Unsupervised Learning of Holistic 3D Scene Understanding

Zhenheng Yang

01/10/2019

University of Southern California

Brief Bio

Zhenheng Yang (杨振恒)

- Education:
  - Bachelor of Engineering (BEng), Tsinghua University (2010-2014)
  - Ph.D. candidate: Computer vision, University of Southern California (2014-current)

- Working Experience:
  - Research Intern: Baidu Research USA. (May – Aug., 2017)
  - Research Intern: Facebook Research (May – Aug., 2018)

Guidelines

- Introduction
  - Problem statement
  - Challenges
  - Our contributions
- Our work
  - Unsupervised depth estimation
  - Unsupervised optical flow estimation
  - Joint unsupervised learning of geometry and motion

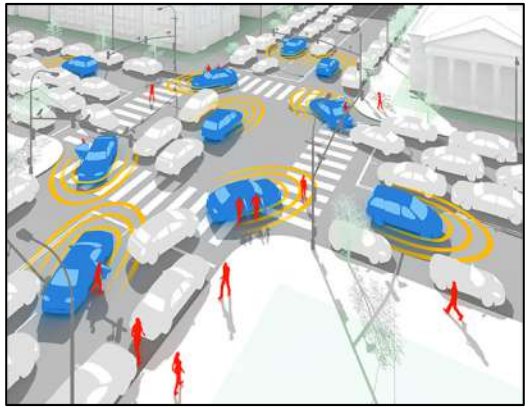
Introduction

- Problem statement
- Challenges
- Our contributions

Problem statement

- Understanding 3D scene layout is a fundamental computer vision problem and has many applications in real life

Augmented reality (AR)

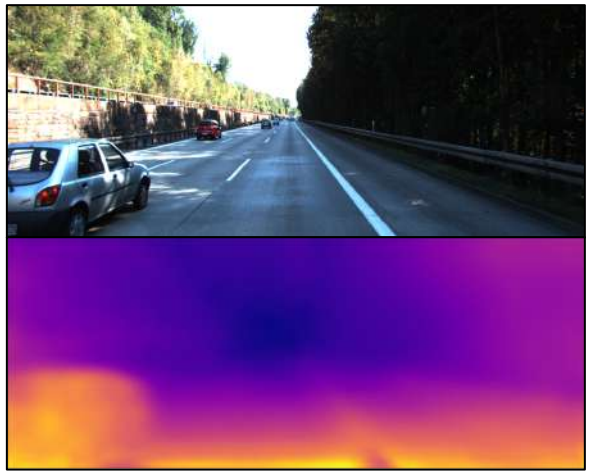


Autonomous driving

Robotics

Problem statement

- 3D scene understanding aims at estimating the 3D geometries of the observed scene
- There are many tasks in 3D scene understanding (static, dynamic)



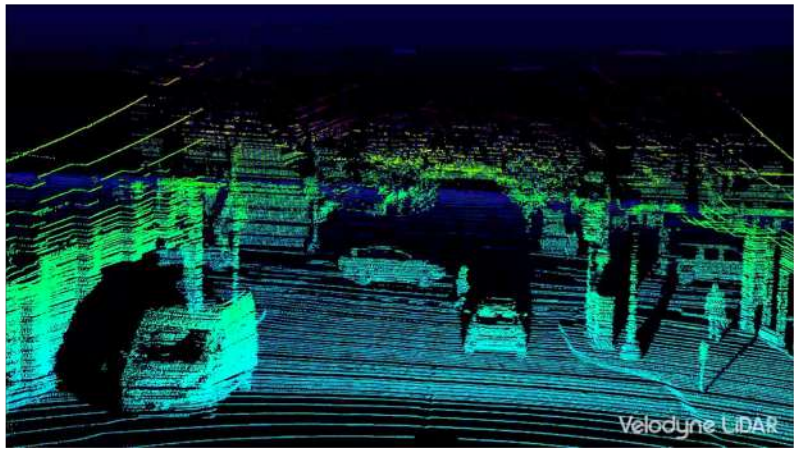
Depth estimation

Segmentation

Motion estimation

Challenges

- Tedious and sometimes impossible annotations



Sparse LiDAR points

- Geometrical cues are coupled
  - Movement of pixels can be caused by moving camera or moving object

Our contributions

- Unsupervised learning of 3D geometry (free of annotations)
- Decomposing geometrical cues and joint learning (coupled geometrical cues)

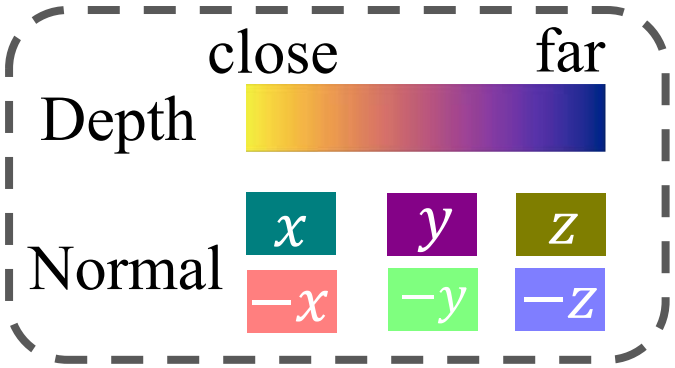
Works

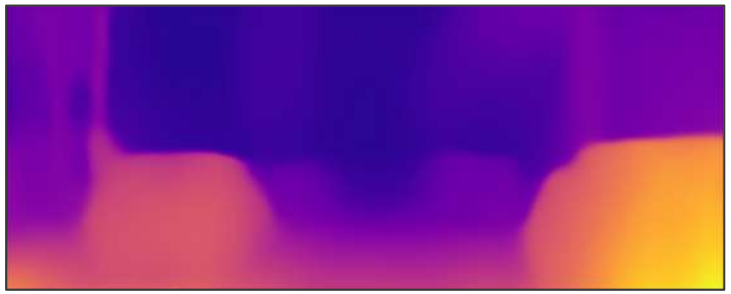
- Unsupervised learning of static 3D cues
- Unsupervised learning of dynamic 3D cues
- Unsupervised joint learning of motion and geometry with holistic 3D understanding

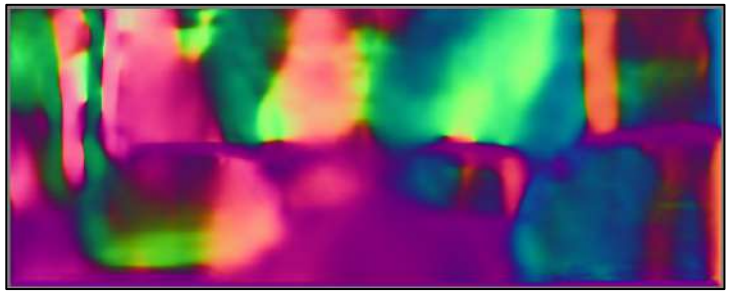
Guidelines

- **Unsupervised depth estimation**
- Unsupervised optical flow estimation
- Joint unsupervised learning of geometry and motion

Unsupervised learning of static 3D cues







Unsupervised Learning of Geometry with Edge-aware Depth-Normal Consistency, AAAI18' (oral)

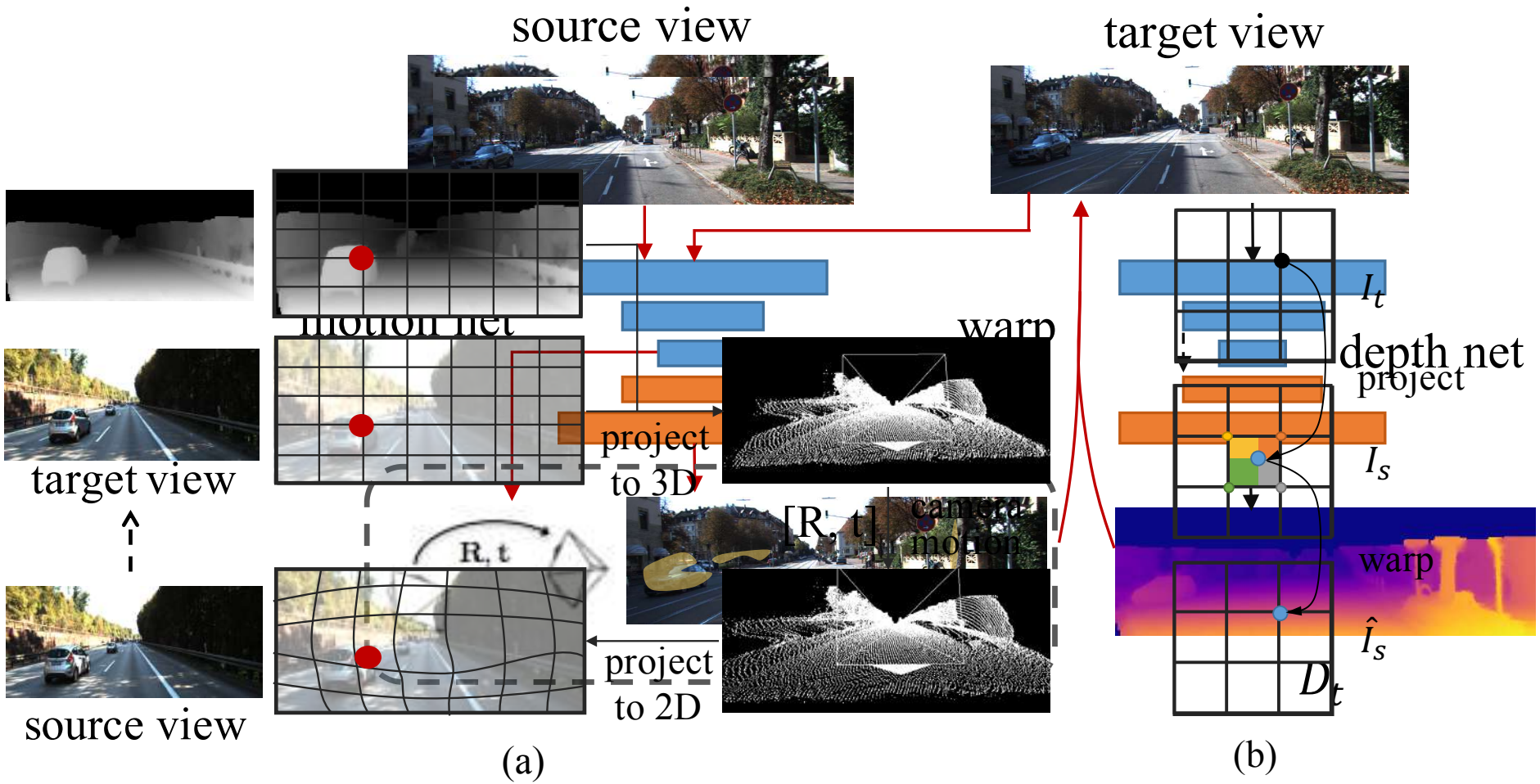
LEGO: Learning Edge with Geometry all at Once by Watching Videos, CVPR18' (spotlight)

