TOYOTA
RESEARCH INSTITUTE

Scaling up ML for Autonomy

Adrien Gaidon (twitter: @adnothing)

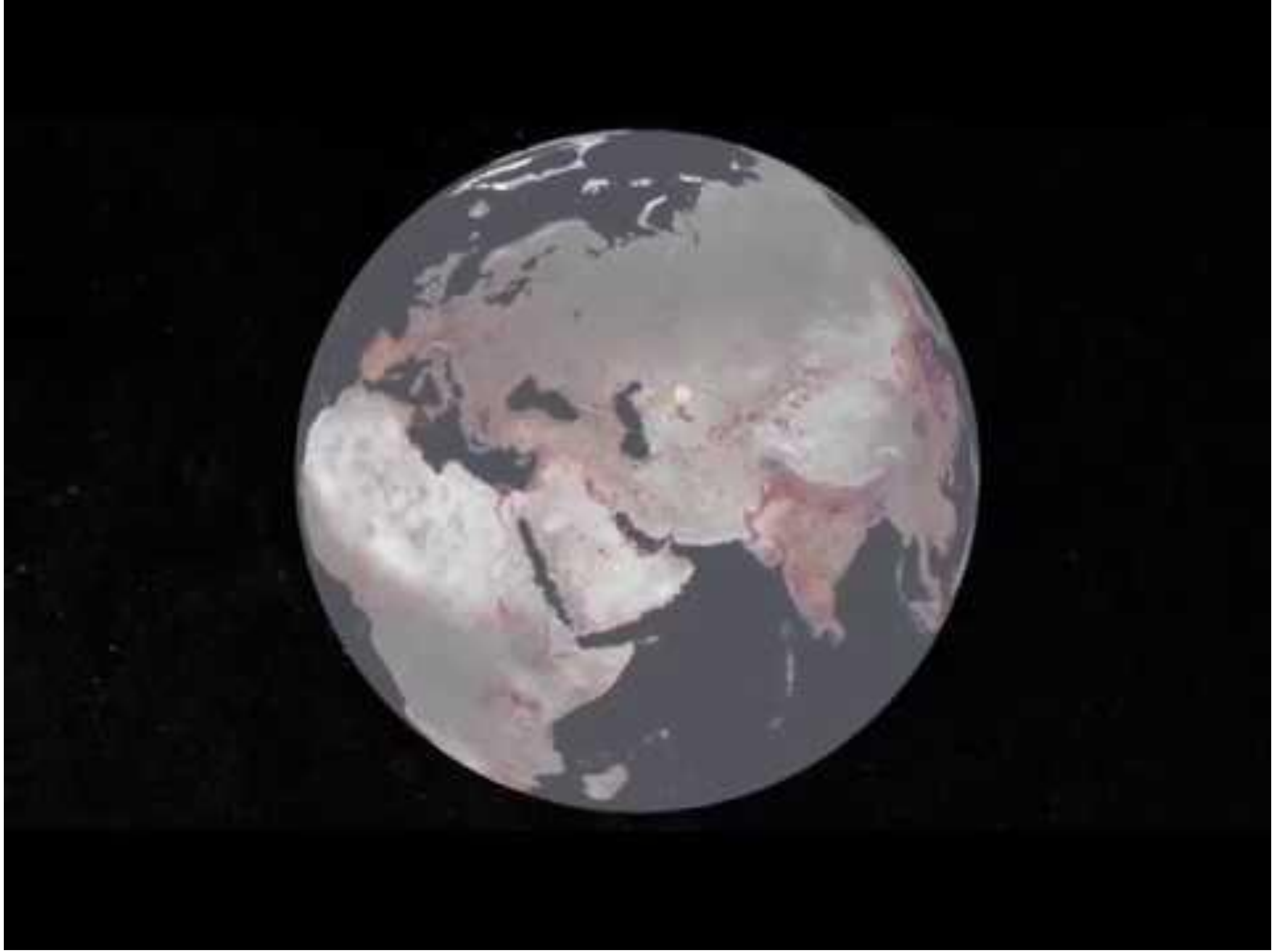
Head of Machine Learning Research
Toyota Research Institute (TRI), CA, USA

Collaborators
ML-Research team
(V. Guizilini, Jie Li,
R. Ambrus, W. Kehl, *et al*)
ML-Engineering team
(S. Pillai, A. Raventos, A.
Bhargava, KH. Lee, *et al*)
Wolfram Burgard
Stanford (SVL, MRL, ASL)

1.35 MILLION

ROAD TRAFFIC DEATHS PER YEAR

World Health Organization: Global status report on road safety 2018



Planet ©

youtu.be/ZqjKDpbtVn0

TOYOTA

100M Cars, 95% Parked

~10s PB

AMOUNT OF DATA PER DAY

World-scale Autonomy?

THE SWE

PROGRAM

←

EVERYTHING

MAPS ?

World-scale Autonomy?

THE SWE

PROGRAM

EVERYTHING

MAPS ?

THE SCIENTIST

LEARN

EVERYTHING

SIM ?

World-scale Autonomy?

THE SWE

PROGRAM

EVERYTHING

MAPS ?

THE MLE

LABEL

EVERYTHING

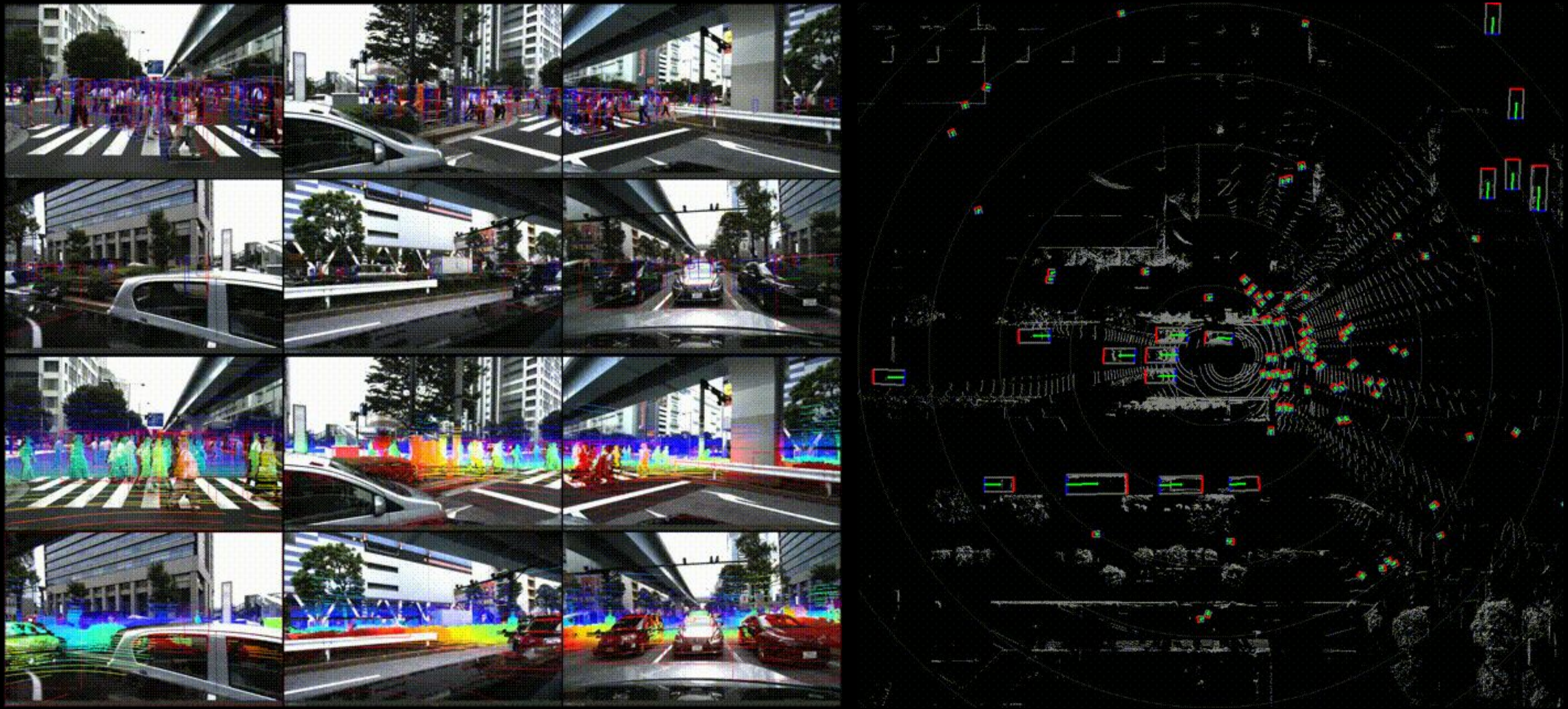
LABELS ?

THE SCIENTIST

LEARN

EVERYTHING

SIM ?



World-scale Autonomy?

THE SWE	THE MLE	TRI	THE SCIENTIST
PROGRAM	LABEL	LEARN	LEARN
EVERYTHING	EVERYTHING	FROM EVERYONE	EVERYTHING
MAPS ?	LABELS ?	STRUCTURE ?	SIM ?

TOYOTA RESEARCH INSTITUTE

Behavior: leverage large scale Demonstrations

***Exploring the Limitations of Behavior Cloning for Autonomous Driving*, ICCV'19 (oral)**
Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction, **RA-L & ICRA'20**
It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction, arXiv:2004.02025
Reinforcement Learning based Control of Imitative Policies for Near-Accident Driving, coming soon
Risk-Sensitive Sequential Action Control with Multi-Modal Human Trajectory Forecasting [...], coming soon
Driving Through Ghosts: Behavioral Cloning with False Positives, coming soon

Supervised Learning: efficiently use available Labels

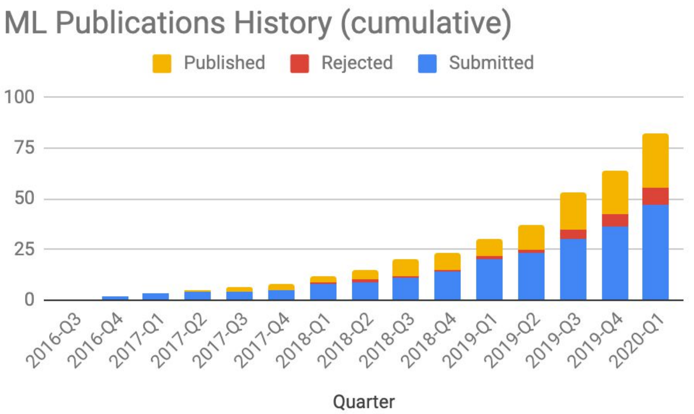
ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape, **CVPR'19**
Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss, **NeurIPS'19**
Learning to Fuse Things and Stuff, arXiv:1812.01192
Spatio-Temporal Graph for Video Captioning with Knowledge Distillation, **CVPR'20**
***Real-Time Panoptic Segmentation from Dense Detections*, CVPR'20 (oral)**
Hierarchical Lovász Embeddings for Proposal-free Panoptic Segmentation, coming soon
Unsupervised Estimation of Segmentation Difficulty, coming soon

Geometry: Self / Semi-Supervised Pseudo-LiDAR and SfM

SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation, **ICRA'19**
Robust Semi-Supervised Monocular Depth Estimation with Reprojected Distances, **CoRL'19**
Two Stream Networks for Self-Supervised Ego-Motion Estimation, **CoRL'19**
Semantically-Guided Representation Learning for Self-Supervised Monocular Depth, **ICLR'20**
Neural Outlier Rejection for Self-Supervised Keypoint Learning, **ICLR'20**
Self-Supervised 3D Keypoint Learning for Ego-motion Estimation, arxiv:1912.03426
***3D Packing for Self-Supervised Monocular Depth Estimation*, CVPR'20 (oral)**
Self-Supervised Neural Camera Models, coming soon

Simulation: Domain Adaptation, Differentiable Rendering, RL

SPIGAN: Privileged Adversarial Learning from Simulation, **ICLR'19**
DeceptionNet: Network-Driven Domain Randomization, **ICCV'19**
Generating Human Action Videos by Coupling 3D Game Engines and Probabilistic Graphical Models, **IJCV'20**
***Autolabeling 3D Objects with Differentiable Rendering of SDF Shape Priors*, CVPR'20 (oral)**
Self-Supervised Differentiable Rendering for Monocular 3D Object Detection, coming soon
Behaviorally Diverse Traffic Simulation via Reinforcement Learning, coming soon
Discovering Avoidable Planner Failures [...] in Behaviorally Diverse Simulation, coming soon



Upcoming workshops co-organized by TRI

ICML: AI for Autonomous Driving (AIAD)
https://sites.google.com/view/aiad2020
ECCV: Perception for Autonomous Driving (PAD)
https://sites.google.com/view/pad2020

Upcoming TRI Dataset Releases

STIP: Stanford-TRI Intent Prediction
http://stip.stanford.edu/
DDAD: Dense Depth for Autonomous Driving
https://github.com/TRI-ML/DDAD

TOYOTA RESEARCH INSTITUTE

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Exploring the Limitations of Behavior Cloning for Autonomous Driving, Codevilla et al, ***ICCV'19 (oral)***

Real-time Panoptic Segmentation

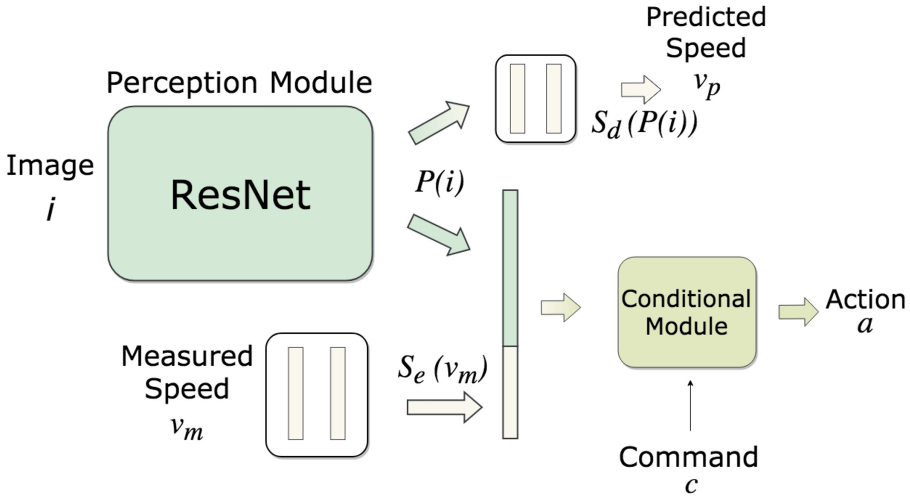
Self-Supervised Pseudo-Lidar Networks

Auto-labeling via Differentiable Rendering

Trillions of Miles driven yearly (3.2T in US alone)

