Self-Supervised Learning for Perception Tasks in Automated Driving

Dr. Wolfram Burgard

Vice President of Automated Driving Technology,
Toyota Research Institute

August 2020

TOYOTA RESEARCH INSTITUTE

Toyota Research Institute

$1B initial budget

321 Employees, secondees, & assignees as of Jan 2019

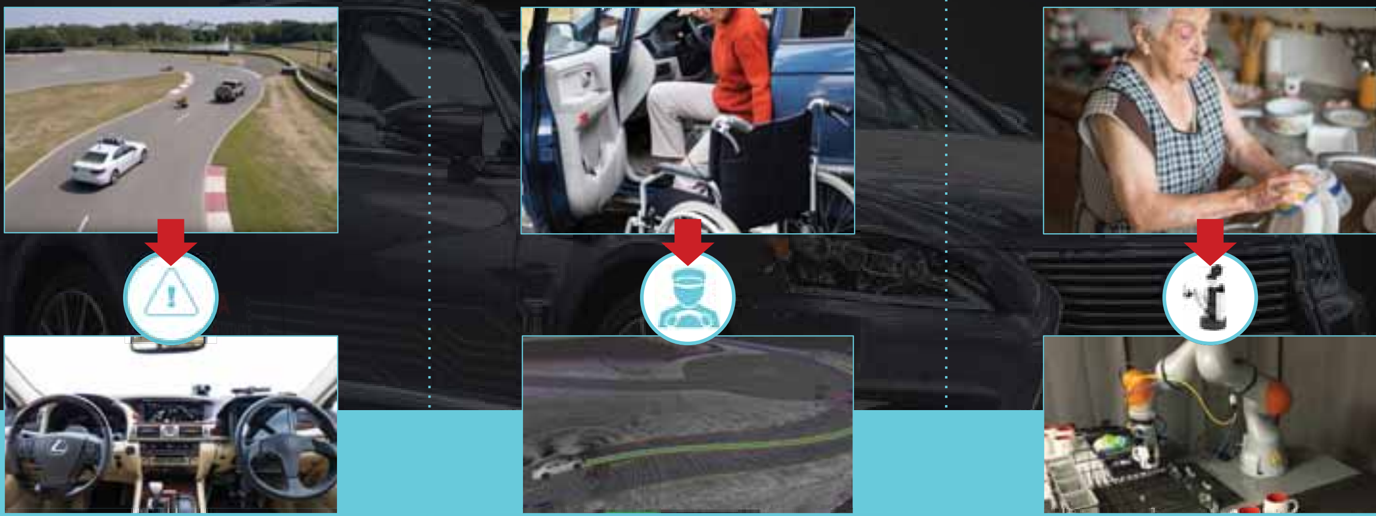
3 sites

- Established in January 2016
  - Leadership with experience from key government agencies & companies (i.e., U.S. DARPA, U.S. Dept. of Transportation, Google, Lyft, Zoox, Ford, U.S.-Japan Council)
  - More than 50% of technical staff hold PhD degrees
- Three facilities in Cambridge, Ann Arbor, & Silicon Valley
- Focus Areas: Automated Driving; Robotics; Advanced Material Design and Discovery; Machine Assisted Cognition
- Working closely with related Toyota Companies:

TOYOTA AI VENTURES · TOYOTA connected · CSRC · TRI-AD

TRI Aims to Transform the Human Condition

Safety | **Access** | **Quality of Life**



Guardian | **Chauffeur** | **Robots**

TRI Automated Driving Approach:
One System, Two Modes

GUARDIAN

- Driver always engaged, but vehicle monitors and intervenes to help prevent collisions
- Builds on similar hardware and software development as fully-autonomous Chauffeur

CHAUFFEUR

- Fully autonomous driving system engaged at all times
- Staged commercial release, likely beginning with shared mobility fleets

Creating an Autonomous Car is Hard

Feb. 27, 2019

The Moore's Law for Self-Driving Vehicles

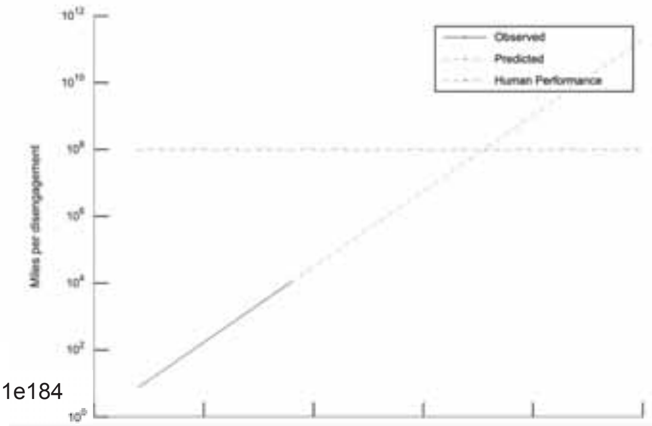
Edwin Olson Follow
Feb 27 · 9 min read

As the CEO of a self-driving car company, I'm constantly asked how long it will be until robo-taxis can take people pretty much anywhere, pretty much any time. We hear wildly different estimates from marketers ("Company X will solve robo-taxis in 2019!") and from engineers ("ugh, it's hard"), so who do we listen to?

For this post, let's measure the performance of a system in terms of the number of *miles per disengagement*. A disengagement, roughly speaking, is when the technology fails and a safety driver must take over. A great self-driving vehicle will have a *big* number—that means that the vehicle can drive a lot of miles and only infrequently fail.

... The number of miles between disengagements will double approximately every 16 months...

In a cosmic coincidence, the Moore's law for self-driving cars is almost the same as the Moore's law for computers—performance doubles every 16 months!



https://medium.com/may-mobility/the-moores-law-for-self-driving-vehicles-b78b8861e184

Creating an Autonomous Car is Hard

Feb. 27, 2019

The Moore's Law for Self-Driving Vehicles

Edwin Olson Follow
Feb 27 · 9 min read

As the CEO of a self-driving car company, I'm constantly asked how long it will be until robo-taxis can take people pretty much anywhere, pretty much any time. We hear wildly different estimates from marketers ("Company X will solve robo-taxis in 2019!") and from engineers ("ugh, it's hard"), so who do we listen to?

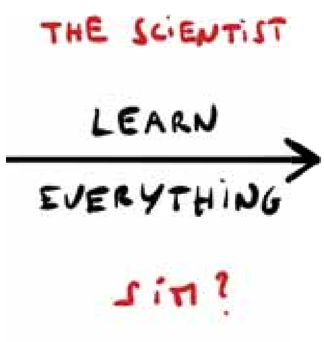
For this post, let's measure the performance of a system in terms of the number of *miles per disengagement*. A disengagement, roughly speaking, is when the technology fails and a safety driver must take over. A great self-driving vehicle will have a *big* number—that means that the vehicle can drive a lot of miles and only infrequently fail.

*Even with performance doubling every 16 months, it will take **16 years** to reach human levels of performance — that's 2035.*

https://medium.com/may-mobility/the-moores-law-for-self-driving-vehicles-b78b8861e184

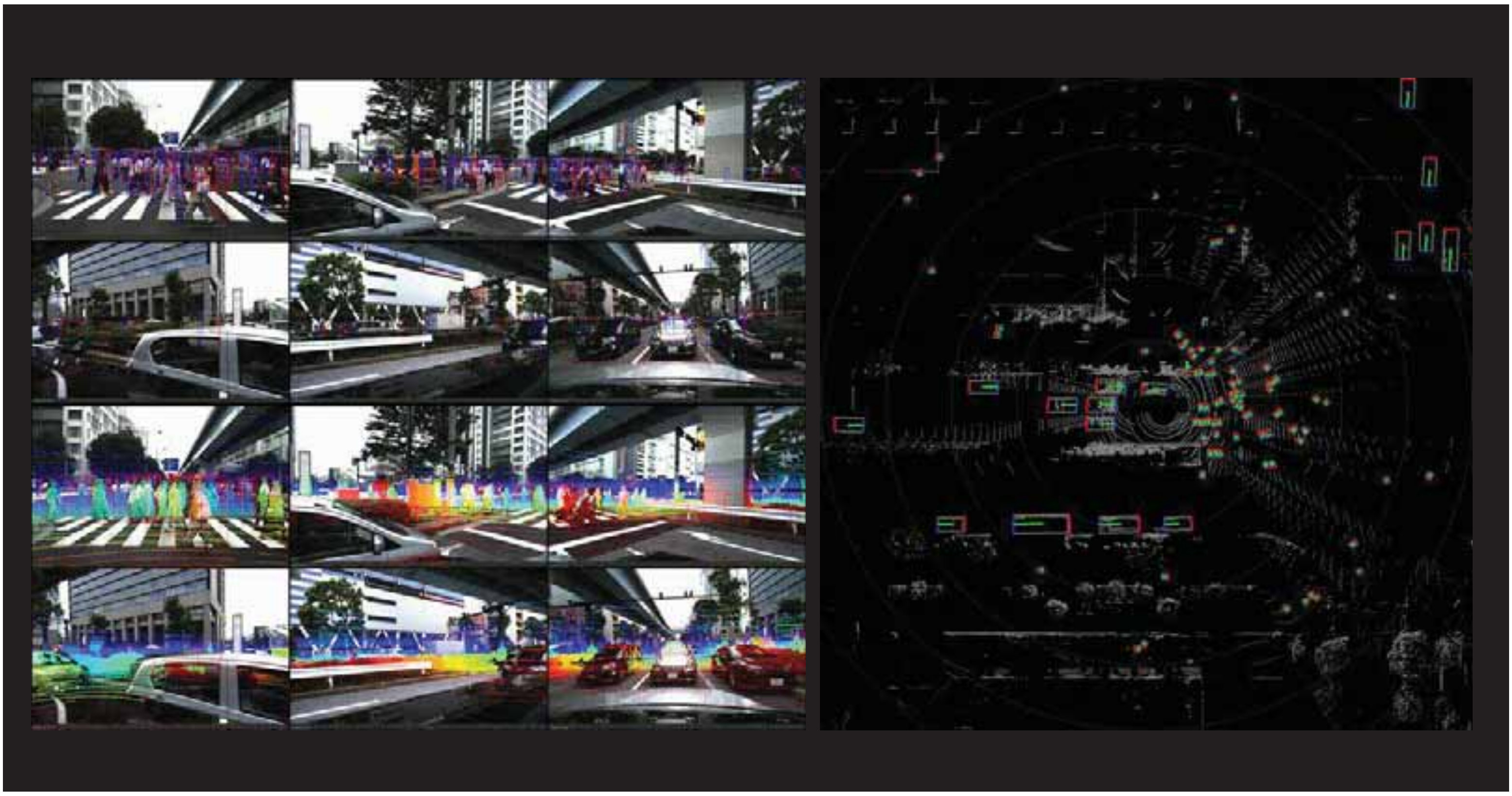
World-scale Autonomy?