Data samples

training

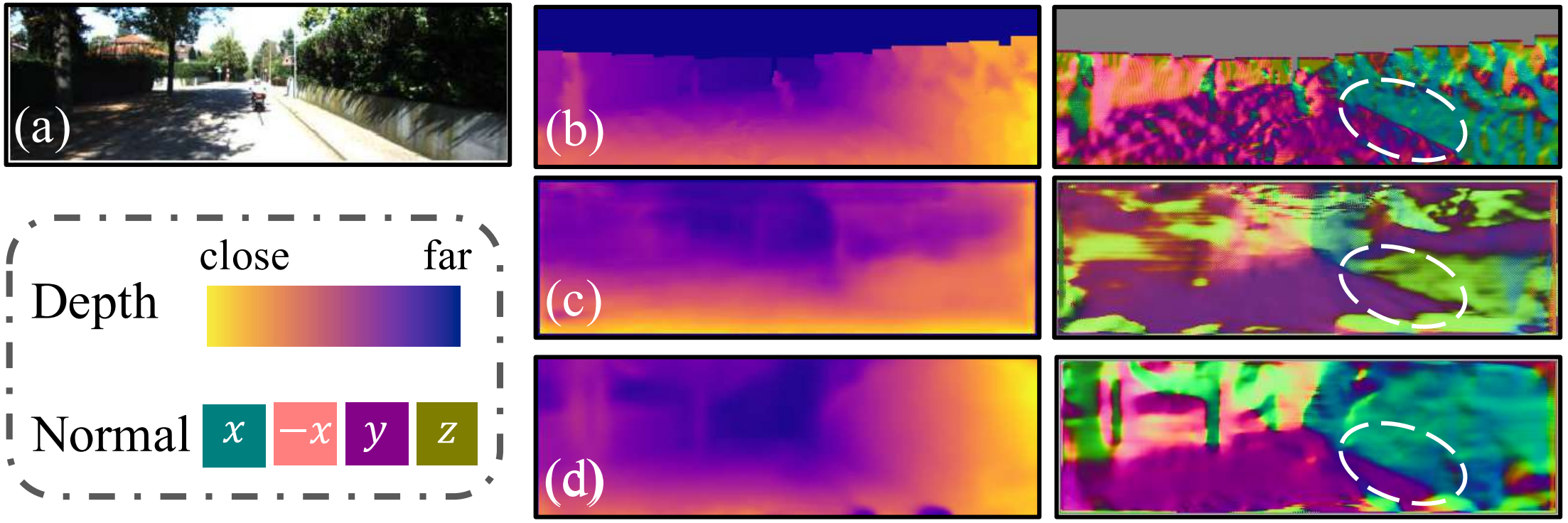
testing

Previous work



Zhou *et al.* CVPR 2017

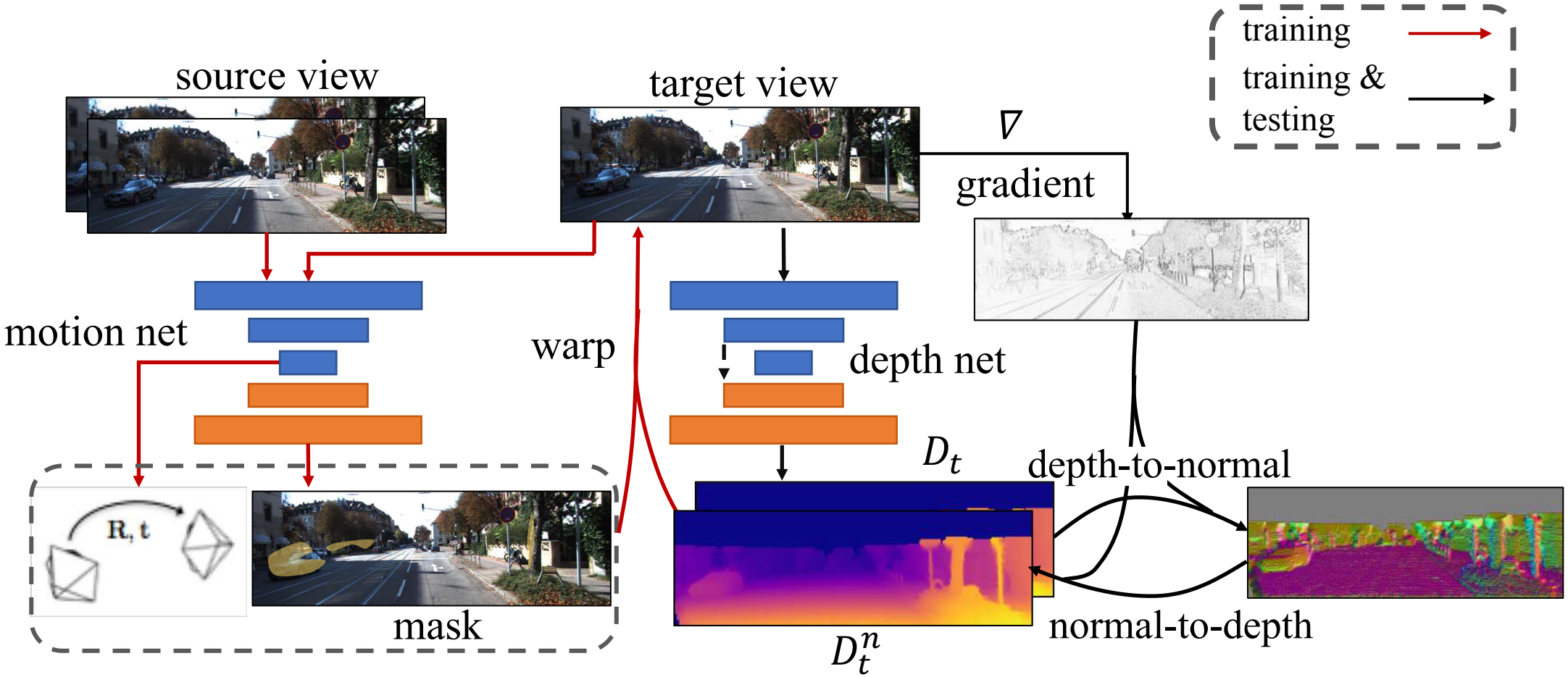
Limitations



(a) Input frame (b) Ground truth (c) Estimation results by Zhou et al. (d) Our results

- Need normal information
- Edge should be incorporated

Approach



Novelties:

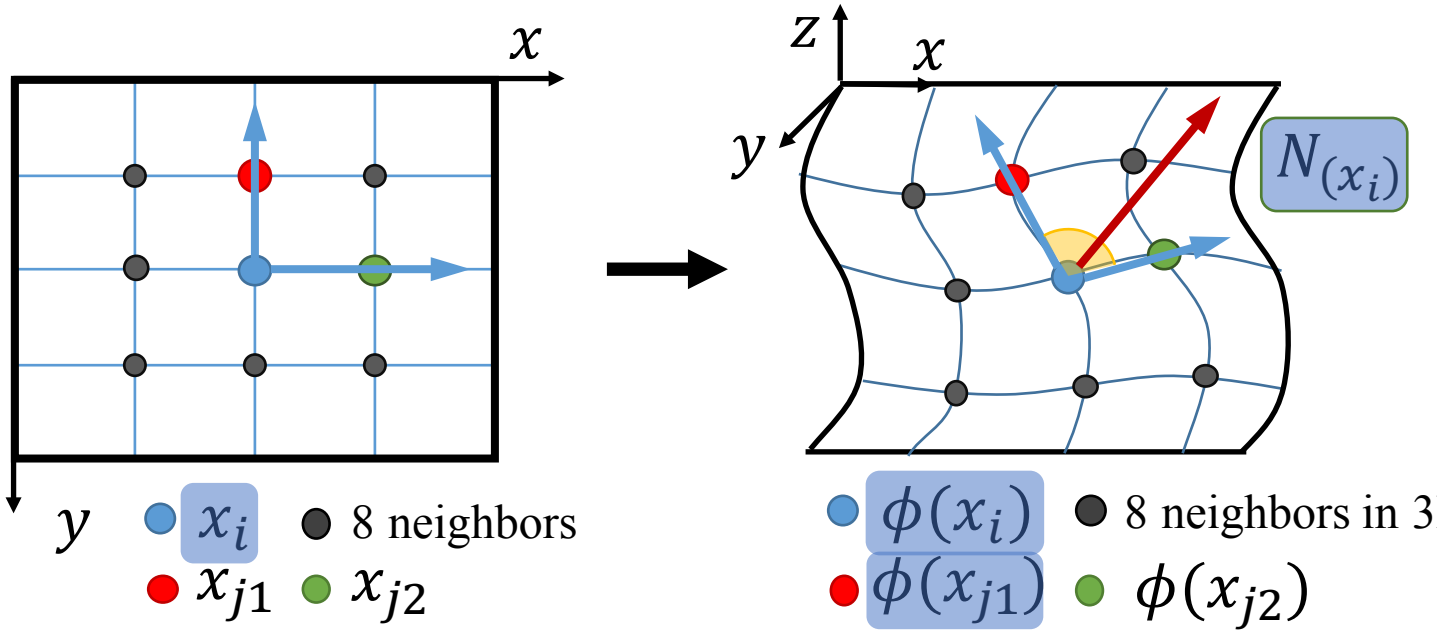
- Depth-normal consistency
- Edge awareness

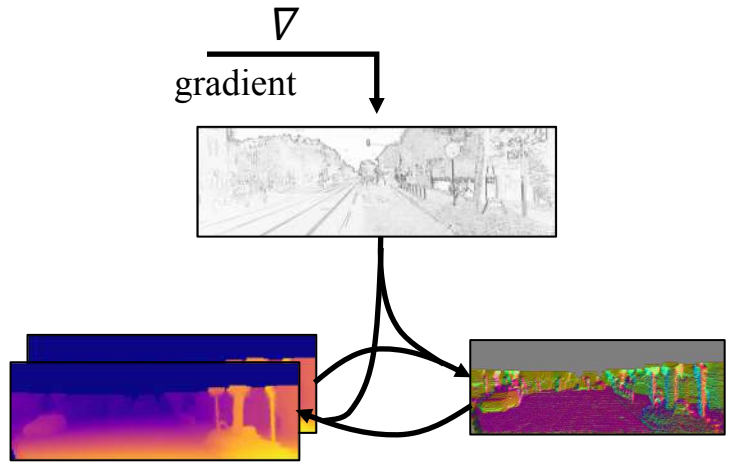
Consistency between depth and surface normal

Edge-aware depth-normal consistency

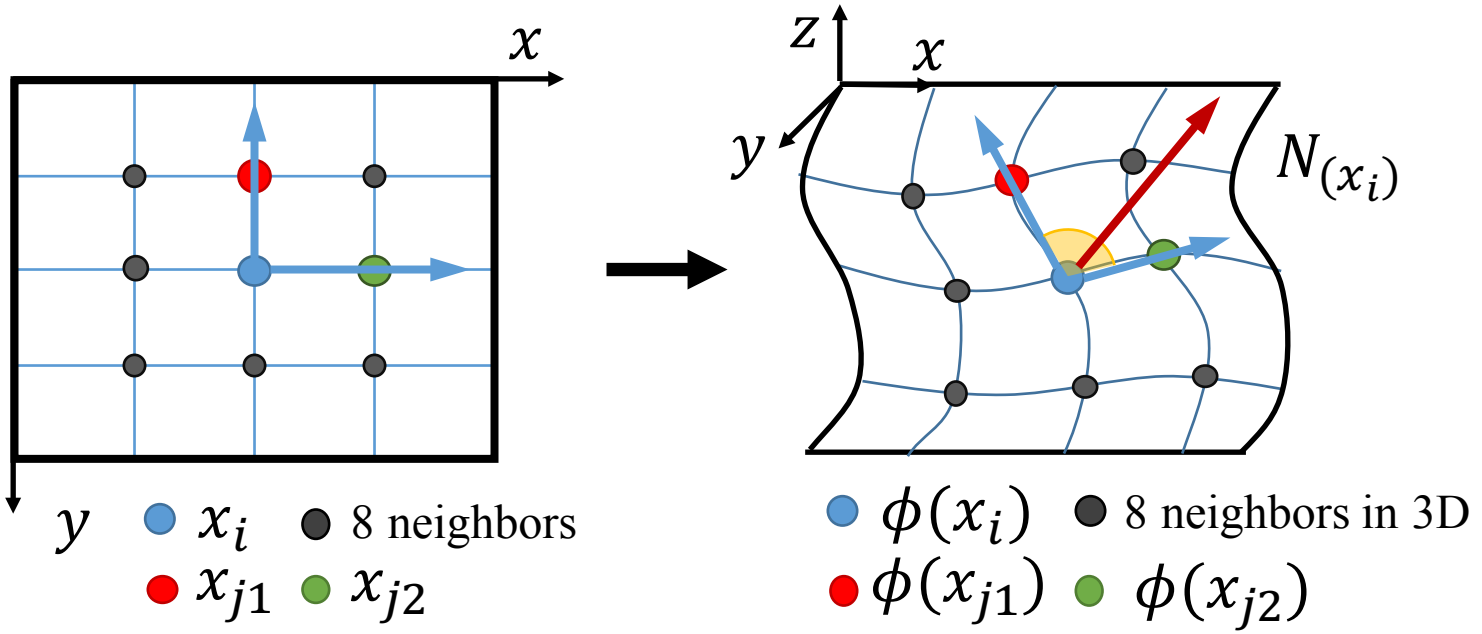
$$\mathcal{E}_{x_i}(D,N) = ||[\cdots, \omega_{ji}(\phi(x_j) - \phi(x_i)), \cdots]^T N(x_i)||^2$$

$$\omega_{ji} = \exp\{-\alpha|I(x_j) - I(x_i)|\}$$





Consistency between depth and surface normal



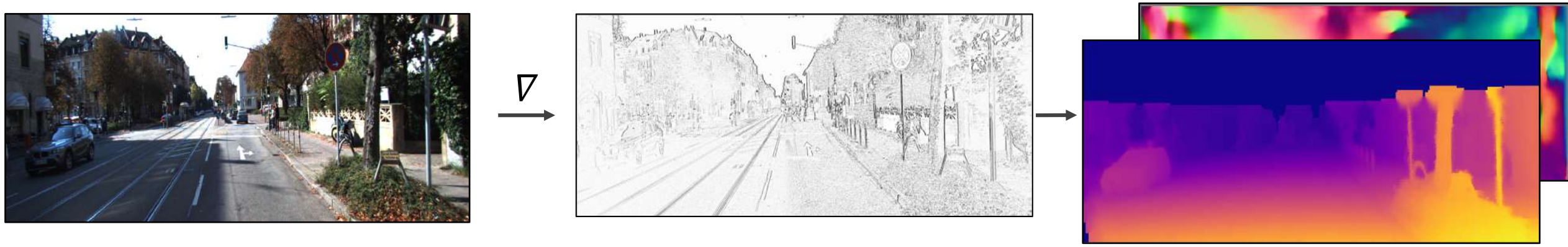
- depth-to-normal: cross-product
- normal-to-depth: dot-product, original depth as reference

Smoothness

Edge-aware smoothness

$$\mathcal{L}_s(D,2) = \sum_{x_t} \sum_{d \in x,y} |\nabla_d^2 D(x_t)| e^{-\alpha |\nabla_d I(x_t)|}$$

$$\mathcal{L}_s(N,1) = \sum_{x_t} \sum_{d \in x,y} |\nabla_d N(x_t)| e^{-\alpha |\nabla_d I(x_t)|}$$



I $\quad e^{\alpha|\nabla I(x_t)|}$ $\quad D/N$

Evaluation (depth)

Method	Test data	Supervision		Lower the better				Higher the better		
		Depth	Pose	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Train set mean	Eigen split	✓		0.403	5.530	8.709	0.403	0.593	0.776	0.878
Eigen *et al.*2016 Coarse		✓		0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen *et al.*2016 Fine		✓		0.203	1.548	6.307	0.282	0.702	0.890	0.958
Kuznietsov *et al.*2017 supervised		✓		0.122	0.763	4.815	0.194	0.845	0.957	0.987
Kuznietsov *et al.*2017 unsupervised			✓	0.308	9.367	8.700	0.367	0.752	0.904	0.952
Godard *et al.*2017			✓	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Zhou *et al.*2017				0.208	1.768	6.856	0.283	0.678	0.885	0.957
Ours				0.182	1.481	6.501	0.267	0.725	0.906	0.963
Train set mean	KITTI split	✓		0.398	5.519	8.632	0.405	0.587	0.764	0.880
Godard *et al.*2017			✓	0.124	1.388	6.125	0.217	0.841	0.936	0.975
Vij.. *et al.*2017				-	-	-	0.340	-	-	-
Zhou *et al.*2017				0.216	2.255	7.422	0.299	0.686	0.873	0.951
Ours				0.1648	1.360	6.641	0.248	0.750	0.914	0.969

