ETH zürich

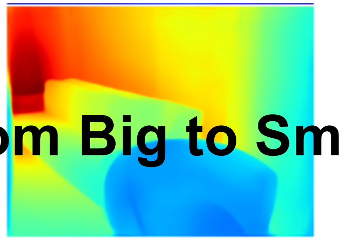
How well does uncertainty estimation actually work?

semantics in robotic perception

Hermann Blum, Cesar Cadena, Roland Siegwart

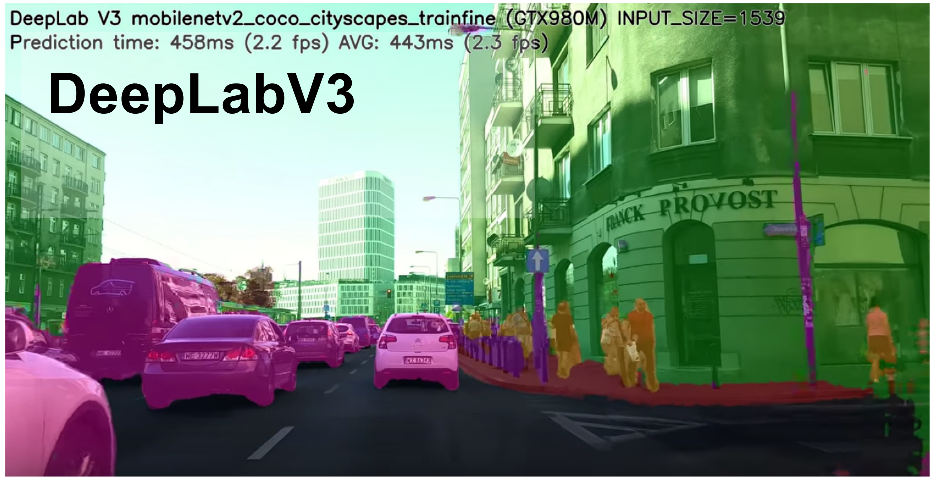
Autonomous Systems Lab

Semantic Scene Understanding

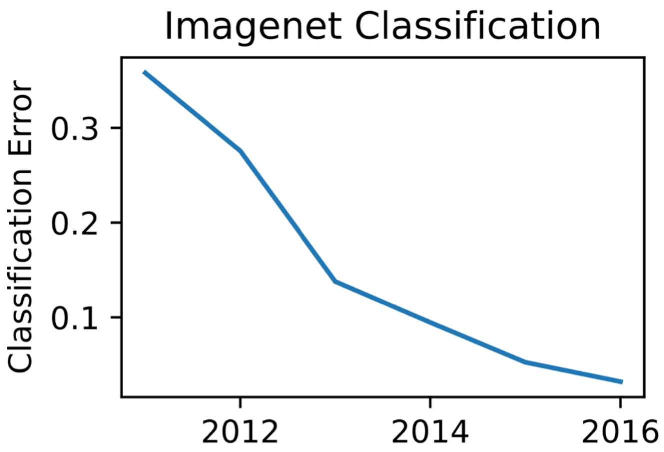


