Beyond Supervised Driving

Adrien Gaidon (twitter: @adnothing)
Machine Learning Lead

Sudeep Pillai (twitter: @sudeeppillai)
Research Scientist

Toyota Research Institute (TRI), CA, USA

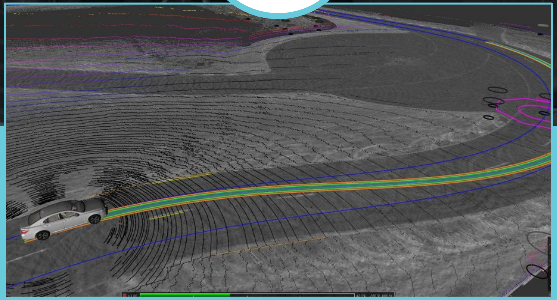

TOYOTA GUARDIAN

TRI Aims to Transform the Human Condition

Safety

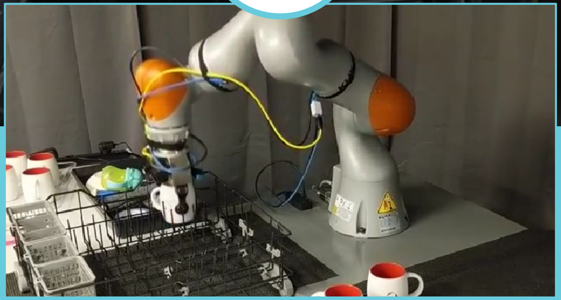
Guardian

Access



Chauffeur

Quality of Life



Robots

Agenda

- **Why Beyond Supervised Driving**
- Self-Supervised Learning: **SuperDepth**
- Sim2Real adaptation: **SPIGAN**

Credit: https://xkcd.com/1897/

"Crowdsourced steering doesn't sound quite as appealing as self driving"

ROI-10D: Monocular Lifting of
2D Detection to 6D Pose and Metric Shape
F Manhardt, W Kehl, A Gaidon
https://arxiv.org/abs/1812.02781

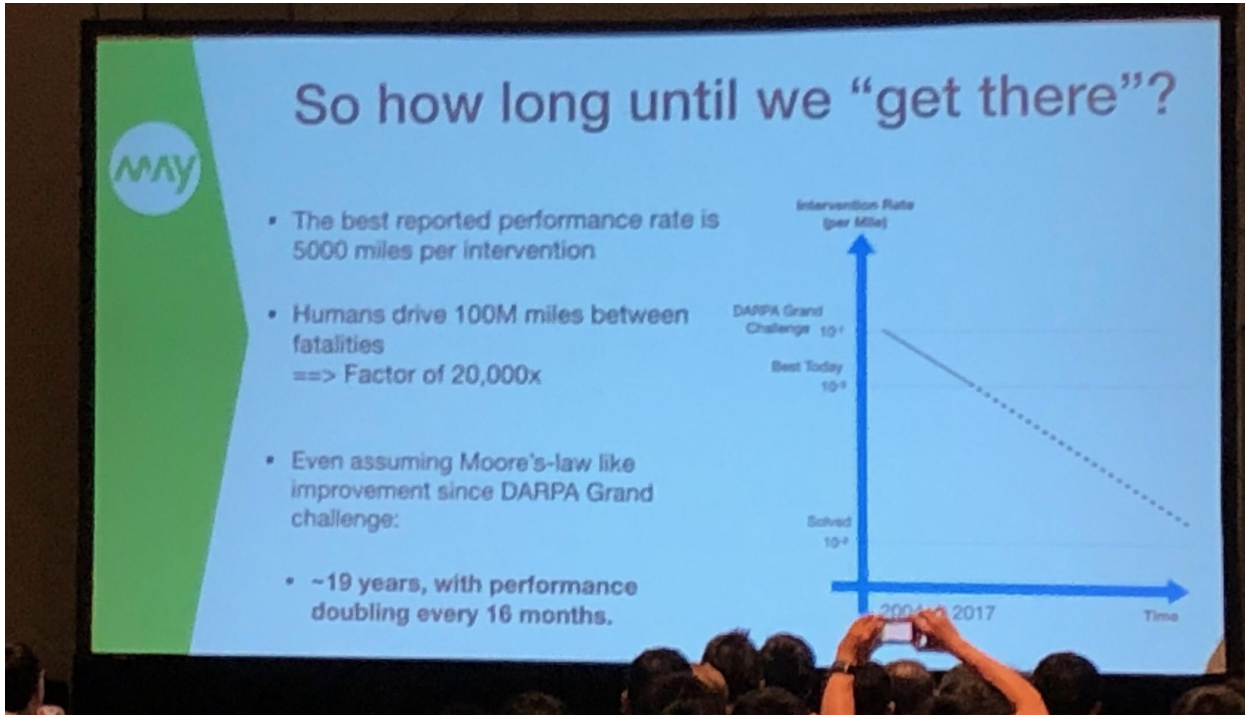
Learning to Fuse Things and Stuff

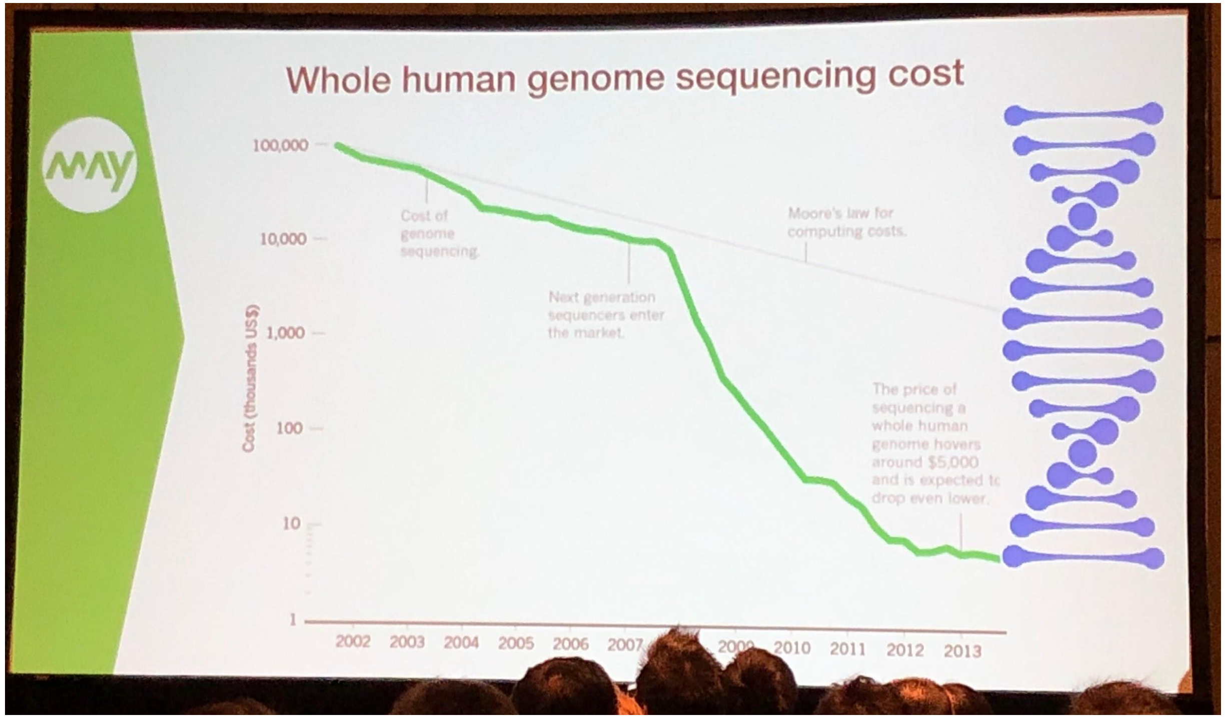
J Li, A Raventos, A Bhargava, T Tagawa, A Gaidon
https://arxiv.org/abs/1812.01192

But, but, … supervised learning Just Works™ !

TOYOTA RESEARCH INSTITUTE

Credit: Ed Olson, May Mobility





Exponential progress with current supervision is not enough.

Why Beyond Supervised Driving?

TOYOTA

> 22PB/day*
(100M cars, 95% parked)

> 10x

YouTube

> 2.5 PB/day*
(400 hours/min HD)

How to learn from all that structured but unlabeled data?

Supervised + Self-Supervised = Win!