World-scale Autonomy?

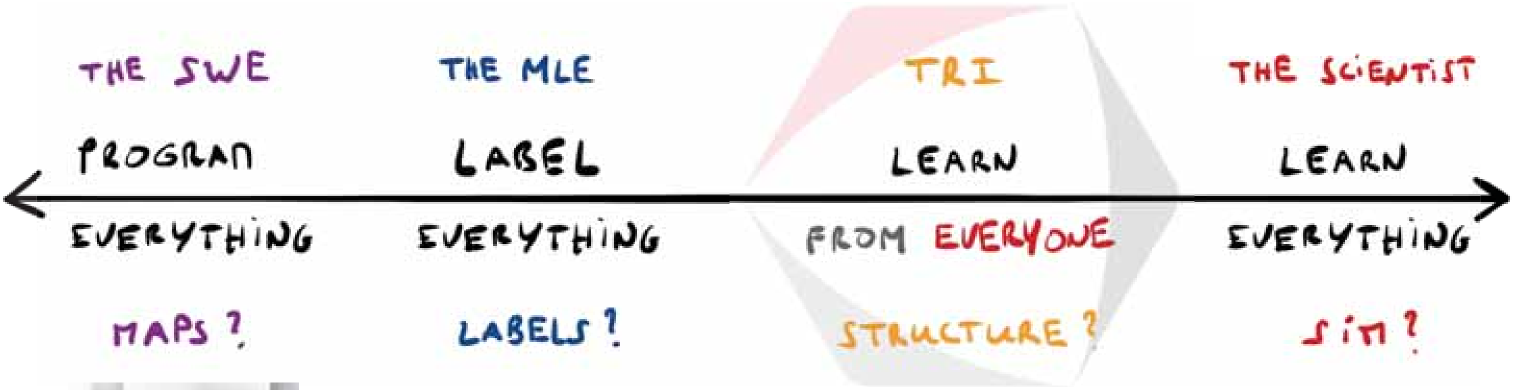


World-scale Autonomy?

THE SWE | THE MLE | THE SCIENTIST
PROGRAM | LABEL | LEARN
EVERYTHING | EVERYTHING | EVERYTHING
MAPS ? | LABELS ? | SIM ?



World-scale Autonomy?



Toyota's Strategic Data Advantage

- Unprecedented scale of data
  - Largest sensor fleet
  - Cover all US roads in under a day: Multiple times!
- Effective use of data
  - Learning from everything is infeasible!
  - Learn from unbiased, diverse and representative data
  - Leverage large volumes of unlabeled, structured data
  - **Data curation, querying, and synthesis**
- Our strategic focus
  - Supervised learning + **Self-supervised learning** from large volumes of structured and unlabeled data

We need to be smart about drinking from the data firehose

Self-Supervised Learning, Learning with Maps,
Transfer Learning, Representation Learning

Agenda

- **Self-Supervised Learning: SuperDepth**
- **Self-Supervised Pseudo-Lidar Networks**
- **Real-time Panoptic Segmentation**

Publication

SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation,

S. Pillai, R. Ambrus, A. Gaidon,

ICRA 2019


SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation

Sudeep Pillai, Rares Ambrus, Adrien Gaidon

Toyota Research Institute (TRI)

arXiv:1810.01849v1 [cs.CV] 3 Oct 2018

Self-Supervised Learning at Toyota-scale

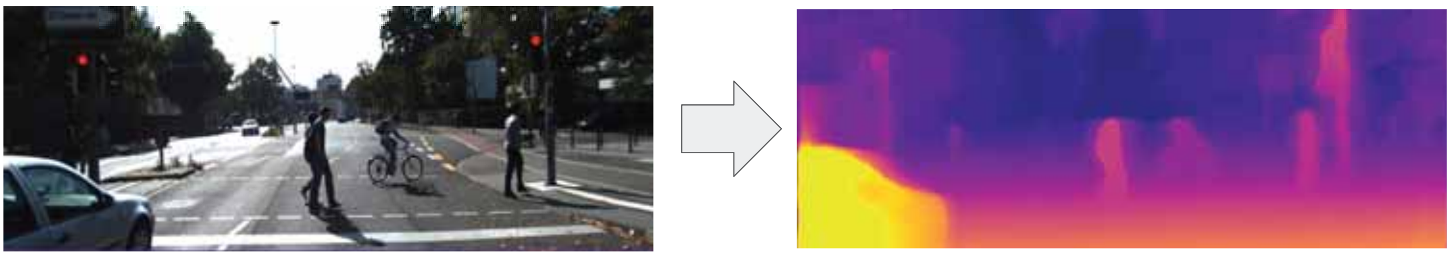
- **SuperDepth: Self-Supervised Monocular Depth**
  - Exploit **large volumes** of **unlabeled**, **structured** camera data
  - Training **only** requires **unlabeled driving video data!**

- Why MonoDepth?
  - LiDARs are expensive and bulky
  - Cameras
    - Rich semantic and geometric sensing
    - Ubiquitous (2019 Toyota models)