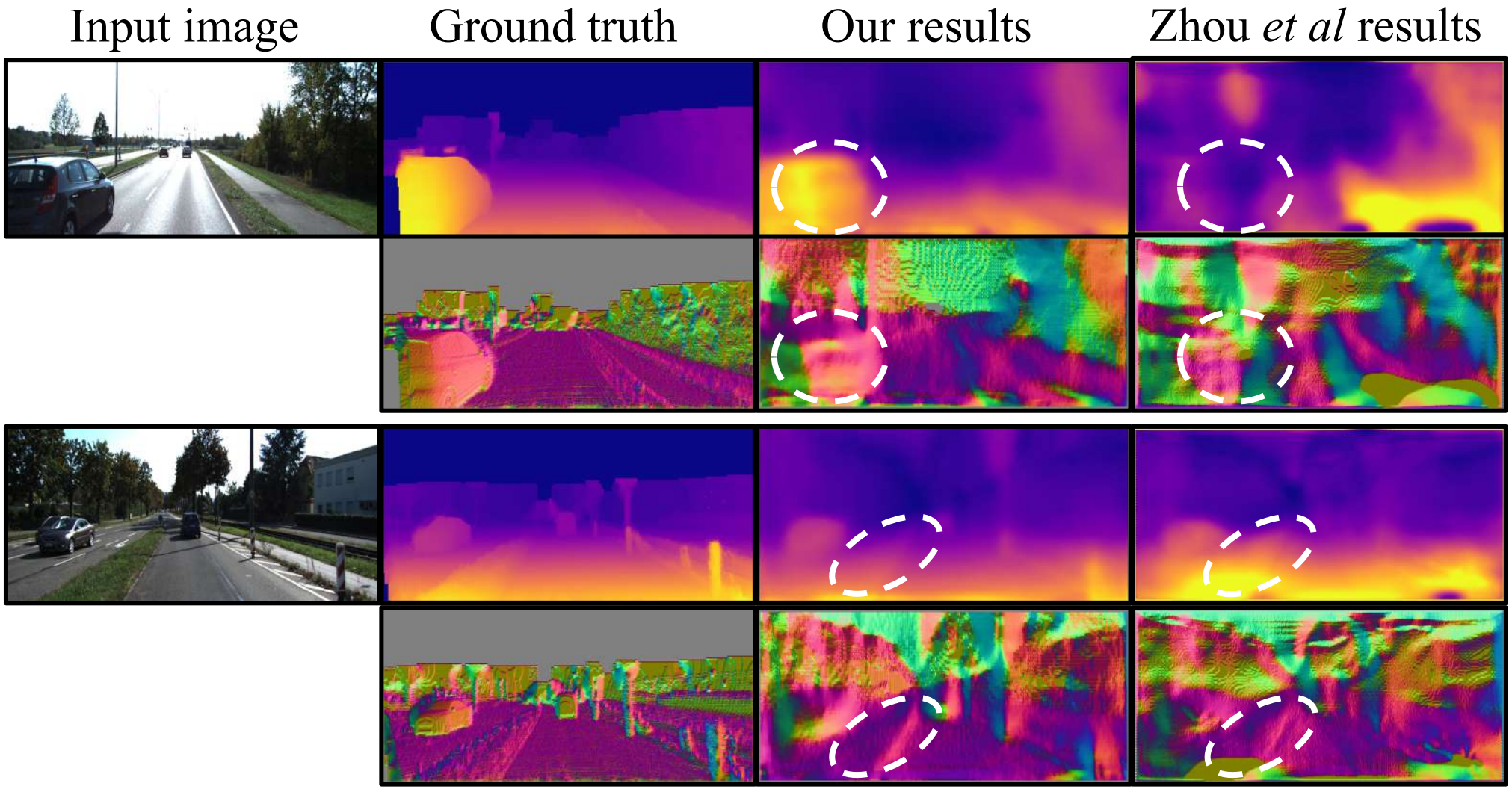
Depth performance comparison with state-of-the-art methods on KITTI dataset

Evaluation (surface normal)

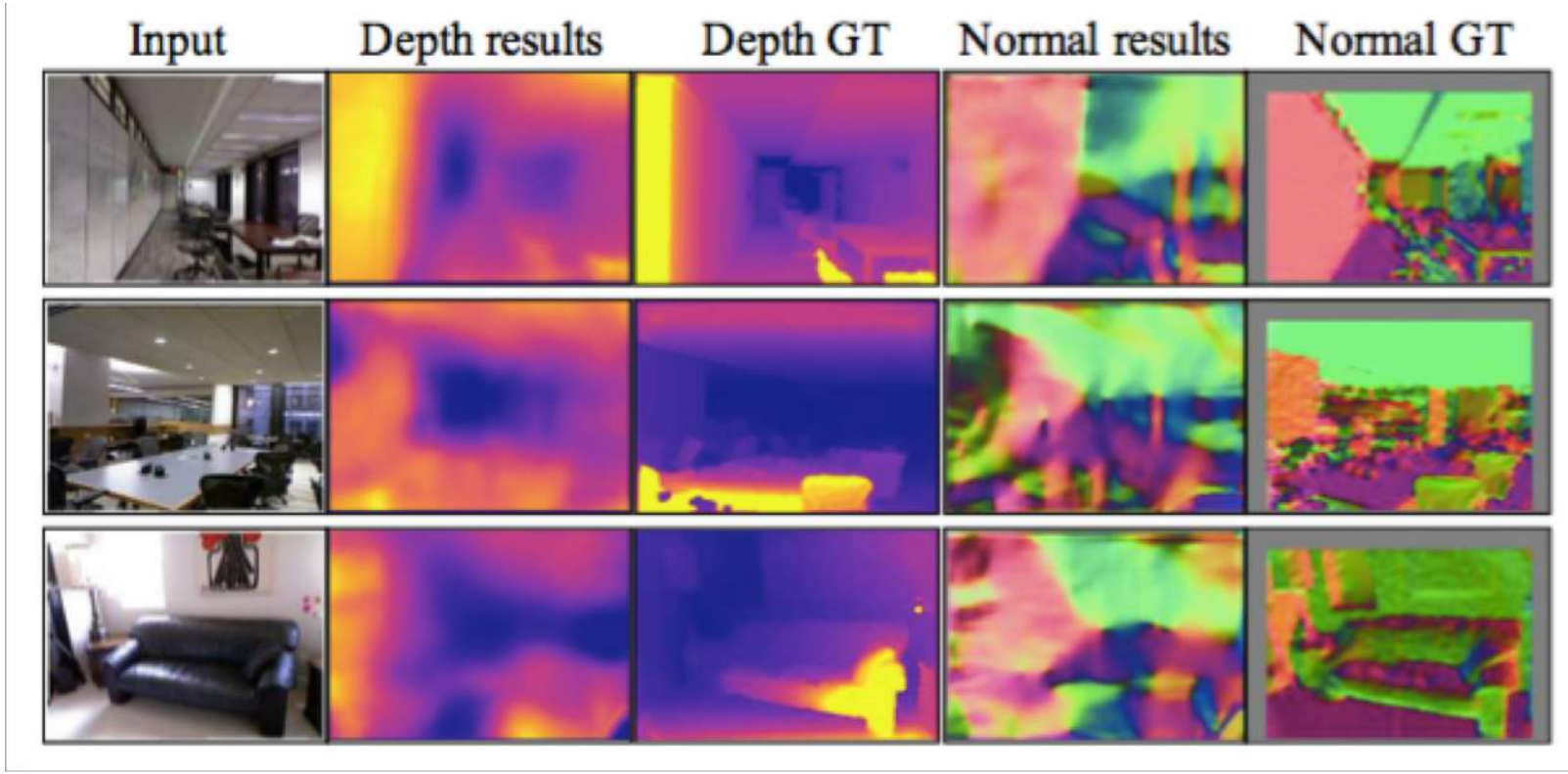
Method	Mean	Median	11.25°	22.5°	30°
Ground truth normal mean	72.39	64.72	0.031	0.134	0.243
Pre-defined scene	63.52	58.93	0.067	0.196	0.302
(Zhou et al. 2017)	50.47	39.16	0.125	0.303	0.425
Ours	47.52	33.98	0.149	0.369	0.473

Normal performance comparison with other methods on KITTI test split

Visual comparison (outdoor)



Visual comparison (indoor)



- Reasonable performance for scene with intersecting planes (first, second row)
- Relatively messy results for scenes with only objects (third row)

Summary

- We incorporate depth-normal consistency and achieved better estimation results

- Depth and normal results are both improved

Guidelines

- **Unsupervised depth estimation**
- Unsupervised optical flow estimation
- Joint unsupervised learning of geometry and motion

