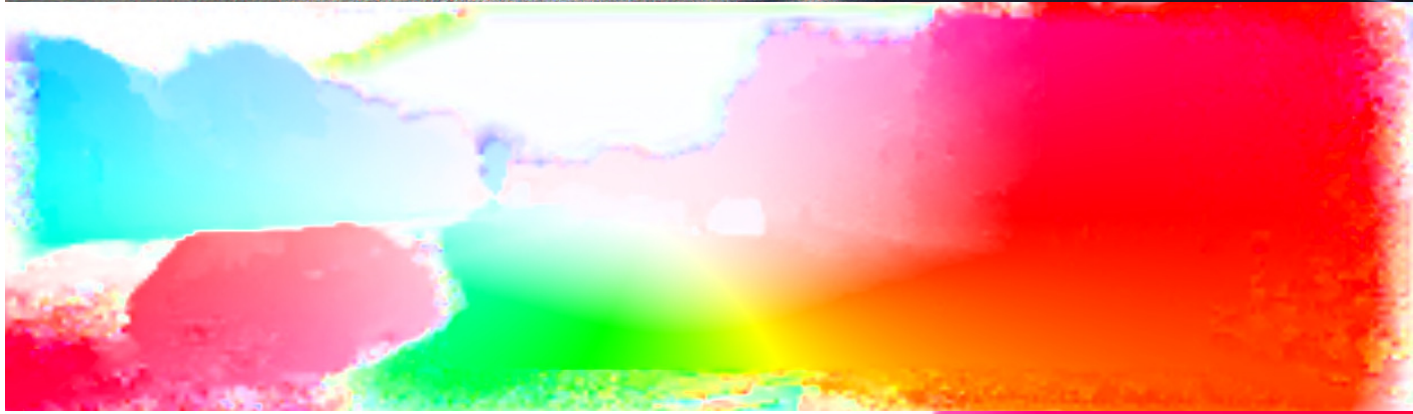
Flow Estimation
with Self-supervision



Experiments: Effect of Self-Supervision

Reference Image

Flow Estimation
without Self-supervision



Flow Estimation
with Self-supervision

Experiments: Effect of Self-Supervision

Reference Image

Flow Estimation
without Self-supervision

Flow Estimation
with Self-supervision

Comparison with State-of-the-art

Reference Image

Flow Estimation using PWC-Net

Flow Estimation using Our Fine-tuned Model

Generalization on Real-World Videos

Reference Image

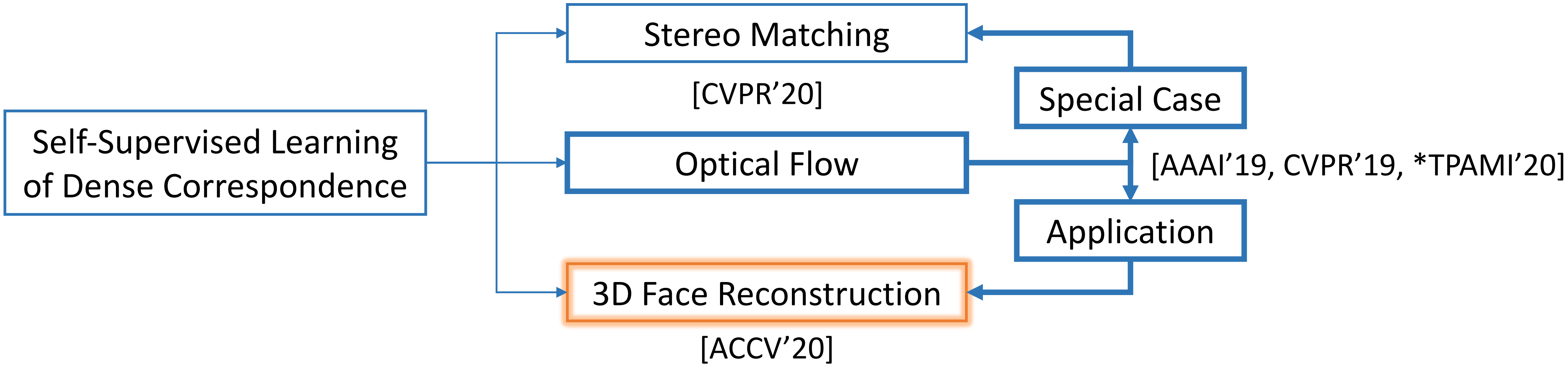
Flow from Our Unsupervised Model

Flow from Our Fine-tuned Model

Summary

- Propose a series of self-supervised learning methods to effectively learn optical flow from unlabeled data, which improve performance >30% than previous methods on average
- Self-supervised learning enables us to utilize more data, and our models have strong generalization capability
- Self-supervised training provides excellent initializations for supervised fine-tuning, which removes the need of synthetic data. This is a new perceptive in supervised flow learning

Thesis Contributions



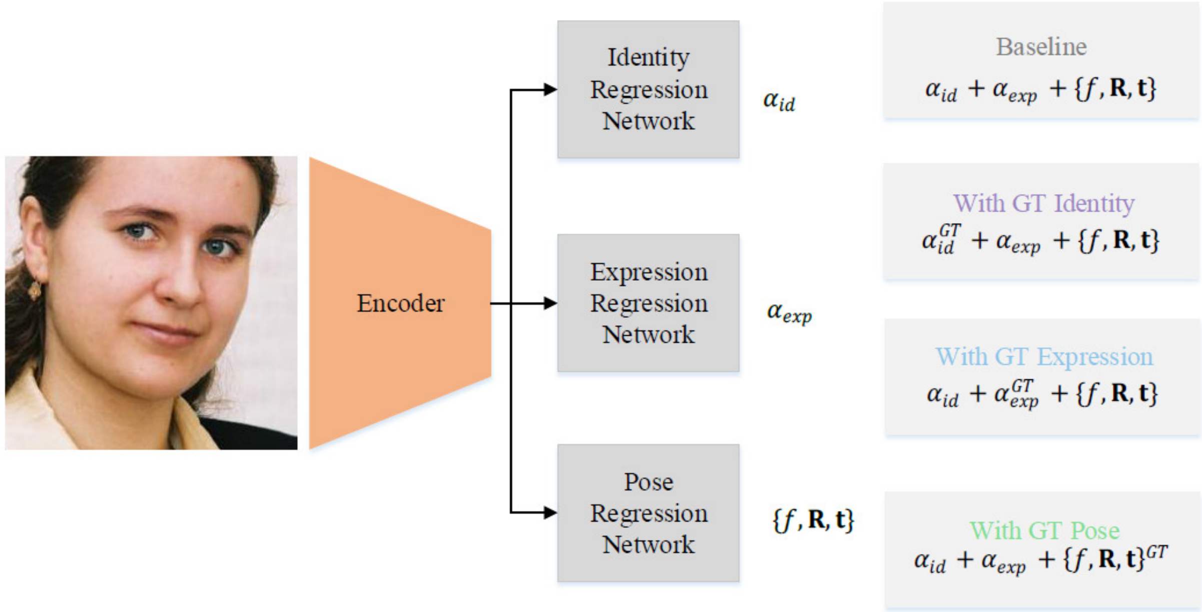
- Optical Flow: a series of self-supervised learning methods to learn optical flow of both occluded and non-occluded pixels
- Stereo Matching: explore the geometric relationship between flow and stereo
- 3D face reconstruction: pose guidance network and multi-image consistency