Toyota Safety Sense 2.0
Camera

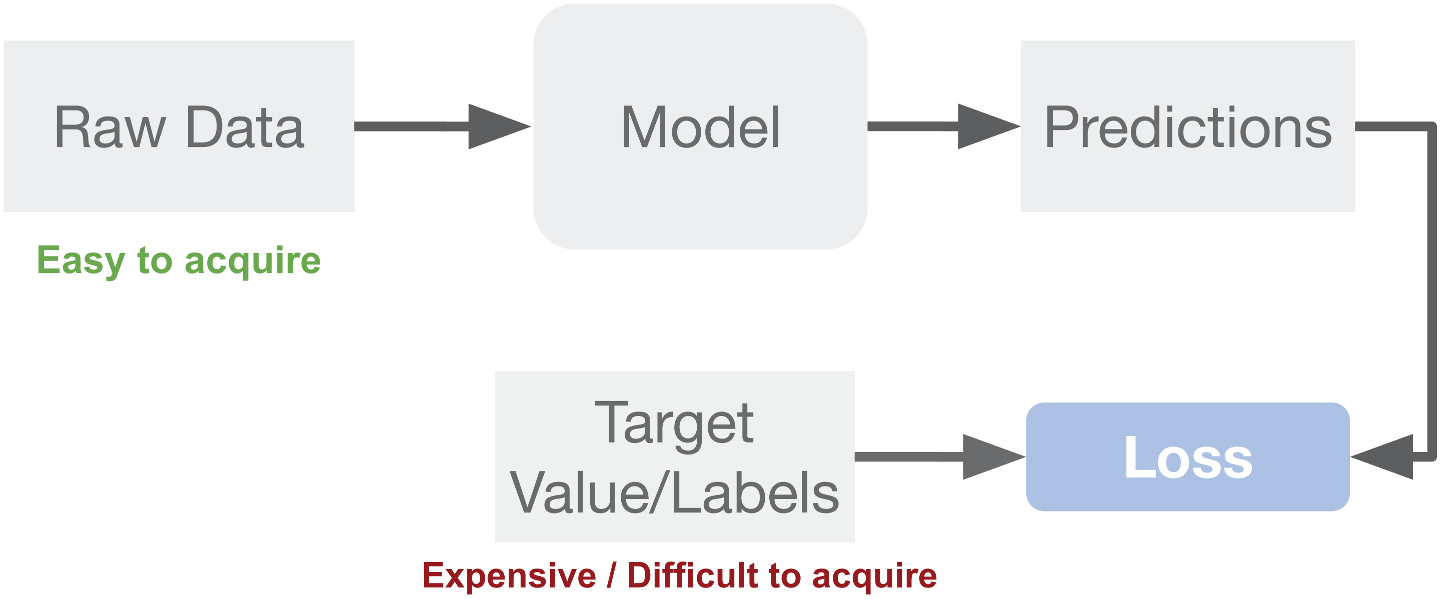
Monocular Depth Estimation

Single RGB Image **Predicted Depth Image**

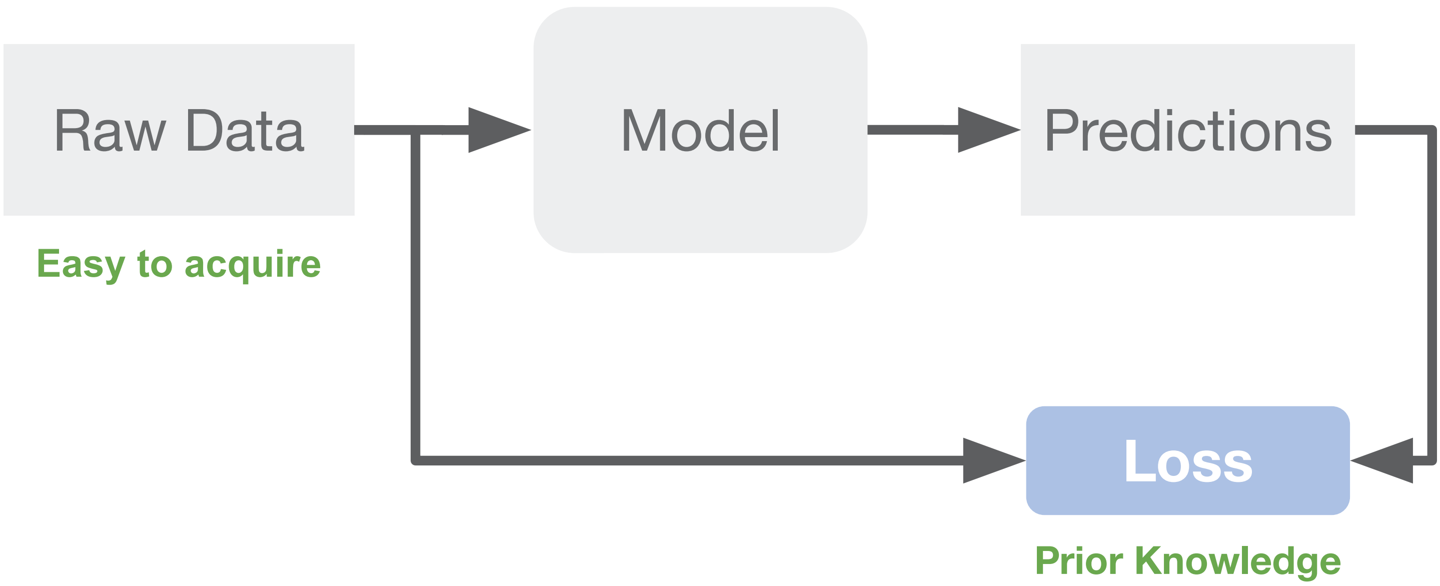


MonoDepth
Network

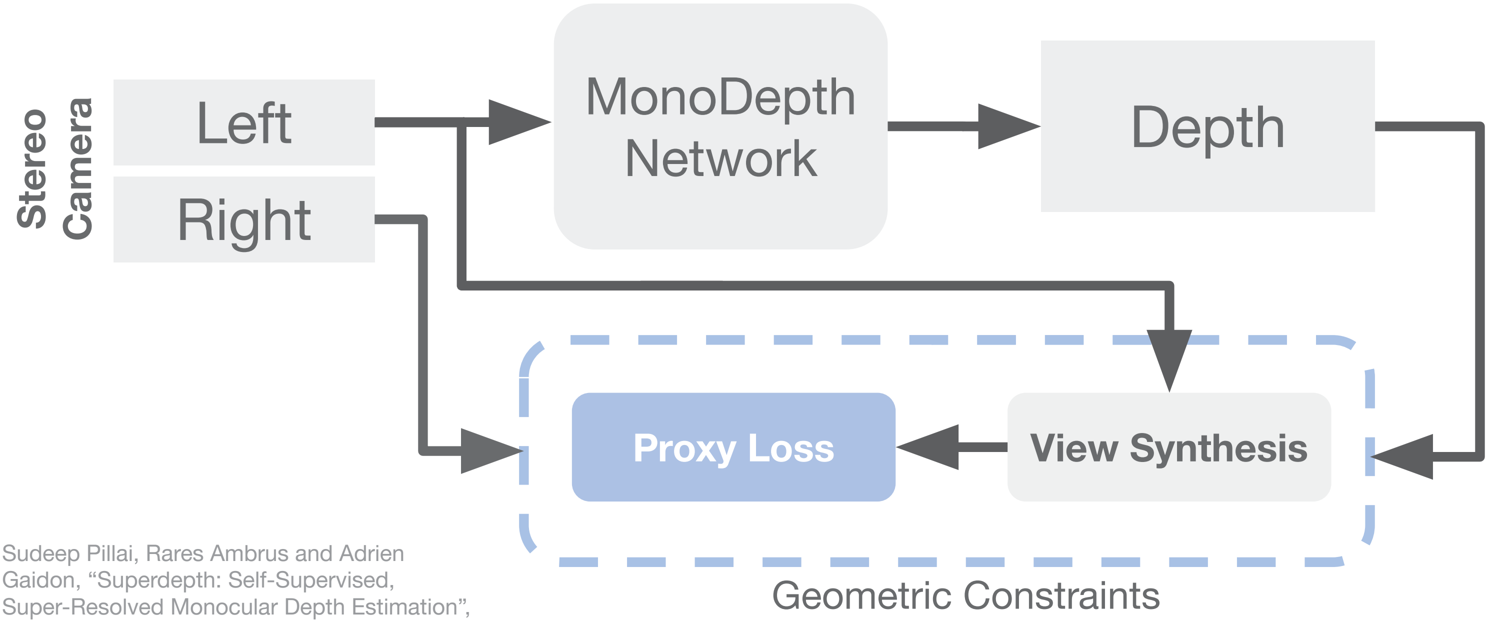
Supervised Learning



Self-Supervised Learning



Self-Supervised Monocular Depth



Sudeep Pillai, Rares Ambrus and Adrien Gaidon, “Superdepth: Self-Supervised, Super-Resolved Monocular Depth Estimation”, ICRA 2019

Self-Supervised Depth Learning Objective

$$\hat{\theta_D} = \arg\min_{\theta_D} \sum_{s \in S} \mathcal{L}_D(I_t, \hat{I}_t; \theta_D)$$

Depth Model Parameters

$$\mathcal{L}_D(I_t, \hat{I}_t) = \mathcal{L}_p(I_t, \hat{I}_t) + \lambda_1 \ \mathcal{L}_s(I_t) + \lambda_2 \ \mathcal{L}_o(I_t)$$

Photometric loss via view-synthesis

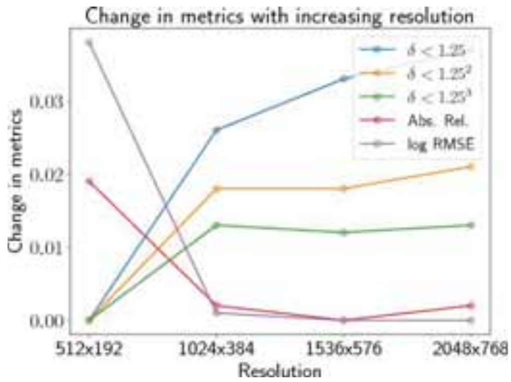
Depth Regularization (edge-aware depth smoothing)

Occlusion Regularization

Photometric Loss ++

- Multi-scale photometric loss is **limited** by resolution
- Super-resolve disparities → **synthesize at high resolutions**

Resolution Matters for View Synthesis!



Depth estimation accuracy **increases** with increasing high-resolution

Abs. Rel, and log RMSE (lower is better)

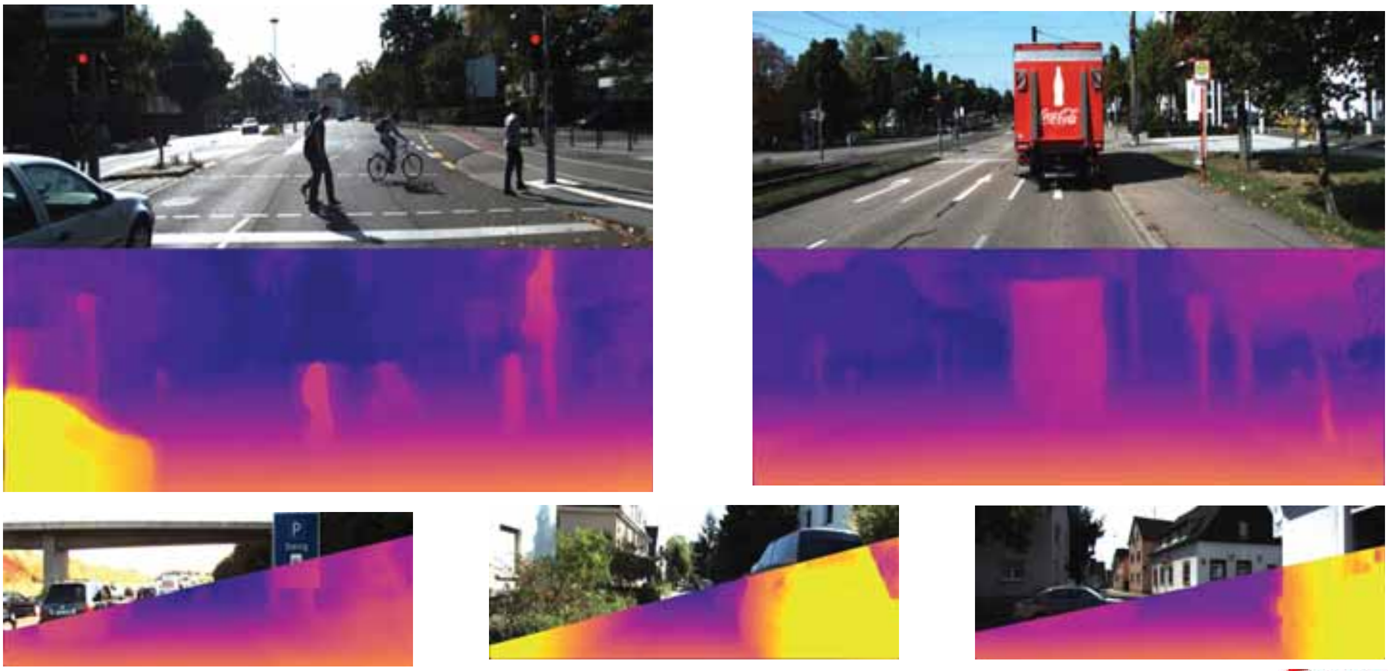
Disparity Estimation Performance

Method	Resolution	Dataset	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
UnDeepVO [25]	416 x 128	K	S	0.183	1.73	6.57	0.268	-	-	-
Godard et al. [6]	640 x 192	K	S	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [6]	640 x 192	CS+K	S	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Godard et al. [8]	640 x 192	K	S	0.115	1.010	5.164	0.212	**0.858**	0.946	0.974
Ours	1024 x 384	K	S	0.116	0.935	5.158	0.210	0.842	0.945	0.977
Ours-SP	1024 x 384	K	S	**0.112**	0.880	4.959	**0.207**	0.850	0.947	0.977
Ours-FA	1024 x 384	K	S	0.115	0.922	5.031	0.206	0.850	0.948	0.978
Ours-SP+FA	1024 x 384	K	S	**0.112**	**0.875**	**4.958**	**0.207**	0.852	**0.947**	**0.977**