Unsupervised joint learning of edge and geometry

Training

Test

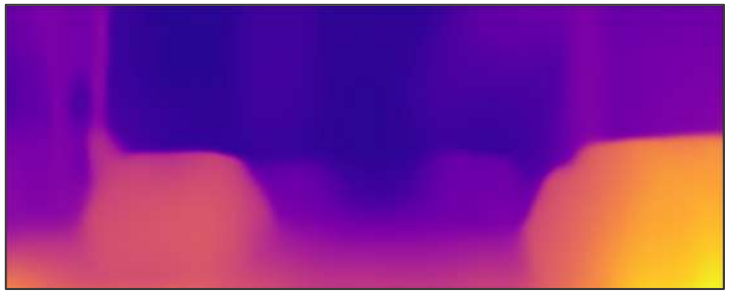
close far

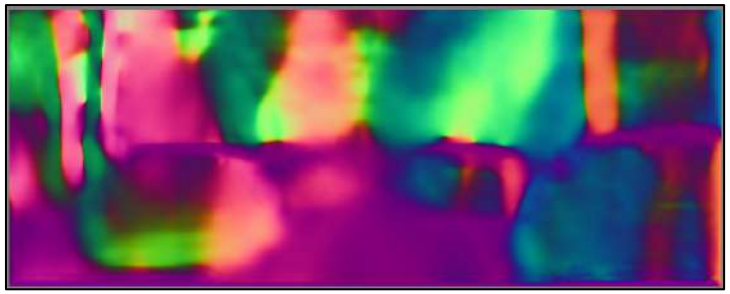
Depth

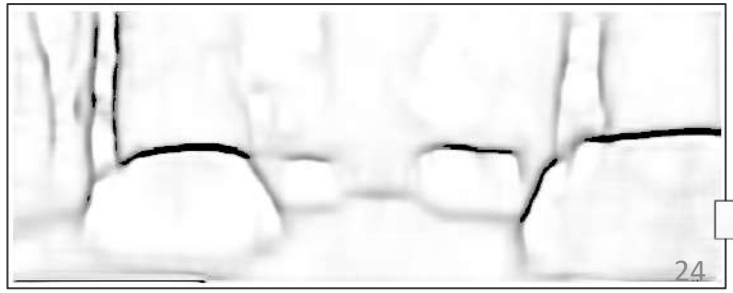
Normal

x y z

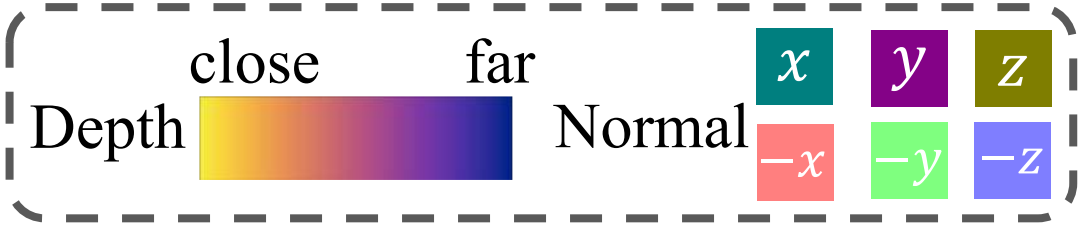
$-x$ $-y$ $-z$

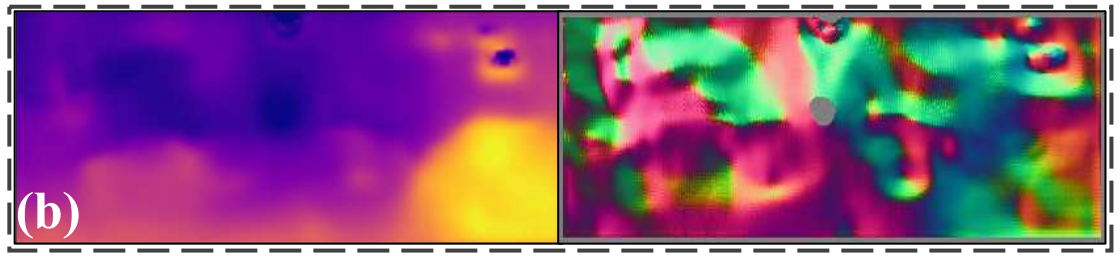




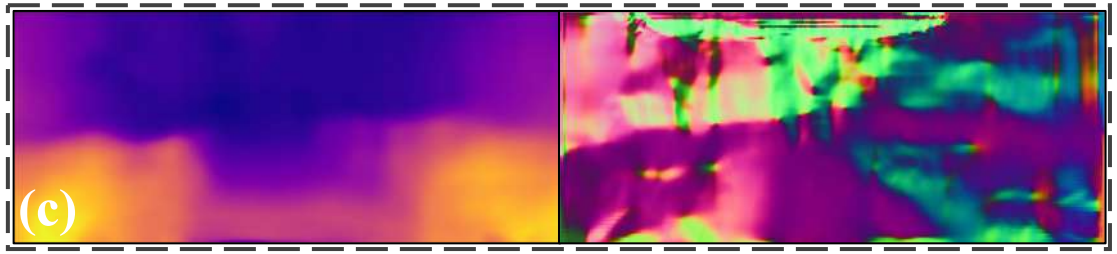


Limitation

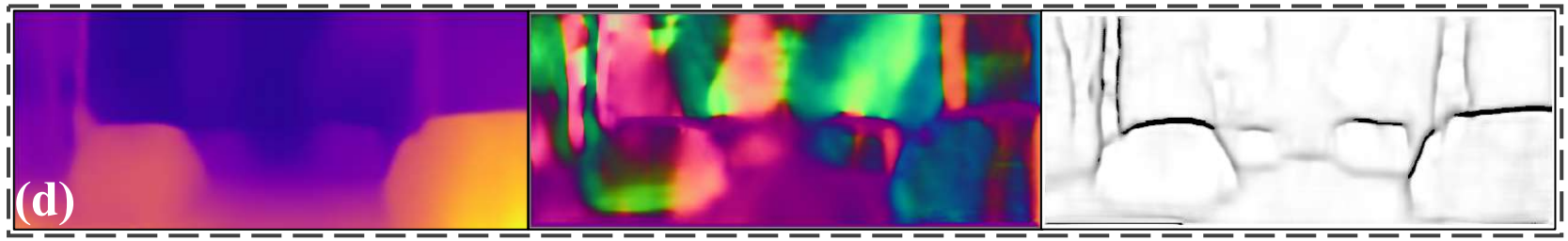




Zhou et al.
CVPR 2017

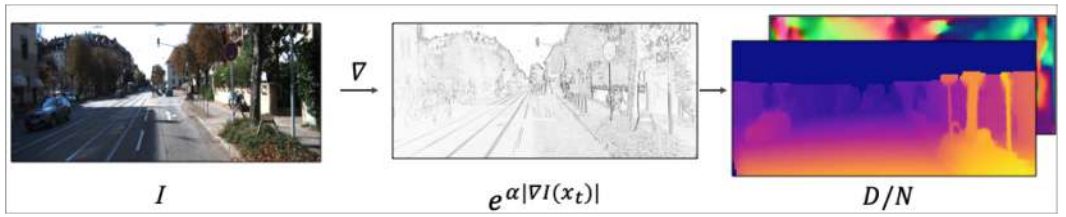


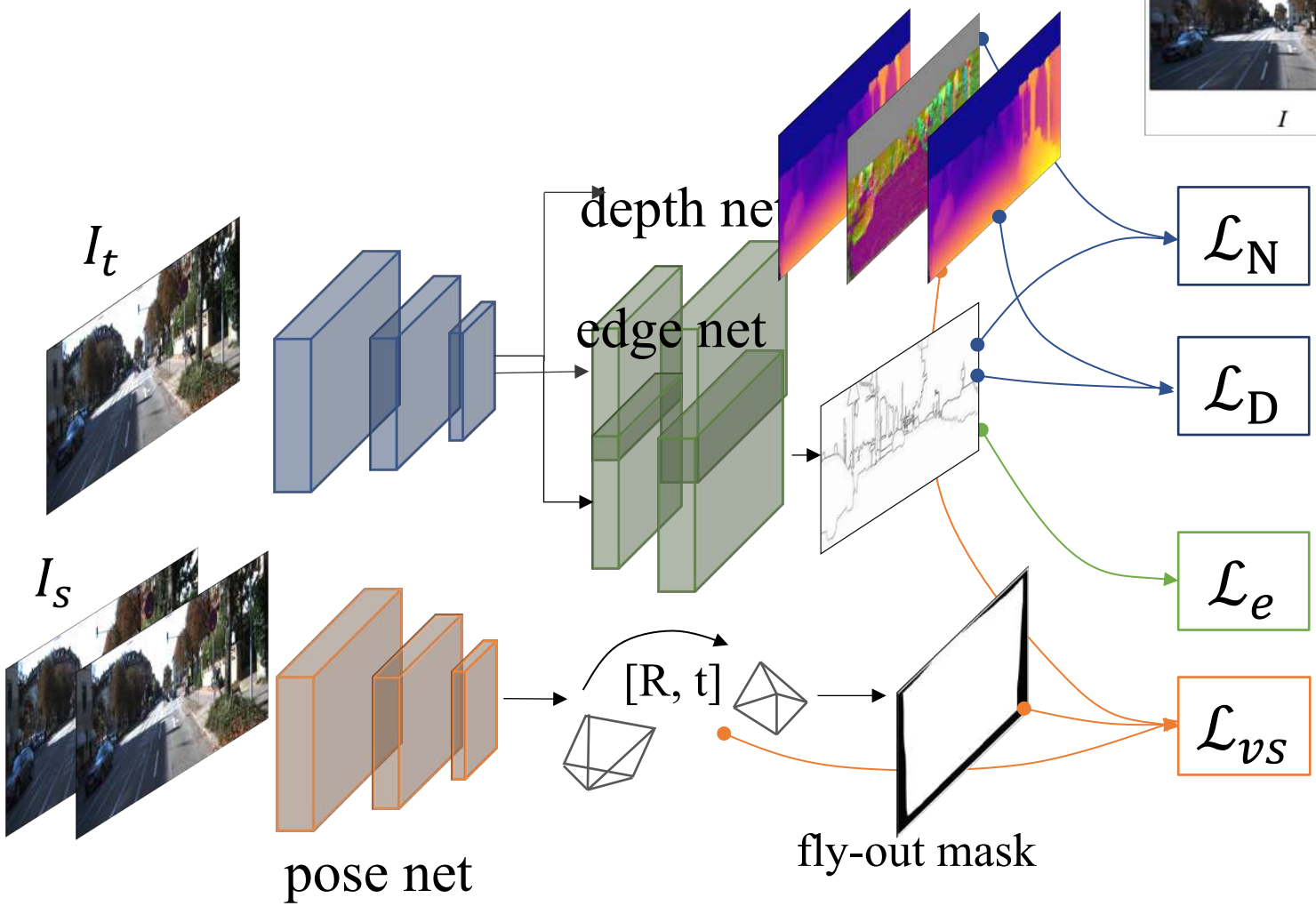
Yang et al.
AAAI 2018



Joint learning of edge and geometry

Approach





$\mathcal{L}_N$

$$\mathcal{L}_N = \|N_t(p_i) - N_t(p_j)\|_1 \kappa(p_i, p_j)$$

$\mathcal{L}_D$

$$\mathcal{L}_D = \frac{D_t(p_i) - D_t(p_k)}{\phi(p_i) - \phi(p_k)} - \frac{D_t(p_j) - D_t(p_k)}{\phi(p_j) - \phi(p_k)}$$

$\mathcal{L}_e$

$$\mathcal{L}_e(E) = ||E_t(p_i)||^2$$

$\mathcal{L}_{vs}$

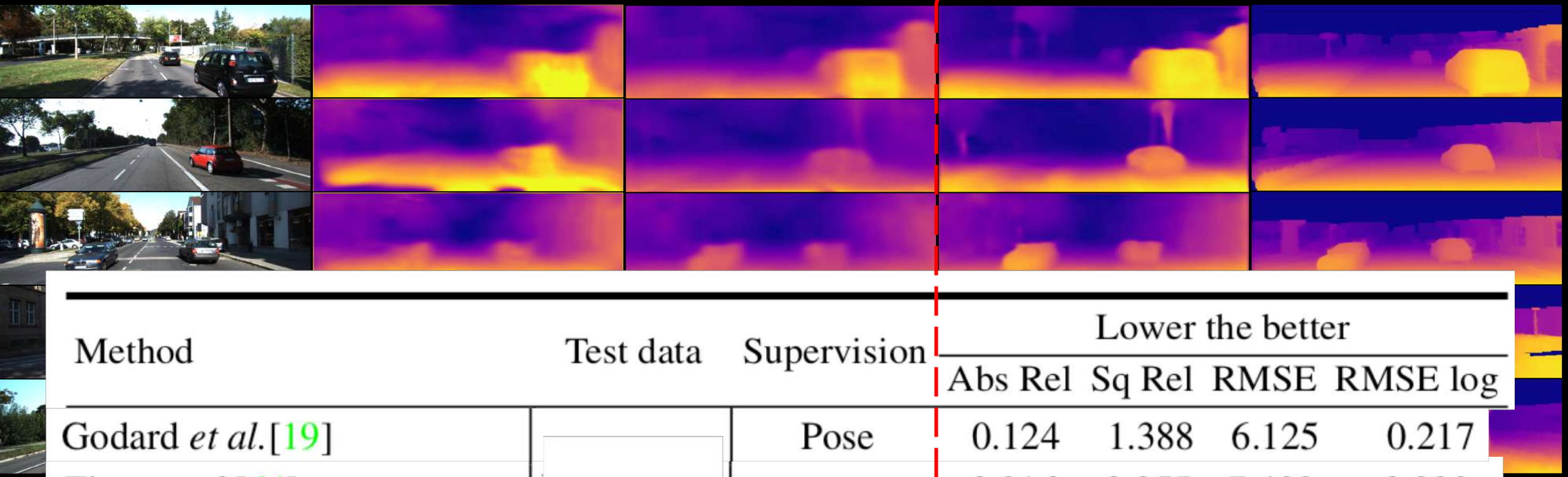
$$\mathcal{L}_{vs}(D, \mathcal{T}) = |I_t(p_t) - \hat{I}_s(p_t)|$$

Training

- Three subnetworks trained from scratch jointly
  - depth net, pose net and edge net

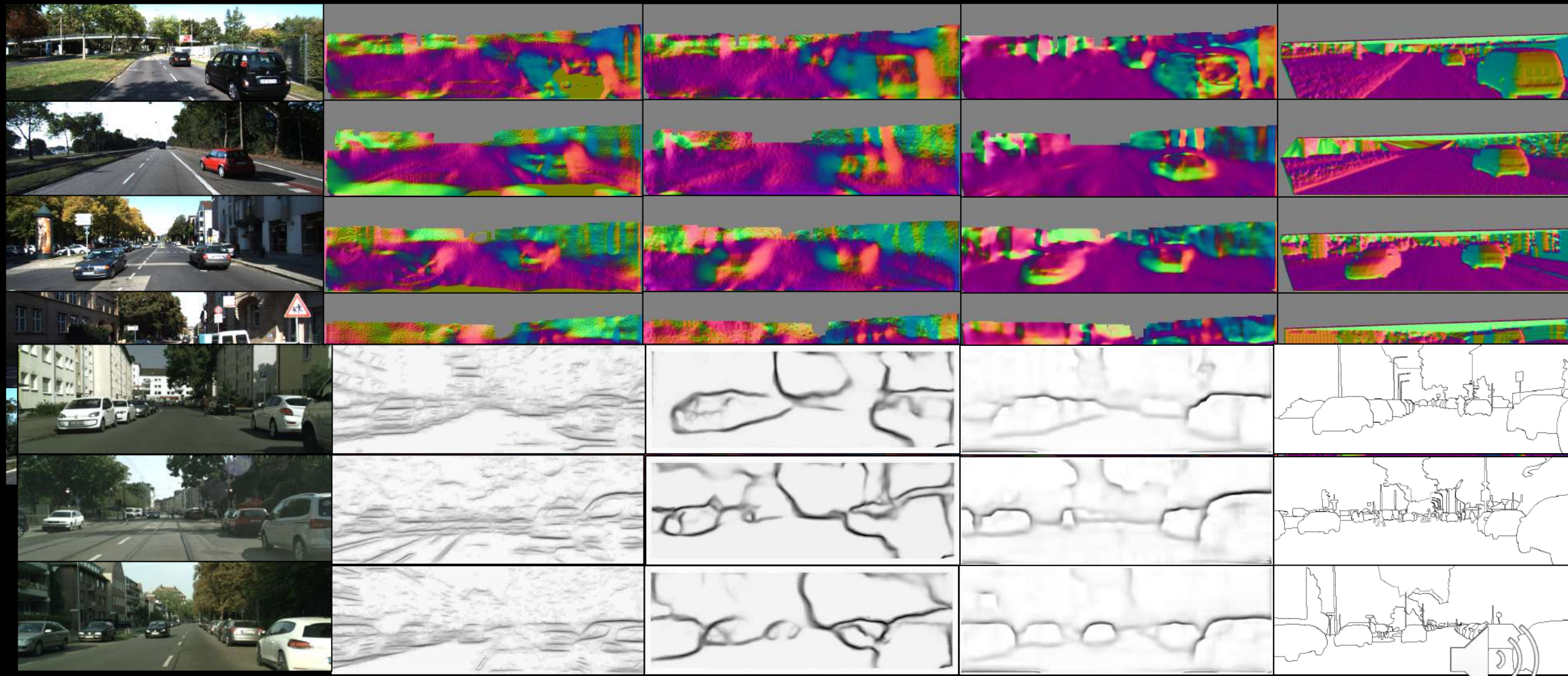
- Trained on KITTI or Cityscapes videos
- Optimizer: Adam, $\beta_1 = 0.9, \beta_2 = 0.009, \epsilon = 10^{-8}$, learning rate 0.002
- Training time: 6 hours (5 epochs) on Titan X (Pascal)

Results (depth)



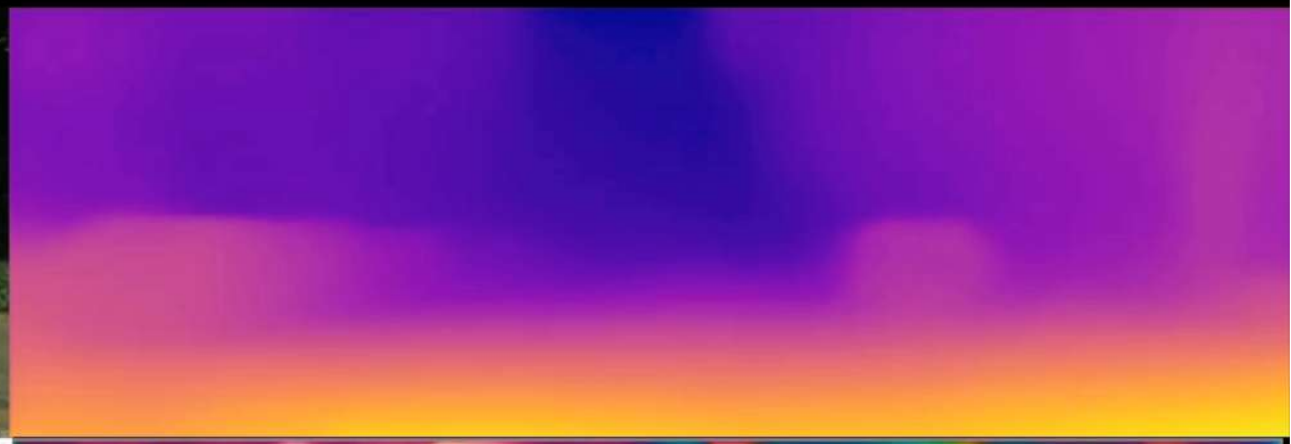
Method	Test data	Supervision	Lower the better			
			Abs Rel	Sq Rel	RMSE	RMSE log
Godard *et al.*[19]	KITTI split	Pose	0.124	1.388	6.125	0.217
Zhou *et al.*[64]			0.216	2.255	7.422	0.299
Yang *et al.*[60]			~~0.165~~	~~1.360~~	~~6.641~~	~~0.267~~
LEGO			0.154	1.272	6.012	0.230
LEGO+CS			0.142	1.237	5.846	0.22[illegible]

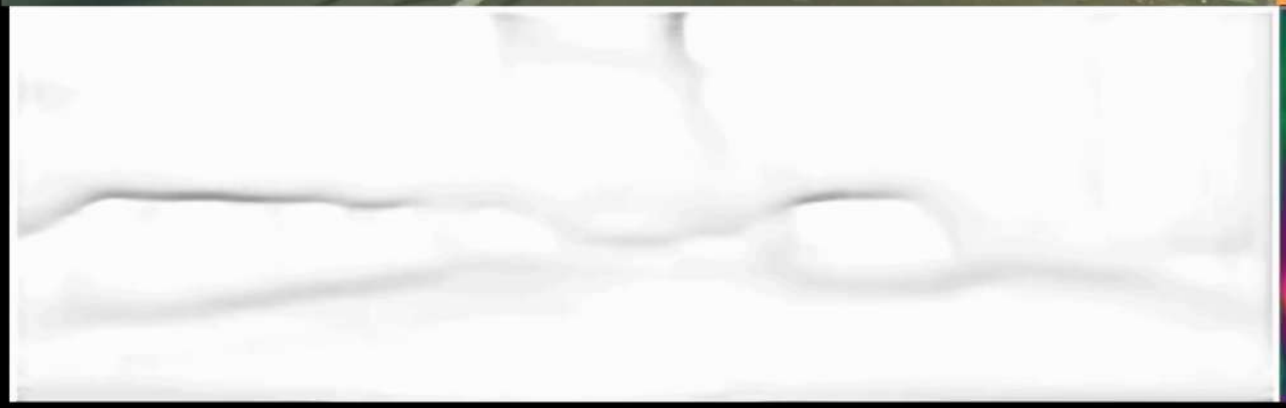
Results (normal & edge)