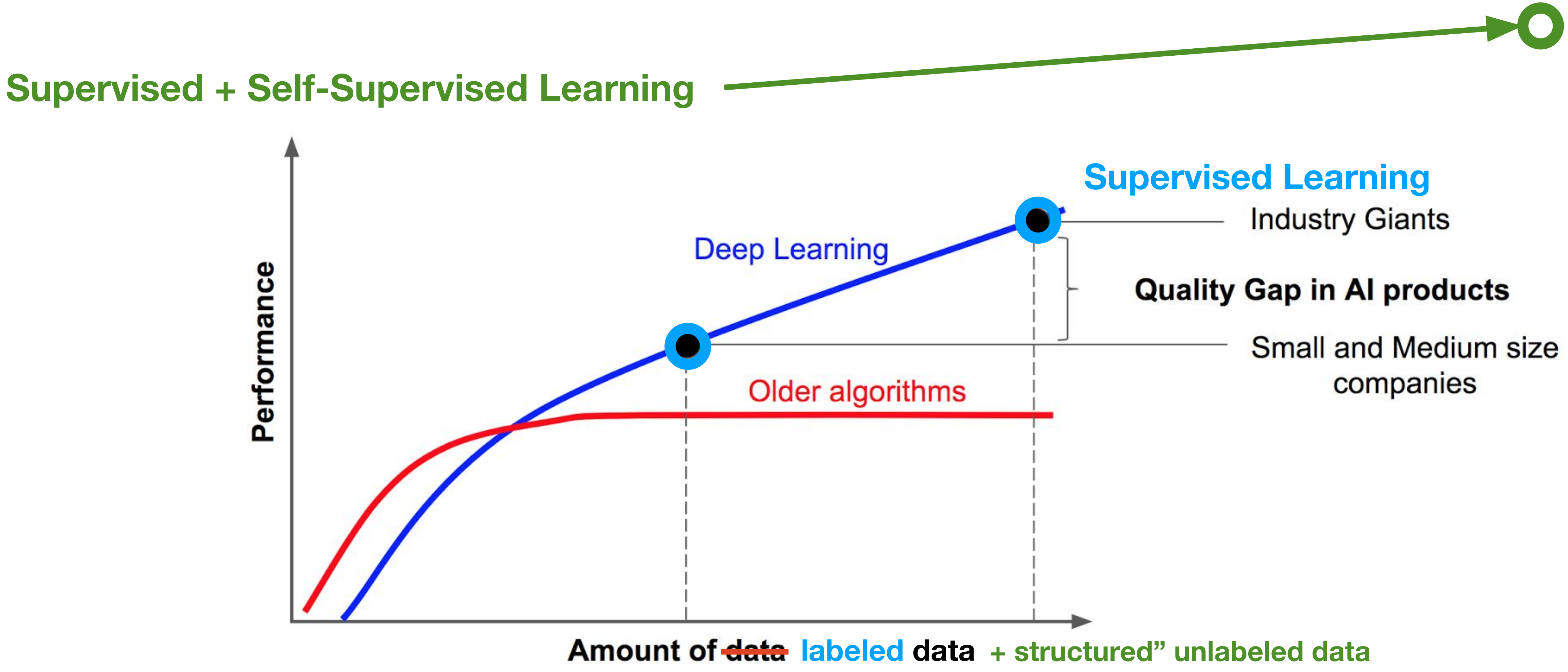


Image courtesy supervise.ly

Learning from labeled data

Learning with labeled and *“structured” unlabeled* data

TOYOTA RESEARCH INSTITUTE

Agenda

- Why Beyond Supervised Driving
- **Self-Supervised Learning: SuperDepth**
- Sim2Real adaptation: **SPIGAN**

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?
  - LiDAR: Expensive, Bulky
  - Cameras
    - Rich semantic and geometric sensing
    - Ubiquitous (2019 Toyota models)

Toyota Safety Sense 2.0
Camera

SuperDepth

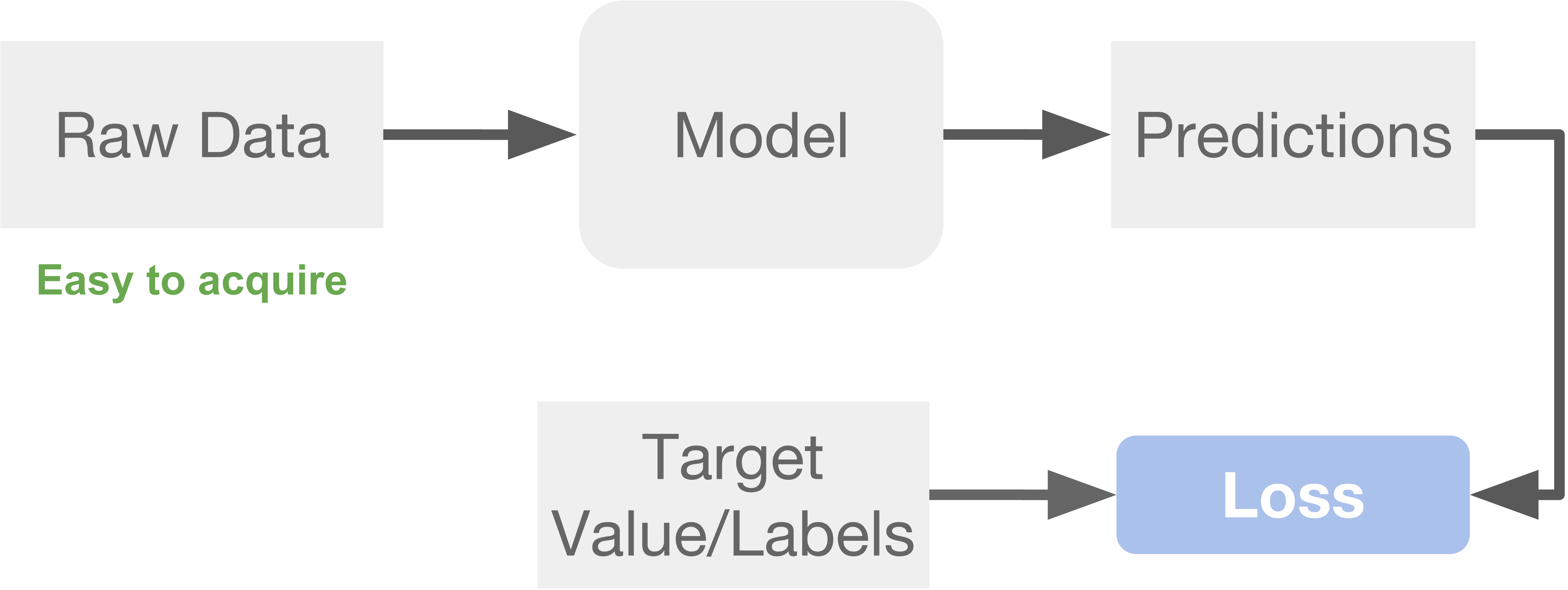
Self-Supervised, Super-Resolved Monocular Depth Estimation

Sudeep Pillai, Rares Ambrus, Adrien Gaidon

ICRA 2019 [arxiv + video]