Behavior Cloning: simplest Imitation Learning

Modernized: deeper ResNet, ImageNet, Data++



module	input dimension	channels
Perception	ResNet 34 [15] outputs 512	
Measured Speed	1	128
	128	128
	128	128
Speed Prediction	512	256
	256	256
	256	1
Joint input	$512 + 128$	512
Control	512	256
	256	256
	256	1

	Training conditions						New town & weather					
Task	CIL[10]	CIRL[26]	CAL[36]	MT[25]	CILR	CILRS	CIL[10]	CIRL[26]	CAL[36]	MT[25]	CILR	CILRS
Straight	98	98	**100**	96	94	96	80	**98**	94	96	92	96
One Turn	89	**97**	**97**	87	92	92	48	80	72	82	**92**	**92**
Navigation	86	93	92	81	88	**95**	44	68	68	78	88	**92**
Nav. Dynamic	83	82	83	81	85	**92**	42	62	64	62	82	**90**

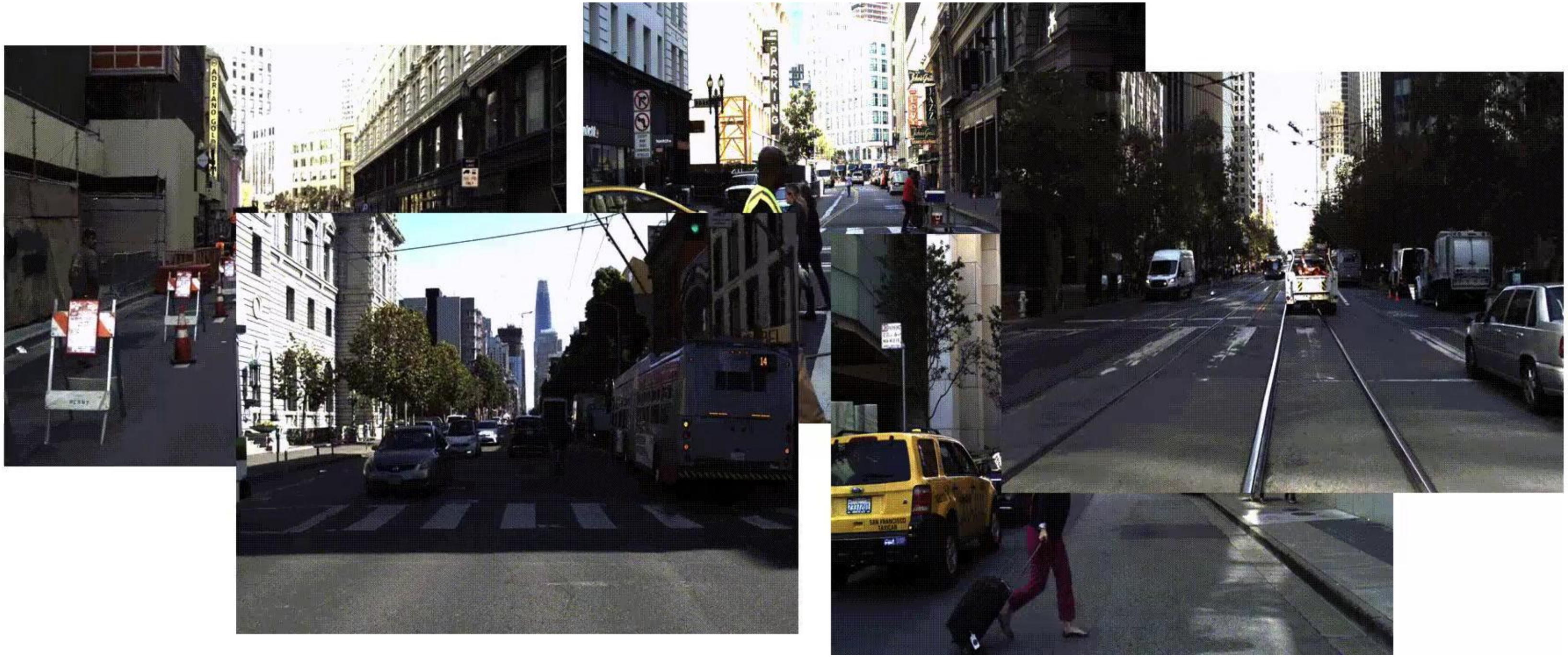
Table 1. Comparison with the state of the art on the original CARLA benchmark. The "CILRS" version corresponds to our CIL-based ResNet using the speed prediction branch, whereas "CILR" is without this speed prediction. These two models and CIL are the only ones that do not use any extra supervision or online interaction with the environment during training. The table reports the percentage of successfully completed episodes in each condition, selecting the best seed out of five runs.

	Training conditions					New Town & Weather				
Task	CIL[10]	CAL[36]	MT[25]	CILR	CILRS	CIL[10]	CAL[36]	MT[25]	CILR	CILRS
Empty	79 ± 1	81 ± 1	84 ± 1	92 ± 1	$\mathbf{97} \pm 2$	24 ± 1	25 ± 3	57 ± 0	66 ± 2	$\mathbf{90} \pm 2$
Regular	60 ± 1	73 ± 2	54 ± 2	72 ± 5	$\mathbf{83} \pm 0$	13 ± 2	14 ± 2	32 ± 2	54 ± 2	$\mathbf{56} \pm 2$
Dense	21 ± 2	$\mathbf{42} \pm 3$	13 ± 4	28 ± 1	$\mathbf{42} \pm 2$	2 ± 0	10 ± 0	14 ± 2	13 ± 4	$\mathbf{24} \pm 8$

Table 2. Results on our *NoCrash* benchmark. Mean and standard deviation on three runs, as CARLA 0.8.4 has significant non-determinism.

Motivation

Trillion miles driven yearly - Requires expensive data annotation



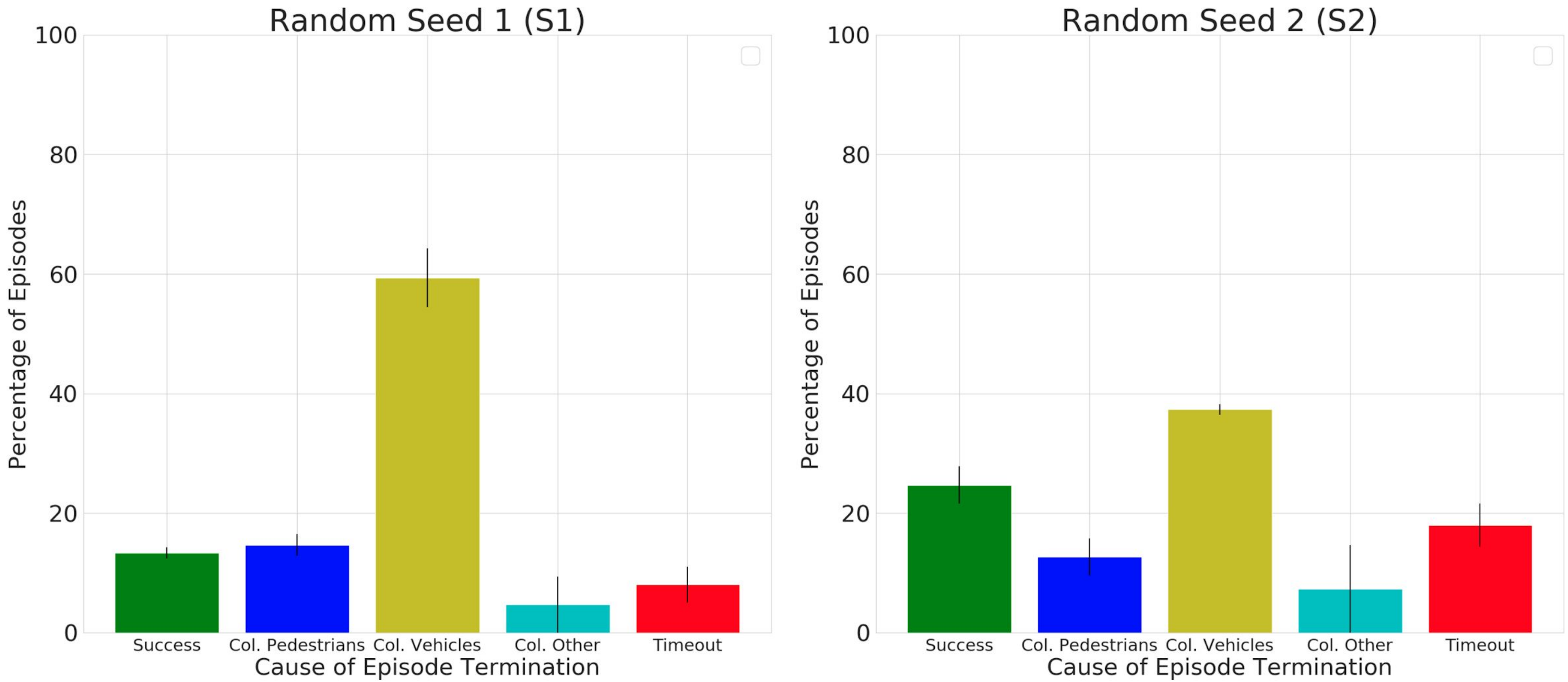


Figure 6. Cause of episode termination on *NoCrash* for two CILRS models (trained on 10 hours with ImageNet initialization) with identical parameters but different random seeds. The episodes were ran under “New Weather & Town” conditions of the “Dense Traffic” task.

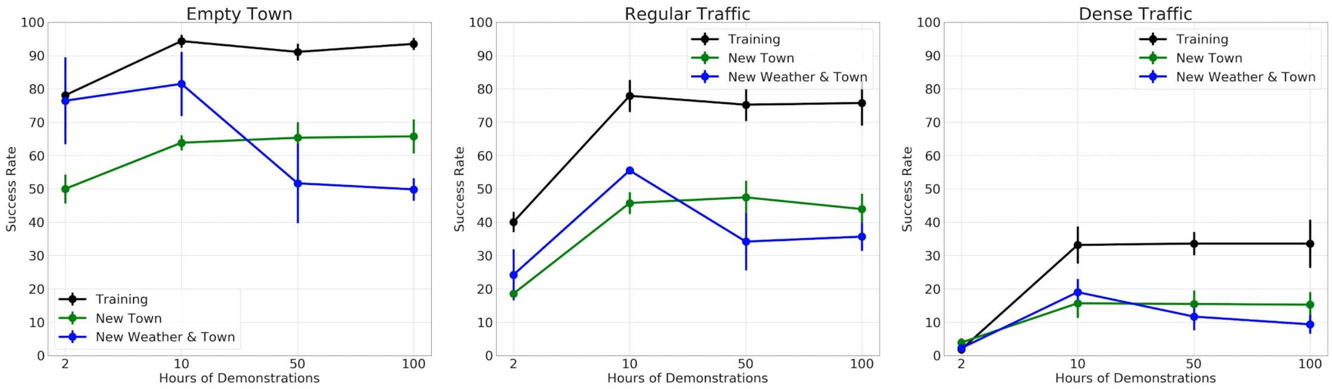


Figure 3. Due to biases in the data, the results may get either saturated or worse with increasing amounts of training data.

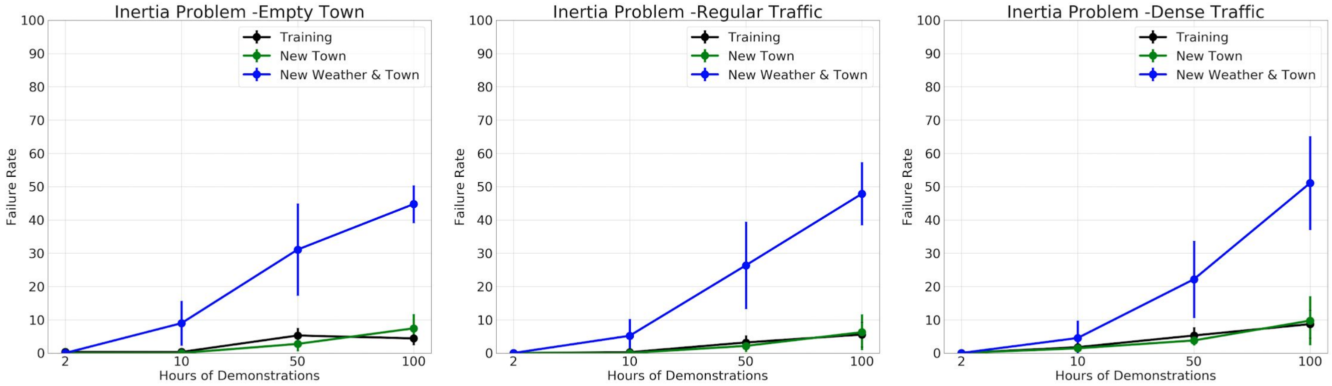


Figure 4. The percentage of episodes that failed due to the inertia problem. We can see that by increasing the amount of data, this bias may further degrade the generalization capabilities of the models.