Depth Estimation Results on the KITTI 2015 Benchmark

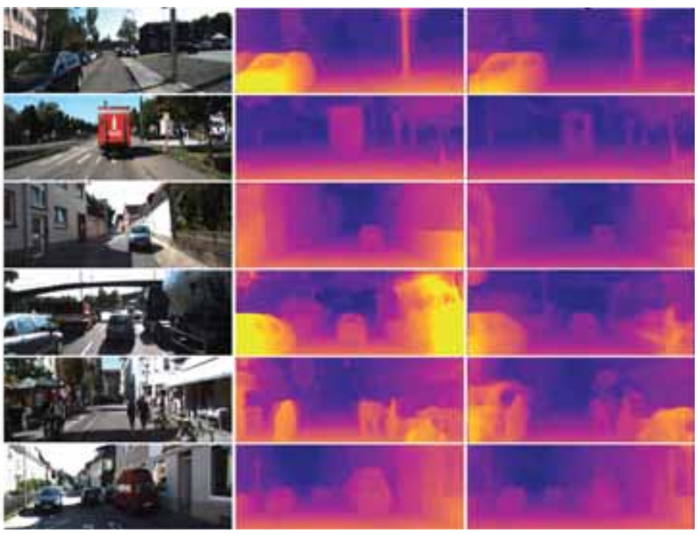
Sub-pixel convolutions (**SP**), Differentiable Flip Augmentation (**FA**)

Qualitative MonoDepth Performance

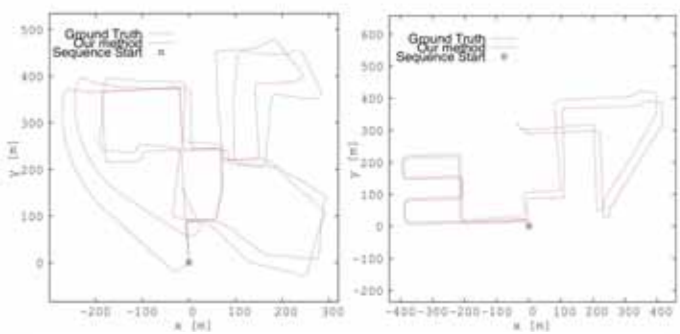


Qualitative Comparison to State-of-the-Art

SuperDepth **MonoDepth [1]**



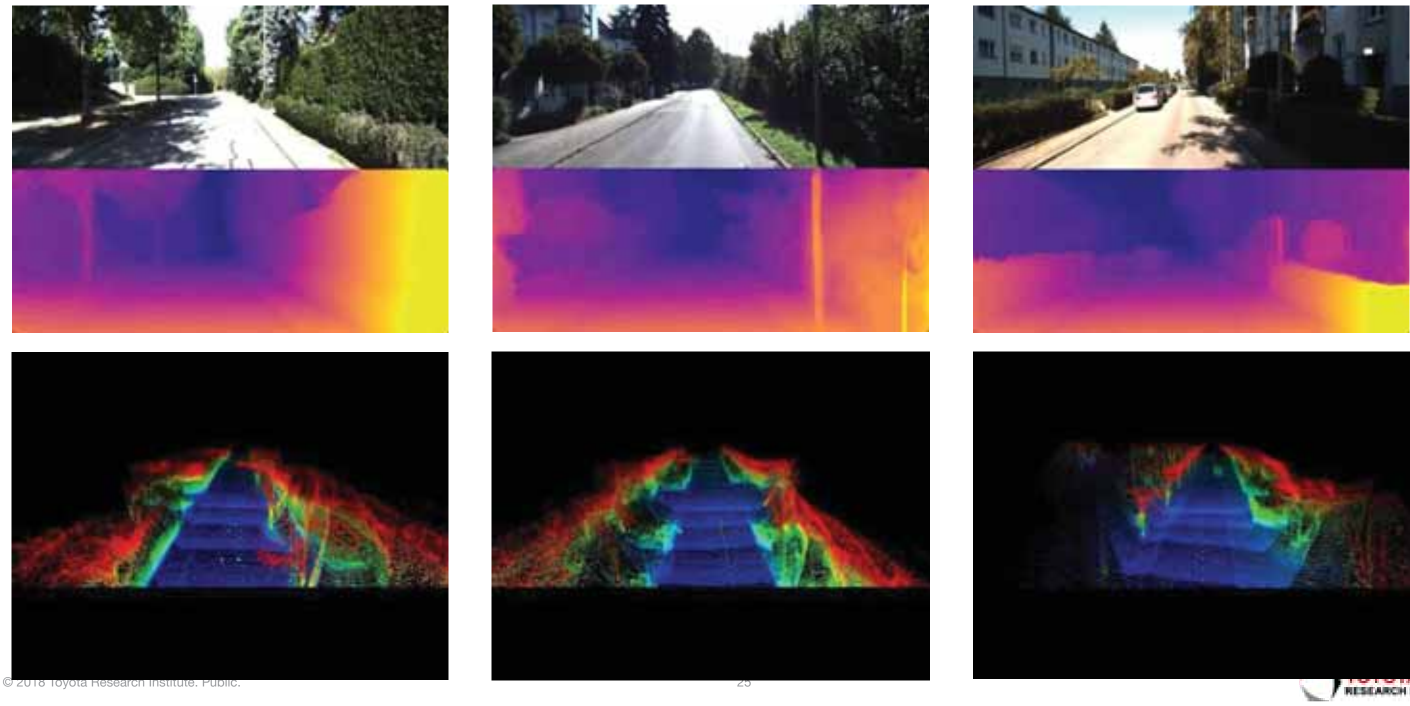
SuperDepth reconstruction is able to capture **fine details**, and **boundaries**



Bonus: We can also recover long-term, **scale-aware camera ego-motion from a single camera!**

[1] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," CVPR, 2017

Dense Monocular 3D Reconstruction



Agenda

- Self-Supervised Learning: SuperDepth
- Self-Supervised Pseudo-Lidar Networks
- Real-time Panoptic Segmentation

Publication

3D Packing for Self-Supervised Monocular Depth Estimation,

V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, A. Gaidon,

CVPR 2020, oral presentation

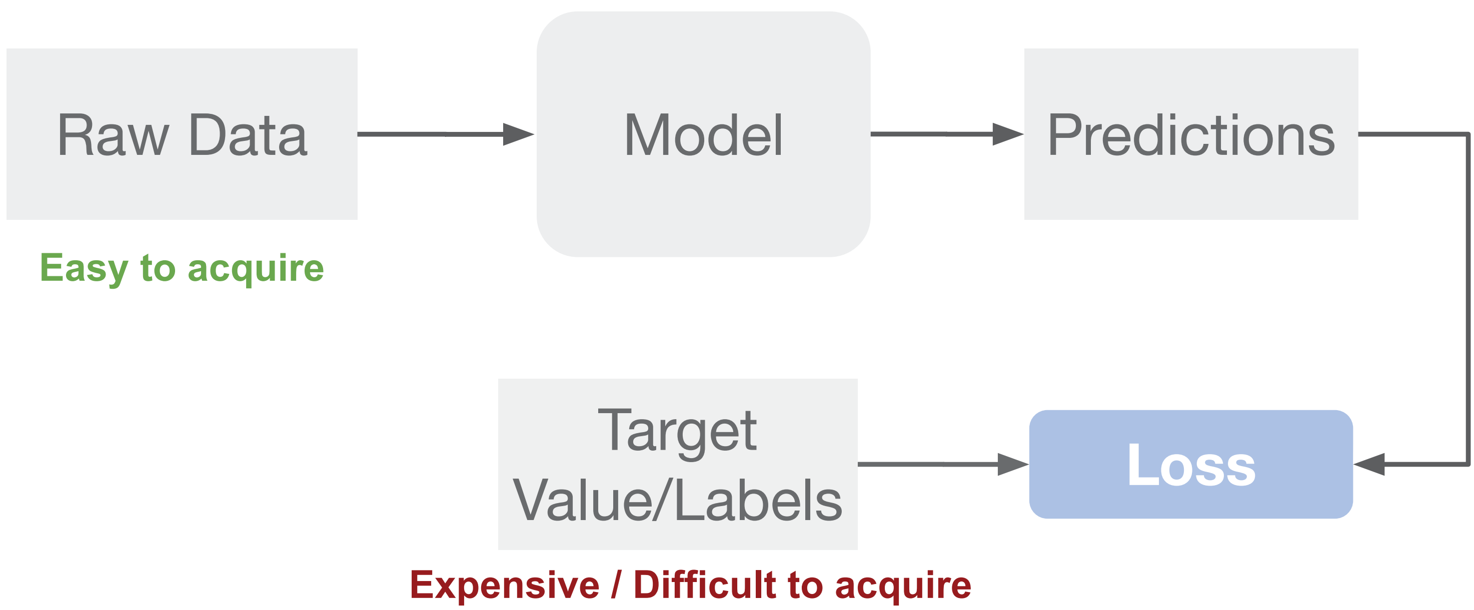

3D Packing for Self-Supervised Monocular Depth Estimation

Vitor Guizilini Rares Ambrus Sudeep Pillai Allan Raventos Adrien Gaidon
Toyota Research Institute (TRI)