* In Submission

Motivation 1

- When predicting pose, identity and expression parameters simultaneously, regressing pose dominates the optimizing procedure, making it hard to obtain accurate 3D face parameters

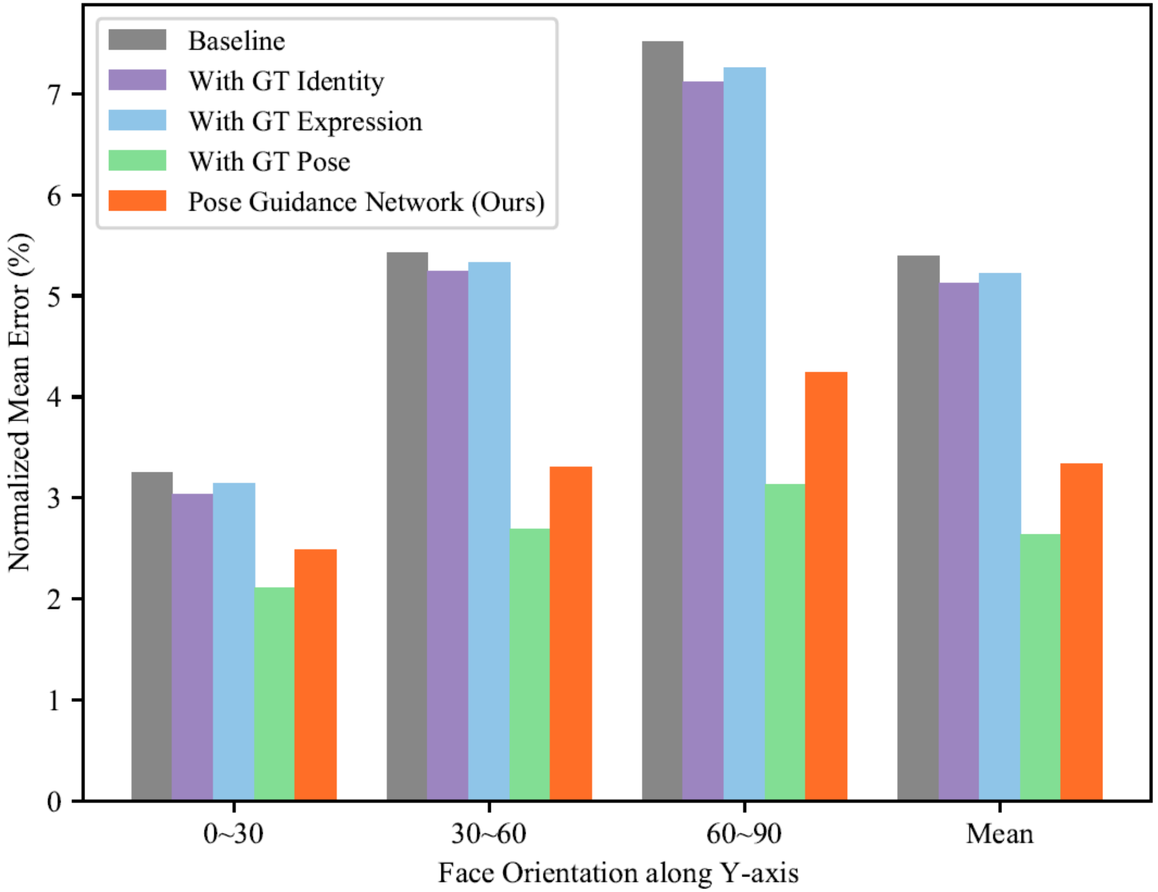
  - ➢Firstly, we train a neural network to simultaneously regress the identity, expression and pose parameters (**Baseline**)
  - ➢Then, we independently replace the predicted identity, expression, and pose parameters with their corresponding ground truth parameters, their errors change to **With GT Identity, Expression, Pose**



Motivation 1

- When predicting pose, identity and expression parameters simultaneously, regressing pose dominates the optimizing procedure, making it hard to obtain accurate 3D face parameters

  - **With GT Pose** reduces the error much more than other two → Regressing pose parameters dominates the optimizing procedure
  - **Pose Guidance Network (Ours)** effectively reduces the error compared to directly regressing the pose parameters and provides informative priors for reconstruct the 3D face



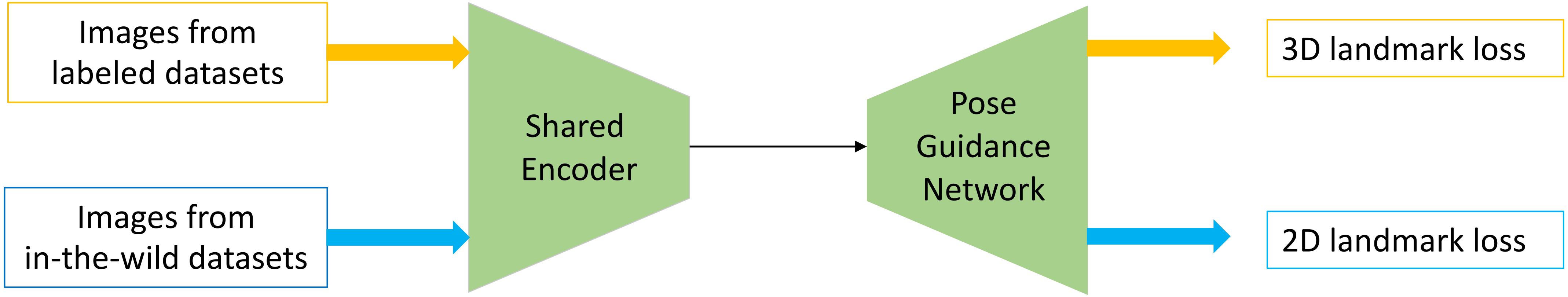
Motivation 2

- 3D face reconstruction from a single 2D image is an ill-posed problem due to depth ambiguity, we propose to learn face reconstruction from multiple frames of the same person

- A novel self-supervised learning scheme built on a visible texture swapping module is introduced:
  - Carefully handle the occlusion and illumination change across frames
  - Self-consistency losses:
    - Photometric space (employ census transform)
    - Optical flow space
    - Semantic space