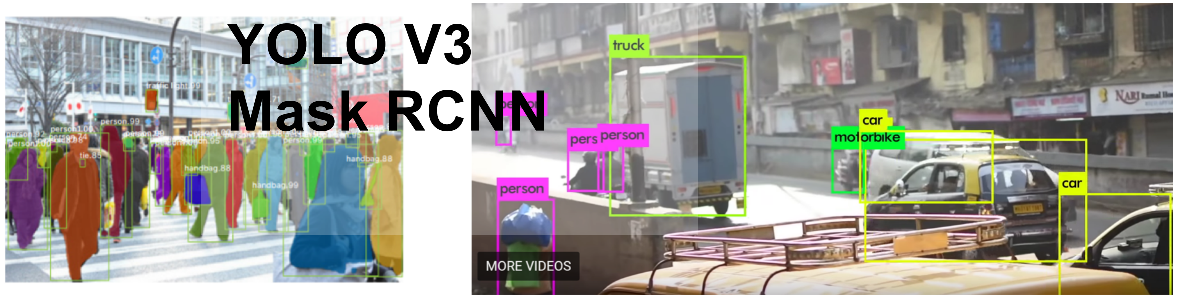
From Big to Small

DeepPerimeter



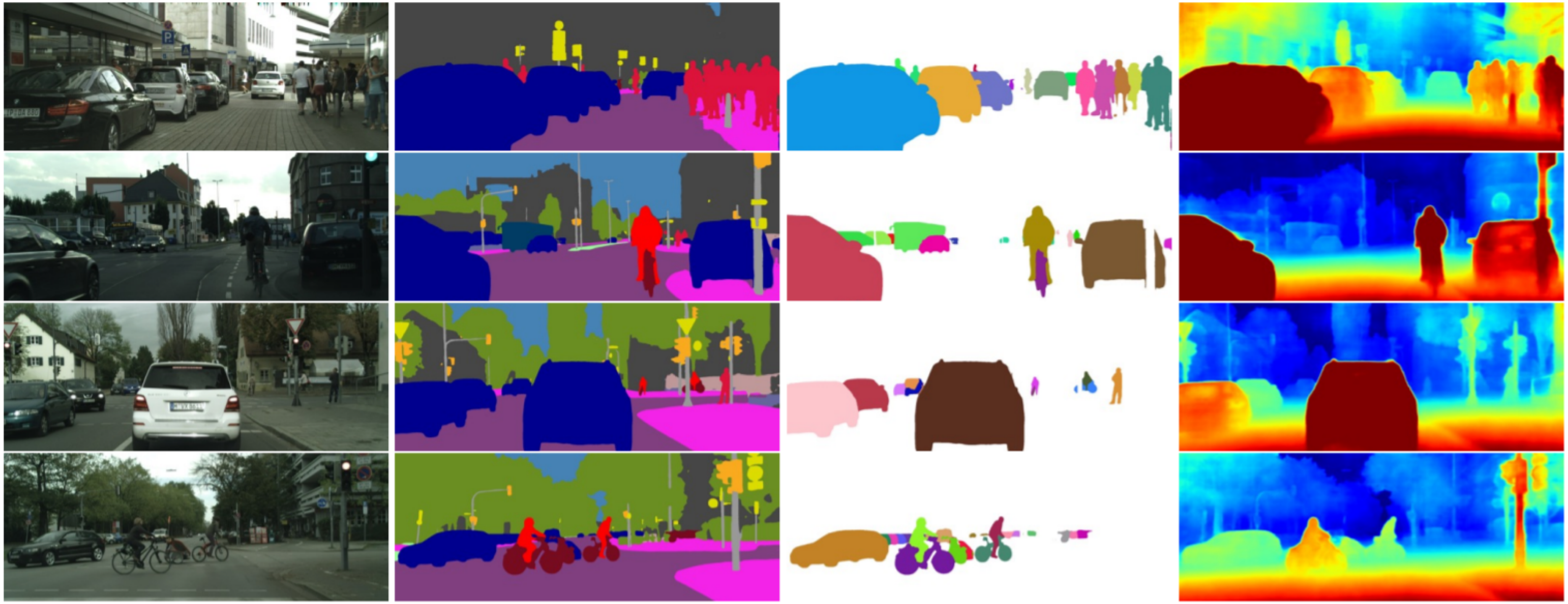
DeepLabV3





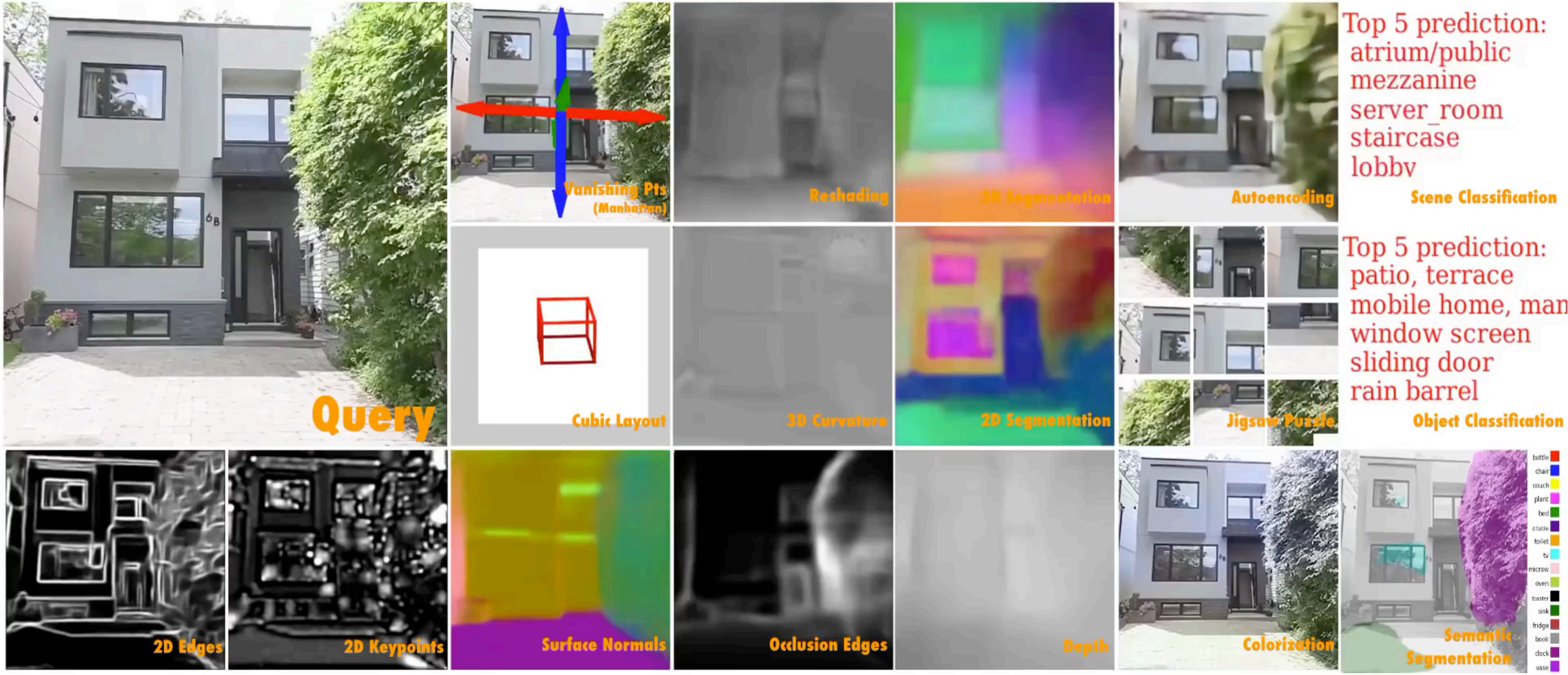
YOLO V3
Mask RCNN

Semantic Scene Understanding



Kendall et al., CVPR '18

Semantic Scene Understanding



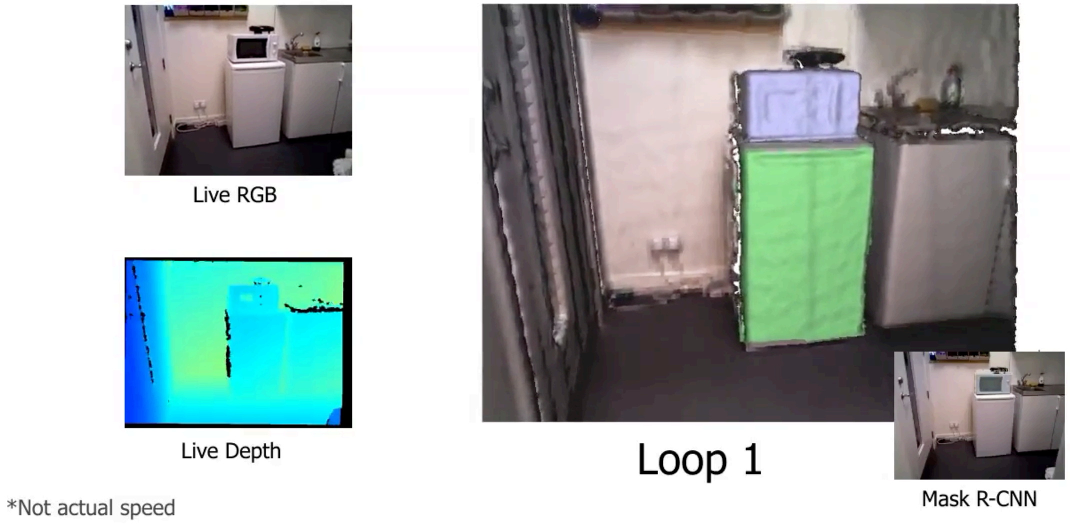
Zamir et al., CVPR '18