TOYOTA RESEARCH INSTITUTE

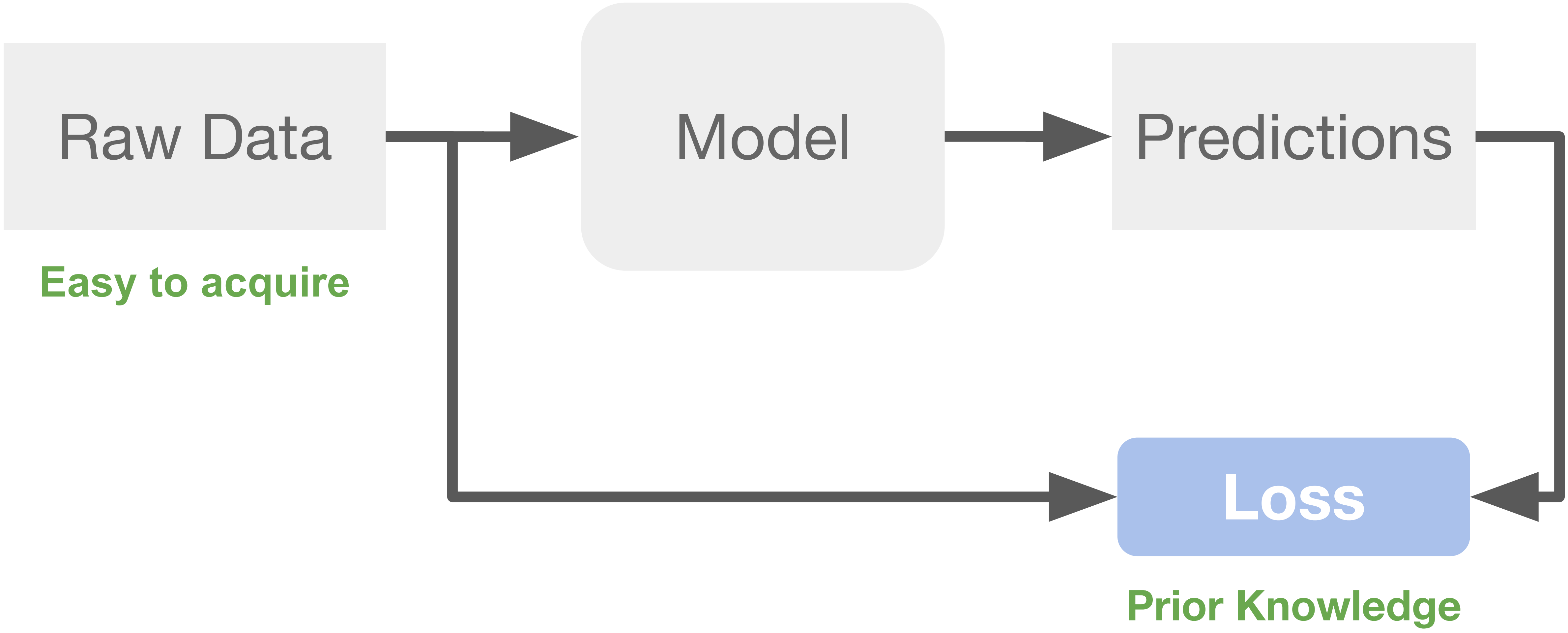
Supervised Learning



Easy to acquire

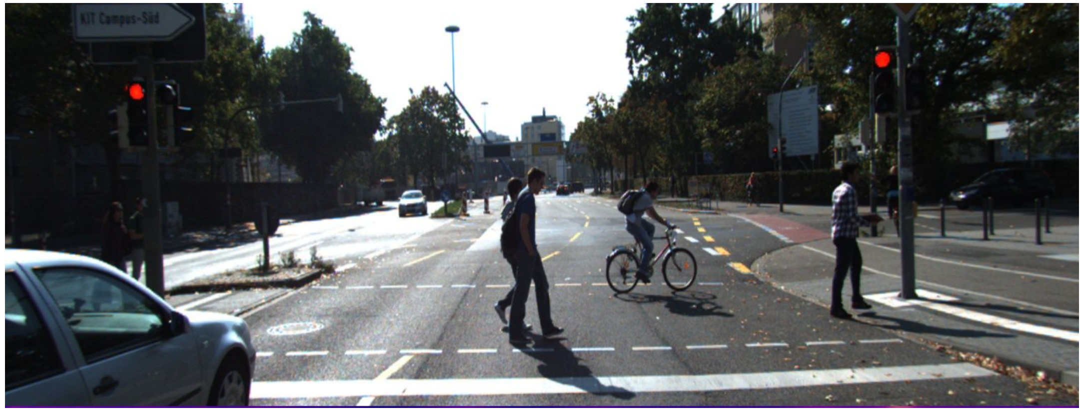
Expensive / Difficult to acquire

Self-Supervised Learning

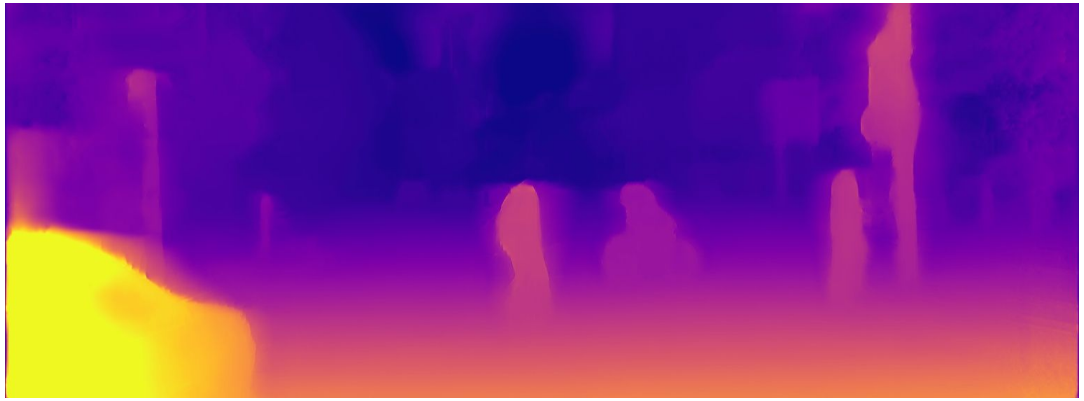


Monocular Depth Estimation

Single RGB Image

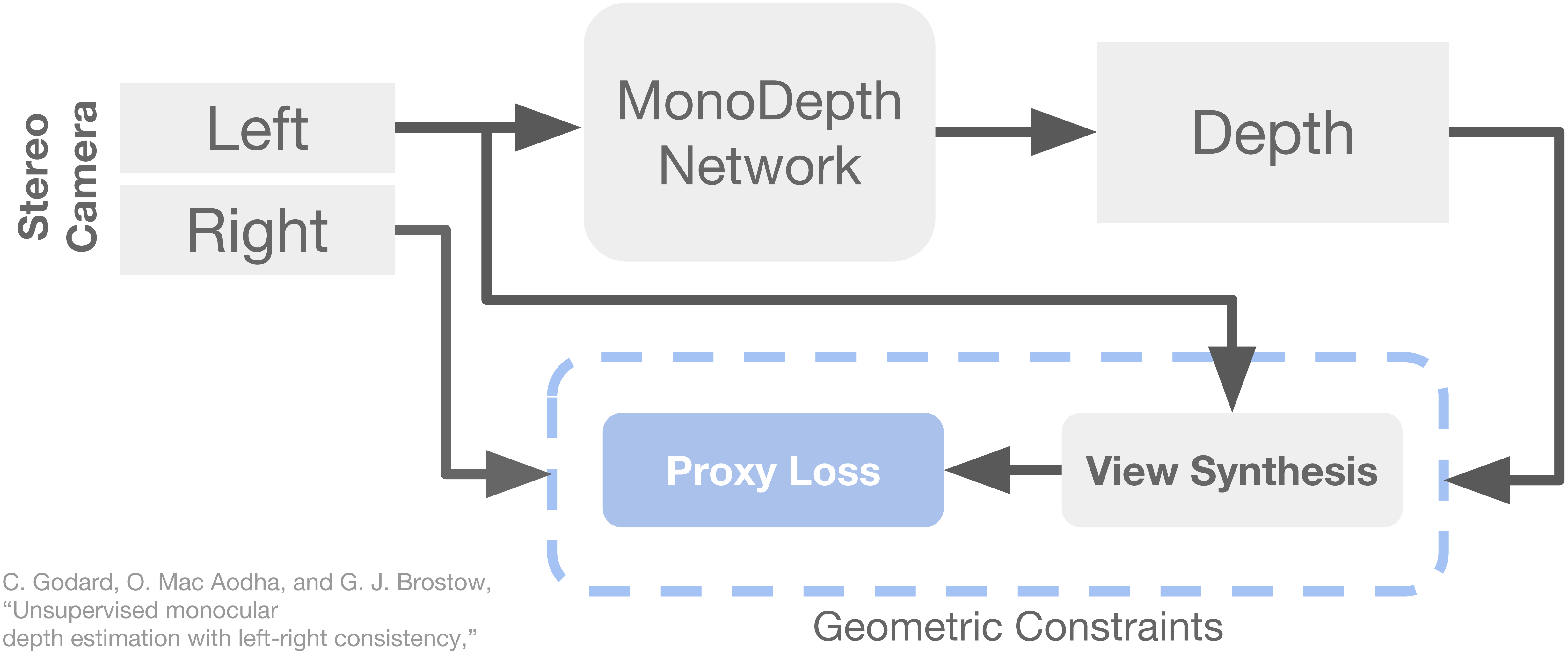


Predicted Depth Image



MonoDepth Network

Self-Supervised Monocular Depth



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

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)