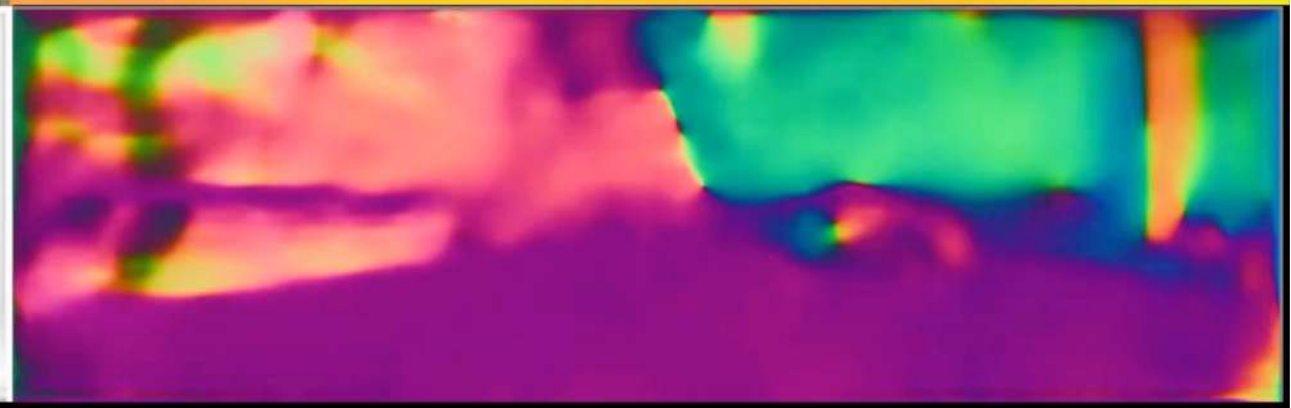


Input image | Li et al. | HED-N | Our edges | Edge GT

Sequence 1







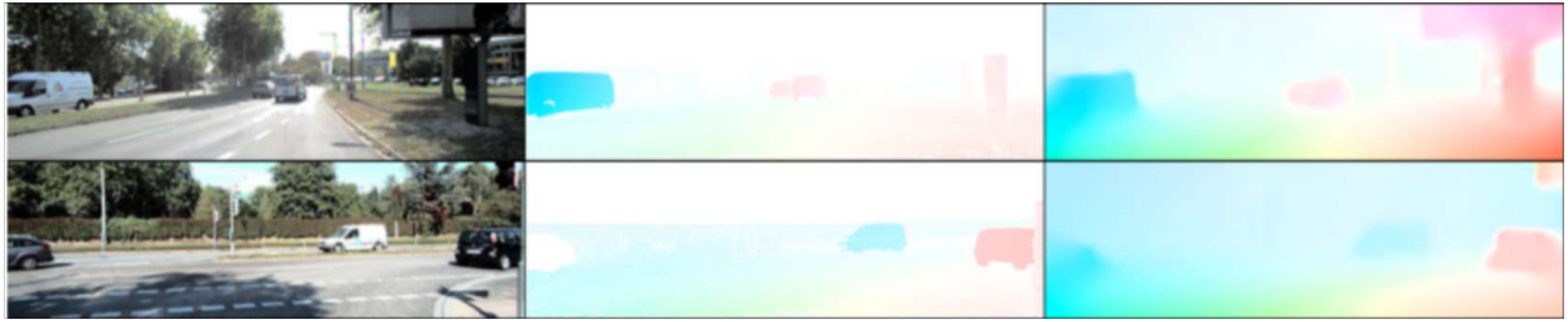
Input frame	Depth
Edge	Normal

Guidelines

- Unsupervised depth estimation
- **Unsupervised optical flow estimation**
- Joint unsupervised learning of geometry and motion

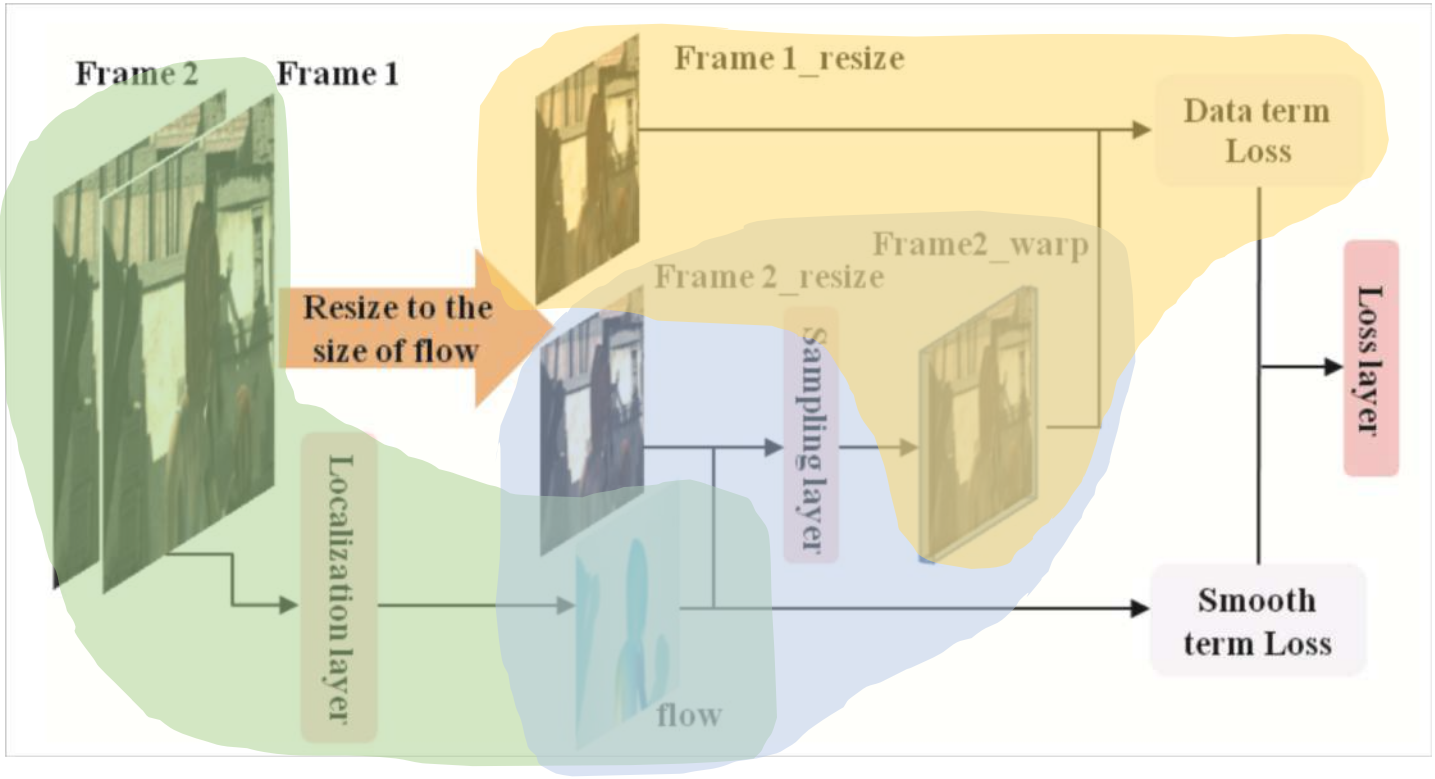
Unsupervised learning of motion

Input frame 1 | Ground truth optical flow | Our estimation results



Occlusion Aware Unsupervised Learning of Optical Flow, CVPR18'

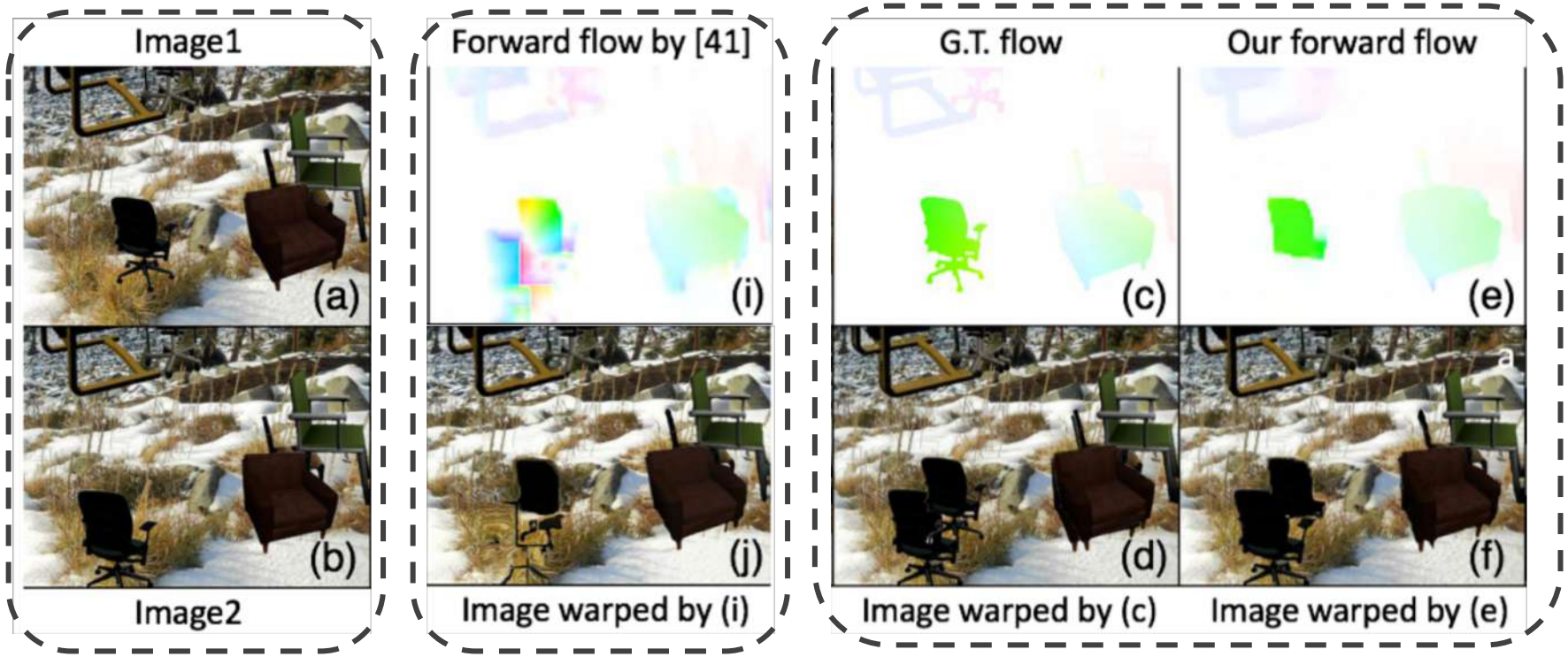
Previous work



Unsupervised optical flow estimation pipeline

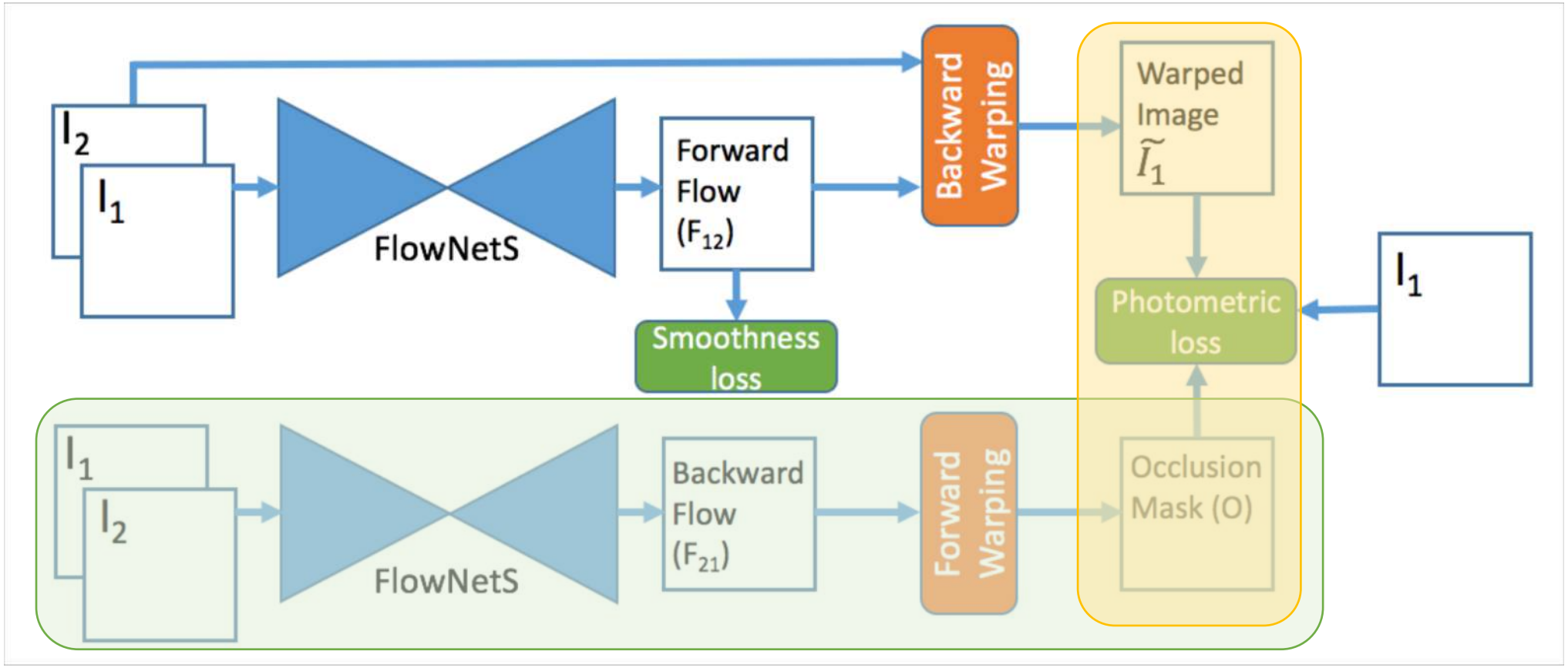
Ren, Zhe, et al. "Unsupervised Deep Learning for Optical Flow Estimation." AAAI. 2017.