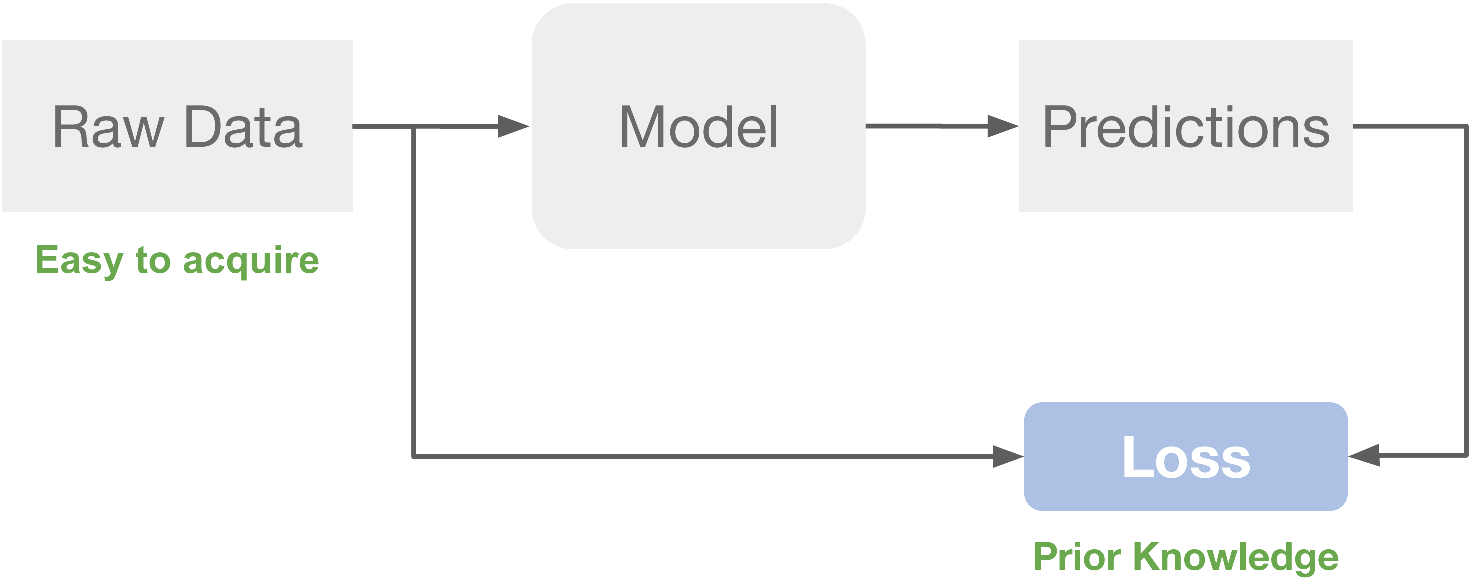
Abstract

1. Introduction

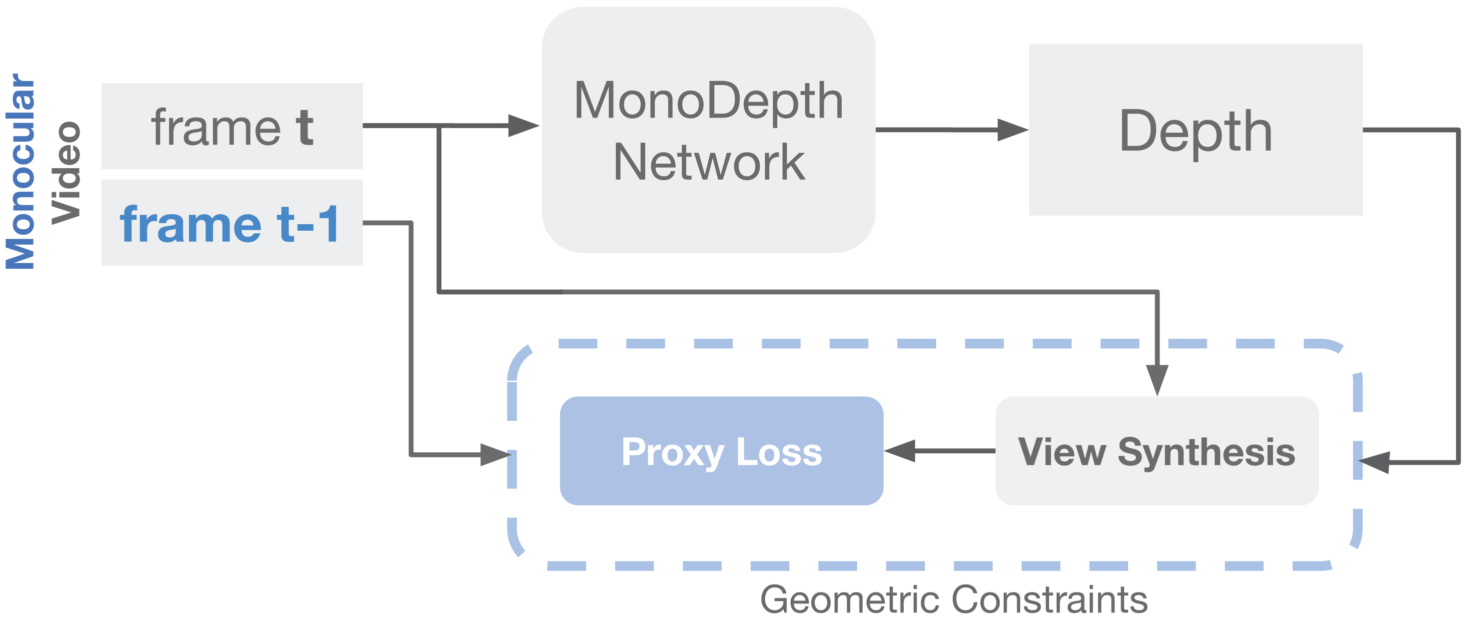
Supervised Learning



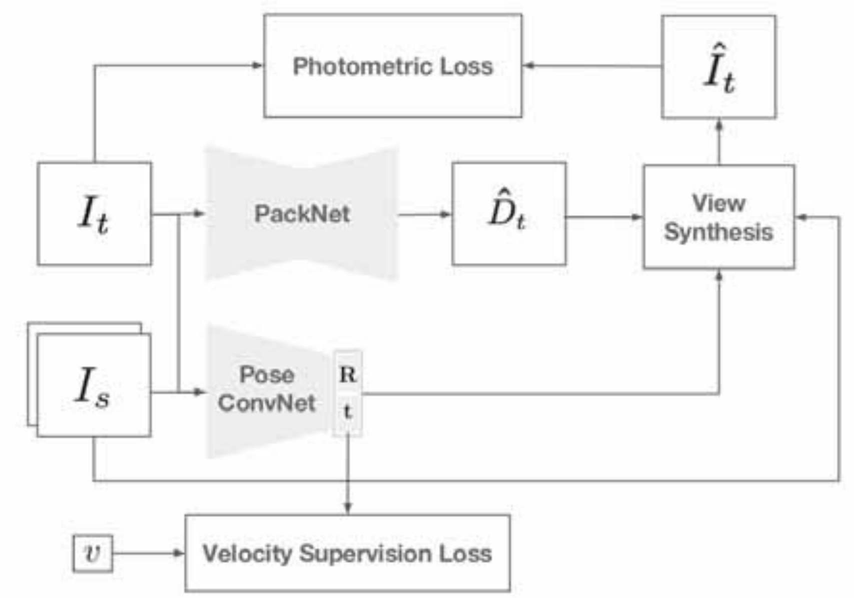
Self-Supervised Learning

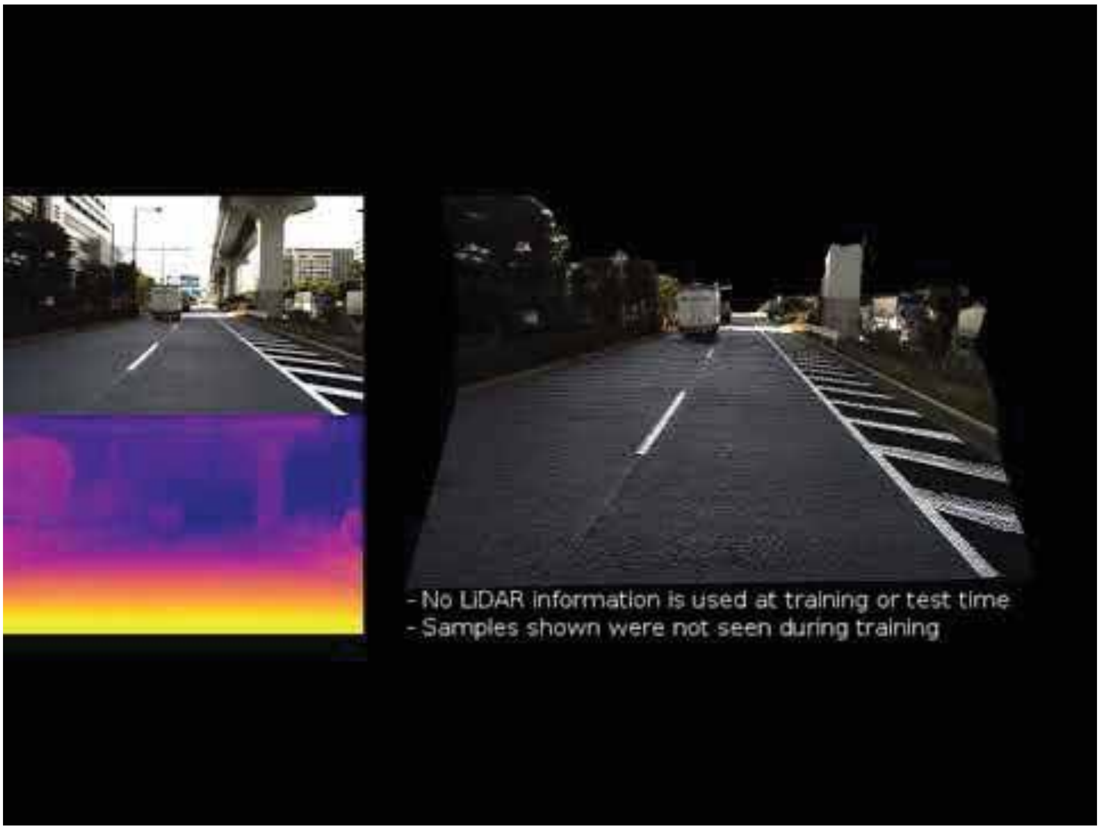


Self-Supervised Structure-from-Motion (SfM)