Applications in Robotics

Object Aware Geometry Estimation

Provides object shape priors

Off-the-shelf perceptual toolbox



J. McCormac, et al. 3DV 2018.

Incorporates geometric and object-based segmentations

M. Grinvald, et al. IROS 2019.

Semantic Aware Geometry Estimation

Object and semantics for estimation and data association

Miksik and Vineet, 2019.

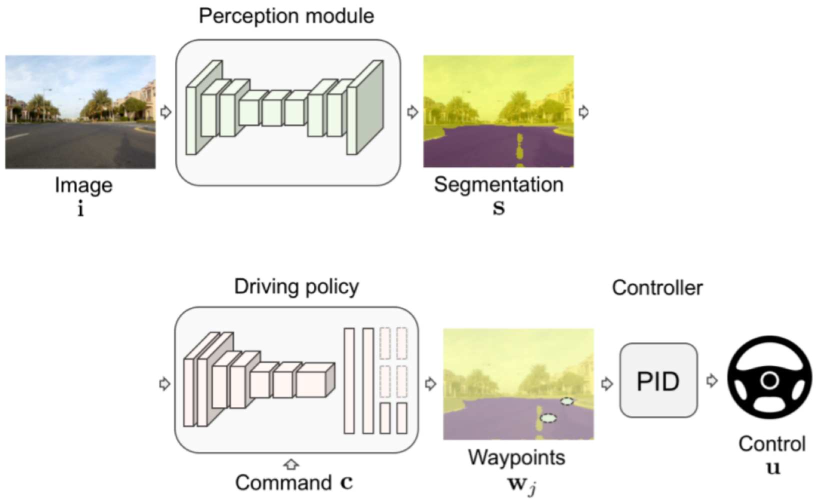
S. Bowman, et al. ICRA 2017.

Semantics for Domain Transfer

Higher level of abstraction

Invariant to illumination and view-point

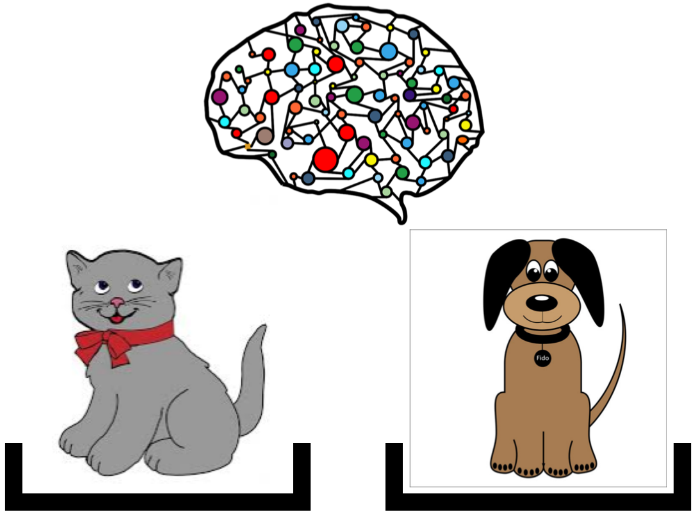
Easier transfer from virtual to real



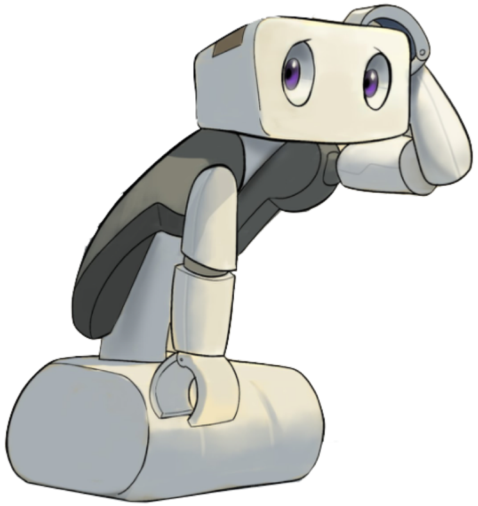
Müller, et al. CoRL 2018.