Method

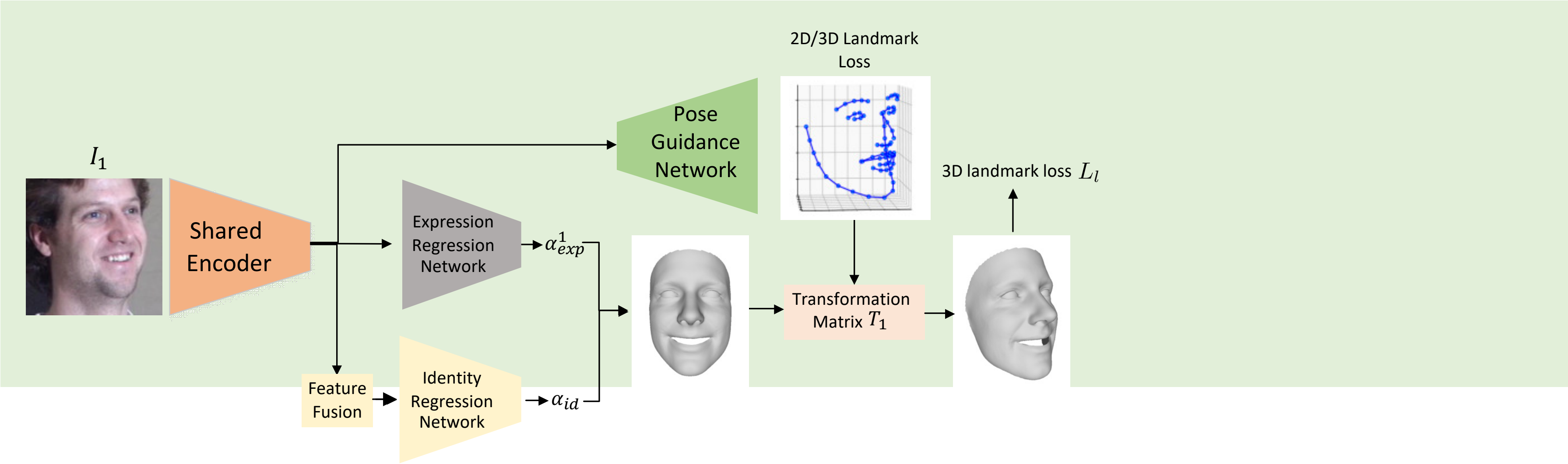
- Step 1: Train shared encoder and pose guidance network, which are fixed during the following steps



We can leverage **unlimited** in-the-wild images

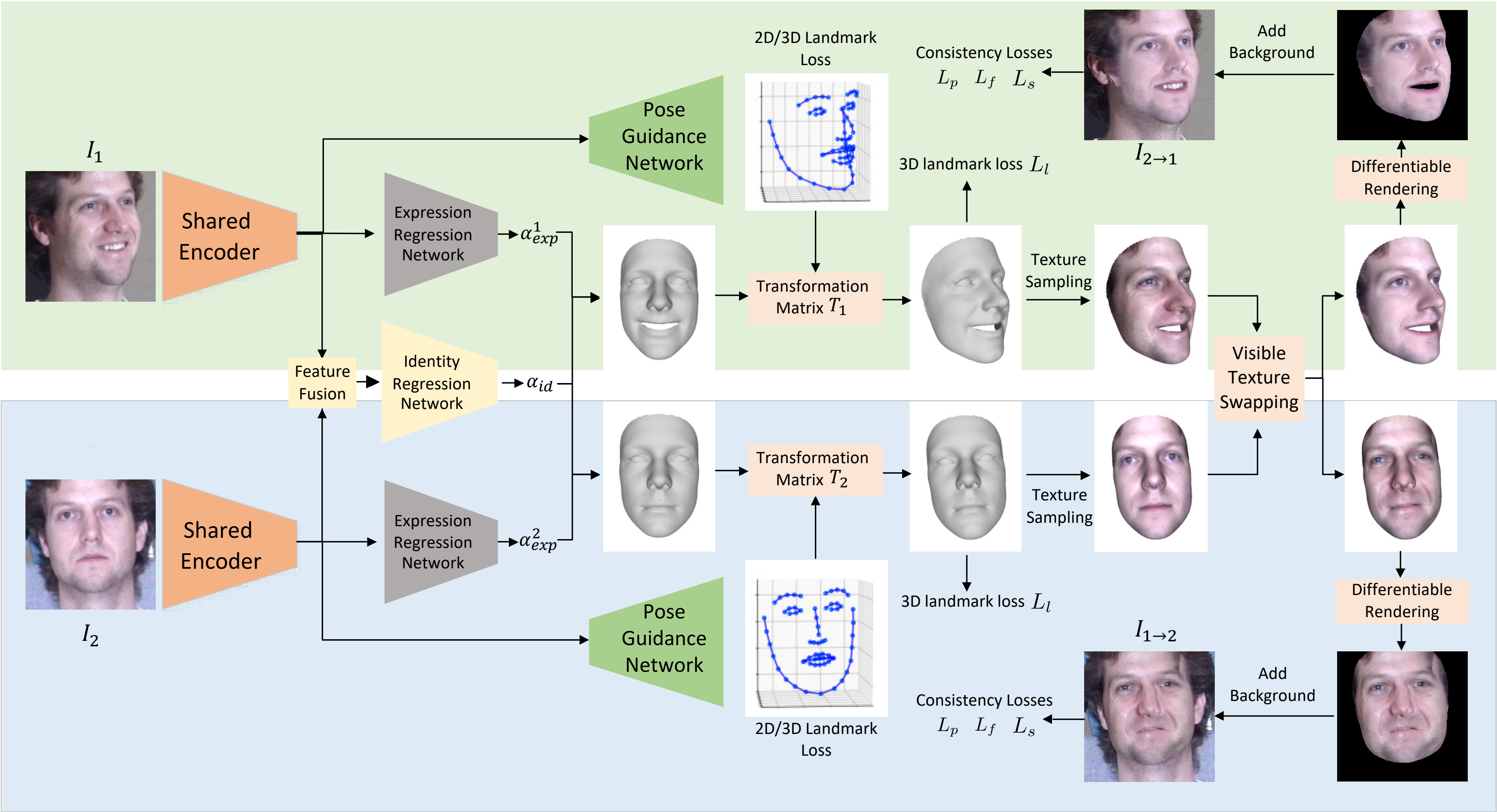
Method

- Step 2: Pre-train using one image with 3D landmark loss L_l and regularization loss L_r