Self-Supervised Structure-from-Motion (SfM)

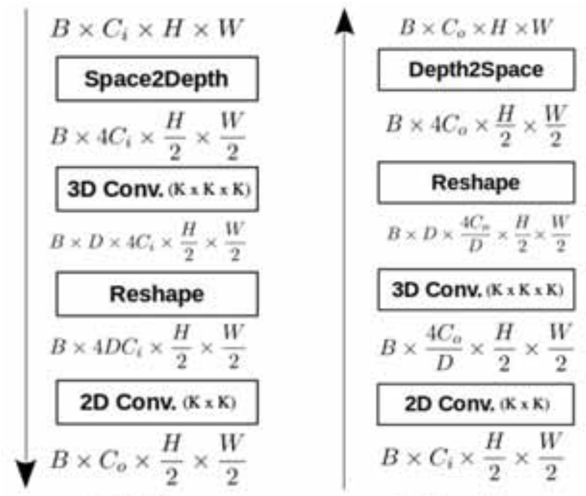




PackNet: Pack it, don't pool it

	Layer Description	K	Output Tensor Dim.
#0	Input RGB image		3×H×W
	Encoding Layers		
#1	Conv2d	5	64×H×W
#2	Conv2d → Packing	7	64×H/2×W/2
#3	ResidualBlock (x2) → Packing	3	64×H/4×W/4
#4	ResidualBlock (x2) → Packing	3	128×H/8×W/8
#5	ResidualBlock (x3) → Packing	3	256×H/16×W/16
#6	ResidualBlock (x3) → Packing	3	512×H/32×W/32
	Decoding Layers		
#7	Unpacking (#6) → Conv2d (⊕ #5)	3	512×H/16×W/16
#8	Unpacking (#7) → Conv2d (⊕ #4)	3	256×H/8×W/8
#9	InvDepth (#8)	3	1×H/8×W/8
#10	Unpacking (#8) → Conv2d (⊕ #3 ⊕ Upsample(#9))	3	128×H/4×W/4
#11	InvDepth (#10)	3	1×H/4×W/4
#12	Unpacking (#10) → Conv2d (⊕ #2 ⊕ Upsample(#11))	3	64×H/2×W/2
#13	InvDepth (#12)	3	1×H/2×W/2
#14	Unpacking (#12) → Conv2d (⊕ #1 ⊕ Upsample(#13))	3	64×H×W
#15	InvDepth (#14)	3	1×H×W

(a) Input Image (b) Max Pooling + Bilinear Upsample (c) Pack + Unpack



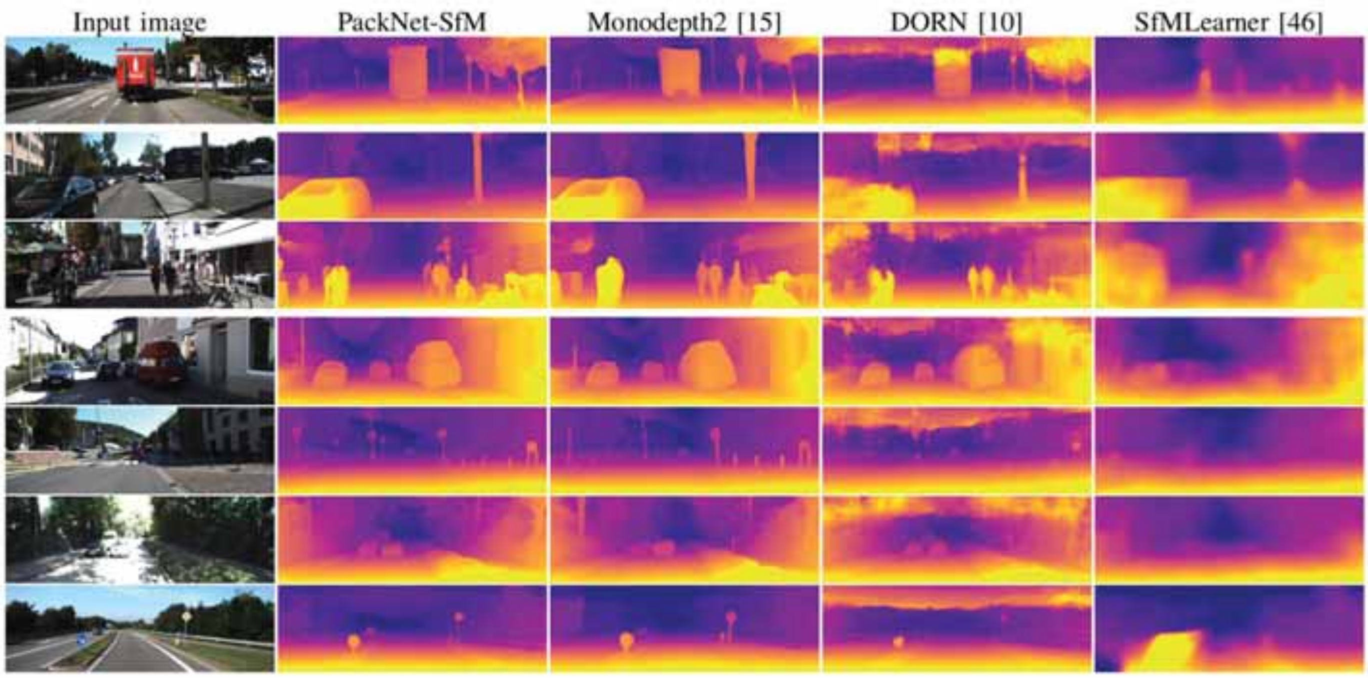
(a) Packing (b) Unpacking

Experimental Results (KITTI)

	Method	Supervision	Resolution	Dataset	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Original [9]	SfMLearner [46]	M	416 x 128	CS + K	0.198	1.836	6.565	0.275	0.718	0.901	0.960
	Klodt et al. [21]	M	416 x 128	CS + K	0.165	1.340	5.764	-	0.784	0.927	0.970
	Vid2Depth [28]	M	416 x 128	CS + K	0.159	1.231	5.912	0.243	0.784	0.923	0.970
	DF-Net [47]	M	576 x 160	CS + K	0.146	1.182	5.215	0.213	0.818	0.943	0.978
	Struct2Depth† [3]	M	416 x 128	K	0.141	1.026	5.291	0.215	0.8160	0.945	0.979
	Monodepth2 [15]	M	640 x 192	K	0.132	1.044	5.142	0.210	0.845	0.948	0.977
	Monodepth2‡ [15]	M	640 x 192	K	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Monodepth2‡ [15]	M	1024 x 320	K	0.115	0.882	4.701	0.190	0.879	0.961	0.982
	PackNet-SfM	M	640 x 192	K	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	PackNet-SfM	M+v	640 x 192	K	0.111	0.829	4.788	0.199	0.864	0.954	0.980
	PackNet-SfM	M	640 x 192	CS + K	0.108	**0.727**	4.426	0.184	0.885	0.963	**0.983**
	PackNet-SfM	M+v	640 x 192	CS + K	0.108	0.803	4.642	0.195	0.875	0.958	0.980
	PackNet-SfM	M	1280 x 384	K	0.107	0.802	4.538	0.186	0.889	0.962	0.981
	PackNet-SfM	M+v	1280 x 384	K	0.107	0.803	4.566	0.197	0.876	0.957	0.979
	PackNet-SfM	M	1280 x 384	CS + K	0.104	0.758	**4.386**	**0.182**	**0.895**	**0.964**	0.982
	PackNet-SfM	M+v	1280 x 384	CS + K	**0.103**	0.796	4.404	0.189	0.881	0.959	0.980
Improved [36]	SfMLeaner [46]	M	416 x 128	CS + K	0.176	1.532	6.129	0.244	0.758	0.921	0.971
	Vid2Depth [28]	M	416 x 128	CS + K	0.134	0.983	5.501	0.203	0.827	0.944	0.981
	GeoNet [42]	M	416 x 128	CS + K	0.132	0.994	5.240	0.193	0.883	0.953	0.985
	DDVO [38]	M	416 x 128	CS + K	0.126	0.866	4.932	0.185	0.851	0.958	0.986
	EPC++ [27]	M	640 x 192	K	0.120	0.789	4.755	0.177	0.856	0.961	0.987
	Monodepth2‡ [15]	M	640 x 192	K	0.090	0.545	3.942	0.137	0.914	0.983	0.995
	Kuznietsov et al.‡ [23]	Sup.	621 x 187	K	0.089	0.478	3.610	0.138	0.906	0.980	0.995
	DORN‡ [10]	Sup.	513 x 385	K	0.072	**0.307**	**2.727**	0.120	0.932	0.984	0.995
	PackNet-SfM	M	640 x 192	K	0.078	0.420	3.485	0.121	0.931	0.986	0.996
	PackNet-SfM	M	1280 x 384	CS + K	**0.071**	0.359	3.153	**0.109**	**0.944**	**0.990**	**0.997**
	PackNet-SfM	M+v	1280 x 384	CS + K	0.075	0.384	3.293	0.114	0.938	0.984	0.995

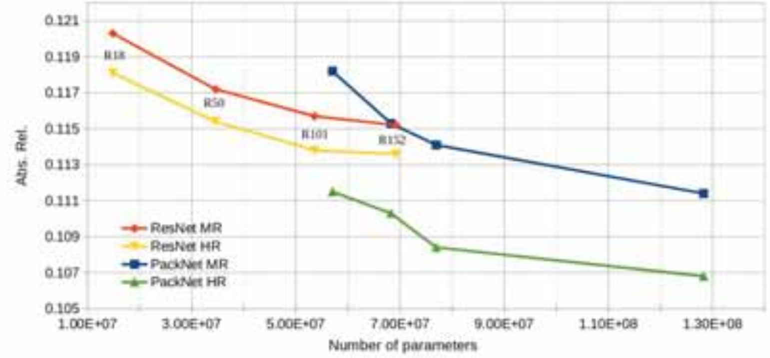
Self-sup. better than sup!