Limitation



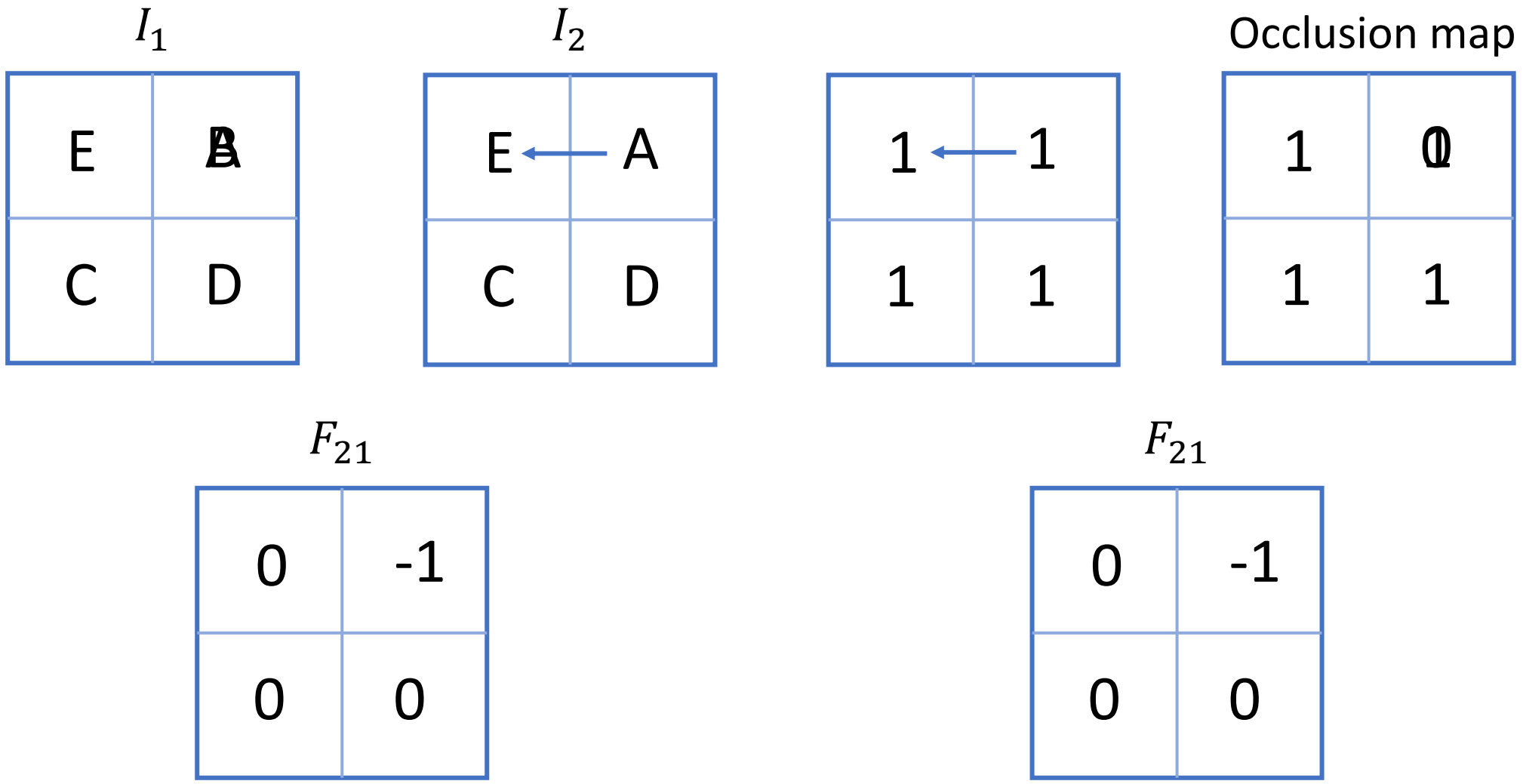
Optical flow confusion at occlusion/de-occlusion regions

Approach



- Explicitly model the occlusion mask to filter out occlusion regions in loss calculation

Occlusion mask

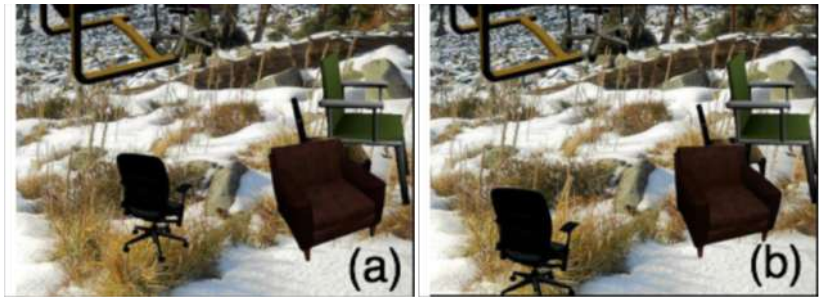


- Occlusion happens in regions in image 1 that are covered in image 2

Evaluation

- Performances on different datasets

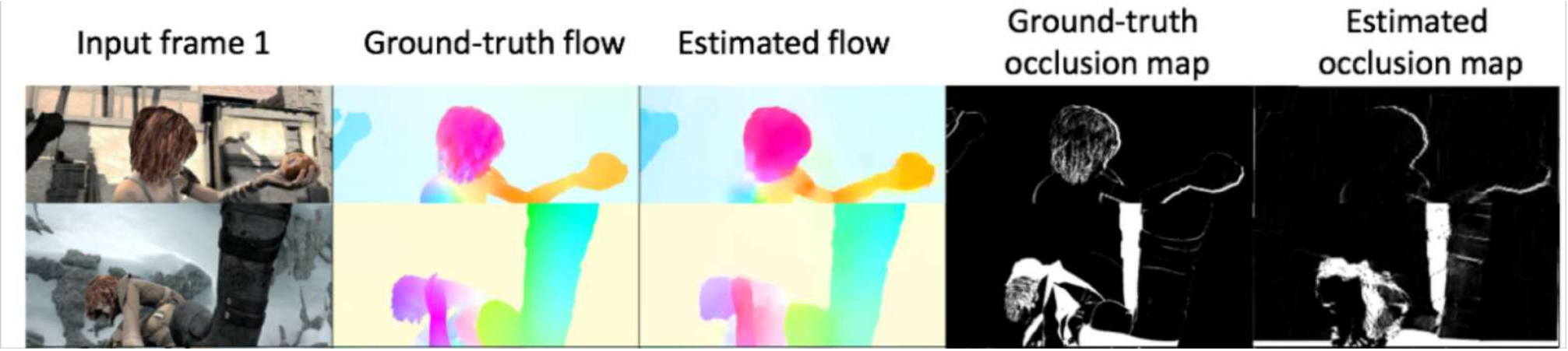
	Methods	Chairs	KITTI 2012		KITTI 2015	
		test	train	test	train	test
Unsupervise	DSTFlow [41]	5.11	16.98	–	24.30	–
	DSTFlow-best [41]	5.11	10.43	12.4	16.79	39%
	BackToBasic [31]	5.3	11.3	9.9	–	–
	Ours	**3.30**	12.95	–	21.30	–
	Ours+ft-Sintel	3.76	12.9	–	22.6	–
	Ours-KITTI	–	**3.55**	**4.2**	**8.88**	**31.2%**



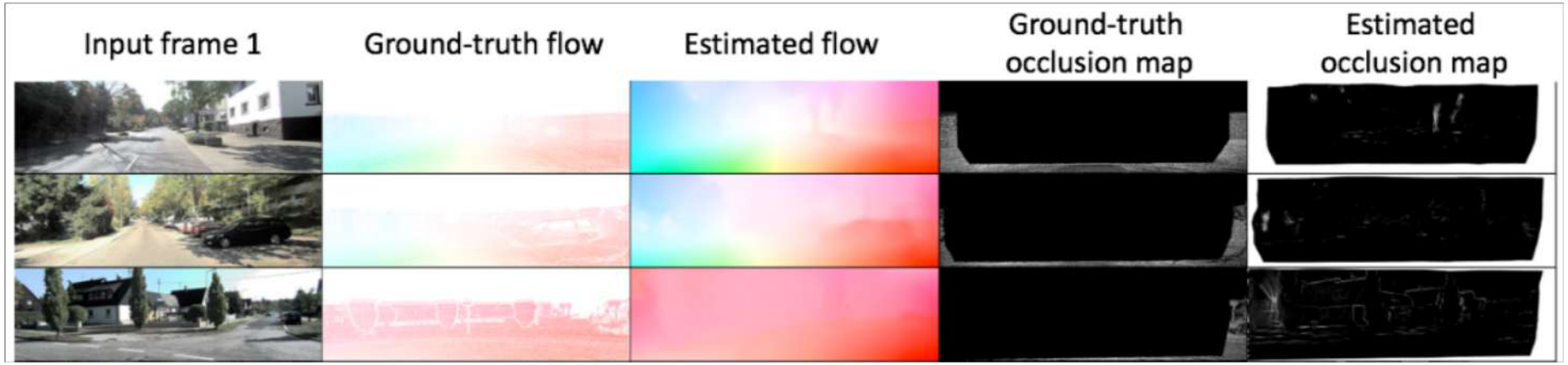
Flying chairs data sample

KITTI data sample

Evaluation



Qualitative results on Sintel dataset



Qualitative results on KITTI2012 dataset

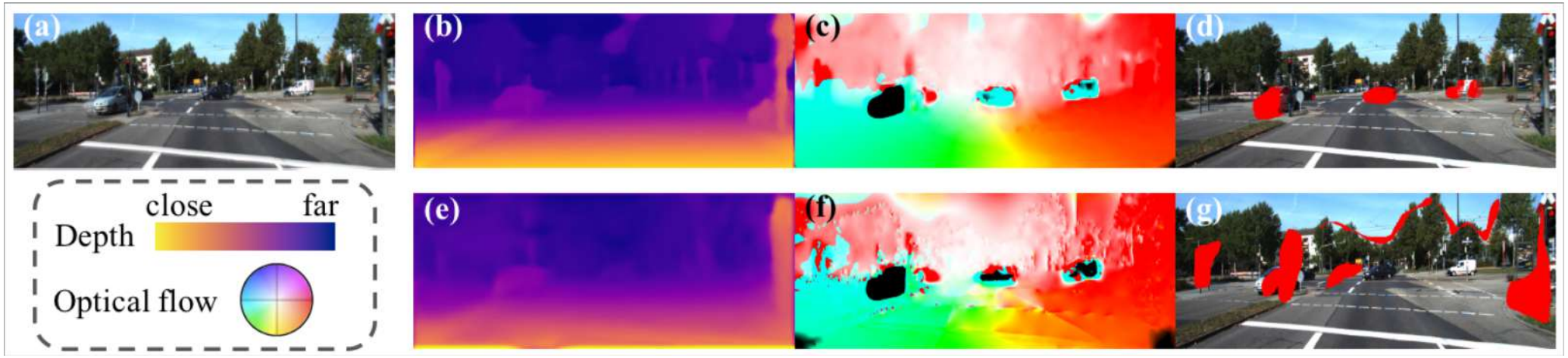
Summary

- The occlusion issue is explicitly modeled in this work
- We evaluated on various benchmarks outperformed previous SOTA methods
- A step-stone for our joint understanding of static and dynamic scenes

Guidelines

- Unsupervised depth estimation
- Unsupervised optical flow estimation
- **Joint unsupervised learning of geometry and motion**

Unsupervised joint learning of geometry and motion



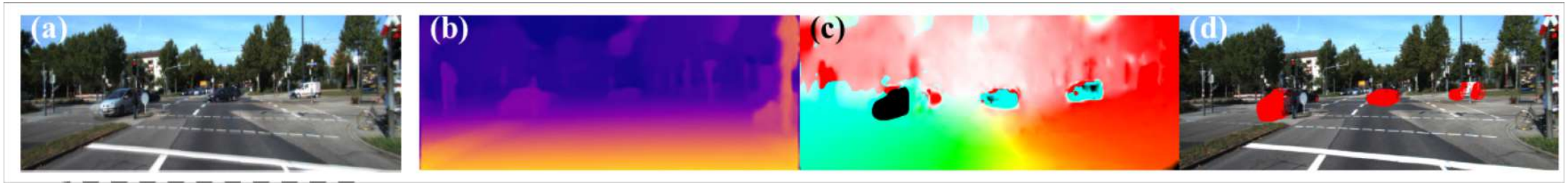
Every Pixel Counts: Unsupervised Geometry Learning with Holistic 3D Motion Understanding*, ECCV workshop 18'*

Every Pixel Counts ++: Joint Geometry and Motion Learning with 3D Holistic Understanding, TPAMI submission

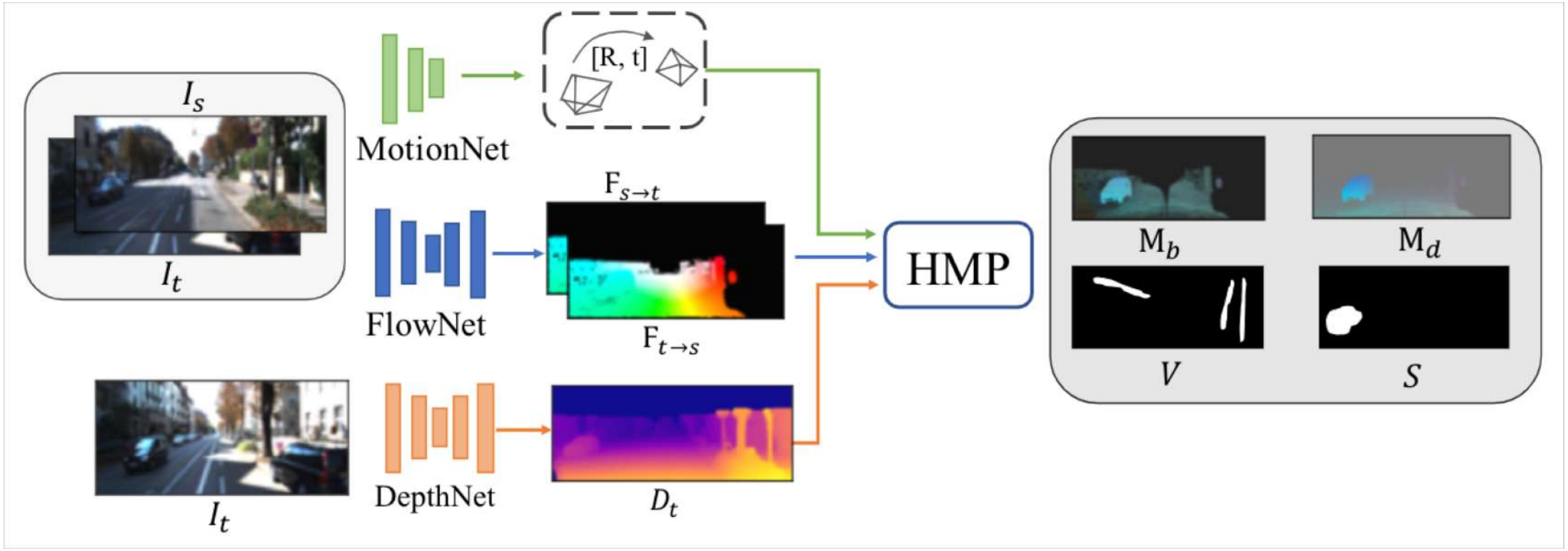
Limitation

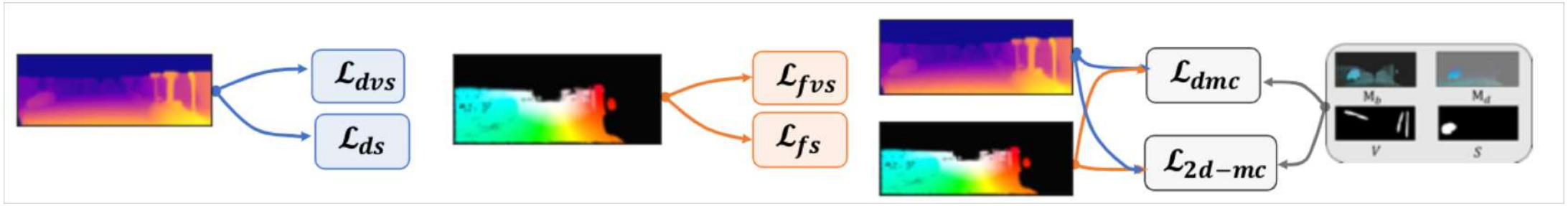
- An important assumption of the unsupervised depth learning is the scene being static.
  - All pixel movement is caused by camera motion