Method

- Step 3: Train using **multiple** images with **full losses**



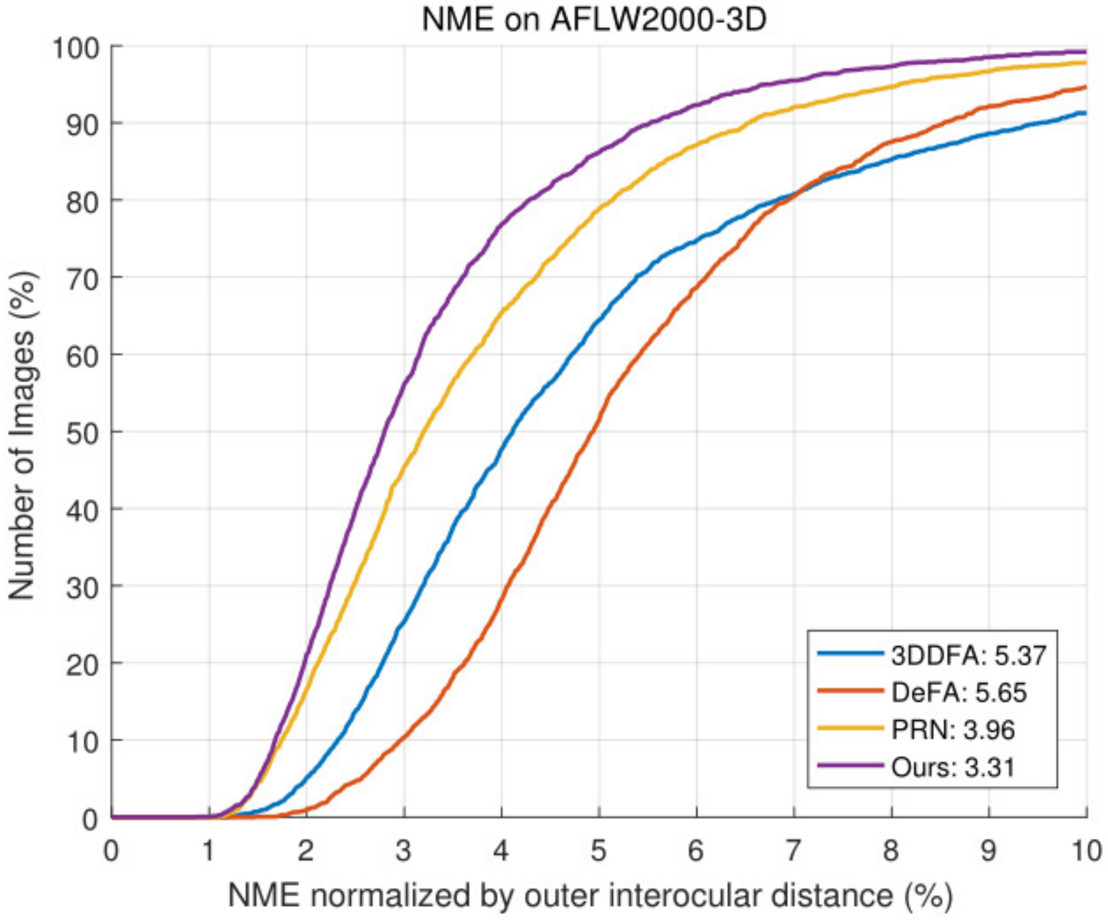
Experiments: Quantitative Results

- We achieve state-of-the-art 2D landmark estimation performance on ALFW2000-3D dataset

Method	NME_{2d}^{68}			
	0 to 30	30 to 60	60 to 90	Mean
SDM[37]	3.67	4.94	9.67	6.12
3DDFA [40]	3.78	4.54	7.93	5.42
3DDFA + SDM [40]	3.43	4.24	7.17	4.94
Yu et al. [39]	3.62	6.06	9.56	-
3DSTN[2]	3.15	4.33	5.98	4.49
DeFA[23]	-	-	-	4.50
Face2Face [34]	3.22	8.79	19.7	10.5
3DFAN [5]	2.77	3.48	4.61	3.62
PRN [12]	2.75	3.51	4.61	3.62
ExpNet [9]	4.01	5.46	6.23	5.23
MMFace-PMN [38]	5.05	6.23	7.05	6.11
MMFace-ICP-128 [38]	2.61	3.65	4.43	3.56
Ours (Pose Guidance Network)	**2.49**	**3.30**	4.24	**3.34**
Ours (3DMM)	2.53	3.32	**4.21**	3.36

Experiments: Quantitative Results

- We achieve state-of-the-art 3D face reconstruction performance on ALFW2000-3D dataset



Experiments: Quantitative Results

- We achieve state-of-the-art 3D shape estimation performance on Florence dataset

Table 2. **Comparison of mean point-to-plane error on the Florence dataset.** Results of other methods are from MVF [36].

Method	Indoor-Cooperative		PTZ-Indoor	
	Mean	Std	Mean	Std
Tran *et al.* [35]	1.443	0.292	1.471	0.290
Tran *et al.* + pool	1.397	0.290	1.381	0.322
Tran *et al.* + [27]	1.382	0.272	1.430	0.306
MoFA [33]	1.405	0.306	1.306	0.261
MoFA + pool	1.370	0.321	1.286	0.266
MoFA + [27]	1.363	0.326	1.293	0.276
Genova *et al.* [13]	1.405	0.339	1.271	0.293
Genova *et al.* + pool	1.372	0.353	1.260	0.310
Genova *et al.* + [27]	1.360	0.346	1.246	0.302
MVF [36] - pretrain	1.266	0.297	1.252	0.285
MVF [36]	1.220	0.247	1.228	0.236
Ours	**1.122**	**0.219**	**1.161**	**0.224**

Experiments: Quantitative Results

- On FaceWarehouse dataset:
  - Single-frame: similar performance with MoFA, Inversefacenet and Tewari *et al.* [34]
  - Multi-frame: outperform FML by 7.5%
  - Pose guidance network and multi-frame self-supervised learning scheme improve the performance

Table 2: **Per-vertex geometric error (measured in mm) on FaceWarehouse dataset.** PGN denotes pose guidance network. Our approach obtains the lowest error, outperforming the best prior art [33] by 7.5%.

Method	MoFA [35]	Inversefacenet [20]	Tewari *et al.* [34]	FML [33]	Ours Single-Frame without PGN	Ours Single-Frame with PGN	Ours Mult-Frame without PGN	Ours Multi-Frame with PGN
Error	2.19	2.11	2.03	2.01	2.18	2.09	1.98	**1.86**

Experiments: Quantitative Results

- Ablation study on Florence dataset demonstrates the effectiveness of photometric consistency loss, census transform, flow consistency loss and semantic consistency loss