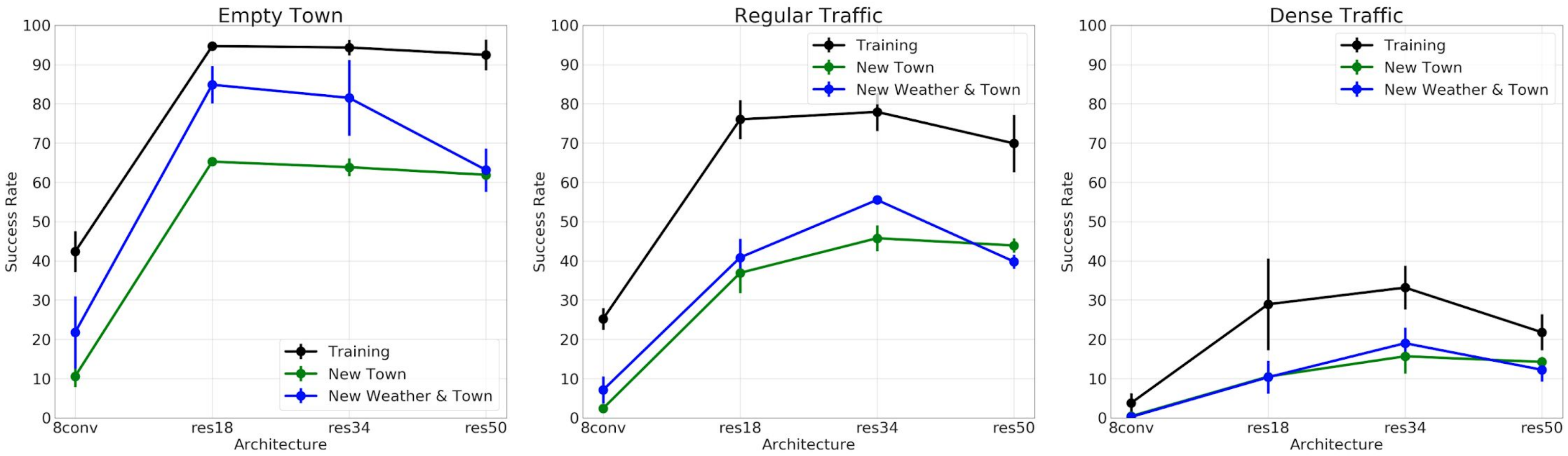
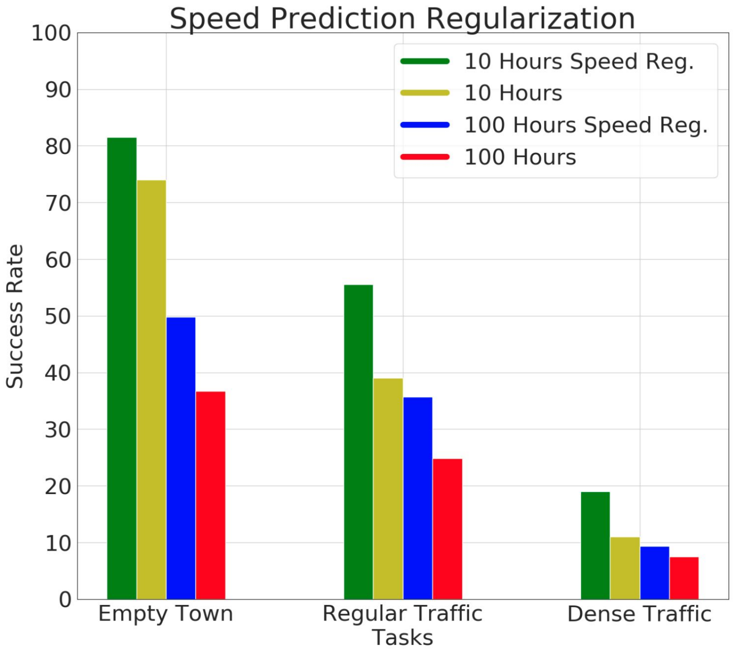


Figure 12. Ablative analysis between different architectures. The eight convolutions architecture,“8conv”, proposed by Codevilla [10] obtained poor results on the more complex CARLA100 benchmark. ResNet based deeper architectures, “res18” and “res34”, were able to improve the results. However, when testing ResNet 50 we notice a significant drop in the quality of the results.

	Task	Variance
CILRS	Empty	23%
	Regular	26%
	Dense	42%
CILRS (ImageNet)	Empty	4%
	Regular	12%
	Dense	38%

Table 3. Estimated variance of the success rate of CILRS on *NoCrash* computed by training 12 times the same model with different random seeds. The variance is reduced by fixing part of the initial weights with ImageNet pre-training.



Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Exploring the Limitations of Behavior Cloning for Autonomous Driving, Codevilla et al, ***ICCV'19 (oral)***

Real-time Panoptic Segmentation

Self-Supervised Pseudo-Lidar Networks

Auto-labeling via Differentiable Rendering

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

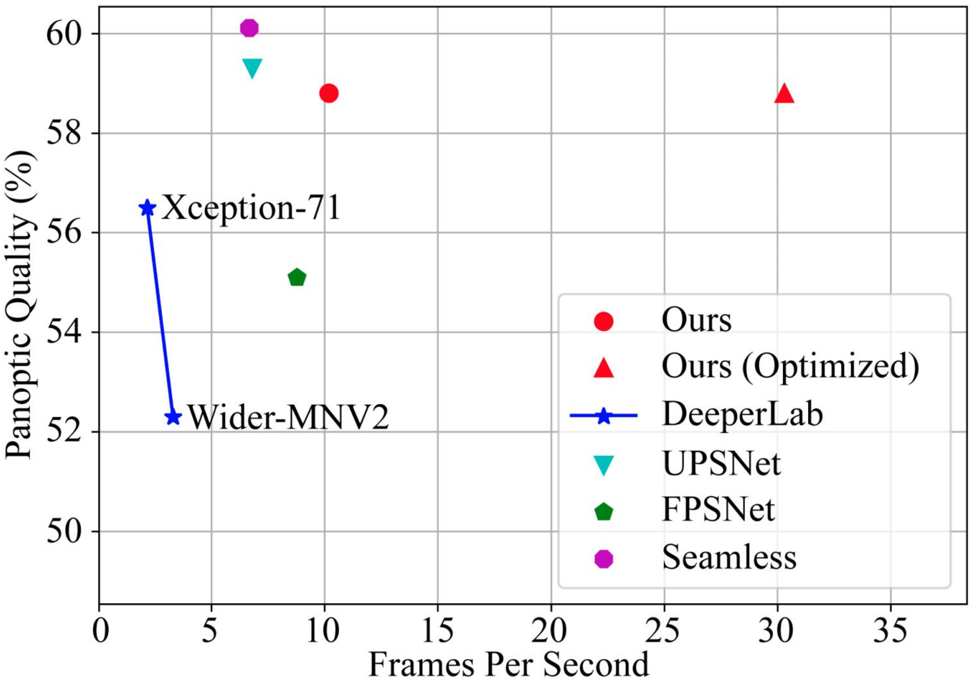
Real-time Panoptic Segmentation

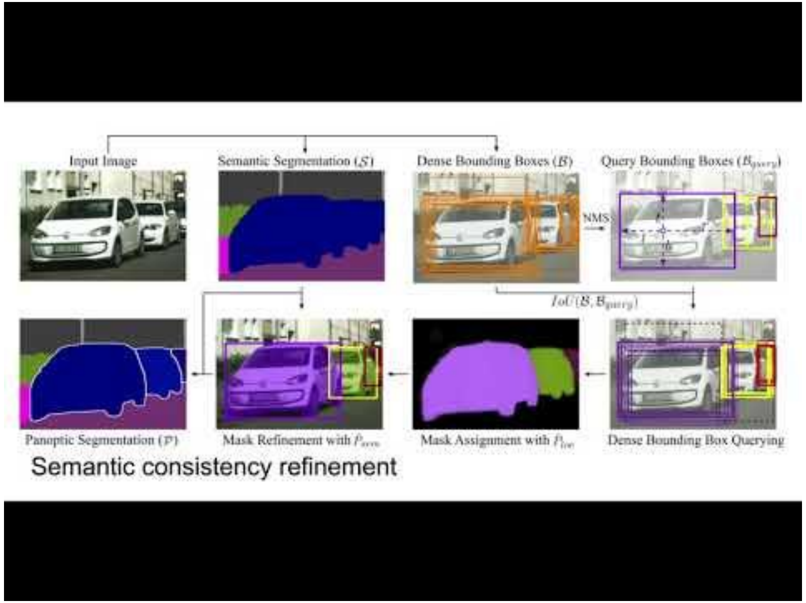
Real-Time Panoptic Segmentation from Dense Detections,
R. Hou, J. Li*, A. Bhargava, A. Raventos et al,* ***CVPR'20 (oral)***

Self-Supervised Pseudo-Lidar Networks

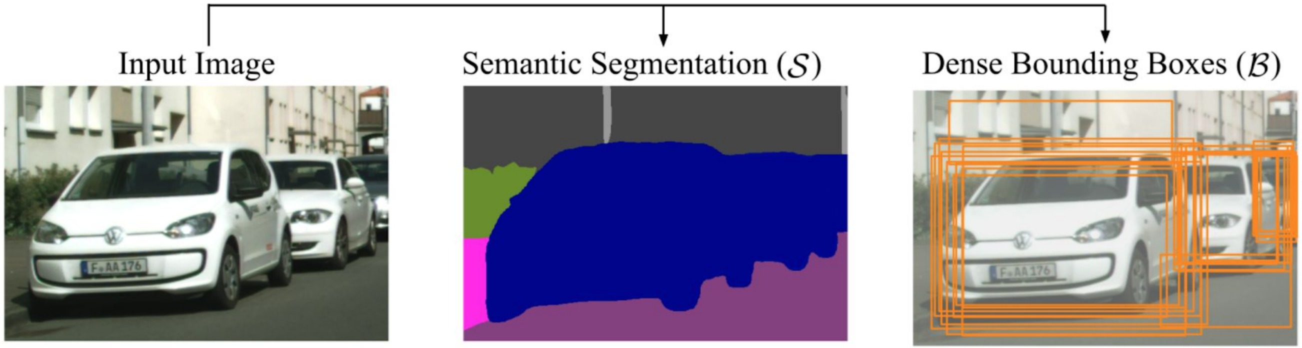
Auto-labeling via Differentiable Rendering