Experimental Results (KITTI)



Experimental Results

Better use of network capacity...

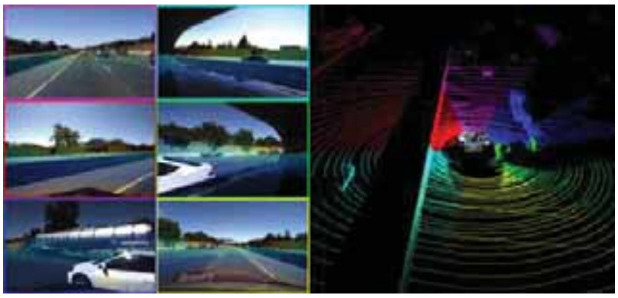


Depth Network	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$
ResNet18	0.133	1.023	5.123	0.211	0.845
ResNet18‡	0.120	0.896	4.869	0.198	0.868
ResNet50	0.127	0.977	5.023	0.205	0.856
ResNet50‡	0.117	0.900	4.826	0.196	0.873
PackNet18	0.118	0.802	4.656	0.194	0.868
PackNet50	0.114	0.818	4.621	0.190	0.875
PackNet-SfM (w/o pack/unpack)	0.122	0.880	4.816	0.198	0.864
PackNet-SfM (w/o 3D convs.)	0.118	0.922	4.831	0.195	0.872
PackNet-SfM	**0.111**	**0.785**	**4.601**	**0.189**	**0.878**

***And** better generalization!*
(KITTI → NuScenes)

Method	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	$\delta < 1.25$
ResNet18	0.218	2.053	8.154	0.355	0.650
ResNet18‡	0.212	1.918	7.958	0.323	0.674
ResNet50	0.216	2.165	8.477	0.371	0.637
ResNet50‡	0.210	2.017	8.111	0.328	0.697
PackNet-SfM	**0.187**	**1.852**	**7.636**	**0.289**	**0.742**

Experimental Results



DDAD: Dense Depth for Autonomous Driving
https://github.com/TRI-ML/DDAD

Frontiers of Monocular 3D Perception @CVPR'20
https://sites.google.com/view/mono3d-workshop

Agenda

- **Self-Supervised Learning: SuperDepth**
- **Self-Supervised Pseudo-Lidar Networks**
- **Real-time Panoptic Segmentation**

Publication

Real-Time Panoptic Segmentation from Dense Detections

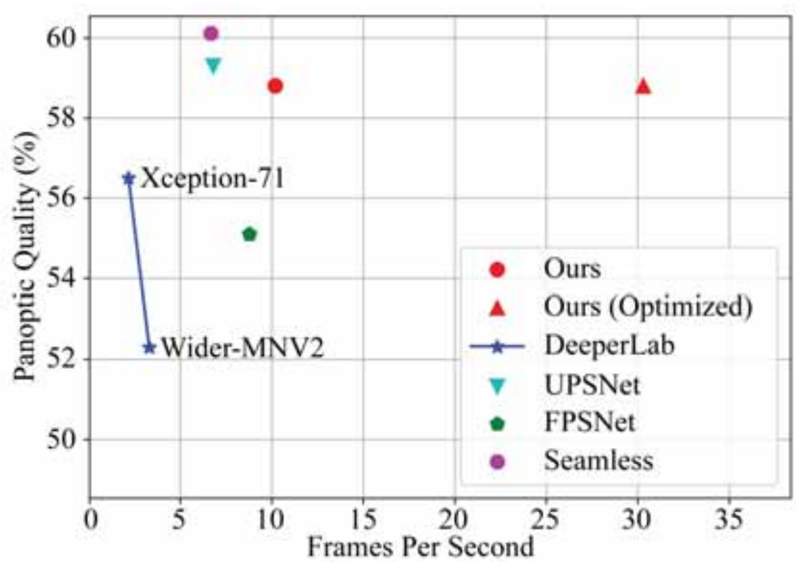
R. Hou, J. Li, A. Bhargava, A. Raventos, F. Guizilini, C. Fang, J. Lynch, A. Gaidon