(a) **Ablation study on Florence.**

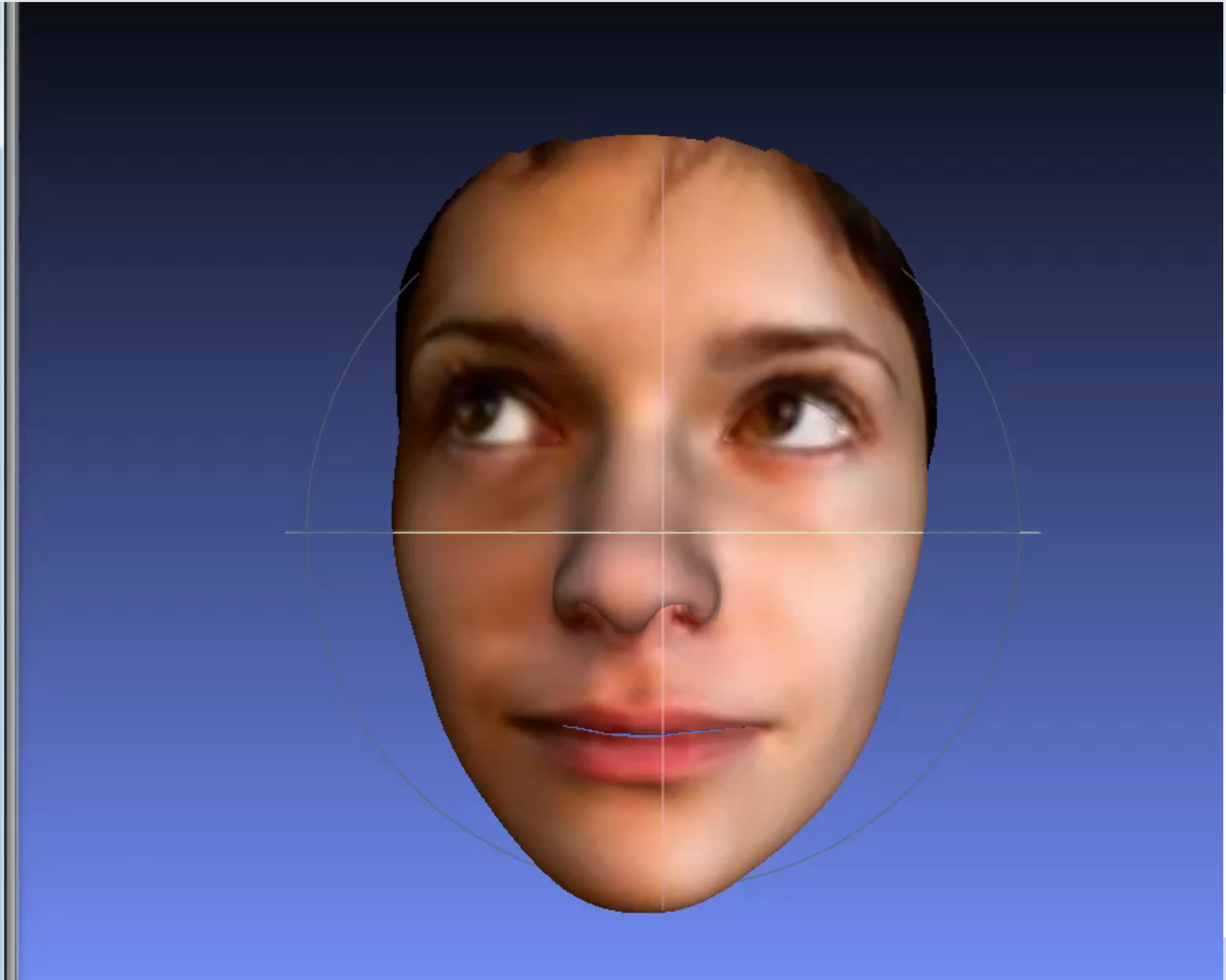
L_{p-}	L_p	L_s	L_f	Indoor-Cooperative Mean	Indoor-Cooperative Std	PTZ-Indoor Mean	PTZ-Indoor Std
✗	✗	✗	✗	1.364	0.352	1.379	0.326
✓	✗	✗	✗	1.263	0.312	1.323	0.251
✗	✓	✗	✗	1.219	0.261	1.255	0.256
✗	✓	✗	✓	1.193	0.230	1.221	0.247
✗	✓	✓	✗	1.161	0.268	1.269	0.276
✗	✓	✓	✓	**1.122**	**0.219**	**1.161**	**0.224**