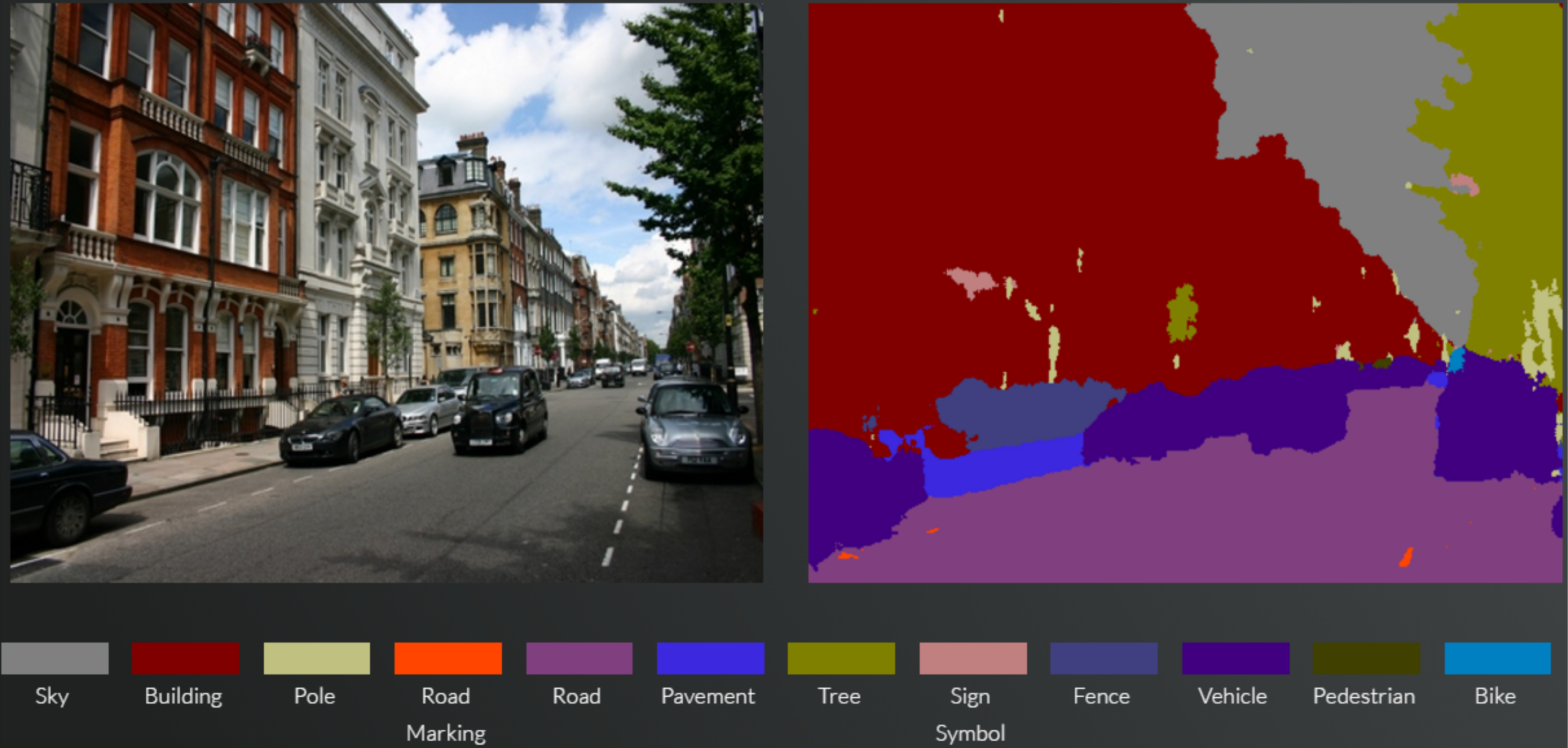
Autonomous Robots operate in an open world.

!