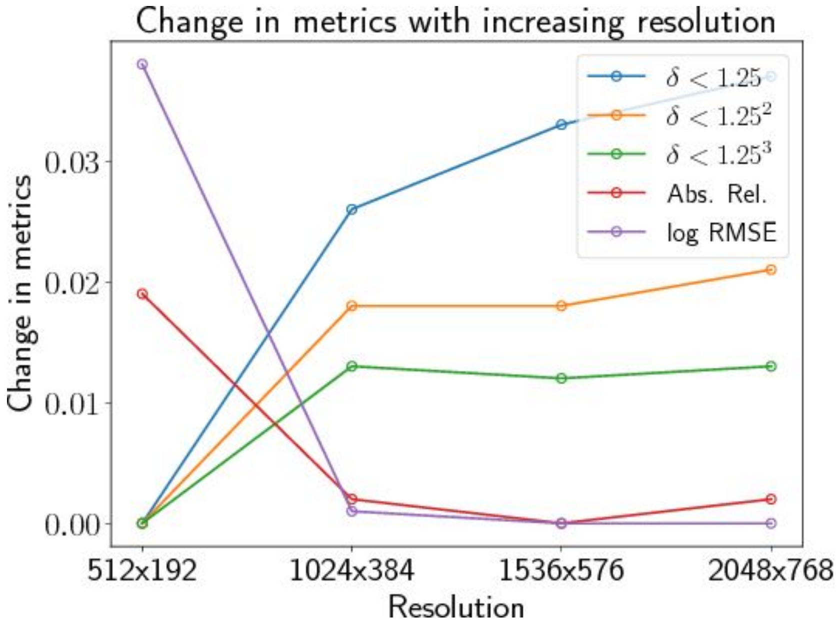
Occlusion Regularization

TOYOTA RESEARCH INSTITUTE 18

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)

Depth Super-Resolution

- **Sub-pixel convolutions for disparity super-resolution (SP)**
  - Replace resize-convolutions [1] with sub-pixel convolutions [2]
  - Improved photometric loss with finer details and crisp boundaries

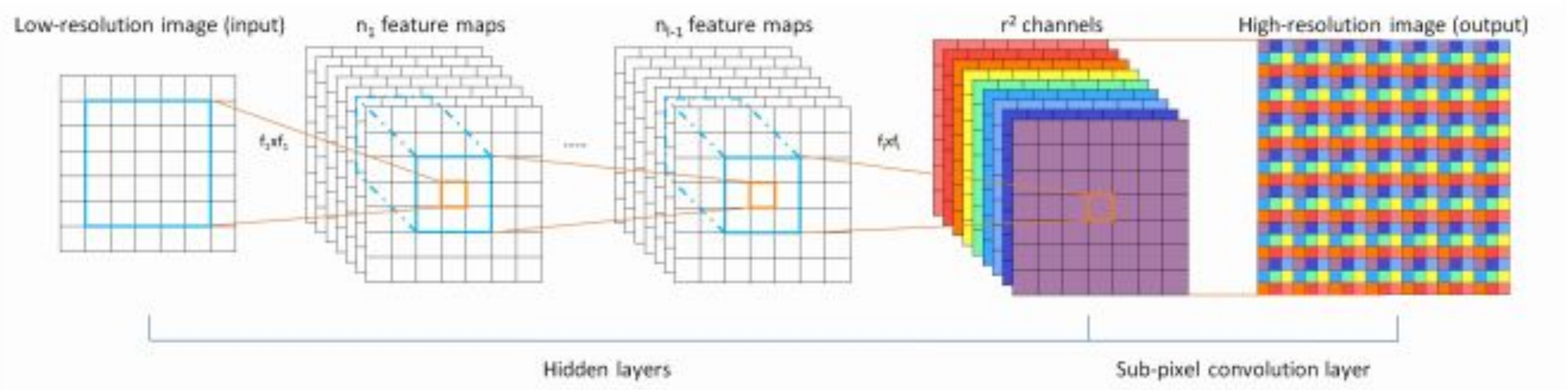


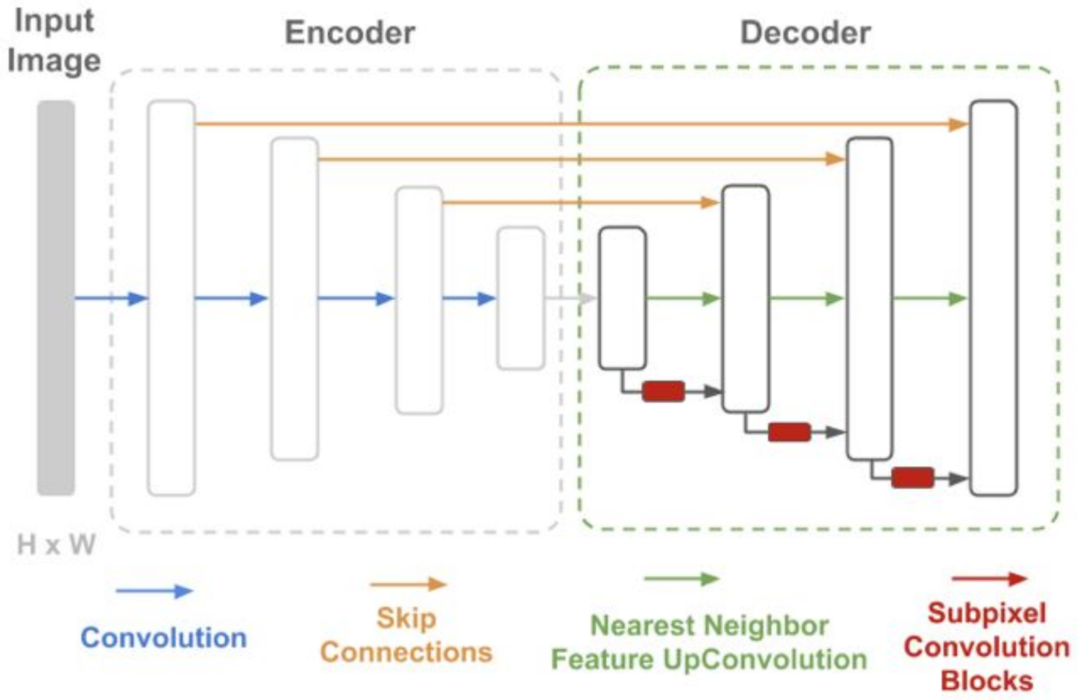
Figure 1. The proposed efficient sub-pixel convolutional neural network (ESPCN), with two convolution layers for feature maps extraction, and a sub-pixel convolution layer that aggregates the feature maps from LR space and builds the SR image in a single step.

A. Odena, V. Dumoulin, and C. Olah, “Deconvolution and checkerboard artifacts,” *Distill*, vol. 1, no. 10, p. e3, 2016.

W. Shi, J. Caballero, F. Husza ́r, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, “Real-time single image and video super- resolution using an efficient sub-pixel convolutional neural network,” CVPR 2016

Depth Super-Resolution

- **Sub-pixel convolutions for disparity super-resolution (SP)**
  - Replace resize-convolutions with sub-pixel convolutions

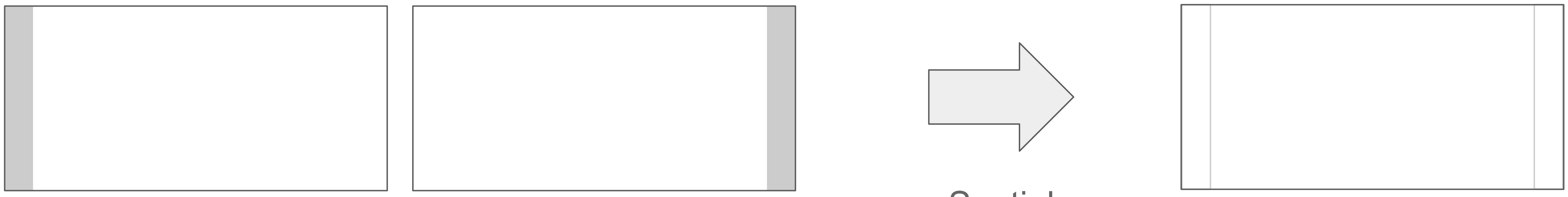


Modified DispNet Architecture

Bonus: Differentiable Flip Augmentation

- **Differentiable flip augmentation (FA)**
    - Differentiable FA using STNs [3] for trainable occlusion handling
    - End-to-end trainable network without boundary artifacts

Priors learned by model due to occluded boundaries
in **fronto-parallel stereo** case



M. Jaderberg, K. Simonyan, A. Zisserman, *et al.*, “Spatial transformer networks,” *NIPS 2015*
C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” *CVPR* 2017