Experiments: Qualitative Results

Input Image

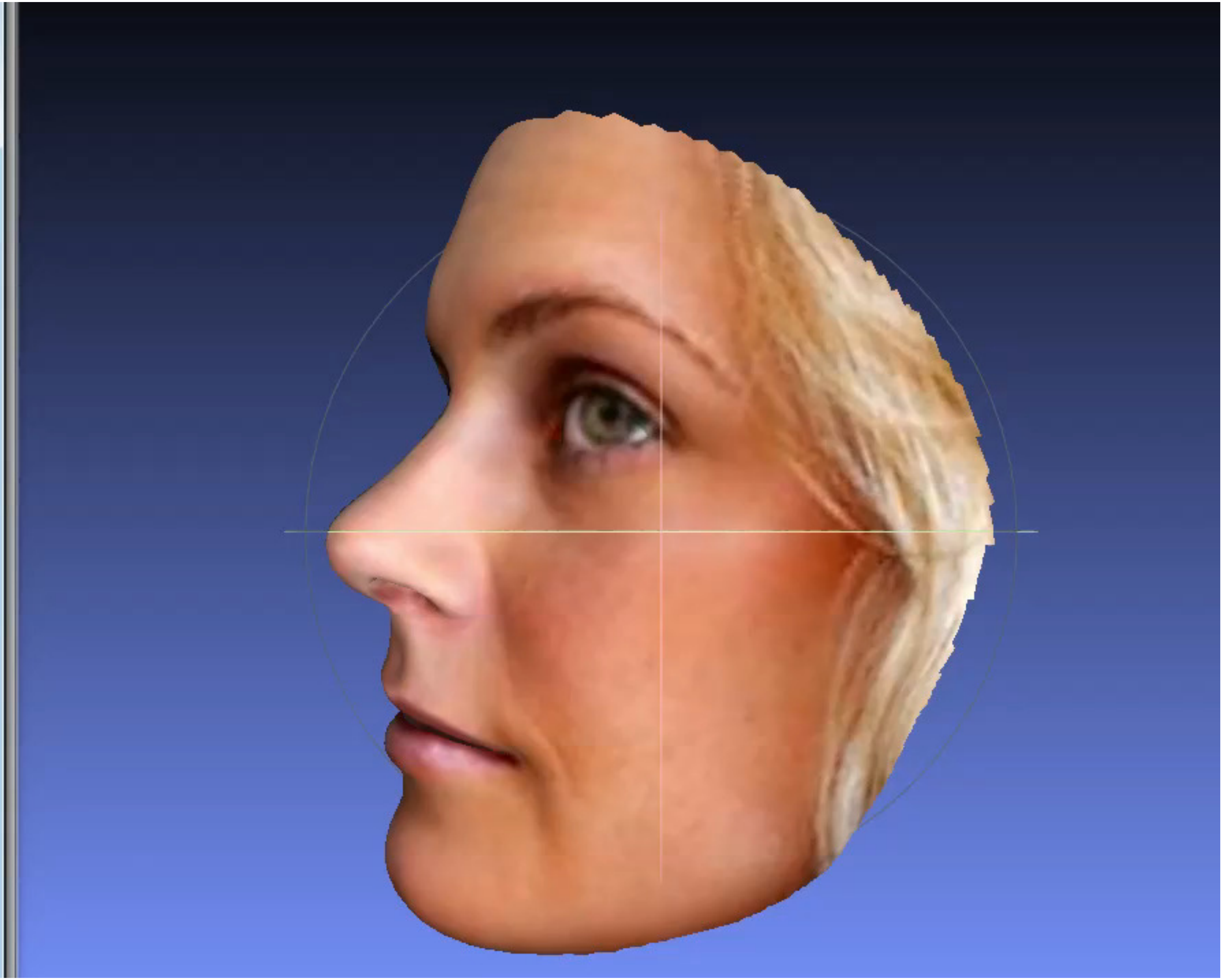
3D Face Geometry

3D Face Texture

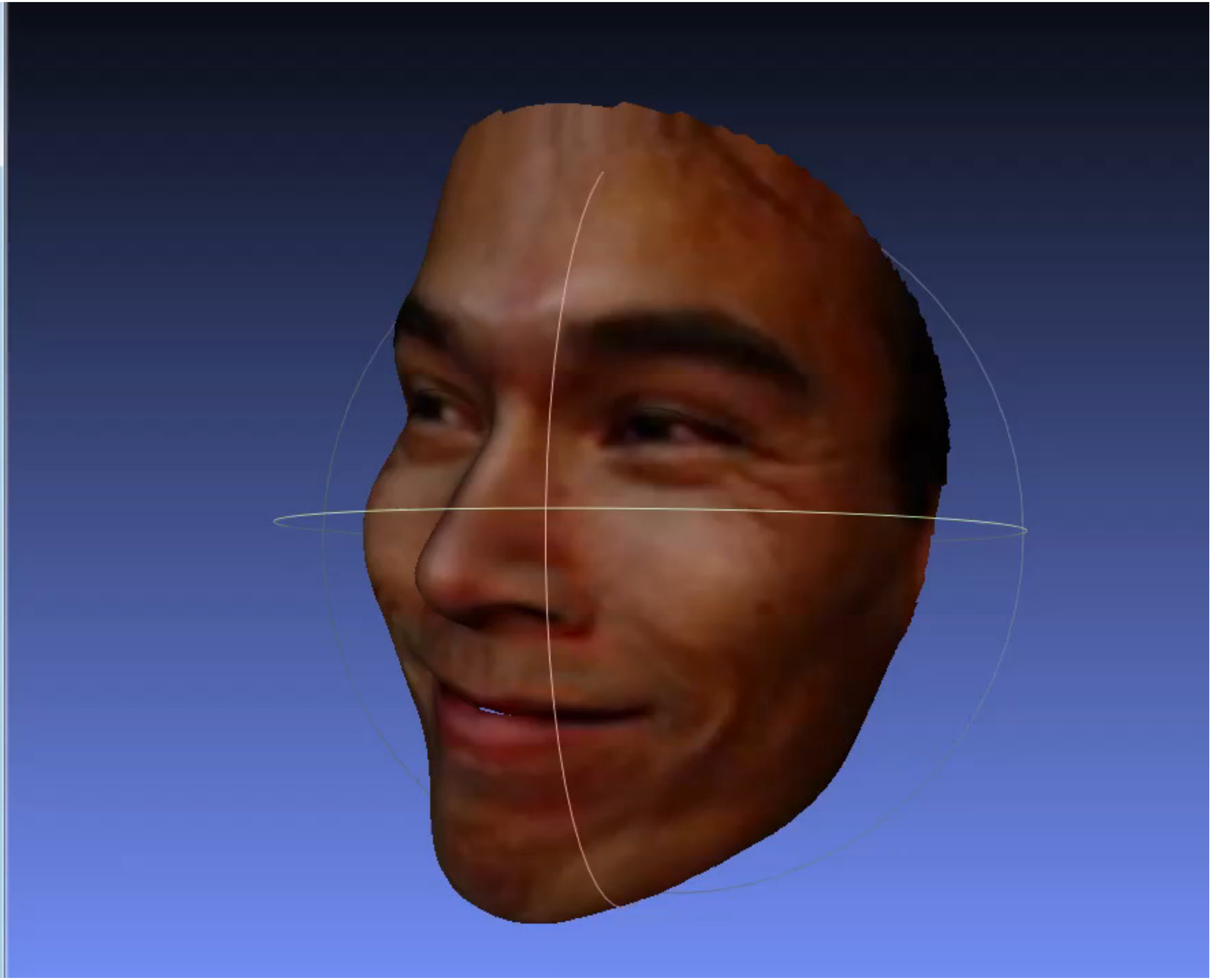


Our model estimates accurate 3D face shape, which fits well with texture.

For profile faces, we can also obtain accurate 3D face reconstruction.

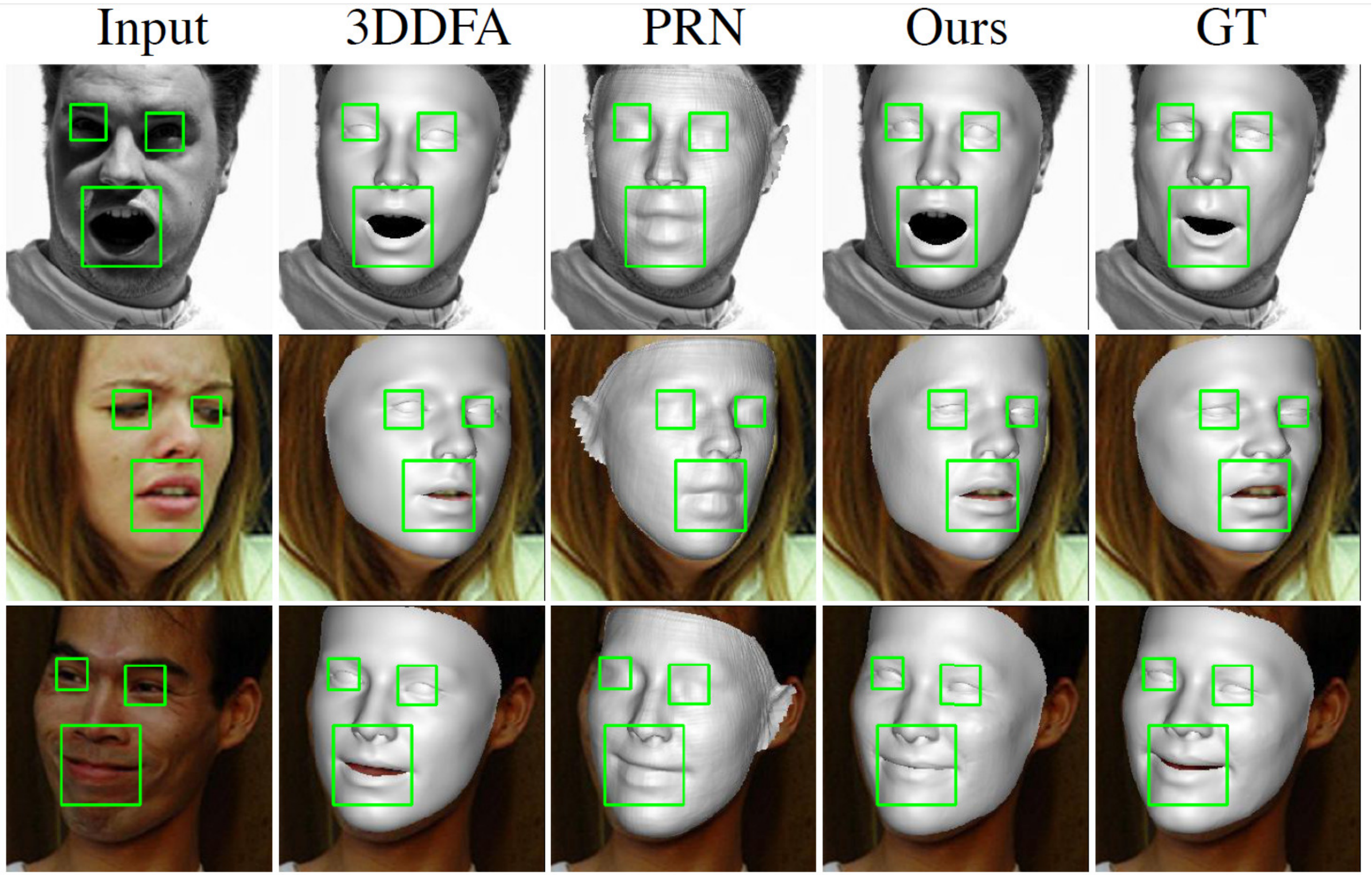


Our model still works well for complicated expressions.