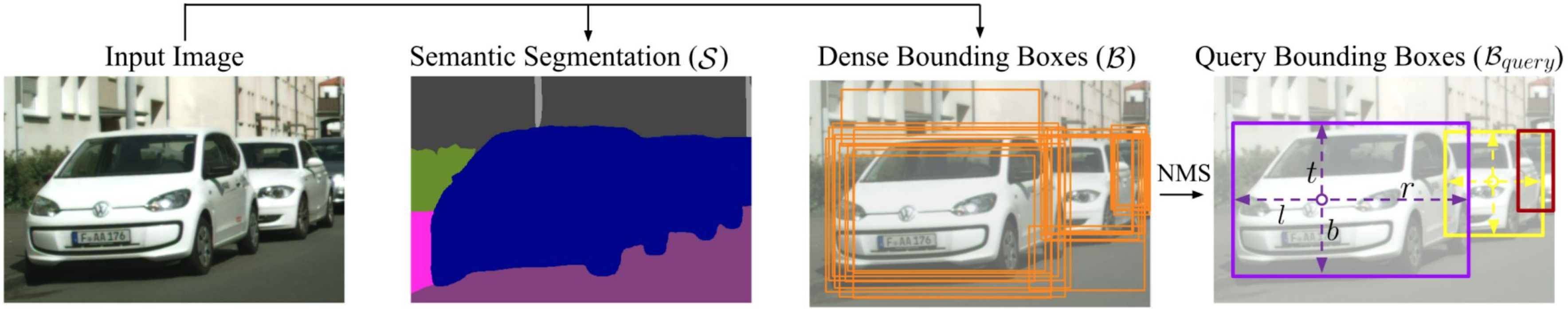
TOYOTA RESEARCH INSTITUTE



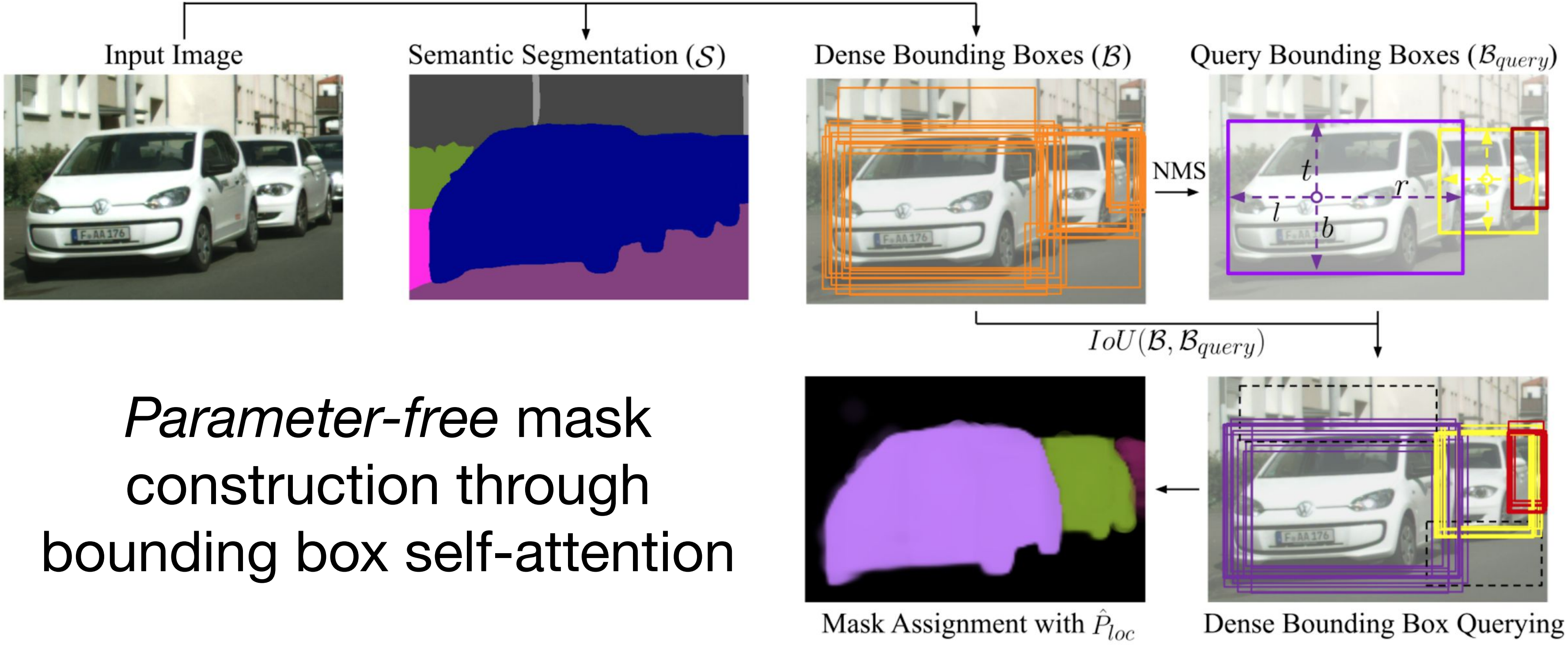


TOYOTA RESEARCH INSTITUTE

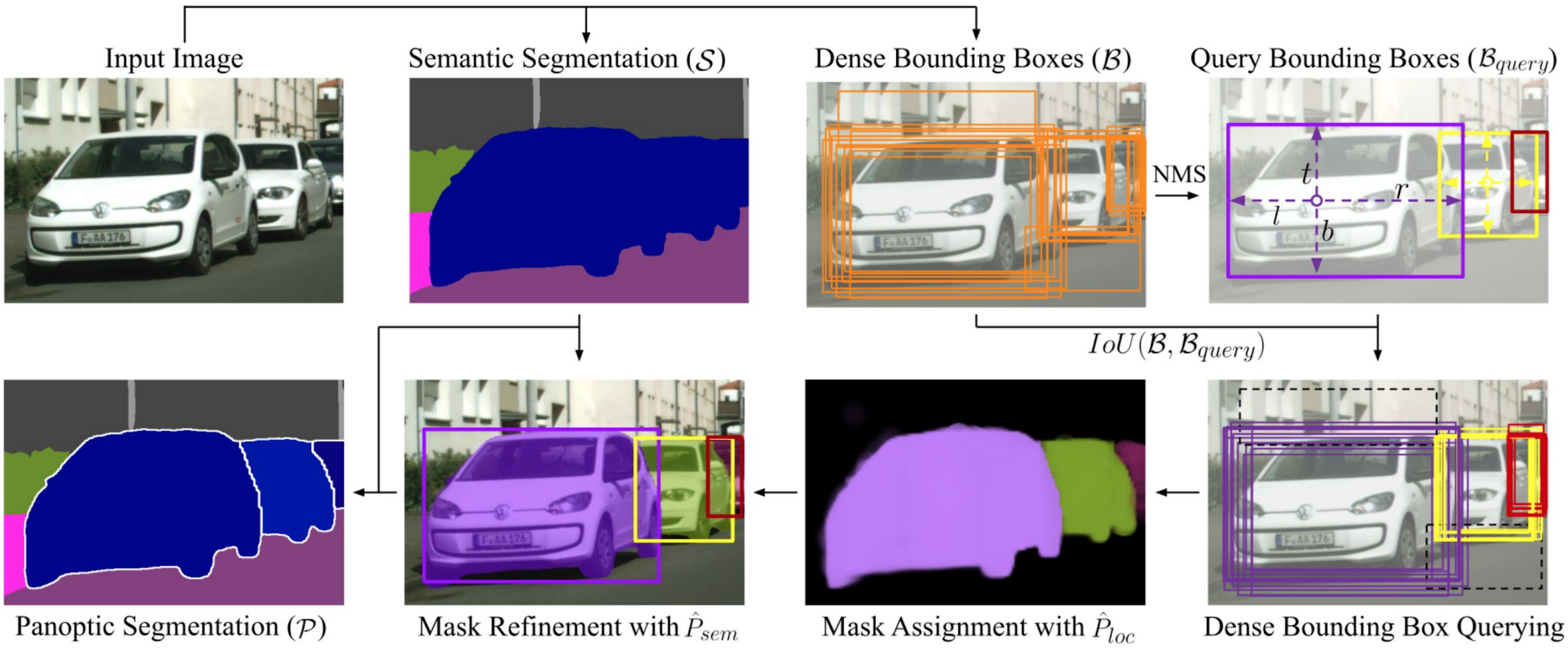

Input Image
Semantic Segmentation ($\mathcal{S}$)
Dense Bounding Boxes ($\mathcal{B}$)



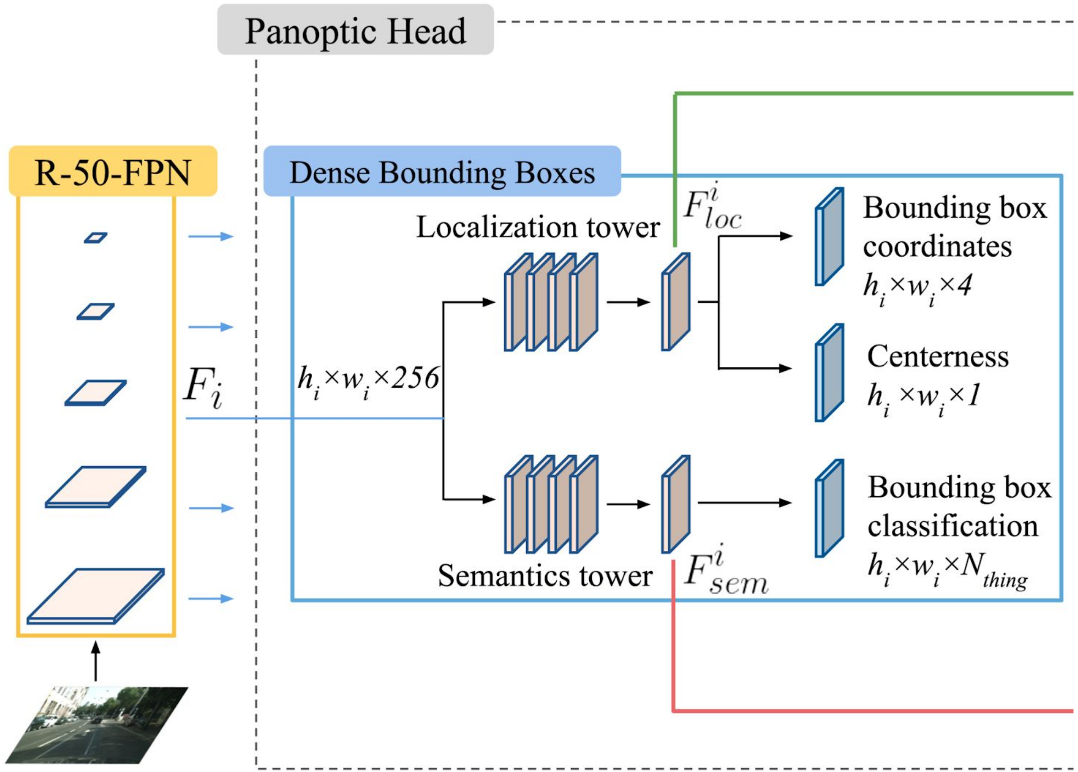
Input Image
Semantic Segmentation ($\mathcal{S}$)
Dense Bounding Boxes ($\mathcal{B}$)
Query Bounding Boxes ($\mathcal{B}_{query}$)
NMS
t
r
l
b



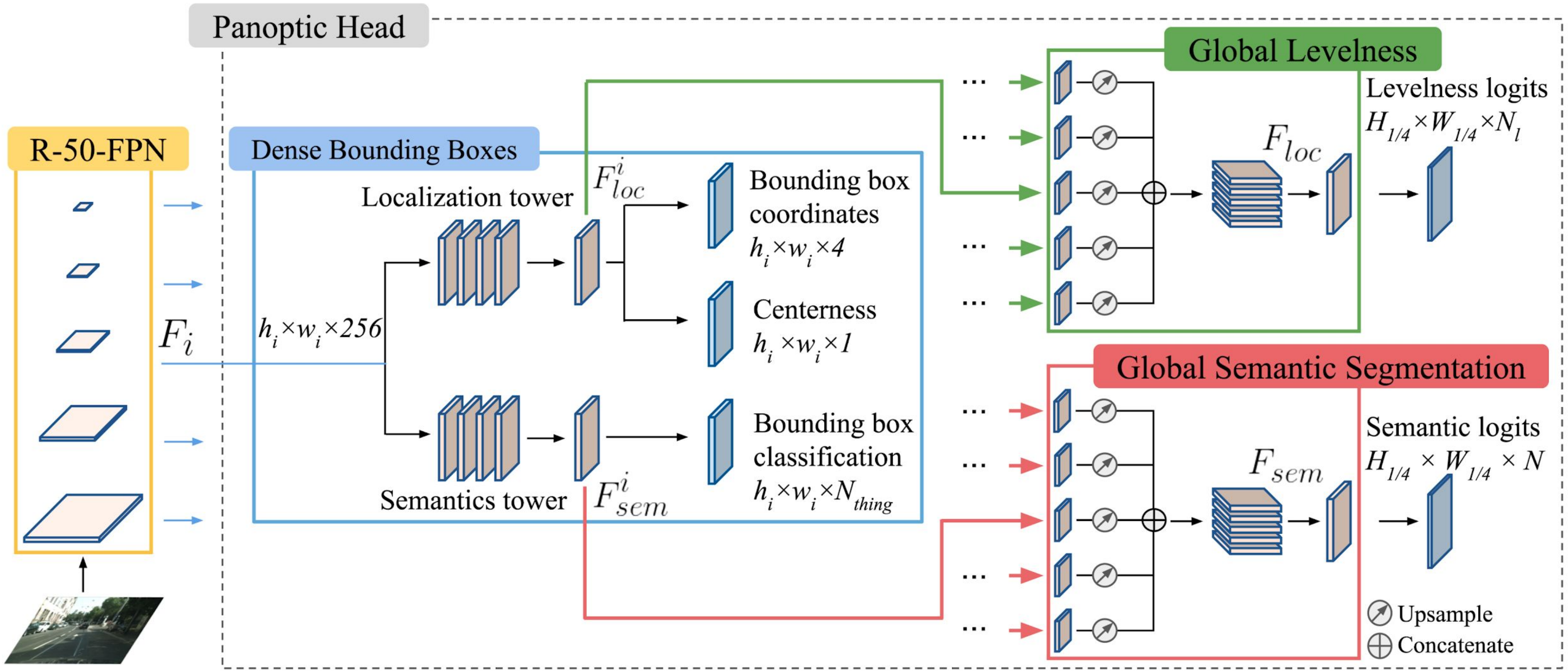
Parameter-free mask construction through bounding box self-attention



TOYOTA RESEARCH INSTITUTE



Panoptic Head
R-50-FPN
Dense Bounding Boxes
Localization tower
F^i_{loc}
Bounding box coordinates
$h_i \times w_i \times 4$
Centerness
$h_i \times w_i \times 1$
F_i
$h_i \times w_i \times 256$
Bounding box classification
$h_i \times w_i \times N_{thing}$
Semantics tower
F^i_{sem}



Panoptic Head
R-50-FPN
Dense Bounding Boxes
Global Levelness
Global Semantic Segmentation
Localization tower
Semantics tower
F_i
$h_i \times w_i \times 256$
F^i_{loc}
F^i_{sem}
Bounding box coordinates
$h_i \times w_i \times 4$
Centerness
$h_i \times w_i \times 1$
Bounding box classification
$h_i \times w_i \times N_{thing}$
F_{loc}
Levelness logits
$H_{1/4} \times W_{1/4} \times N_l$
F_{sem}
Semantic logits
$H_{1/4} \times W_{1/4} \times N$
Upsample
Concatenate

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	99ms

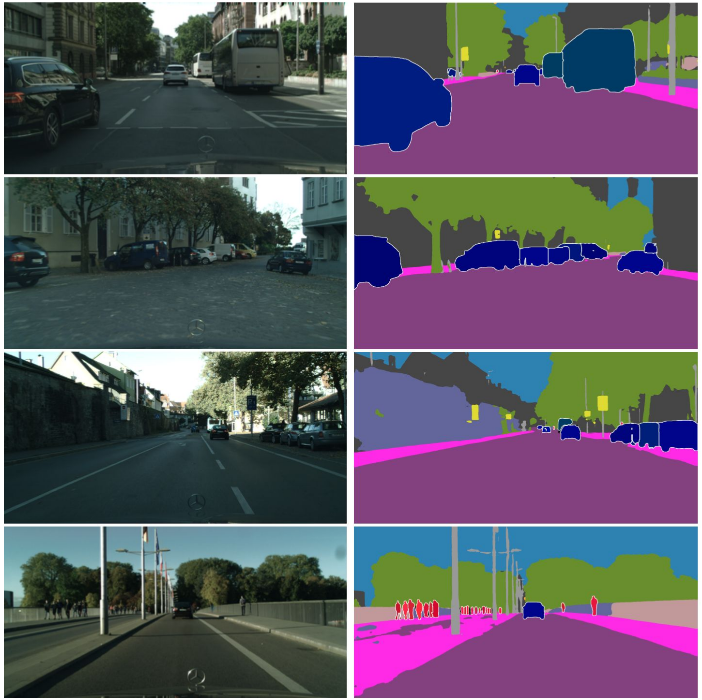
COCO (val)

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**	63ms

TOYOTA RESEARCH INSTITUTE

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



Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Real-time Panoptic Segmentation

Real-Time Panoptic Segmentation from Dense Detections,
R. Hou, J. Li*, A. Bhargava, A. Raventos et al,* ***CVPR'20 (oral)***

Self-Supervised Pseudo-Lidar Networks

Auto-labeling via Differentiable Rendering

TOYOTA RESEARCH INSTITUTE

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Real-time Panoptic Segmentation

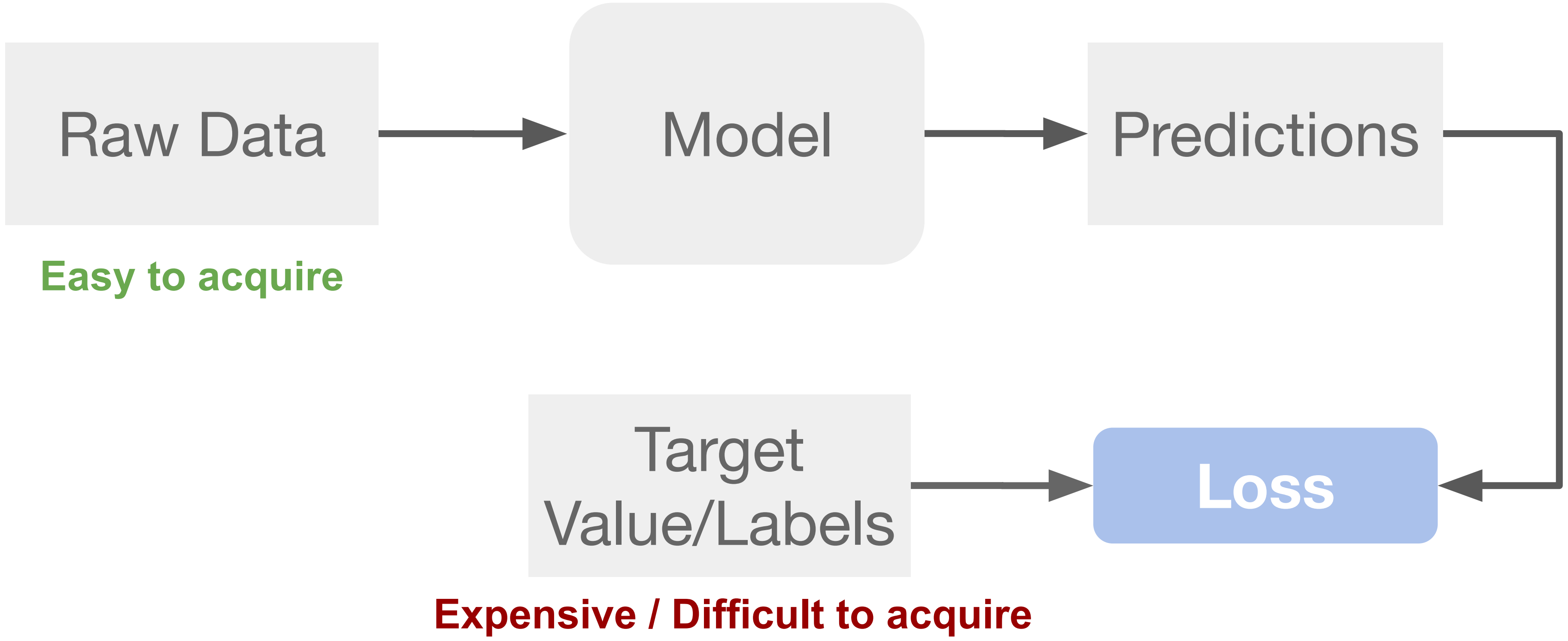
Self-Supervised Pseudo-Lidar Networks

3D Packing for Self-Supervised Monocular Depth Estimation
V. Guizilini, R. Ambrus, S. Pillai et al, ***CVPR'20 (oral)***