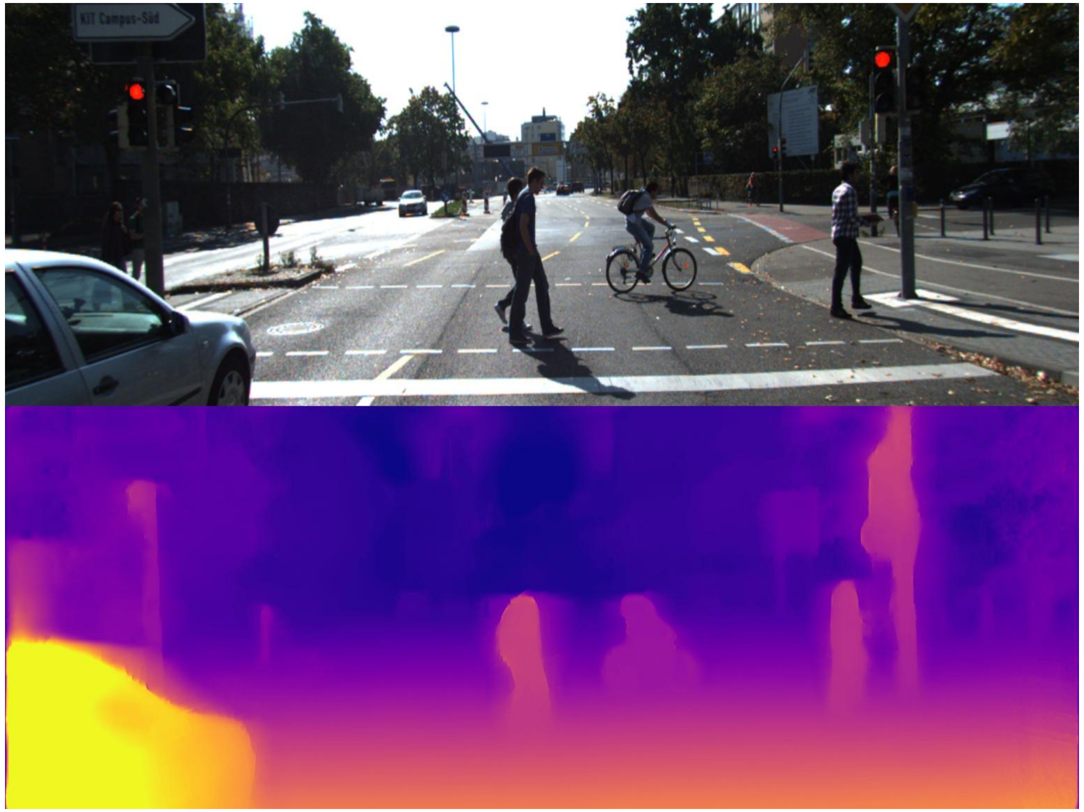
Disparity Estimation Performance

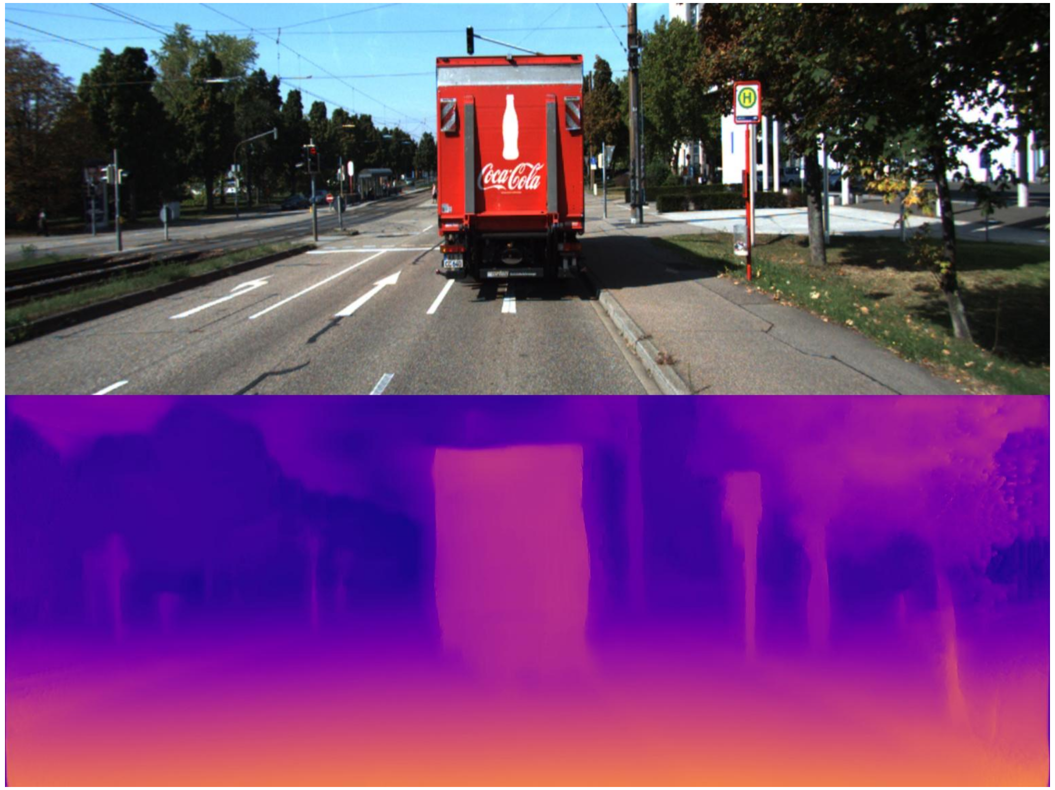
Method	Resolution	Dataset	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
UnDeepVO [25]	416 x 128	K	S	0.183	1.73	6.57	0.268	-	-	-
Godard et al. [6]	640 x 192	K	S	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [6]	640 x 192	CS+K	S	0.124	1.076	5.311	0.219	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**

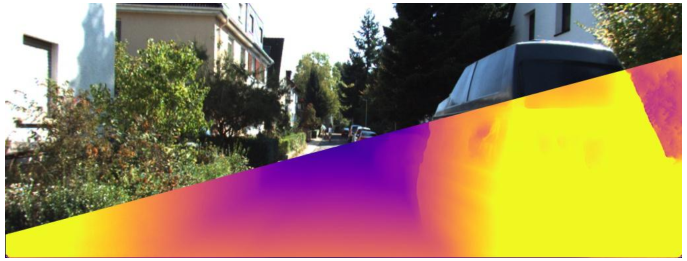
Depth Estimation Results on the KITTI 2015 Benchmark