Experiments: Qualitative Results

- Comparison with other methods on ALFW2000-3D dataset



Experiments: Qualitative Results

Our multi-image face reconstruction method is based on **texture sampling**, therefore texture quality shall have a big impact. To verify this, we fine-tune our model on a **high-quality video** from Youtube.

Our model can generate very accurate shape and expression, such as the challenging expression of complete eye-closing.

Summary

- Propose a pose guidance network to predict the 3D landmarks for estimating the pose parameters
- Utilize both annotated images with 3D landmarks and unlabeled images with pseudo 2D landmarks
- Explore multi-frame consistency based on a visible texture swapping module

Future Work

- More accurate optical flow estimation
  - **Occlusion detection:** soft mask vs. hard mask
  - **Robust transform:** learned transforms vs. hand-crafted transforms
  - **Network architecture:** quarter resolution vs. full resolution
  - **Multi-task learning:** joint learn optical flow and depth
  - **External guidance:** utilize dense annotations in synthetic data

Future Work

- Optical flow-based applications
  - Optical flow as fixed features: straightforward
  - Optical flow with **task-specific** patterns
    - TV-Net [Fan L .et CVPR 2018] for video action recognition.

TV-L1 is extracted optical flow features

TV-Net with training is the learned flow-like features.

With training, TVNet generates more abstractive motion features than TV-L1.

Publications

[1] **Pengpeng Liu,** Xintong Han, Irwin King, Michael Lyu, Jia Xu. *Unsupervised Domain Adaptation for Optical Flow Estimation.* **(CVPR 2021)** *

[2] **Pengpeng Liu,** Irwin King, Michael R. Lyu and Jia Xu. *Learning by Distillation: A Self-Supervised Learning Framework for Optical Flow Estimation.* (**TPAMI 2020**) *

[3] **Pengpeng Liu,** Xintong Han, Michael Lyu, Irwin King, Jia Xu. *Learning 3D Face Reconstruction with a Pose Guidance Network.* (**ACCV 2020, Oral**)

[4] **Pengpeng Liu,** Michael Lyu, Irwin King, Jia Xu. *Flow2Stereo: Effective Self-Supervised Learning of Optical Flow and Stereo Matching.* (**CVPR 2020**)

[5] **Pengpeng Liu,** Michael Lyu, Irwin King, Jia Xu. *SelfFlow: Self-Supervised Learning of Optical Flow.* (**CVPR 2019, Oral, Best Paper Finalist**)

[6] **Pengpeng Liu,** Irwin King, Michael Lyu, Jia Xu. *DDFlow: Learning Optical Flow with Unlabeled Data Distillation.* (**AAAI 2019, Oral**)

[7] **Pengpeng Liu,** Xiaojuan Qi, Pinjia He, Yikang Li, Michael Lyu and Irwin King. *Semantically Consistent Image Completion with Fine-grained Details.* (ArXiv Technical Report 2018)

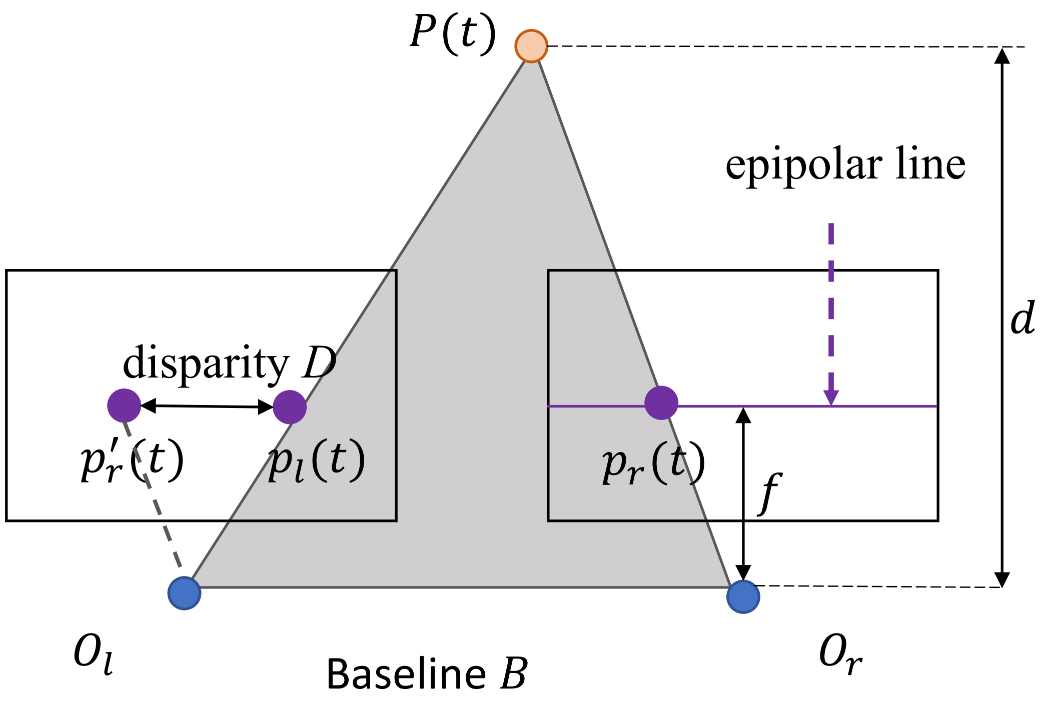
* denotes in submission

Thanks!

Back up slides

Correspondence is Crucial

- Stereo matching for rectified image pairs



- Epipolar line is horizontal.
- $D = p_l(t) - p'_r(t)$
- Suppose f is focal length, d is depth, B is the distance between two cameras, then $d = fB/D$.

Disparity is inversely proportional to depth!

Motivation

- Unsupervised Learning Methods
  - How to effectively learn optical flow of **occluded** pixels?
  - How to reduce the **performance gap** compared with supervised learning methods?

- Supervised Learning Methods
  - Can we **remove** the reliance of **synthetic data**?
  - Can we **simplify** the training procedure?