?





http://mi.eng.cam.ac.uk/projects/segnet/demo.php#demo

http://mi.eng.cam.ac.uk/projects/segnet/demo.php#demo

http://mi.eng.cam.ac.uk/projects/segnet/demo.php#demo

The Guardian

This article is more than 3 years old

Tesla driver dies in first fatal crash while using autopilot mode

The autopilot sensors on the Model S failed to di white tractor-trailer crossing the highway agair sky

Danny Yadron *and* **Dan Tynan** *in San Francisco*

Fri 1 Jul 2016 00.14 BST

The first known death caused by a self-driving car by Tesla Motors on Thursday, a development that i cause consumers to second-guess the trust they pu booming autonomous vehicle industry.

The 7 May accident occurred in Williston, Florida, driver, Joshua Brown, 40, of Ohio put his Model S i autopilot mode, which is able to control the car du driving.

Against a bright spring sky, the car's sensors systen distinguish a large white 18-wheel truck and trailer highway, Tesla said. The car attempted to drive ful the trailer, "with the bottom of the trailer impactin of the Model S", Tesla said in a blogpos

Hyperdrive

Uber Halts Autonomous Car Tests After Fatal Crash in Arizona

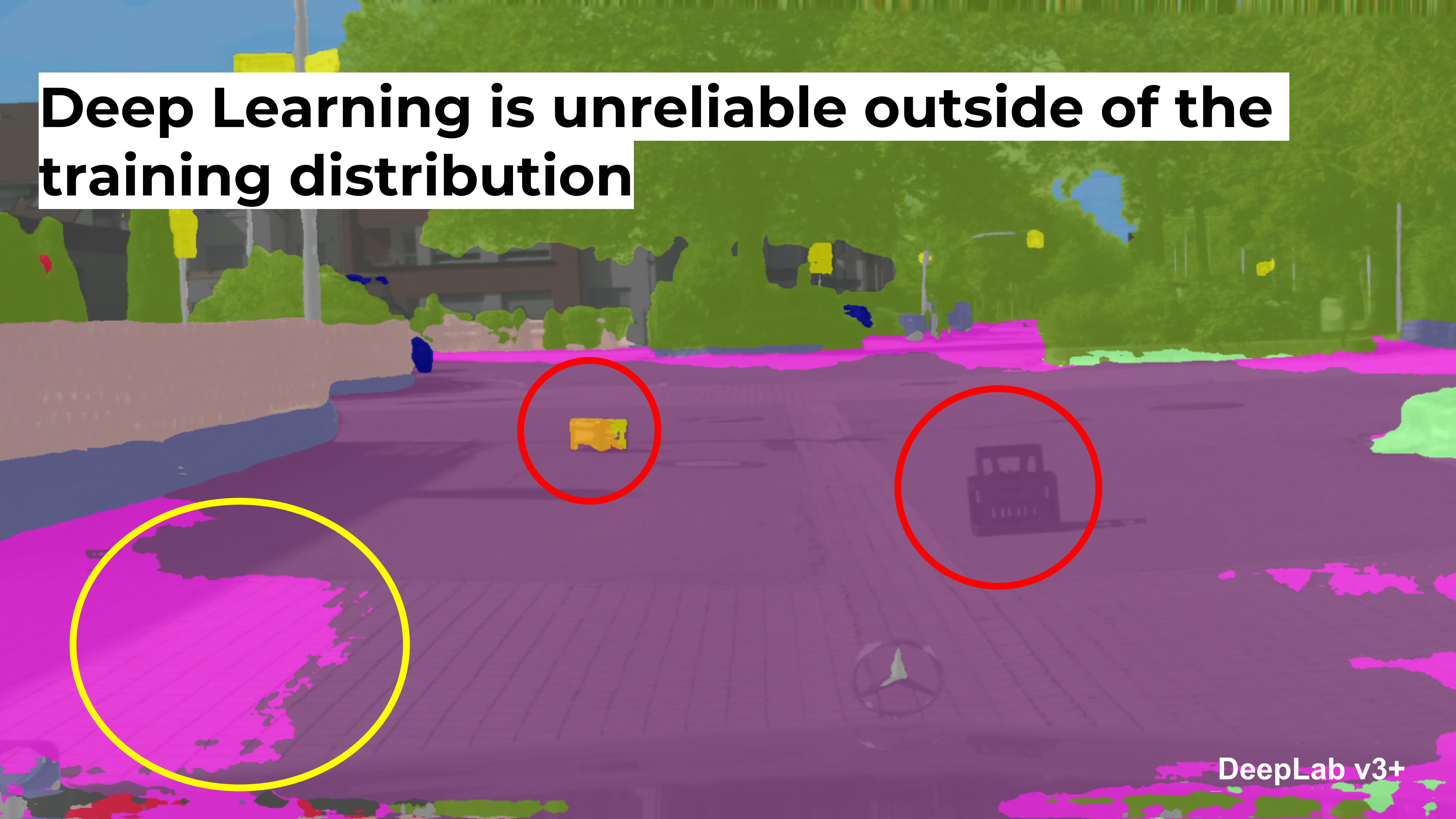
By Mark Bergen and Eric Newcomer

20. März 2018, 00:56 GMT+8

Updated on 20. März 2018, 04:31 GMT+8

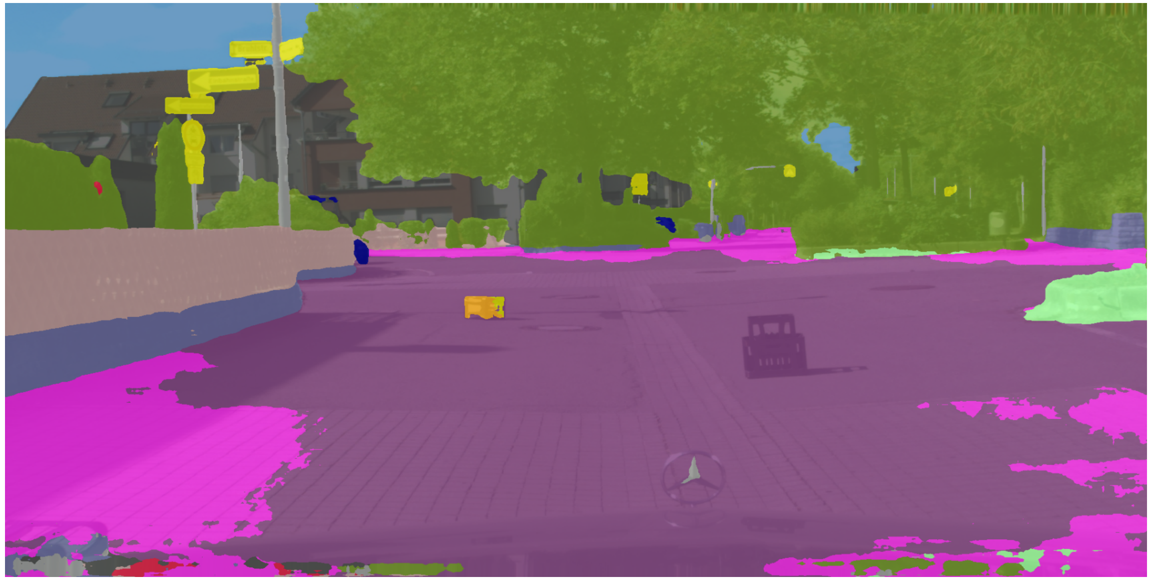
- First known pedestrian death involving a self-driving vehicle
- Incident may raise questions about safety of technology