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

Qualitative MonoDepth Performance

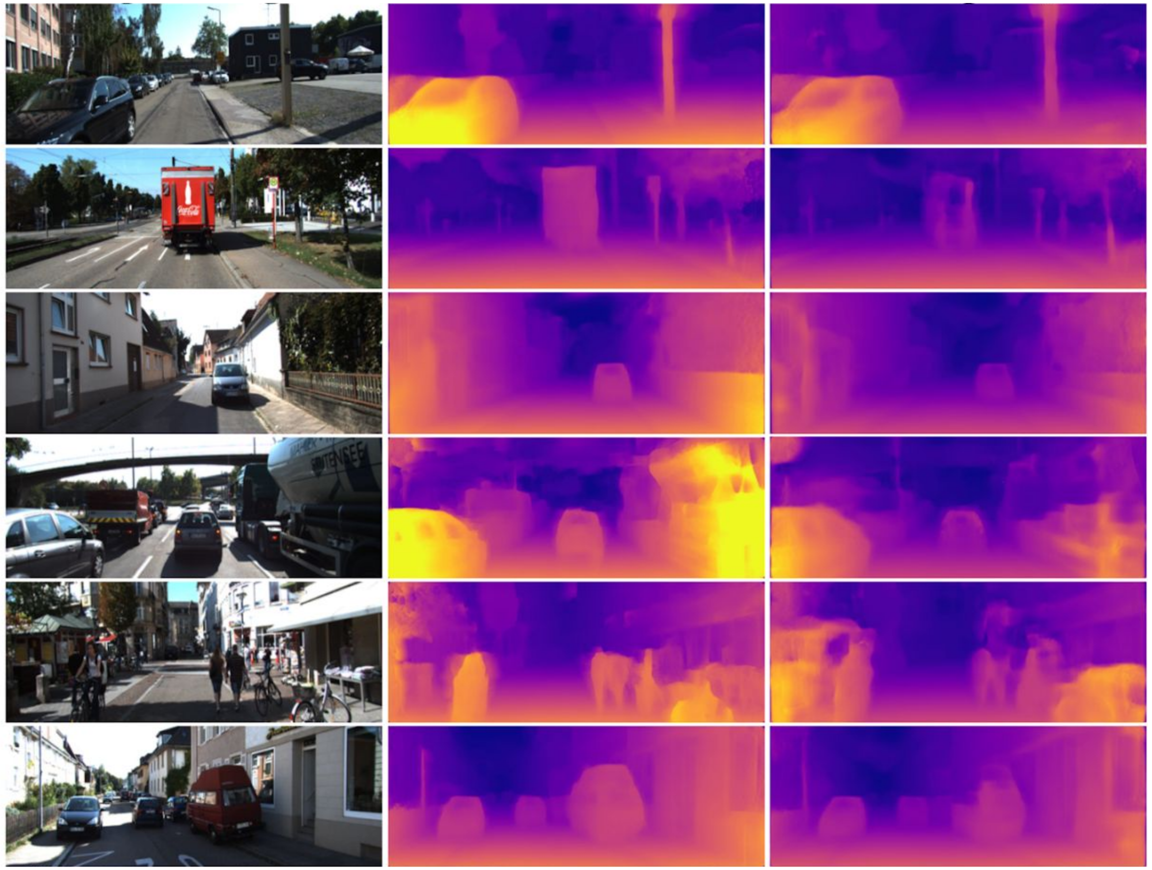




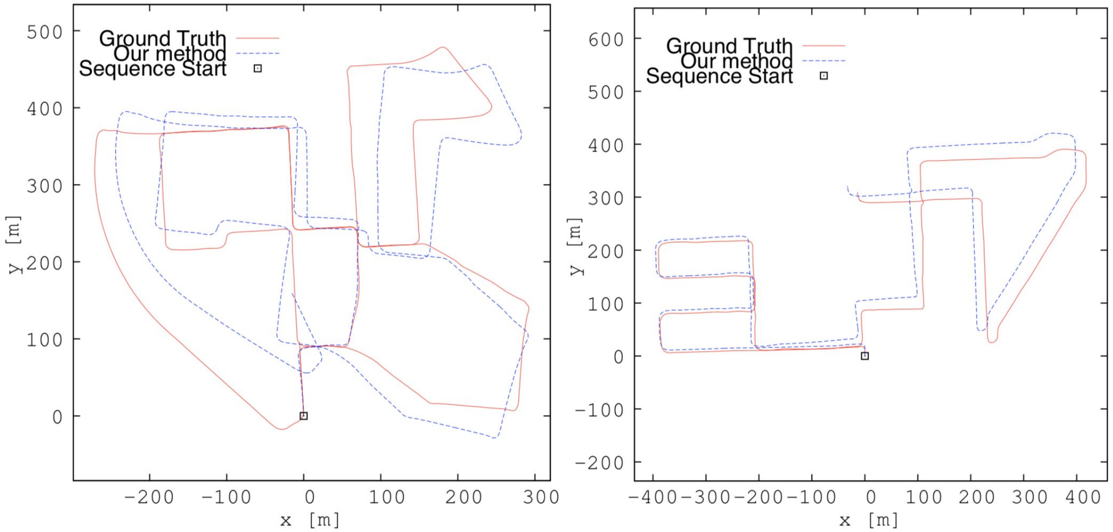


Qualitative Comparison to State-of-the-Art

SuperDepth MonoDepth [1]



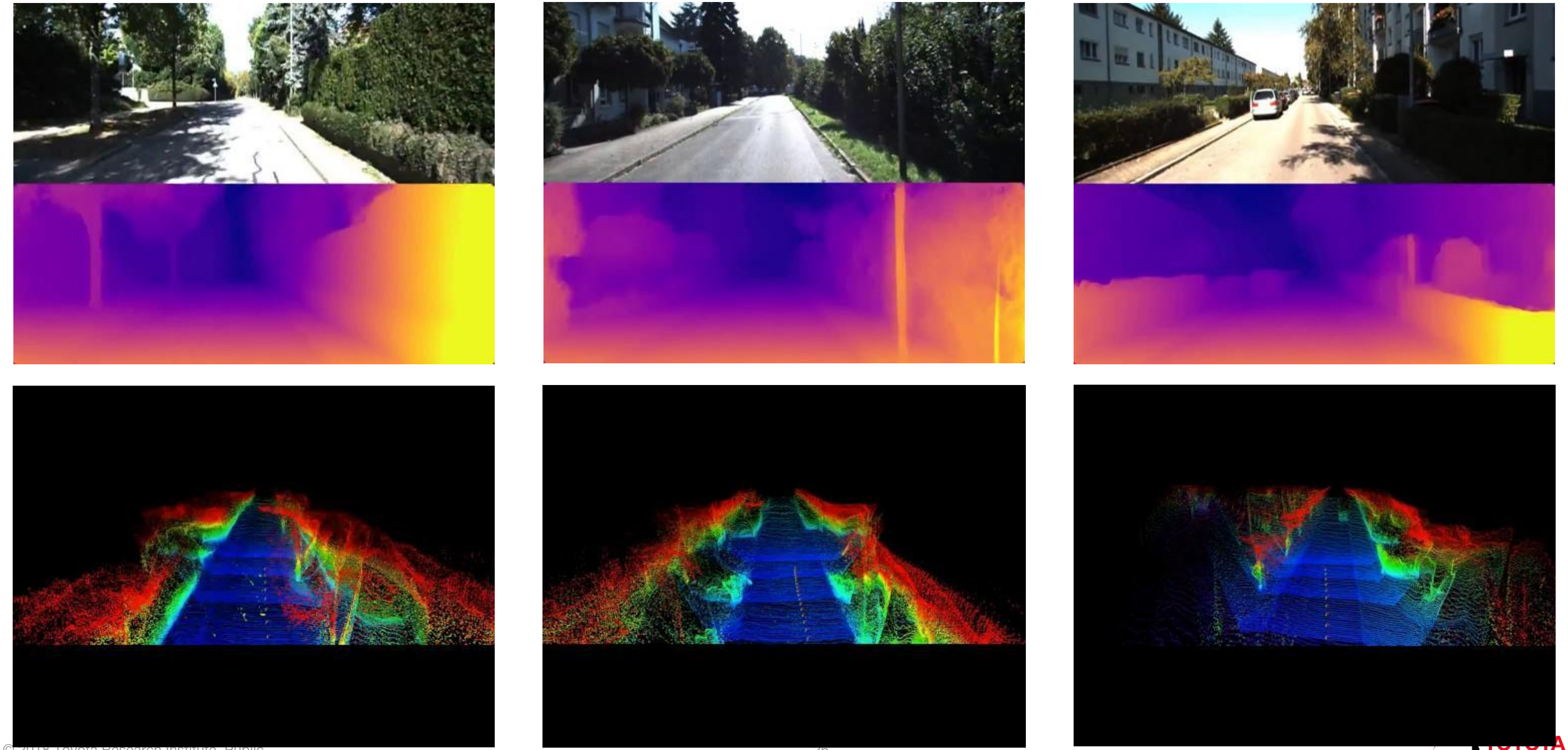
SuperDepth reconstruction is able to capture **fine details**, and **boundaries**



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

- Why Beyond Supervised Driving
- Self-Supervised Learning: **SuperDepth**
- **Sim2Real adaptation: SPIGAN**

TOYOTA RESEARCH INSTITUTE


TOYOTA GUARDIAN

SPIGAN

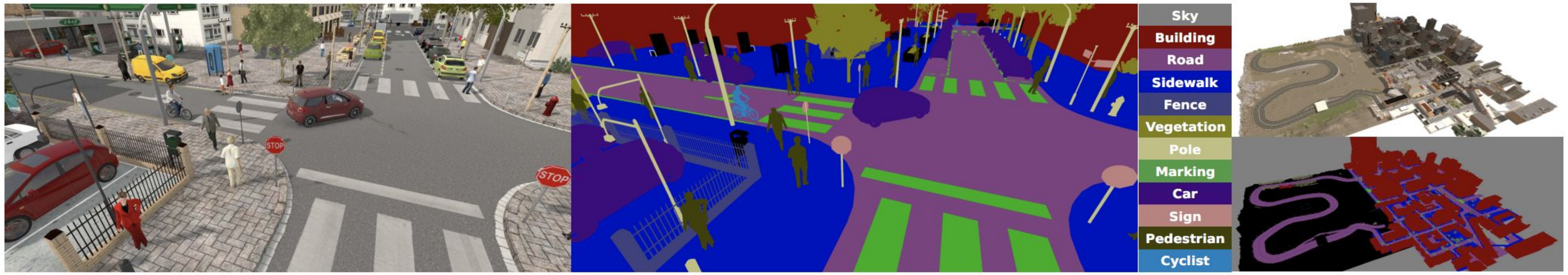
Privileged Adversarial Learning from Simulation