- Optical flow represents both camera motion and object motion



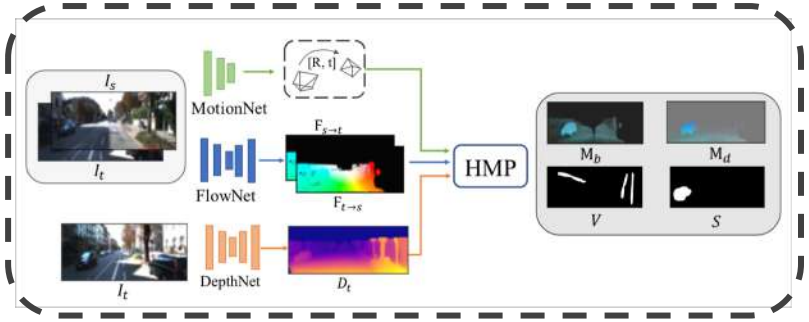
Approach

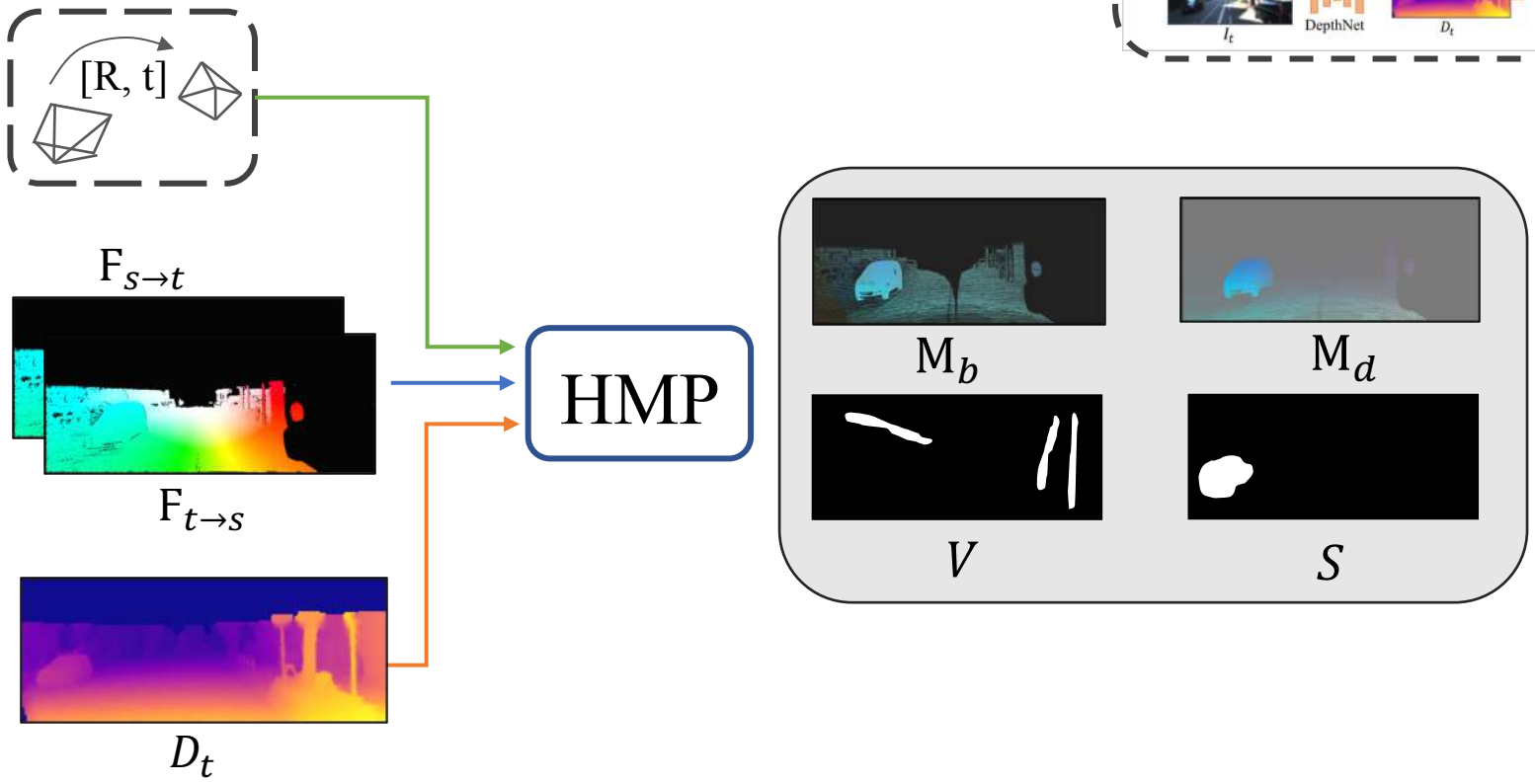




Every Pixel Counts ++: Joint Geometry and Motion Learning with 3D Holistic Understanding, TPAMI submission

Approach (HMP)





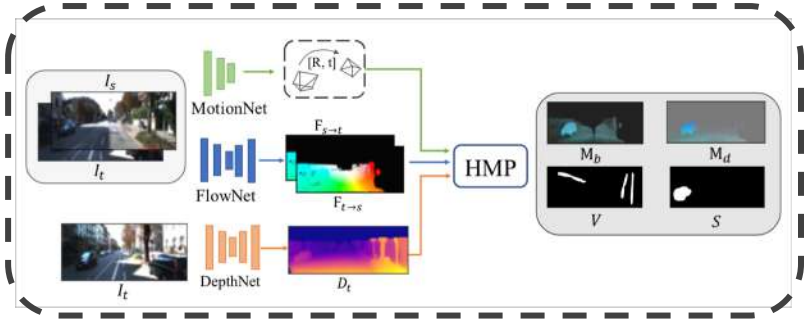
$$\mathbf{M}_b(p_t) = \mathbf{V}(p_t)[\mathbf{T}_{t\to s}\phi(p_t|\mathbf{D}_t) - \phi(p_t|\mathbf{D}_t)],$$

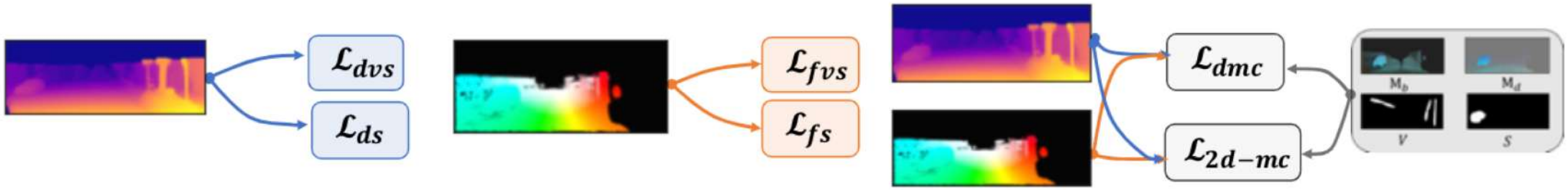
$$\mathbf{M}_d(p_t) = \mathbf{V}(p_t)[\phi(p_t + \mathbf{F}_{t\to s}(p_t)|\mathbf{D}_s) - \phi(p_t|\mathbf{D}_t)] - \mathbf{M}_b(p_t)$$

$$\mathbf{V}(p_t) = \mathbb{1}(\sum\nolimits_p (1 - |p_t - (p + \mathbf{F}_{s\to t})|) > 0),$$

$$\mathbf{S}_t(p_t) = 1 - \exp\{-\alpha(\mathbf{M}_d(p_t)/\mathbf{D}_t)\}$$

Approach (loss terms)





View synthesis:

$$\mathcal{L}_{vs}(\mathbf{O}) = \sum_{p_t} \mathbf{V}(p_t) * s(I_t(p_t), \hat{I}_t(p_t))$$

$$s(I(p), \hat{I}(p)) = (1-\beta) * |I(p) - \hat{I}(p)| + \beta * (1 - \frac{1}{2}\text{SSIM}(I(p), \hat{I}(p)))$$

Smoothness:

$$\mathcal{L}_s(\mathbf{O}, \mathbf{W}, o) = \sum_{p_t} \sum_{d \in x,y} \mathbf{W}(p_t) |\nabla_d^o \mathbf{O}(p_t)| e^{-\alpha |\nabla_d^2 I(p_t)|}$$

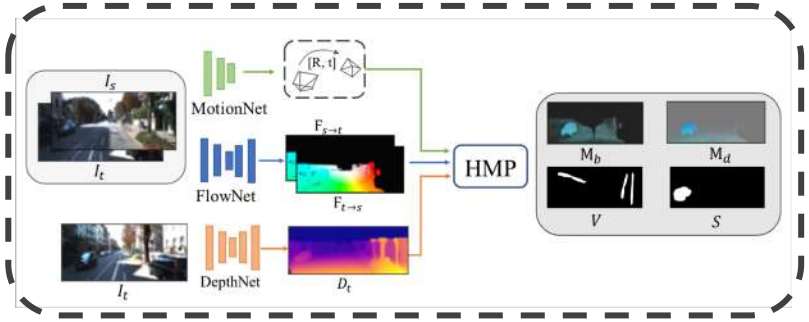
HMP consistency:

$$\mathcal{L}_{mc} = \sum_{p_t} (1 - \mathbf{S}(p_t)) |\mathbf{M}_d(p_t)|_1,$$

$$\mathcal{L}_{ms} = \mathcal{L}_s(\mathbf{M}_d, \mathbf{S}, 1),$$

$$\mathcal{L}_{dmc} = \sum_{p_t} \mathbf{V}(p_t)(1 - \mathbf{S}(p_t))(|\mathbf{D}_s(p_{sf}) - \hat{\mathbf{D}}_s(p_{st})|$$

Training



- Iterative training:
  1. Train DepthNet and MotionNet in an unsupervised approach
  2. Train FlowNet in an unsupervised approach
  3. Iteratively do:

     Fix DepthNet and MotionNet, add HMP loss, train FlowNet

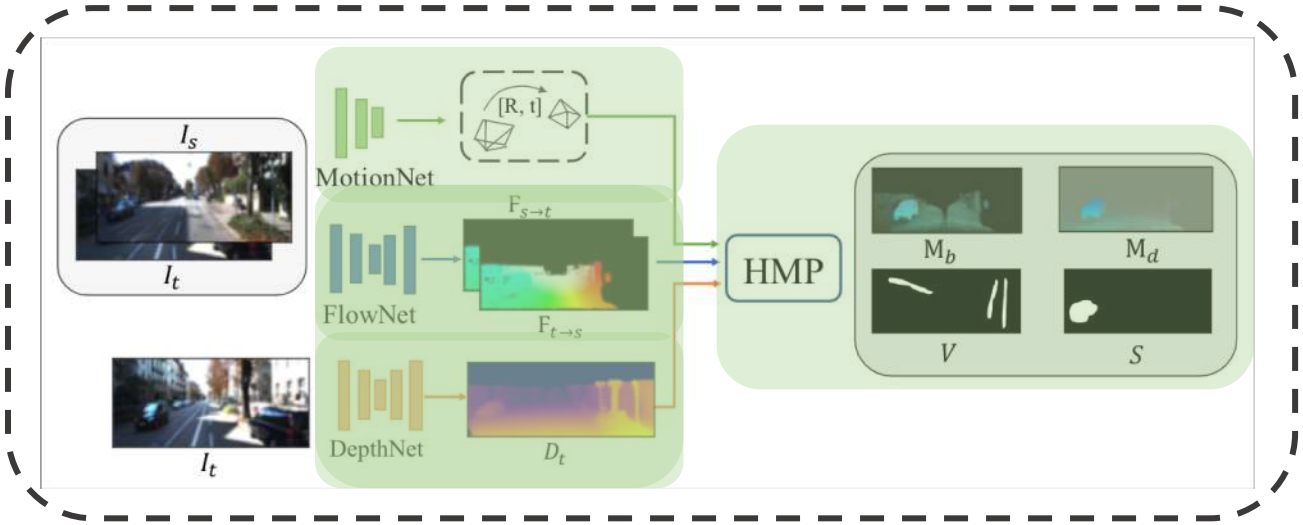
     Fix FlowNet, add HMP loss, train DepthNet and MotionNet

  4. Jointly train all three networks
- Both pre-training and finetuning on unlabeled KITTI videos

Evaluation

Five tasks to evaluate:

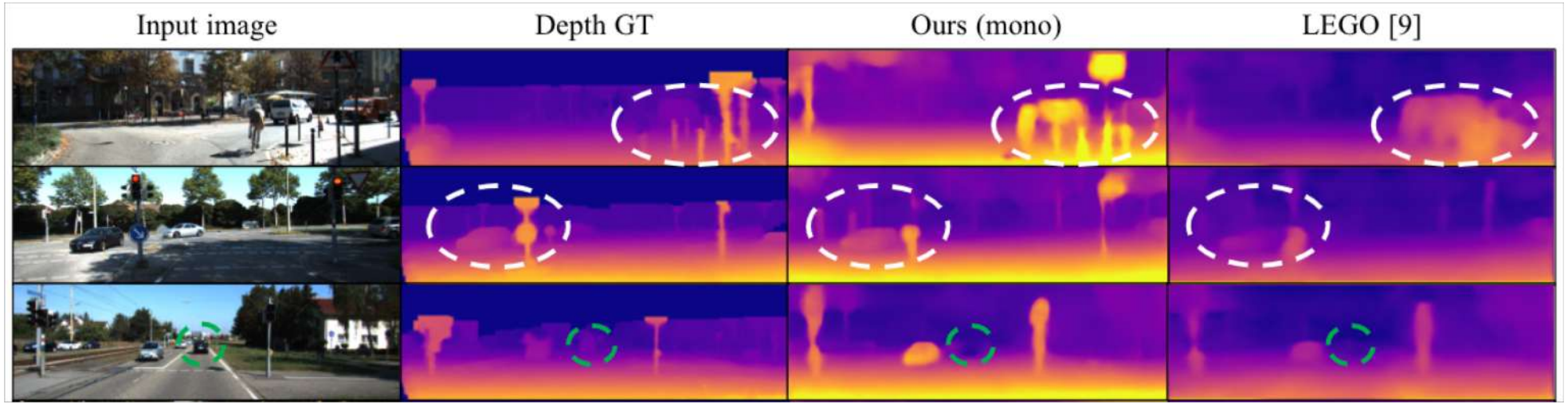
1. Depth evaluation (DepthNet)
2. Optical flow evaluation (FlowNet)
3. Odometry evaluation (MotionNet)
4. Motion segmentation (HMP)
5. Scene flow evaluation (Depth + Flow)