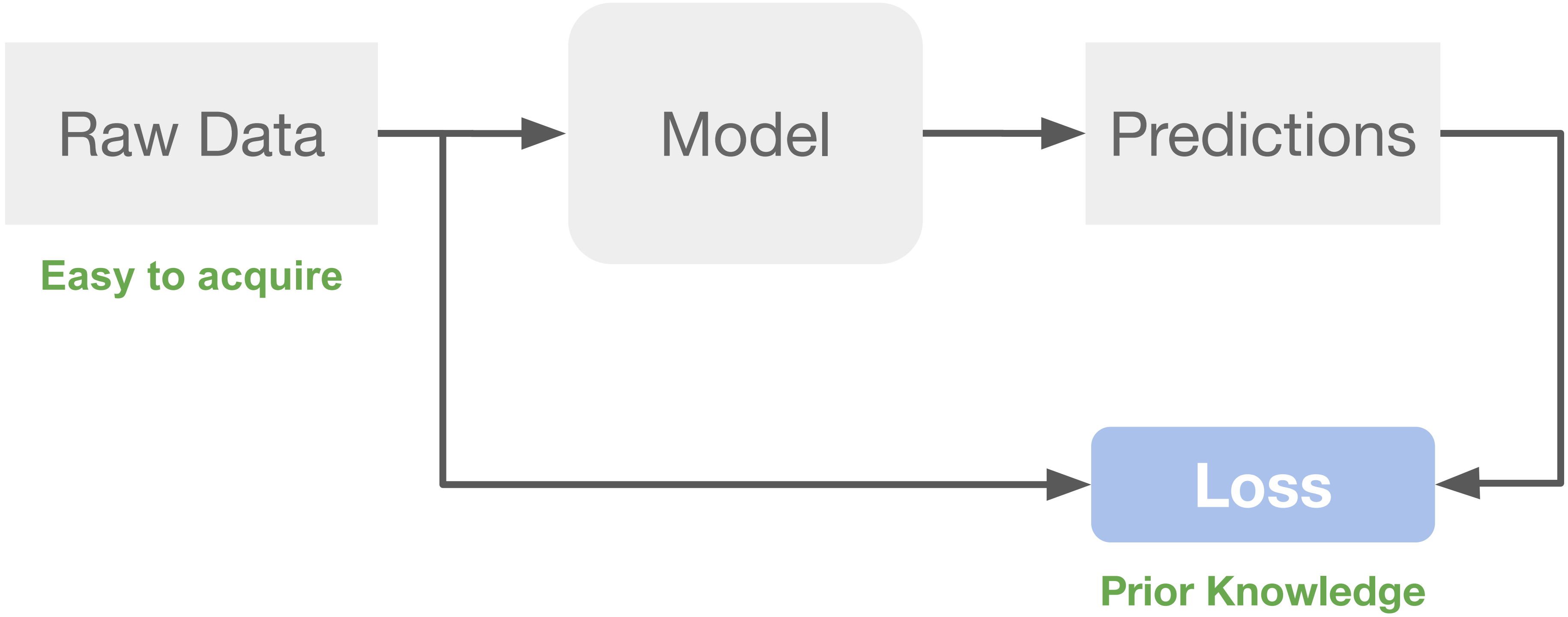
Auto-labeling via Differentiable Rendering

TOYOTA RESEARCH INSTITUTE

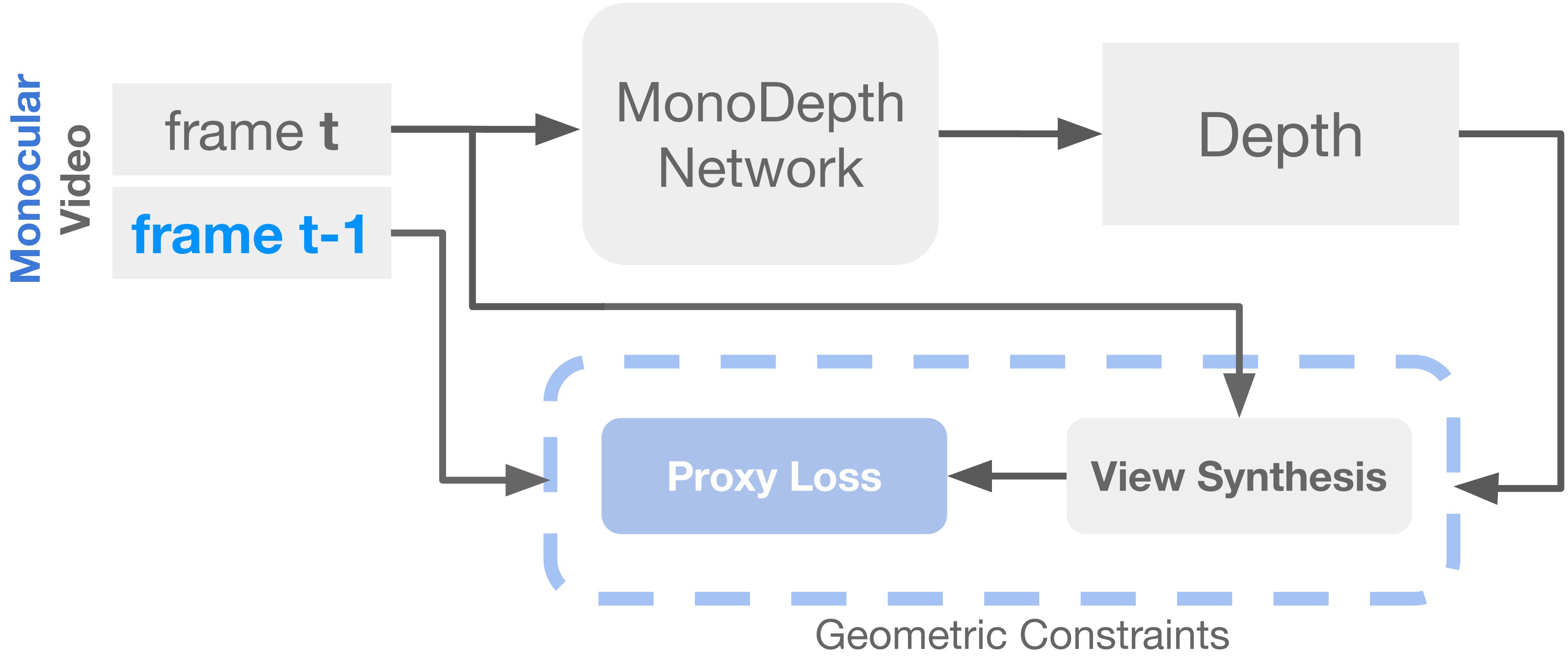
Supervised Learning



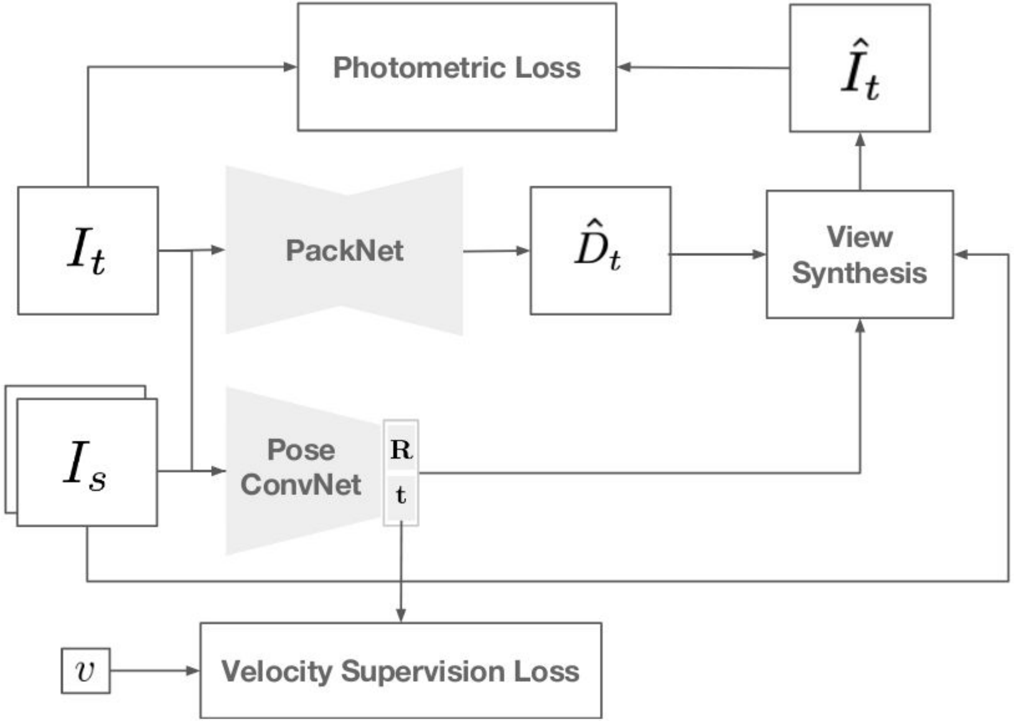
Self-Supervised Learning

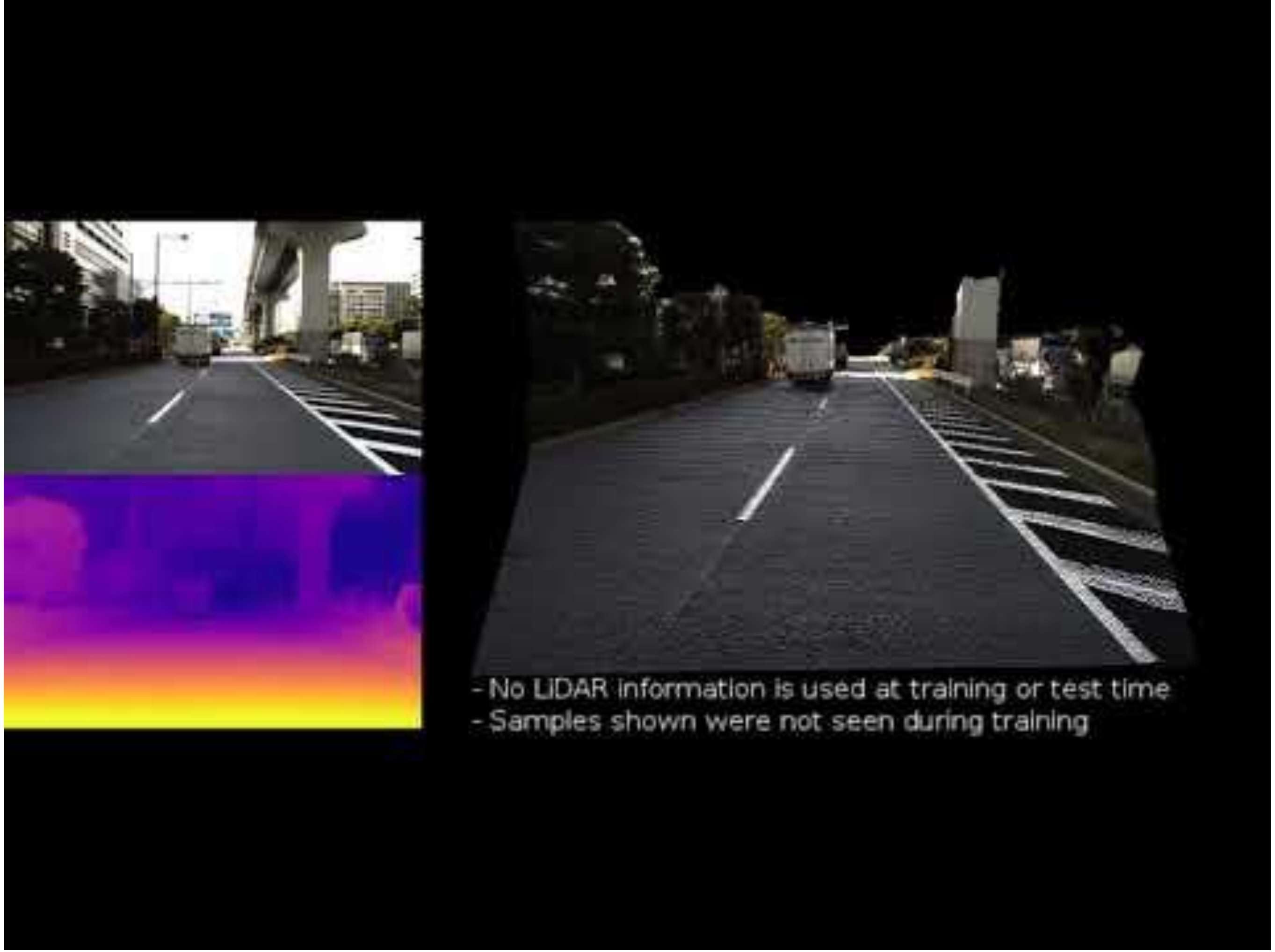


Self-Supervised Structure-from-Motion (SfM)