Kuan Lee, German Ros, Jie Li, Adrien Gaidon

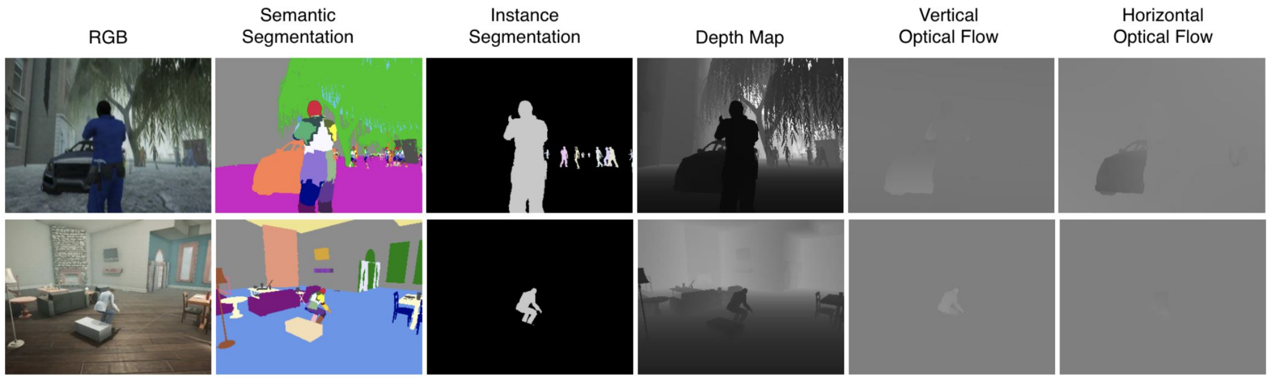
ICLR 2019 [arxiv]

Learning Using Simulator Privileged Information

Gaidon et al, "Virtual worlds as proxy for multiobject tracking analysis.", CVPR'16

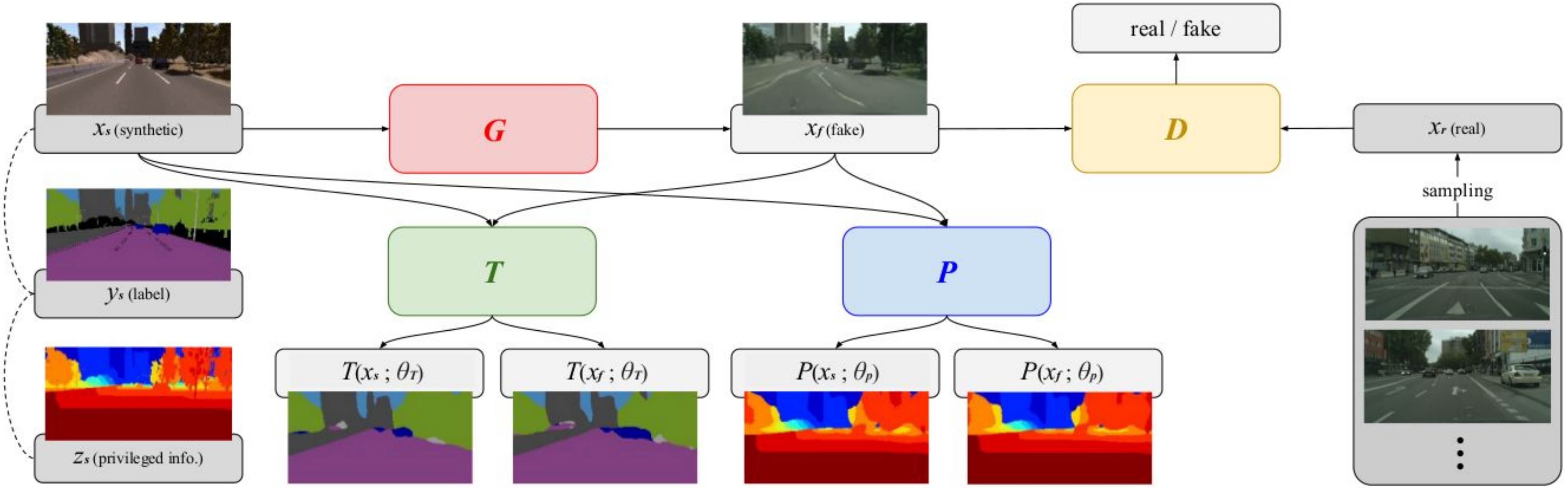


Ros et al, "The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes", CVPR'16



de Souza et al, "Procedural Generation of Videos to Train Deep Action Recognition Networks.", CVPR'17

Network Architecture



Minimax Learning Objective

$$\min_{\boldsymbol{\theta}_G,\boldsymbol{\theta}_T,\boldsymbol{\theta}_P} \max_{\boldsymbol{\theta}_D} \alpha\mathcal{L}_{\text{GAN}} + \beta\mathcal{L}_T + \gamma\mathcal{L}_P + \delta\mathcal{L}_{\text{perc}} \qquad (1)$$

adversarial loss

task loss
(this is what we care about)

privileged
regularization

perceptual regularization
(self-regularization)

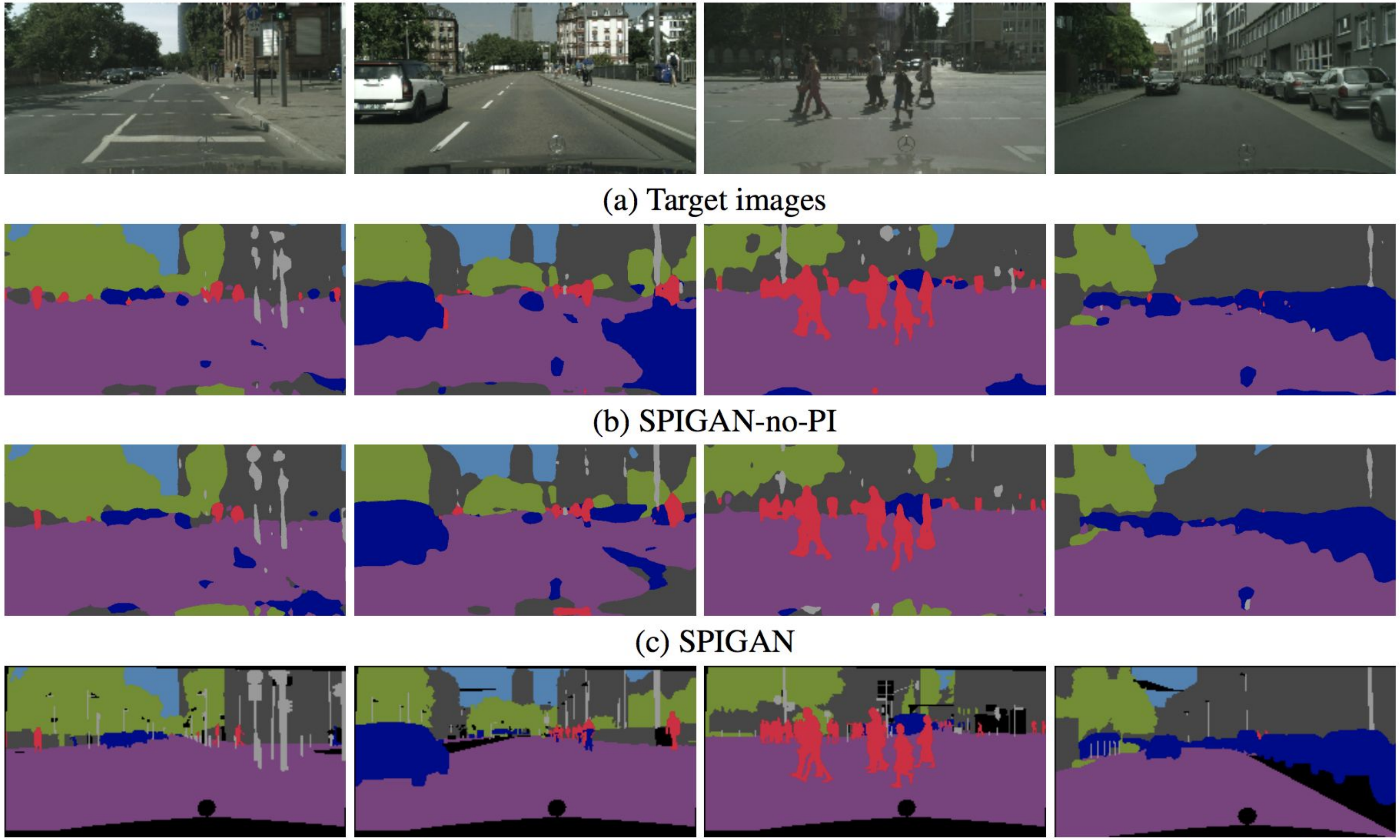
Experiments: Synthia → Cityscapes+Vistas

Method	flat	const.	object	nature	sky	human	vehicle	mIoU	Neg. Rate
FCN Source (Cityscapes)	79.6	51.0	8.7	29.0	50.9	3.0	31.6	36.3	–
SPIGAN-no-PI (Cityscapes)	90.3	58.2	6.8	35.8	69.0	9.5	52.1	46.0	0.16
SPIGAN (Cityscapes)	**91.2**	**66.4**	**9.6**	**56.8**	**71.5**	**17.7**	**60.3**	**53.4**	**0.09**
FCN Source (Vistas)	61.5	40.8	10.4	53.3	65.7	16.6	30.4	39.8	–
SPIGAN-no-PI (Vistas)	53.0	30.8	3.6	14.6	53.0	5.8	26.9	26.8	0.80
SPIGAN (Vistas)	**74.1**	**47.1**	**6.8**	**43.3**	**83.7**	**11.2**	**42.2**	**44.1**	**0.42**

Table 2: Semantic Segmentation results (per category and mean IoUs, higher is better) for SYNTHIA adapting to Cityscapes and Vistas. The last column is the ratio of images in the validation set for which we observe negative transfer (lower is better).