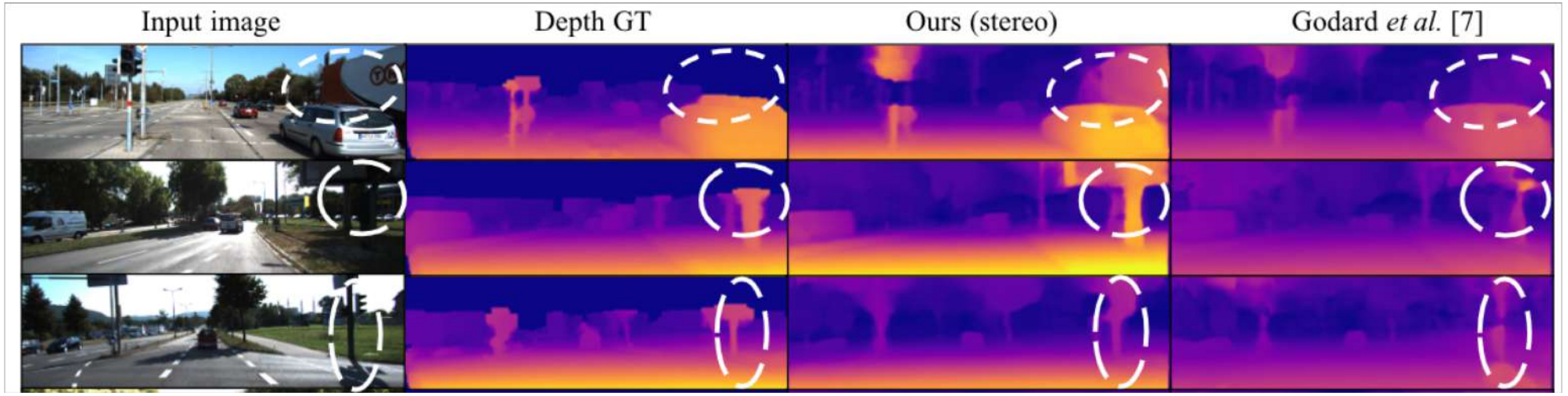


Evaluation (depth)

Method	Stereo	Lower the better				Higher the better		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Train mean		0.403	5.530	8.709	0.403	0.593	0.776	0.878
Zhou *et al.* [9]		0.208	1.768	6.856	0.283	0.678	0.885	0.957
LEGO [5]		0.162	1.352	6.276	0.252	0.783	0.921	0.969
Ours (mono depth only)		0.151	1.448	5.927	0.233	0.809	0.933	0.971
Ours (mono depth consist)		0.146	1.065	5.405	0.220	0.812	0.939	0.975
Ours (mono flow consist)		0.148	1.034	5.546	0.223	0.802	0.938	0.975
Ours (mono vis flow consist)		0.144	1.042	5.358	0.218	0.813	0.941	0.976
Ours (mono)		**0.141**	**1.029**	**5.350**	**0.216**	**0.816**	**0.941**	0.976
UnDeepVO [43]	✓	0.183	1.730	6.570	0.268	-	-	-
Godard *et al.* [8]	✓	0.148	1.344	**5.927**	0.247	0.803	0.922	0.964
Ours (stereo depth only)	✓	0.141	1.224	5.548	0.229	0.811	0.934	0.972
Ours (stereo depth consist)	✓	0.134	1.063	5.353	0.218	0.826	0.941	0.975
Ours (stereo)	✓	**0.128**	**0.935**	**5.011**	**0.209**	**0.831**	**0.945**	**0.979**

Evaluation (depth)

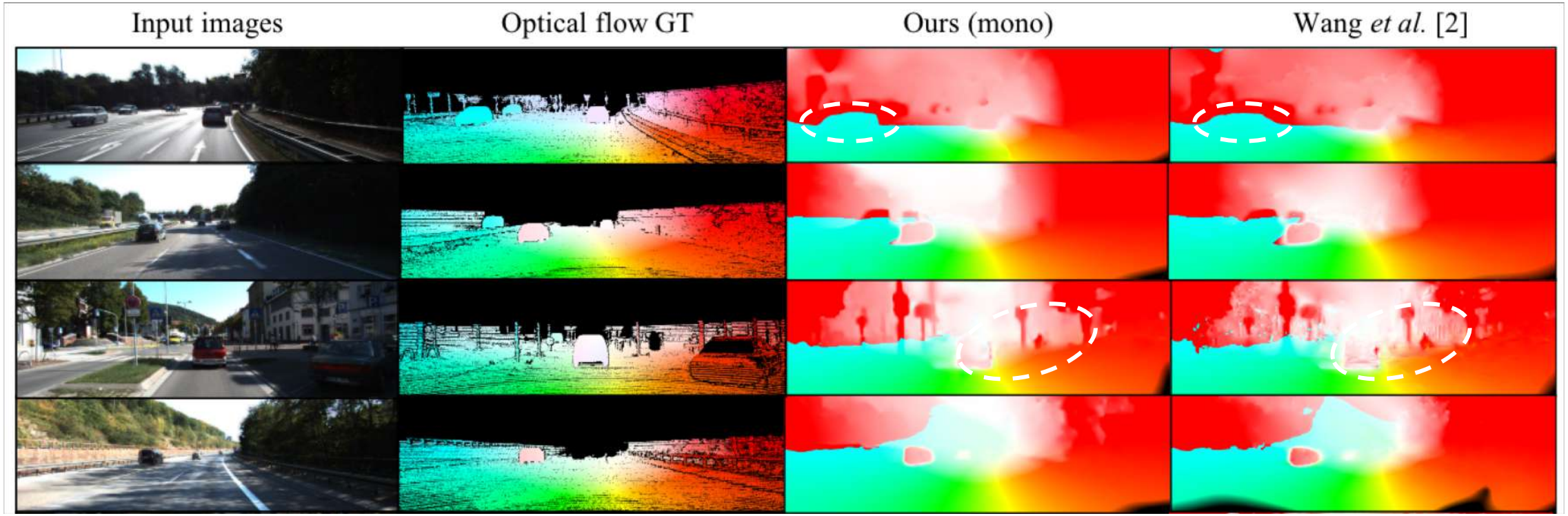




Evaluation (optical flow)

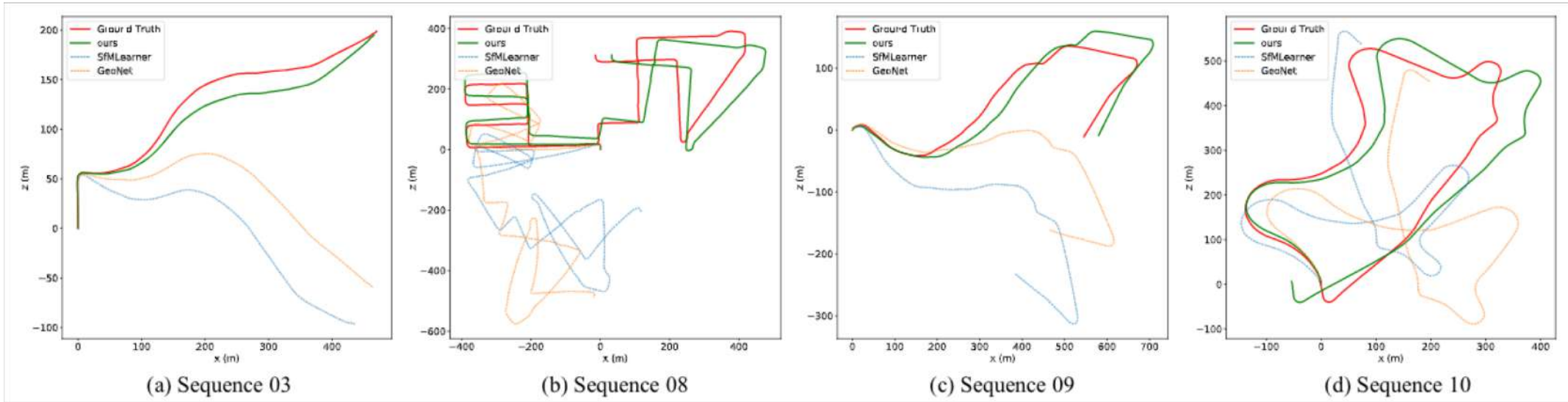
Method	KITTI 2012		KITTI 2015			
	Train	Test	Train	Test		
	all	all	all	bg	fg	all
DSTFlow [14]	10.43	12.40	16.79	-	-	39.00%
Unflow-CSS [54]	3.29	-	8.10	-	-	-
OccAwareFlow [6]	3.55	4.20	8.88	-	-	31.20%
Multi-frame [80]	-	-	6.59	22.67%	24.27%	22.94%
GeoNet [45]	-	-	10.81	-	-	-
DF-Net [78]	3.54	4.40	8.98	-	-	25.70%
Adversarial-Collaboration [62]	-	-	7.76	-	-	-
Ours (mono)	**2.30**	**2.6**	5.84	**20.61%**	**26.32%**	**21.56%**
Ours (stereo)	-	-	**5.66**	-	-	-

Evaluation (optical flow)



Evaluation (odometry)

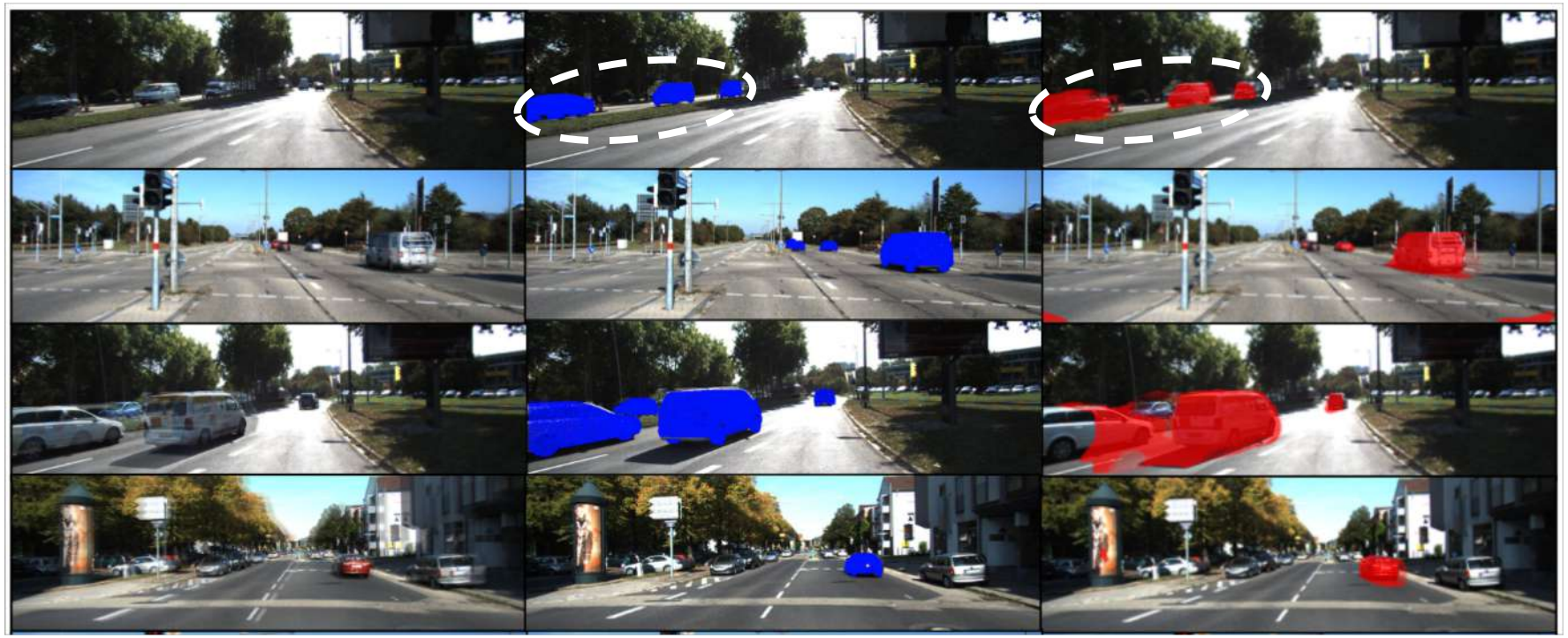
Methods	Seq. 09		Seq. 10	
	$t_{err}\%$	$r_{err}(°/100)$	$t_{err}\%$	$r_{err}(°/100)$
Zhou *et al.* [9]	30.75	11.41	44.22	12.42
GeoNet [46]	39.43	14.30	28.99	8.85
EPC++ (mono)	**3.72**	**1.60**	**6.06**	**2.22**



(a) Sequence 03 (b) Sequence 08 (c) Sequence 09 (d) Sequence 10

Evaluation (motion segmentation)

	pixel acc.	mean acc.	mean IoU	f.w. IoU
Explainability mask [9]	0.61	0.54	0.38	0.64
Yang *et al.* [7]	0.89	0.75	0.52	0.87
Ours(mono)	**0.94**	0.40	0.36	**0.92**
Ours(stereo)	0.90	**0.65**	**0.47**	0.84



Evaluation (scene flow)

Method	D1			D2			FL		
	bg	fg	bg+fg	bg	fg	bg+fg	bg	fg	bg+fg
Yang *et al.* [7]	23.62	27.38	26.81	18.75	70.89	60.97	25.34	28.00	25.74
Ours (mono)	30.67	34.38	32.73	18.36	84.64	65.63	**17.57**	27.30	19.78
Ours (stereo)	**22.76**	**26.63**	**23.84**	**16.37**	**70.39**	**60.32**	17.58	**26.89**	**19.64**

Summary

- Add more geometrical constrains
- Decouple different geometrical cues (depth, normal ,edge)
- Joint learning of multiple tasks is very helpful
- Decompose the background and dynamic motion

Thank you!