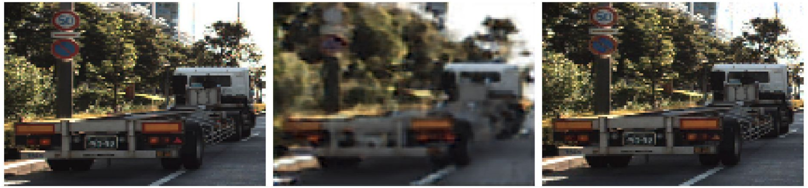
Self-Supervised Structure-from-Motion (SfM)



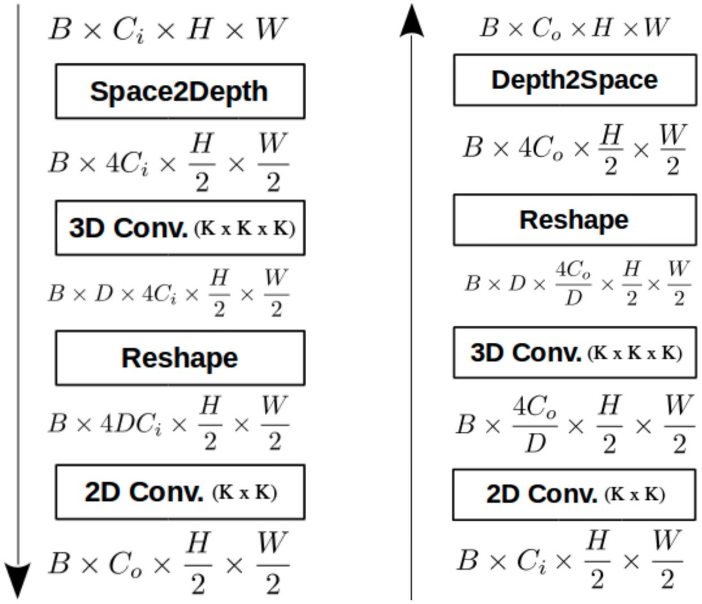

- No LiDAR information is used at training or test time
- Samples shown were not seen during training

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



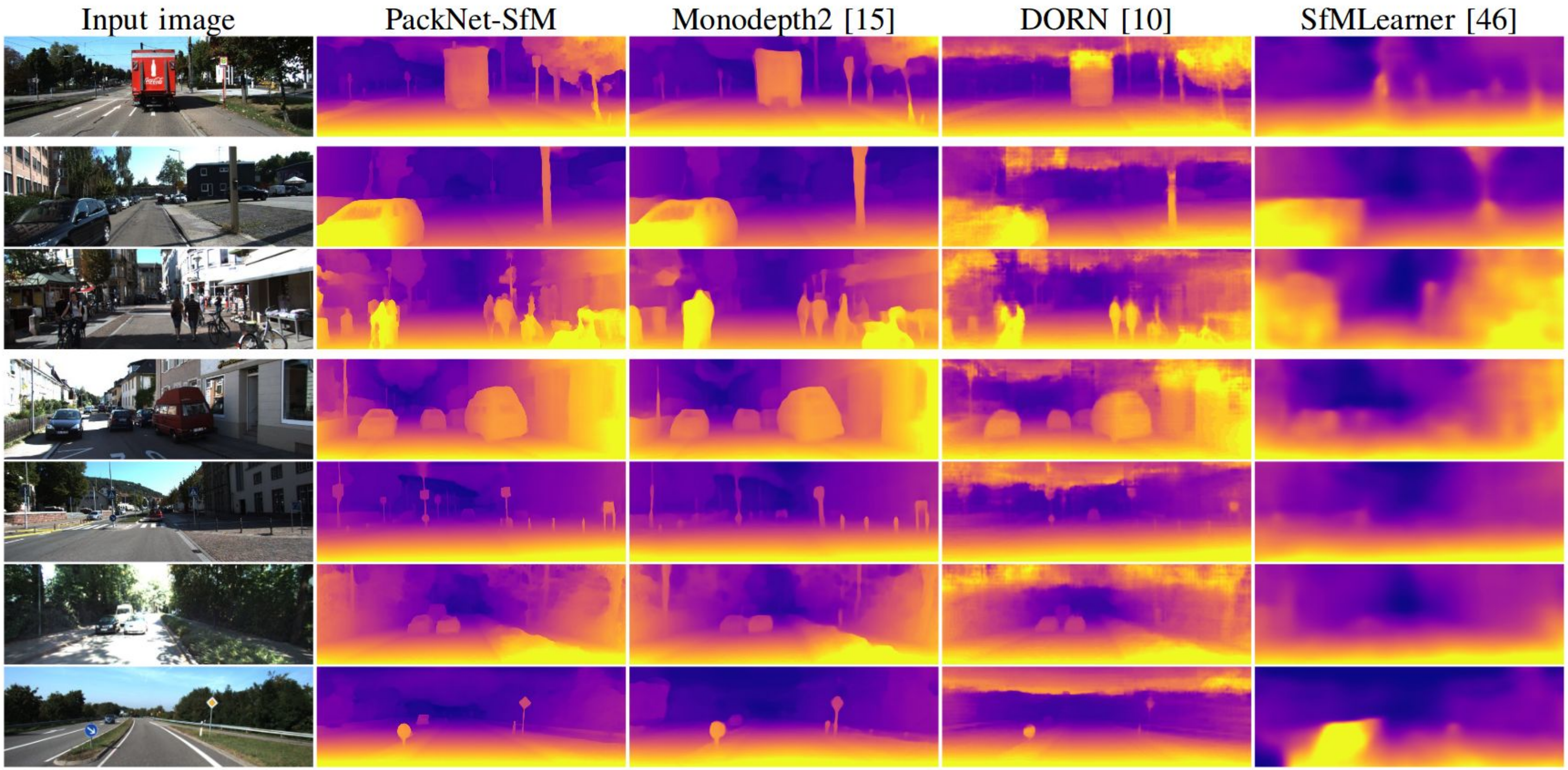
(a) Packing (b) Unpacking

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!

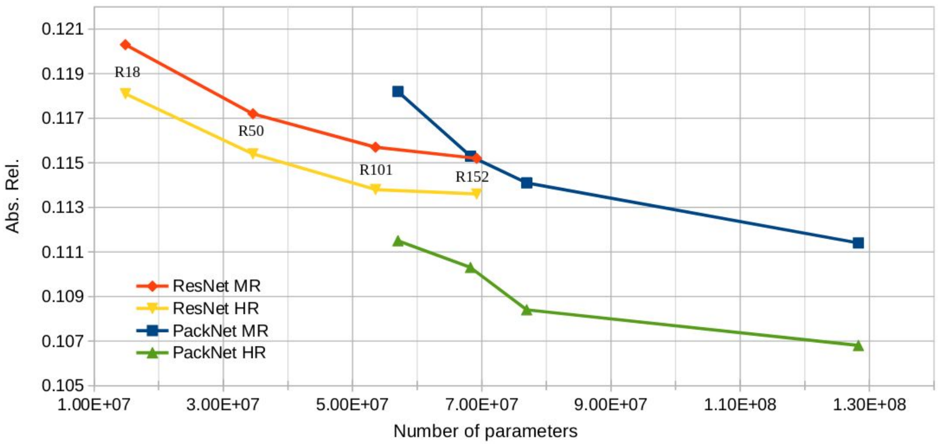
Experimental Results (KITTI)



TOYOTA RESEARCH INSTITUTE

Experimental Results

Better use of network capacity...



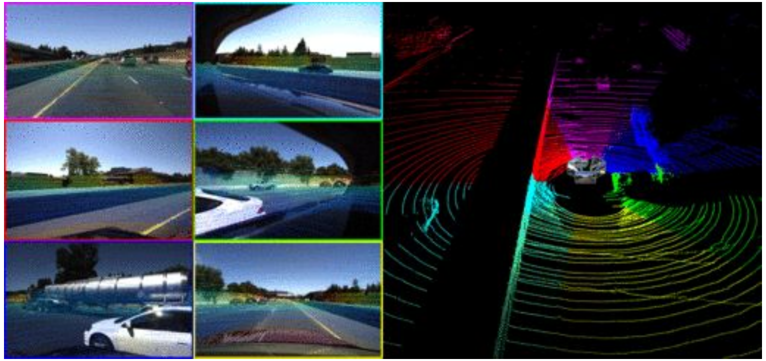
Depth Network	Abs Rel	Sq Rel	RMSE	RMSE_{log}	$\delta < 1.25$
ResNet18	0.133	1.023	5.123	0.211	0.845
ResNet18‡	0.120	0.896	4.869	0.198	0.868
ResNet50	0.127	0.977	5.023	0.205	0.856
ResNet50‡	0.117	0.900	4.826	0.196	0.873
PackNet18	0.118	0.802	4.656	0.194	0.868
PackNet50	0.114	0.818	4.621	0.190	0.875
PackNet-SfM (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**

TOYOTA RESEARCH INSTITUTE

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

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Real-time Panoptic Segmentation

Self-Supervised Pseudo-Lidar Networks

3D Packing for Self-Supervised Monocular Depth Estimation
V. Guizilini, R. Ambrus, S. Pillai et al, ***CVPR'20 (oral)***

Auto-labeling via Differentiable Rendering

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Real-time Panoptic Segmentation

Self-Supervised Pseudo-Lidar Networks

Auto-labeling via Differentiable Rendering

Autolabeling 3D Objects with Differentiable Rendering of SDF Shape Priors, S. Zakharov, W. Kehl et al,* ***CVPR'20 (oral)***

TOYOTA RESEARCH INSTITUTE

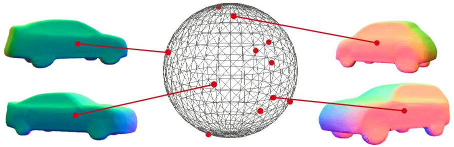
Auto-labeling in 3D

Input: image, point cloud, 2d bounding boxes

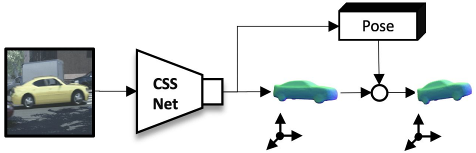
Output: 3d boxes with pose + shape

Goal: use auto-labels instead of manual ones

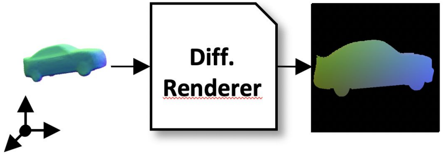
Shape Representation

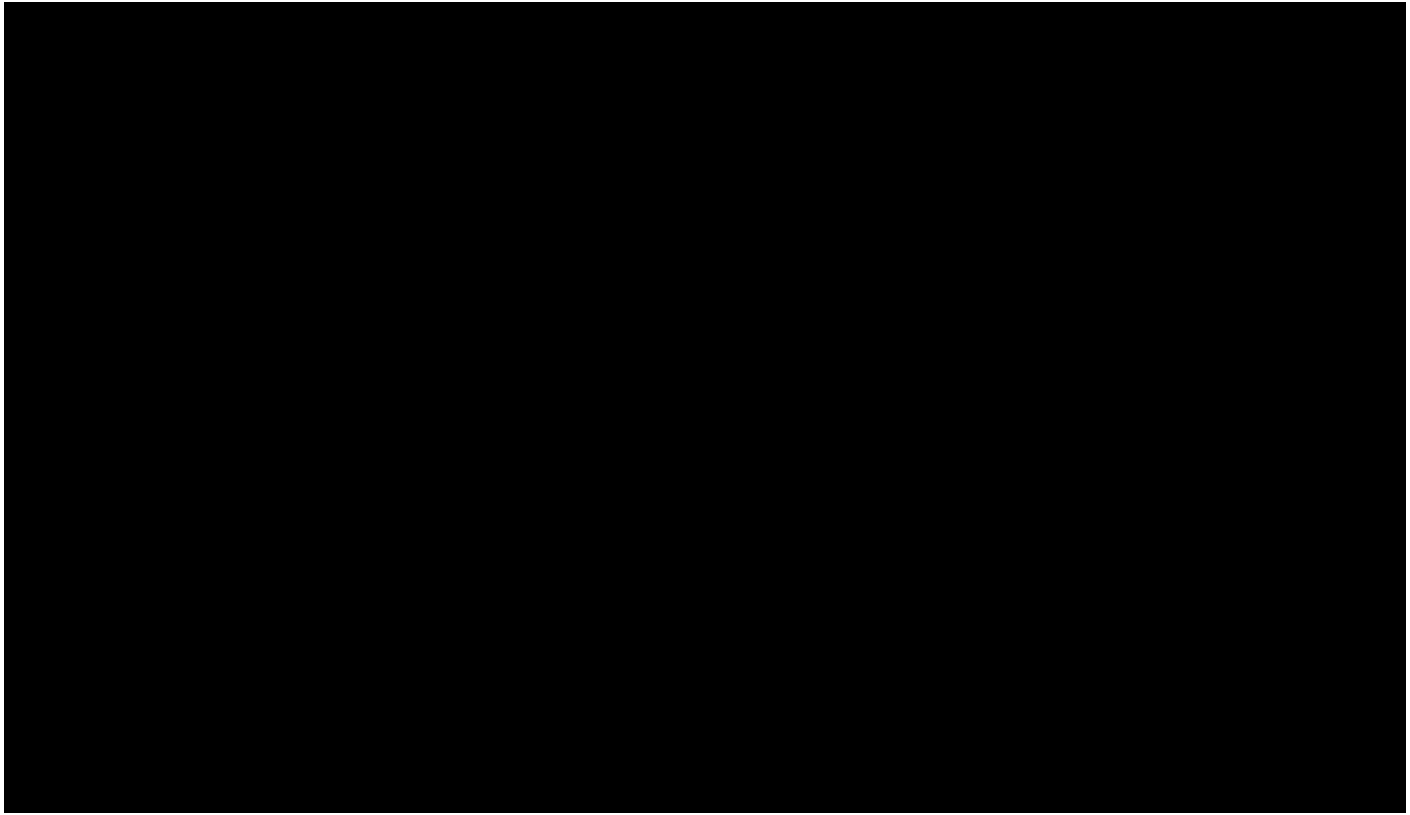


Pose/Shape Estimator



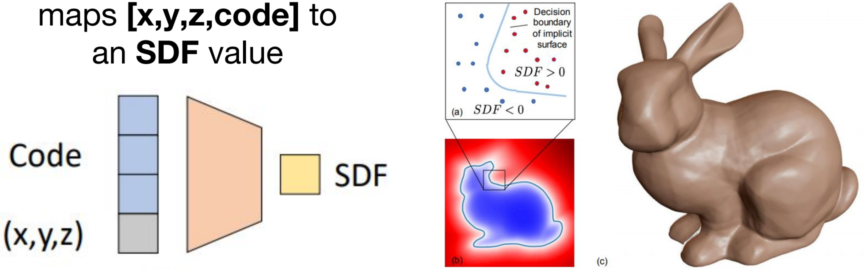
Differentiable Renderer



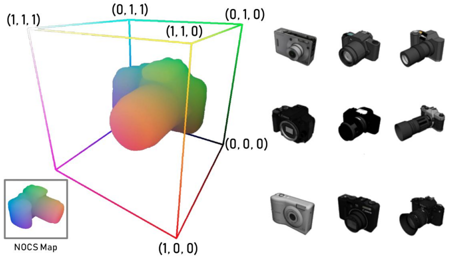


Coordinate Shape Space (CSS): DeepSDF + NOCS

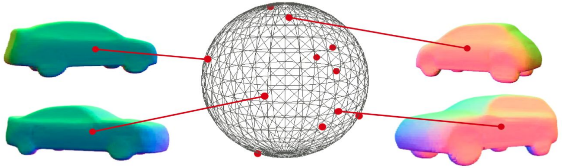
DeepSDF network maps **[x,y,z,code]** to an **SDF** value