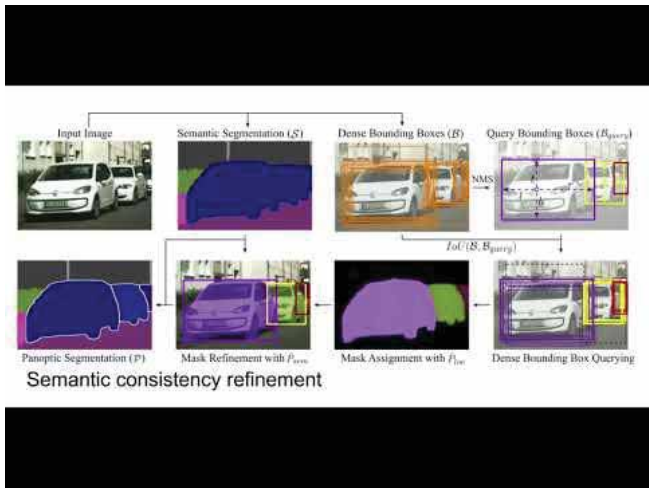
CVPR 2020, oral presentation

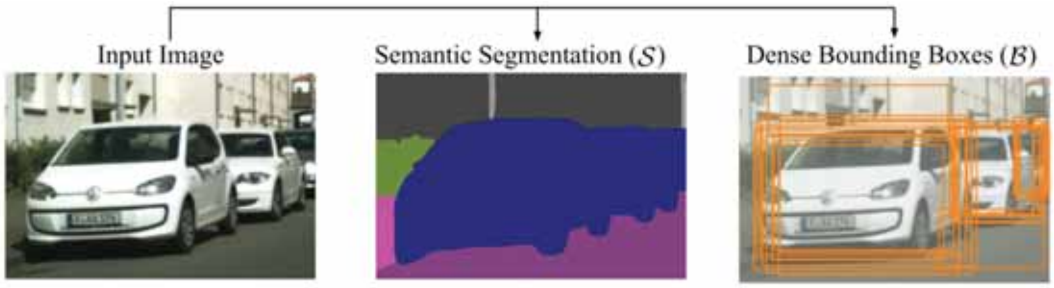

Real-Time Panoptic Segmentation from Dense Detections

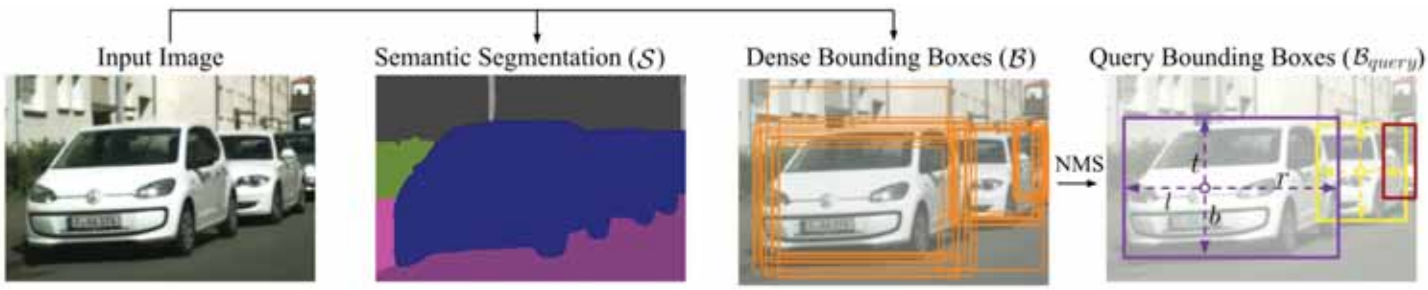
Abstract

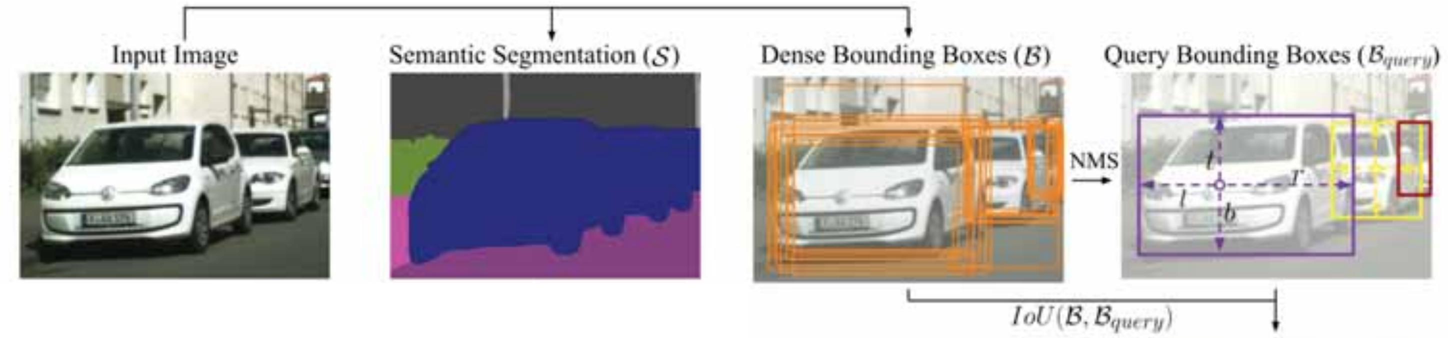
1. Introduction



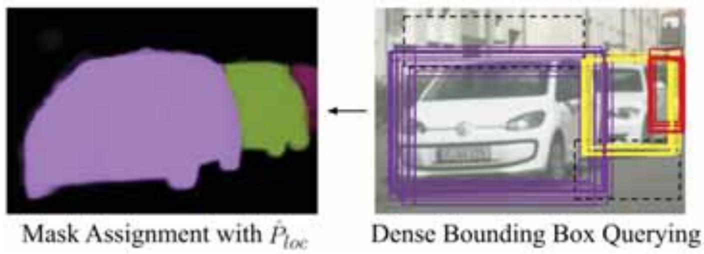


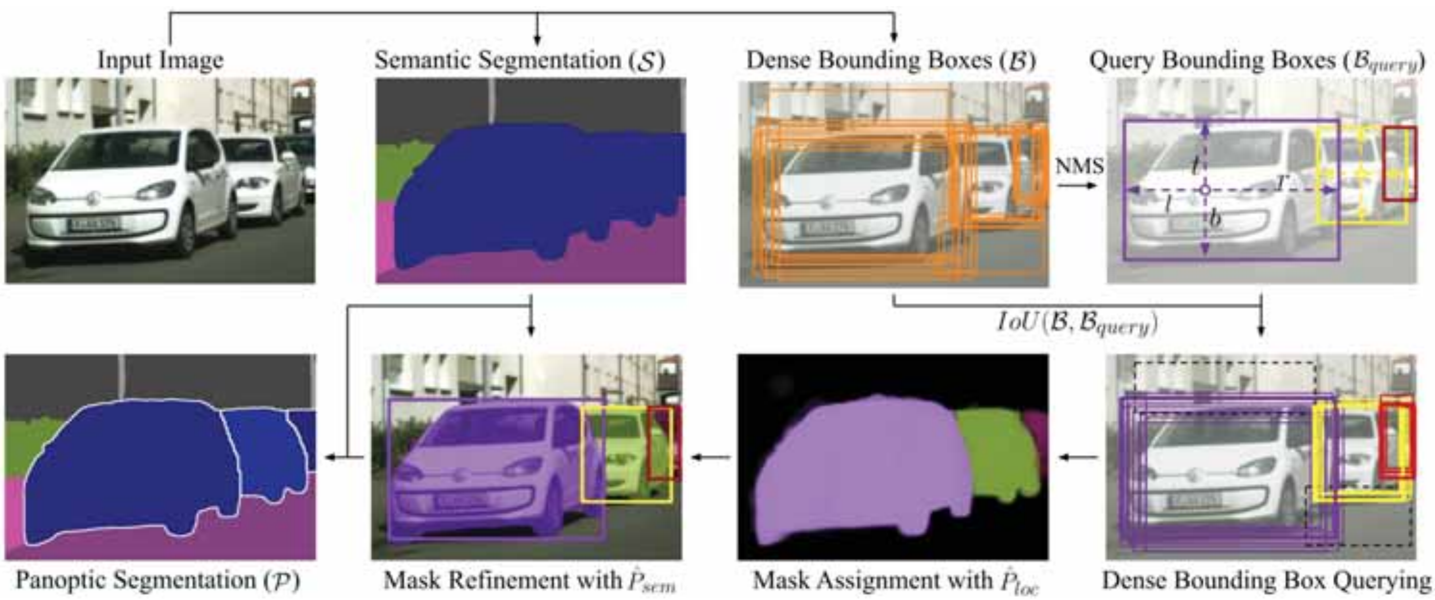

Input Image
Semantic Segmentation ($\mathcal{S}$)
Dense Bounding Boxes ($\mathcal{B}$)

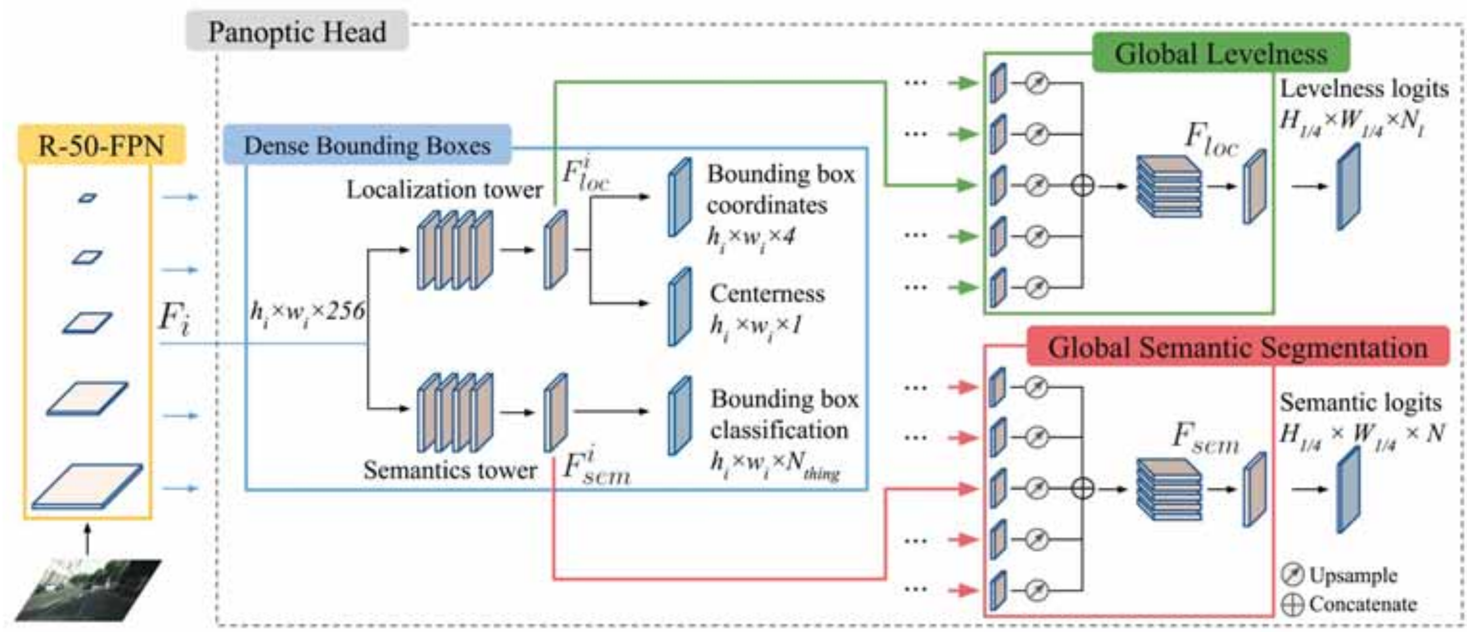

Input Image
Semantic Segmentation ($\mathcal{S}$)
Dense Bounding Boxes ($\mathcal{B}$)
NMS
Query Bounding Boxes ($\mathcal{B}_{query}$)
t
l
r
b



Parameter-free mask construction through bounding box self-attention







Cityscapes (val)

Method	Backbone	PQ	PQ^{th}	PQ^{st}	mIoU	AP	GPU	Inference Time
Two-Stage								
TASCNet [15]	ResNet-50-FPN	55.9	50.5	59.8	-	-	V100	160ms
AUNet[16]	ResNet-50-FPN	56.4	52.7	59.0	73.6	33.6	-	-
Panoptic-FPN [13]	ResNet-50-FPN	57.7	51.6	62.2	75.0	32.0	-	-
AdaptIS† [30]	ResNet-50	59.0	55.8	61.3	75.3	32.3	-	-
UPSNet [36]	ResNet-50-FPN	59.3	54.6	62.7	75.2	33.3	V100	140ms*
Seamless Panoptic [28]	ResNet-50-FPN	60.2	55.6	63.6	74.9	33.3	V100	150ms*
Single-Stage								
DeeperLab [38]	Wider MNV2	52.3	-	-	-	-	V100	251ms
FPSNet [7]	ResNet-50-FPN	55.1	48.3	60.1	-	-	TITAN RTX	114ms
SSAP [8]	ResNet-50	56.6	49.2	-	-	**31.5**	1080Ti	>260ms
DeeperLab [38]	Xception-71	56.5	-	-	-	-	V100	312ms
Ours	ResNet-50-FPN	**58.8**	**52.1**	63.7	77.0	29.8	V100	**99**ms

COCO (val)

Method	Backbone	PQ	PQ^{th}	PQ^{st}	Inf. Time
Two-Stage					
Panoptic-FPN [13]	ResNet-50-FPN	33.3	45.9	28.7	-
AdaptIS† [30]	ResNet-50	35.9	40.3	29.3	-
AUNet [16]	ResNet-50-FPN	39.6	49.1	25.2	-
UPSNet [36]	ResNet-50-FPN	42.5	48.5	33.4	110ms*
Single-Stage					
DeeperLab [38]	Xcep-71	33.8	-	-	94ms
SSAP [8]	ResNet-50	36.5	-	-	-
Ours	ResNet-50-FPN	**37.1**	**41.0**	**31.3**	**63**ms

☆ Ours

Supervision: Weak = 95% Strong

Two towers	Levelness	Mask loss	PQ	PQ^{th}	PQ^{st}
Fully Supervised					
			56.8	48.1	63.1
✓			57.1	47.8	**63.8**
✓	✓		58.1	50.4	63.7
✓	✓	✓	**58.8**	**52.1**	63.7
Weakly Supervised (No mask label)					
✓	✓		55.7	45.2	63.3

Conclusion

- Building truly autonomous cars requires machine learning
- The supervised learning approach does not scale
- We need to go beyond supervised learning and be able to learn from structured, unlabeled data


Thank You!