Loss Functions

- Occlusion estimation: based on the forward-backward consistency prior

$$\begin{cases} |\mathrm{w}_f + \hat{\mathrm{w}}_f|^2 < \alpha_1(|\mathrm{w}_f|^2 + |\hat{\mathrm{w}}_f|^2) + \alpha_2, \\ \mathbf{p} + \mathrm{w}_f(\mathbf{p}) \in \Omega, \end{cases}$$

- Photometric loss

$$L_p = \sum \psi(I_1 - I_2^w) \odot (1 - O_f) / \sum (1 - O_f) \\ + \sum \psi(I_2 - I_1^w) \odot (1 - O_b) / \sum (1 - O_b)$$

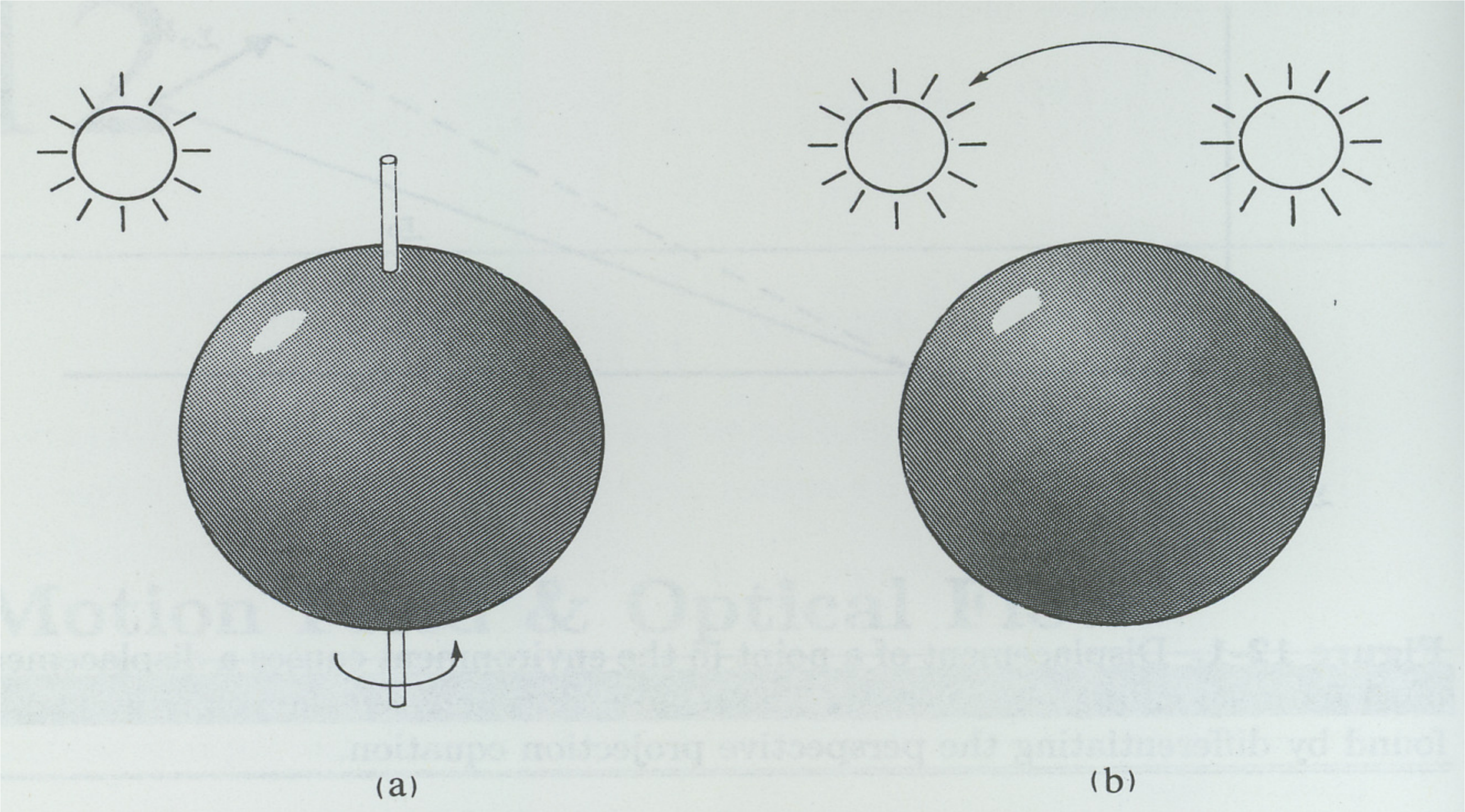
- Loss for occluded pixels

$$M_f = \mathrm{clip}(\widetilde{O}_f - O_f^p, 0, 1)$$

$$L_o = \sum \psi(\mathrm{w}_f^p - \widetilde{\mathrm{w}}_f) \odot M_f / \sum M_f \\ + \sum \psi(\mathrm{w}_b^p - \widetilde{\mathrm{w}}_b) \odot M_b / \sum M_b$$

- $\psi(x)$ is a robust loss function.

Optical Flow ≠ Motion Field



Motion field exists but no optical flow

No motion field but shading changes

Background

- 3DMM: represents 3D faces with linear combination of PCA vectors.
- 3 types of parameters: identity, expression and pose parameters.
- Face geometry:

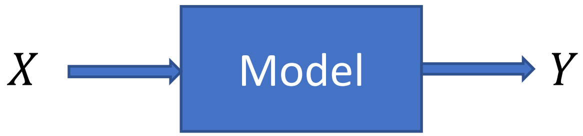
$$\mathbf{S}(\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}) = \overline{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha}_{id} + \mathbf{B}_{exp}\boldsymbol{\alpha}_{exp}.$$

- Projection:

$$\mathbf{v}(\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot (f \cdot \mathbf{R} \cdot \mathbf{s} + \mathbf{t}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \cdot [f \cdot \mathbf{R} \quad \mathbf{t}] \cdot \begin{bmatrix} \mathbf{s} \\ 1 \end{bmatrix}$$

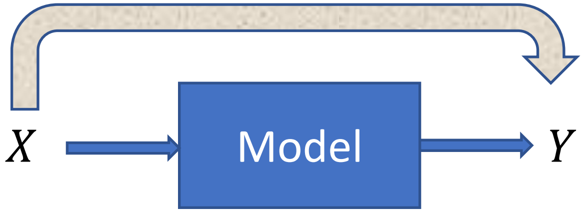
Self-Supervised Learning

- Definition: a form of unsupervised learning where the supervision signal is purely generated from the data itself (no manual labeling)



Supervised Learning

Automatically generated



Self-Supervised Learning

- In computer vision, it usually contains two stages:
  - Design a pre-text task to learn representative features or generate pseudo labels
  - Employ the learned features or labels to train deep learning models in a supervised manner

Transformation Matrix

$$\min_{\mathbf{T}} ||\mathbf{T} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix} - \mathbf{X}_{UV}||_2$$

$$\mathbf{T} = \mathbf{X}_{UV} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix}^T \cdot \left(\begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix}^T \right)^{-1}$$