TOYOTA RESEARCH INSTITUTE 33

Experiments: Synthia → Cityscapes



(a) Target images

(b) SPIGAN-no-PI

(c) SPIGAN

(d) Ground truth

Experiments: Synthia → Cityscapes

(a) Source images

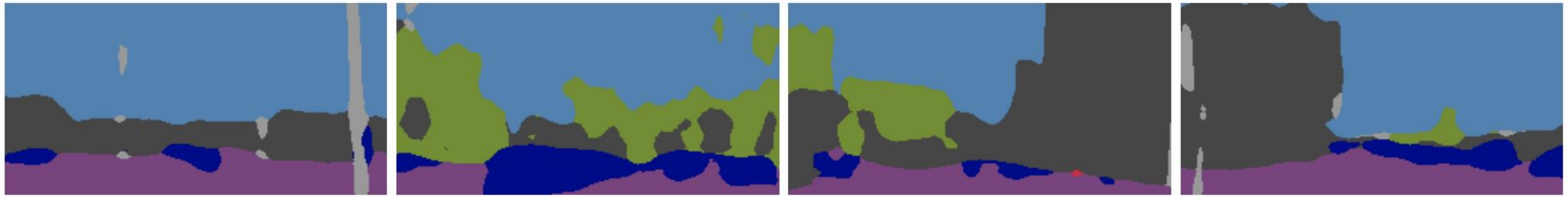
(b) SPIGAN-no-PI

(c) SPIGAN

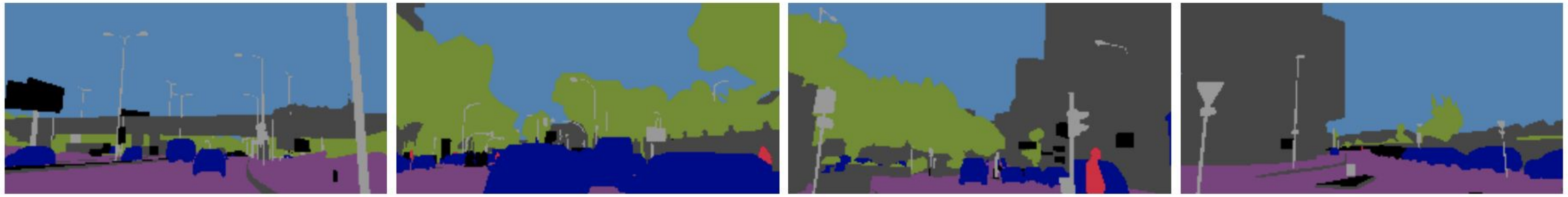
Experiments: Synthia → Vistas

(a) Target images

(b) SPIGAN-no-PI



(c) SPIGAN

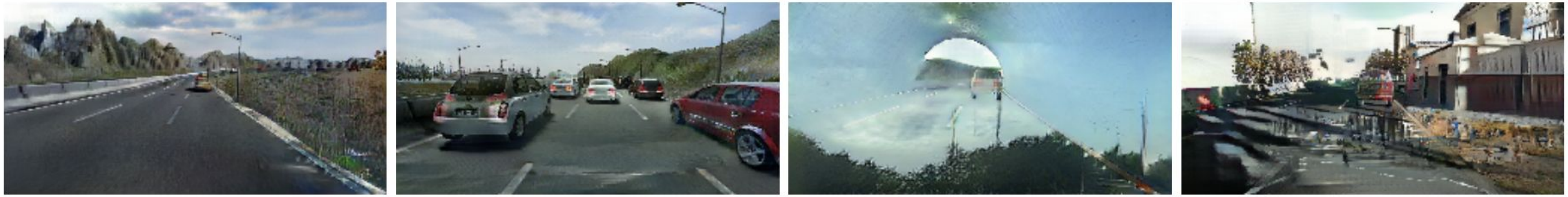


(d) Ground truth labels

36

Experiments: Synthia → Vistas

(a) Source images

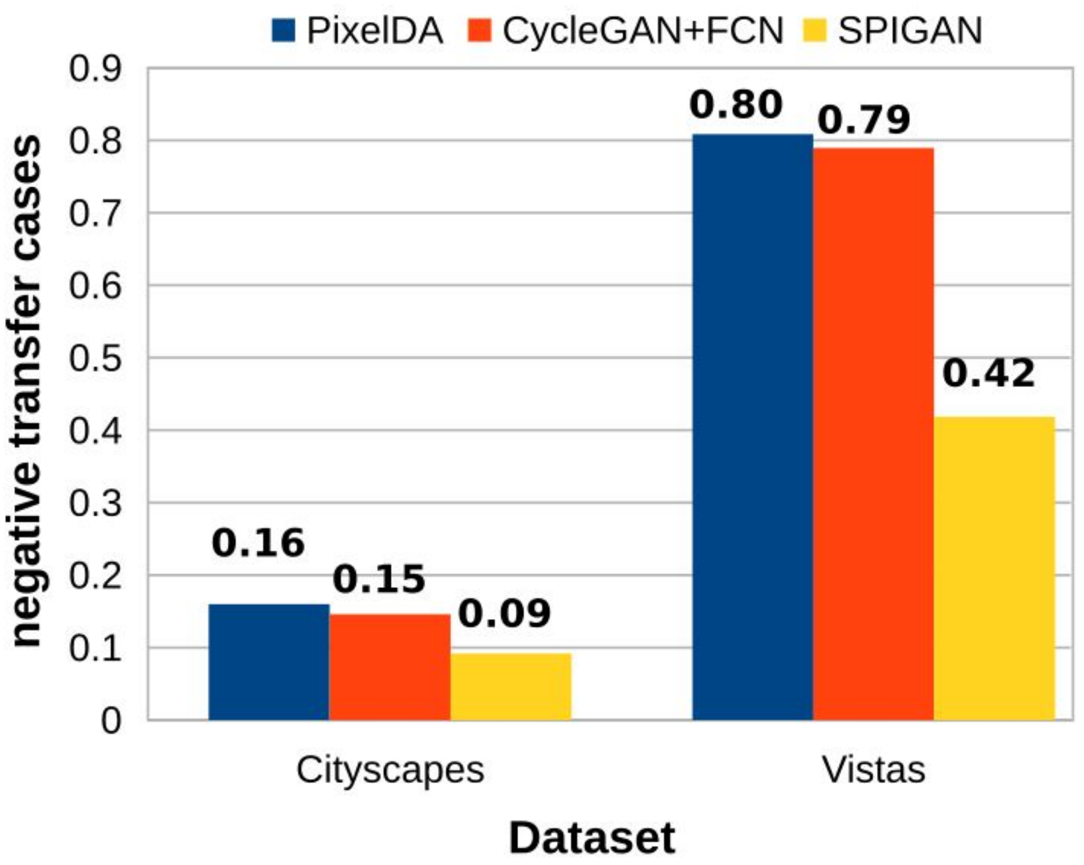


(b) SPIGAN-no-PI

(c) SPIGAN

37

Experiments: negative transfer



Conclusion

Beyond Supervised Driving at TRI

- **Why?** Need ***all*** the data for ***true*** autonomy
- **SuperDepth**: ***Self-Supervised,*** Super-Resolved Monocular Depth Estimation
- **SPIGAN**: ***Unsupervised sim2real*** adaptation using privileged information from the simulator

Learning from Structured Unlabeled Data


Thank You!