Normalized Object Coordinate Space (NOCS)



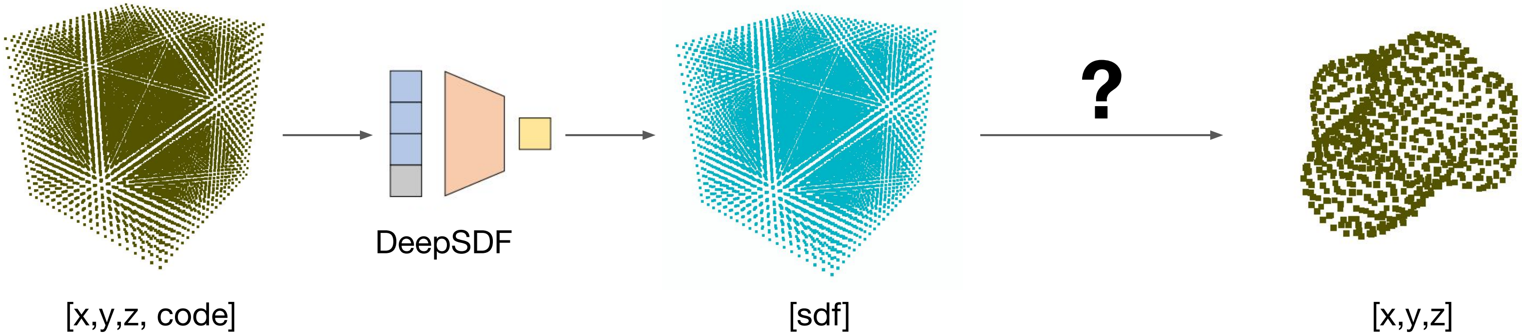
3D models represented in the shape space



3D latent code represents a unique car shape

TOYOTA RESEARCH INSTITUTE

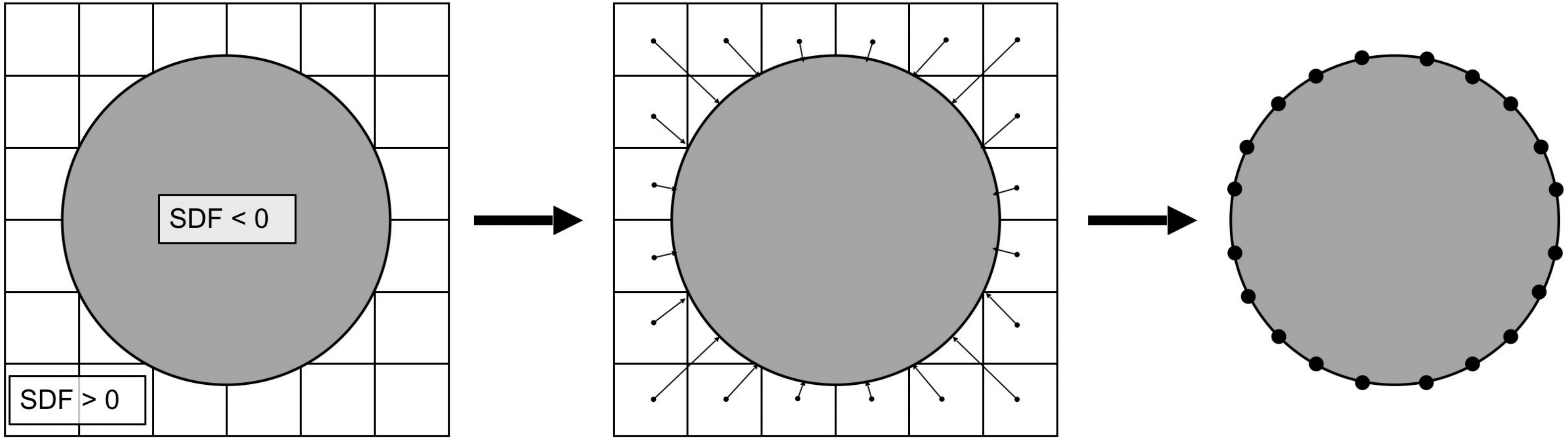
From SDF to Coordinates?



How to render SDF?

- Marching Cubes (differentiable?)
- Raycasting (slow)

Zero-Isosurface Projection



$$\mathbf{s} = f(\boldsymbol{grid})$$

$$\boldsymbol{n} = \frac{\partial f(\boldsymbol{grid})}{\partial \boldsymbol{grid}}$$

$$\boldsymbol{grid} - \frac{\partial f(\boldsymbol{grid})}{\partial \boldsymbol{grid}} f(\boldsymbol{grid})$$

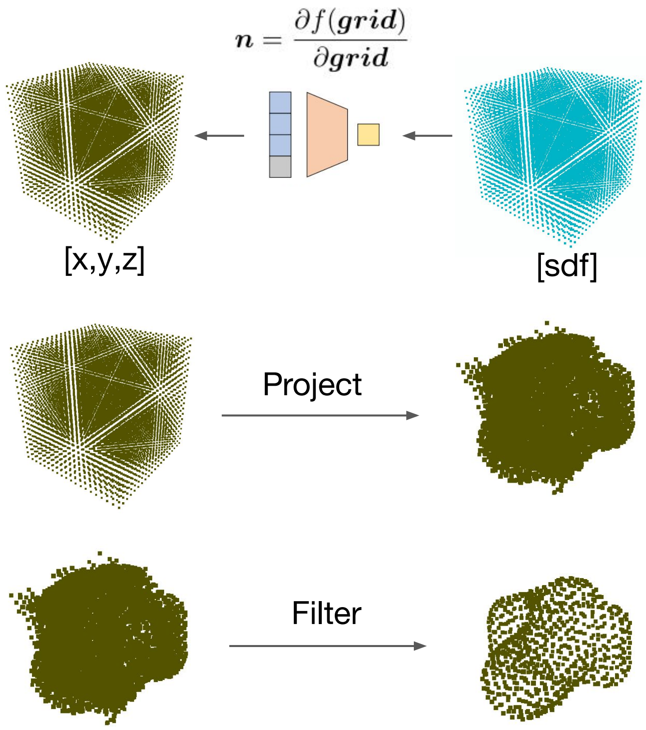
Zero-Isosurface Projection

1. Project the grid points to the surface using the *SDF values* and the *analytically estimated normals*:

$$p = grid - \frac{\partial f(grid)}{\partial grid} f(grid)$$

2. Mask the points that are far from the surface:

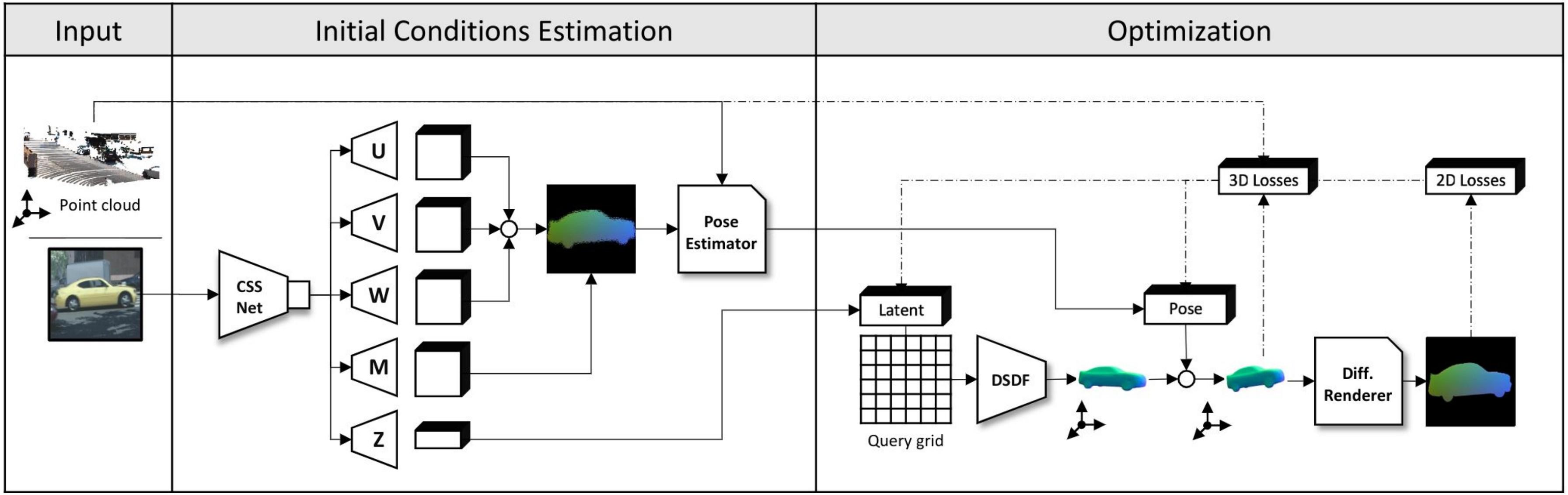
$$p_{masked} = p, \text{where} |f(grid)| < 0.1$$



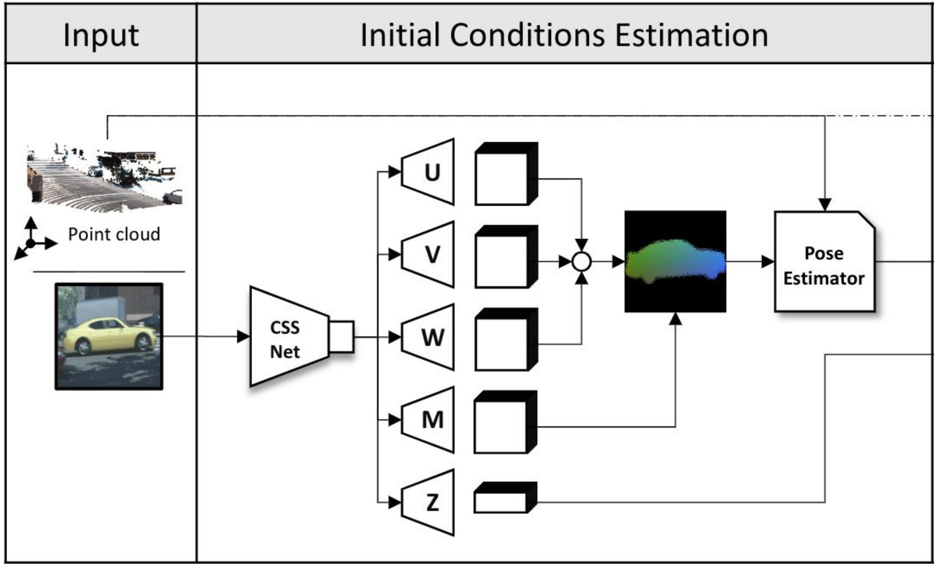
Differentiable all the way back to the latent vector!

TOYOTA RESEARCH INSTITUTE

Auto-labeling Pipeline



Auto-labeling Pipeline



CSS Network (R18-based):
U, V, W: Normalized Object Coordinates (NOCS) for each pixel
M: Object mask
z: Latent shape vector

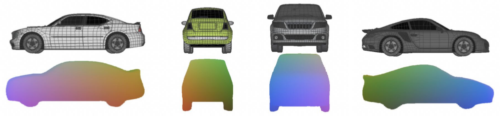
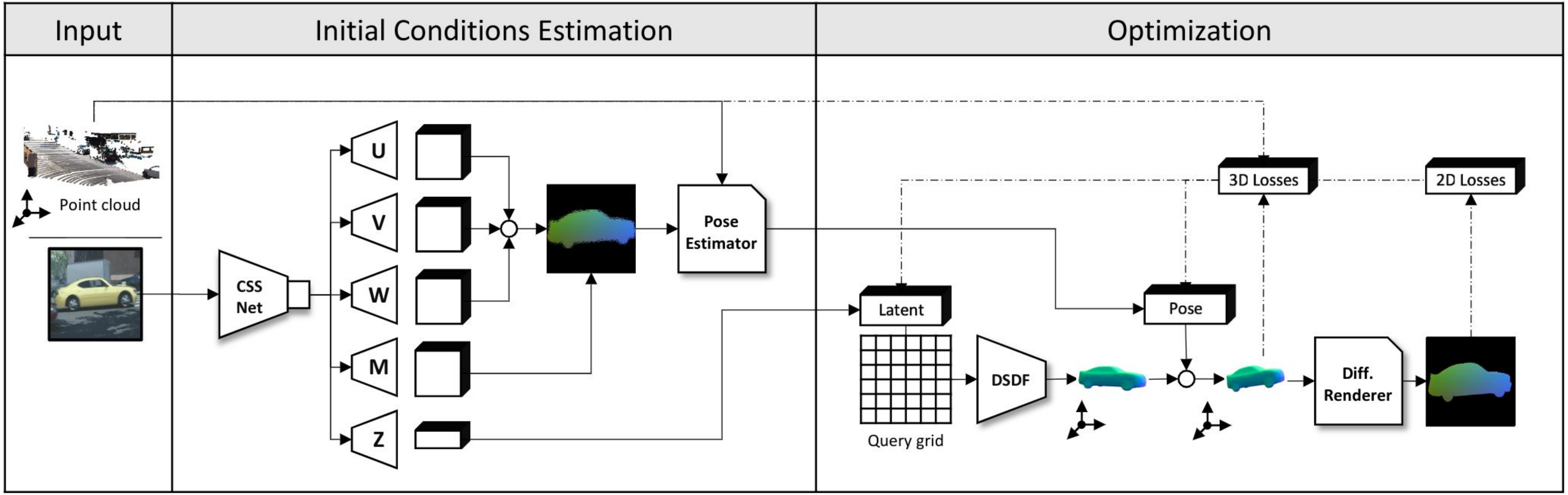
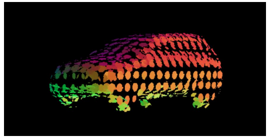


Figure 2: CSS representation. Top: Car models from the PD dataset [1]. Bottom: The same cars in the CSS representation: decoded shape vector **z** colored with NOCS.

Pose estimator uses either:
PnP (2D-based)
Kabsch / Procrustes (3D-based)

Auto-labeling Pipeline





Training: Domain Adaptation

Parallel Domain (Training data)

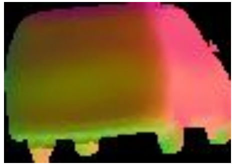
KITTI (Testing data)

Training: Domain Randomization

- **2D:**
  - Transforms:
    - Random rotation
    - Random horizontal flip
    - Random resized crop
  - Noise:
    - Color jitter: brightness, contrast, saturation, hue

- **3D:**
  - Phong Lighting:
    - Random ambient, diffuse and specular lights from pre-computed normals

RGB

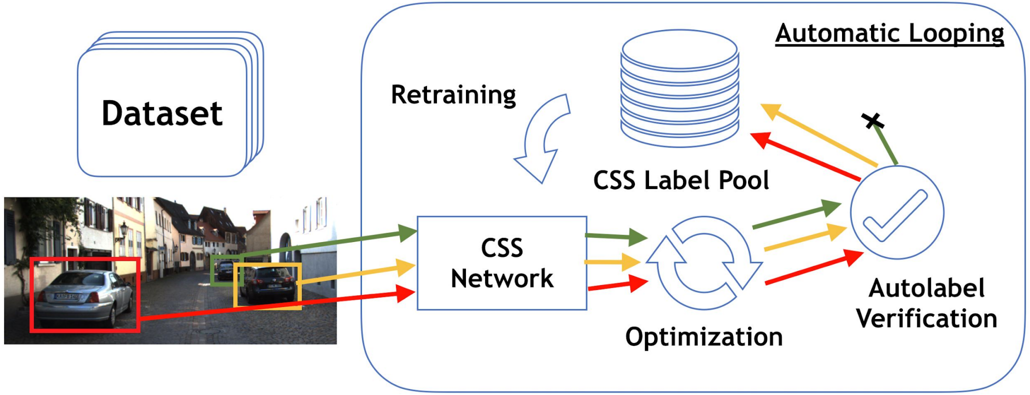


Normals

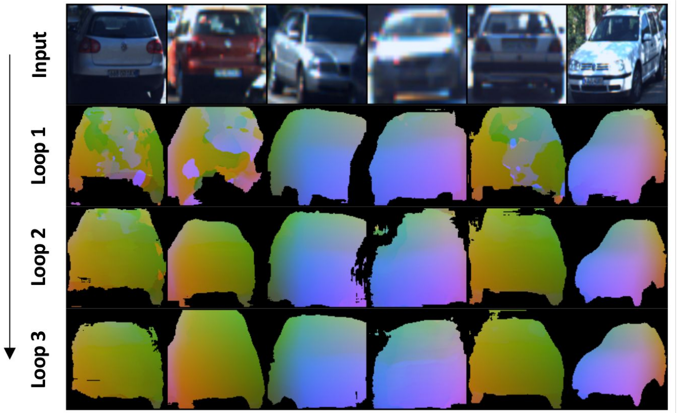
Color jitter

+Random transforms

Curriculum Learning

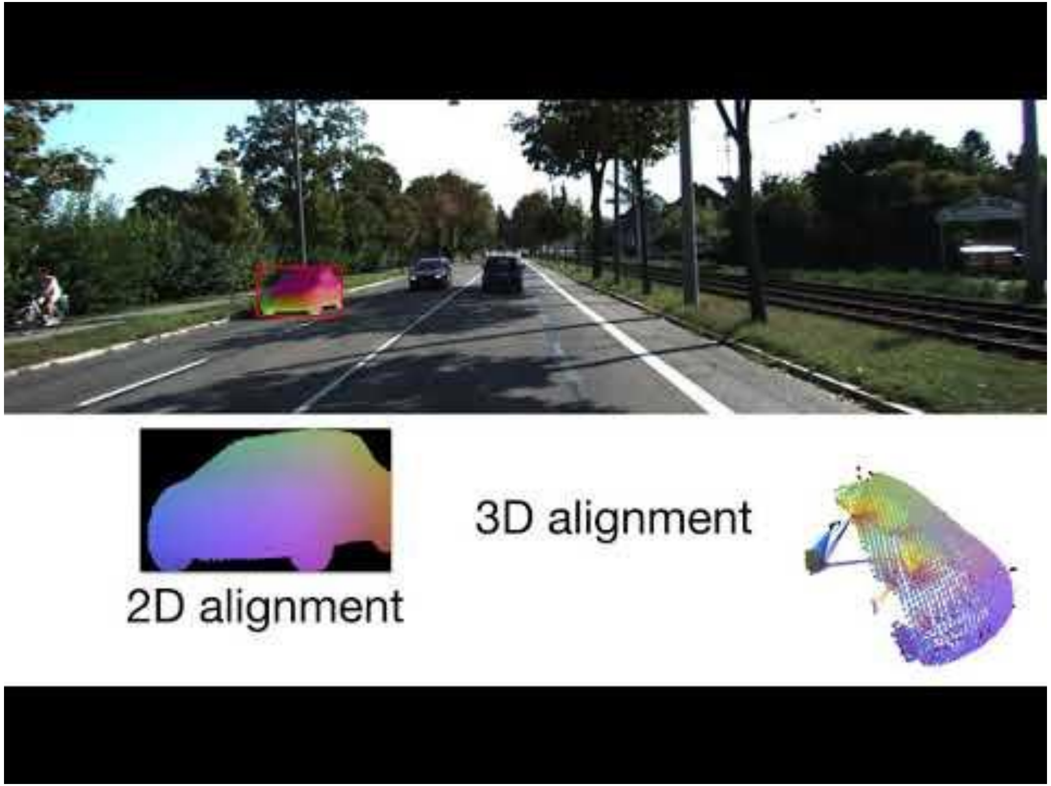


Automatic annotation pipeline



CSS Network prediction quality of our network over consecutive loops for the same patch

Qualitative Results

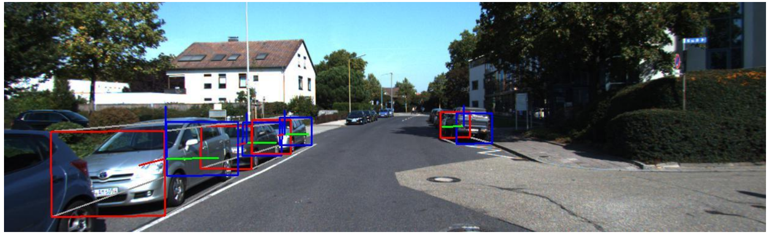


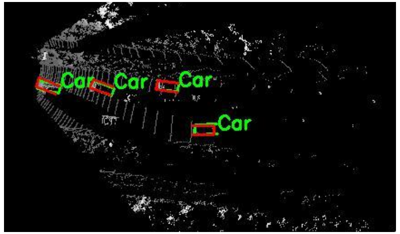
TOYOTA RESEARCH INSTITUTE

Quantitative Results

Method	2D AP @ 0.5/0.7 Easy	2D AP @ 0.5/0.7 Moderate	3D AP @ 0.5/0.7 Easy	3D AP @ 0.5/0.7 Moderate	BEV AP @ 0.5/0.7 Easy	BEV AP @ 0.5/0.7 Moderate
PointPillars [20] (Original Labels)	- / -	- / -	94.8 / 81.1	92.4 / 68.2	95.1 / 92.1	95.1 / 84.7
PointPillars [20] (Autolabels)	- / -	- / -	90.7 / 22.4	71.1 / 13.3	94.9 / 81.0	88.5 / 59.8
MonoDIS [35] (Original Labels)	96.1 / 95.5	92.6 / 86.5	45.7 / 11.0	32.9 / 7.1	52.4 / 17.7	37.2 / 11.9
MonoDIS [35] (Autolabels)	96.7 / 85.8	86.2 / 67.6	32.9 / 1.23	22.1 / 0.54	51.1 / 15.7	34.5 / 10.52

Table 2: We compare the performance of 3D object detectors trained on true KITTI labels vs. our autolabels. On the BEV metric, **the detectors trained on autolabels alone achieve results equal to the current state-of-the-art**. On the 3D AP metric, both autolabel trained detectors achieve competitive results at the IoU 0.5 threshold.





Scaling up ML for Autonomy

Behavior Cloning and its Limitations

Real-time Panoptic Segmentation

Self-Supervised Pseudo-Lidar Networks

Auto-labeling via Differentiable Rendering

Autolabeling 3D Objects with Differentiable Rendering of SDF Shape Priors, S. Zakharov, W. Kehl et al,* ***CVPR'20 (oral)***

TOYOTA RESEARCH INSTITUTE

World-scale Autonomy?

THE SWE	THE MLE	TRI	THE SCIENTIST
PROGRAM	LABEL	LEARN	LEARN
EVERYTHING	EVERYTHING	FROM EVERYONE	EVERYTHING
MAPS ?	LABELS ?	STRUCTURE ?	SIM ?

TOYOTA RESEARCH INSTITUTE

Behavior: leverage large scale Demonstrations

***Exploring the Limitations of Behavior Cloning for Autonomous Driving*, ICCV'19 (oral)**
Spatiotemporal Relationship Reasoning for Pedestrian Intent Prediction, **RA-L & ICRA'20**
It Is Not the Journey but the Destination: Endpoint Conditioned Trajectory Prediction, arXiv:2004.02025
Reinforcement Learning based Control of Imitative Policies for Near-Accident Driving, coming soon
Risk-Sensitive Sequential Action Control with Multi-Modal Human Trajectory Forecasting [...], coming soon
Driving Through Ghosts: Behavioral Cloning with False Positives, coming soon

Supervised Learning: efficiently use available Labels

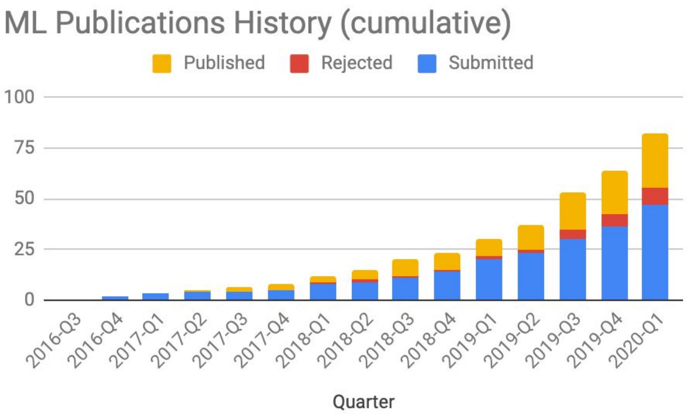
ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape, **CVPR'19**
Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss, **NeurIPS'19**
Learning to Fuse Things and Stuff, arXiv:1812.01192
Spatio-Temporal Graph for Video Captioning with Knowledge Distillation, **CVPR'20**
***Real-Time Panoptic Segmentation from Dense Detections*, CVPR'20 (oral)**
Hierarchical Lovász Embeddings for Proposal-free Panoptic Segmentation, coming soon
Unsupervised Estimation of Segmentation Difficulty, coming soon

Geometry: Self / Semi-Supervised Pseudo-LiDAR and SfM

SuperDepth: Self-Supervised, Super-Resolved Monocular Depth Estimation, **ICRA'19**
Robust Semi-Supervised Monocular Depth Estimation with Reprojected Distances, **CoRL'19**
Two Stream Networks for Self-Supervised Ego-Motion Estimation, **CoRL'19**
Semantically-Guided Representation Learning for Self-Supervised Monocular Depth, **ICLR'20**
Neural Outlier Rejection for Self-Supervised Keypoint Learning, **ICLR'20**
Self-Supervised 3D Keypoint Learning for Ego-motion Estimation, arxiv:1912.03426
***3D Packing for Self-Supervised Monocular Depth Estimation*, CVPR'20 (oral)**
Self-Supervised Neural Camera Models, coming soon

Simulation: Domain Adaptation, Differentiable Rendering, RL

SPIGAN: Privileged Adversarial Learning from Simulation, **ICLR'19**
DeceptionNet: Network-Driven Domain Randomization, **ICCV'19**
Generating Human Action Videos by Coupling 3D Game Engines and Probabilistic Graphical Models, **IJCV'20**
***Autolabeling 3D Objects with Differentiable Rendering of SDF Shape Priors*, CVPR'20 (oral)**
Self-Supervised Differentiable Rendering for Monocular 3D Object Detection, coming soon
Behaviorally Diverse Traffic Simulation via Reinforcement Learning, coming soon
Discovering Avoidable Planner Failures [...] in Behaviorally Diverse Simulation, coming soon



Upcoming workshops co-organized by TRI

ICML: AI for Autonomous Driving (AIAD)
https://sites.google.com/view/aiad2020
ECCV: Perception for Autonomous Driving (PAD)
https://sites.google.com/view/pad2020

Upcoming TRI Dataset Releases

STIP: Stanford-TRI Intent Prediction
http://stip.stanford.edu/
DDAD: Dense Depth for Autonomous Driving
https://github.com/TRI-ML/DDAD

Scaling up ML for Autonomy

Behavior Cloning and its Limitations

More data & params → SotA policy but...

Real-time Panoptic from Bounding Boxes

SotA, 4x faster, weak sup. = 95% strong

Self-Supervised Pseudo-Lidar Networks

Self sup > sup! Don't pool it: Pack it.

Auto-labeling via Differentiable Rendering

Diff. shape priors + geometry ~ labels

TOYOTA RESEARCH INSTITUTE