Deep Learning is unreliable outside of the training distribution

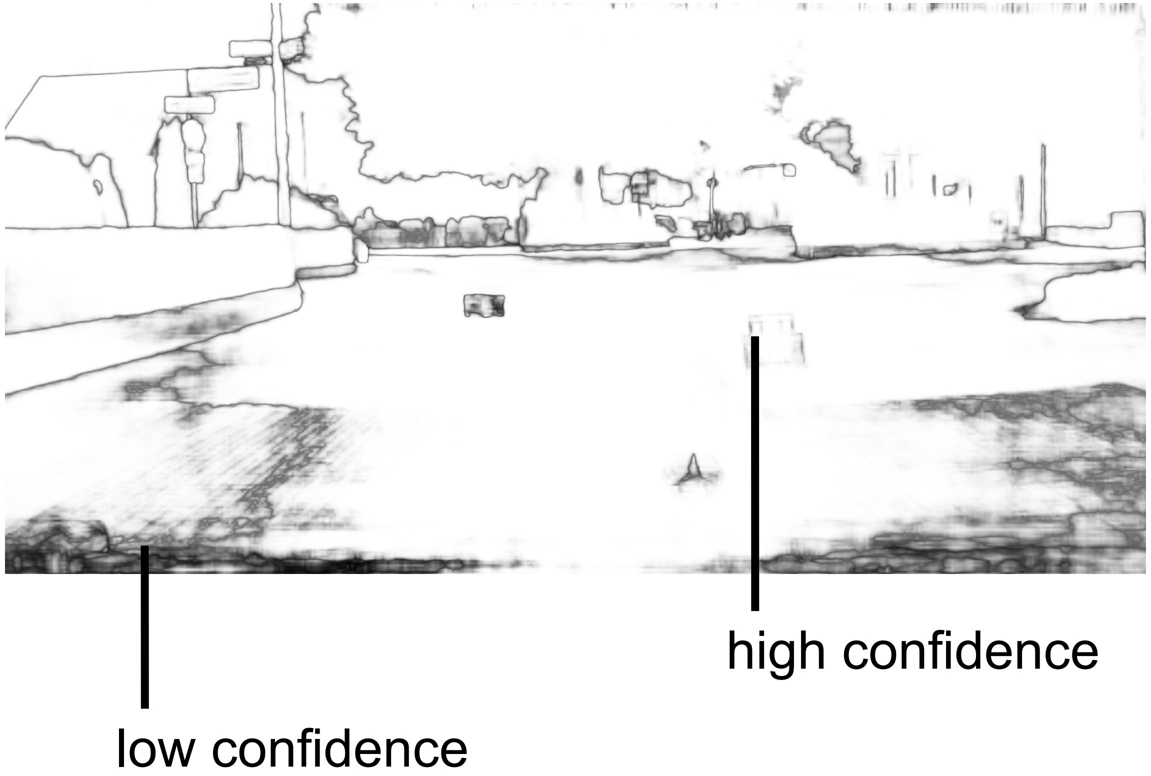


Softmax output is overconfident

Semantic Segmentation

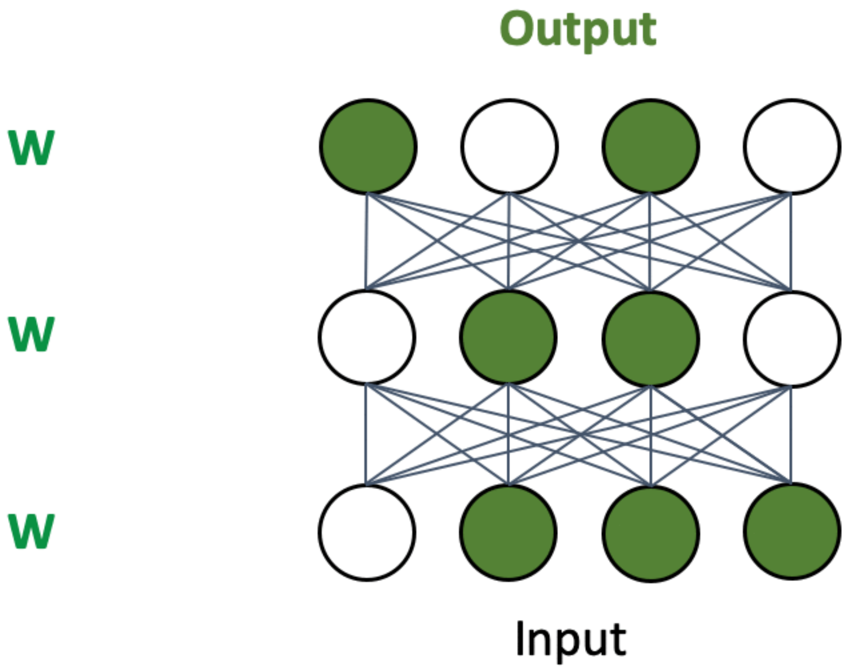


Softmax Confidence



Hendrycks, D., & Gimpel, K. (ICLR 2016). A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks.

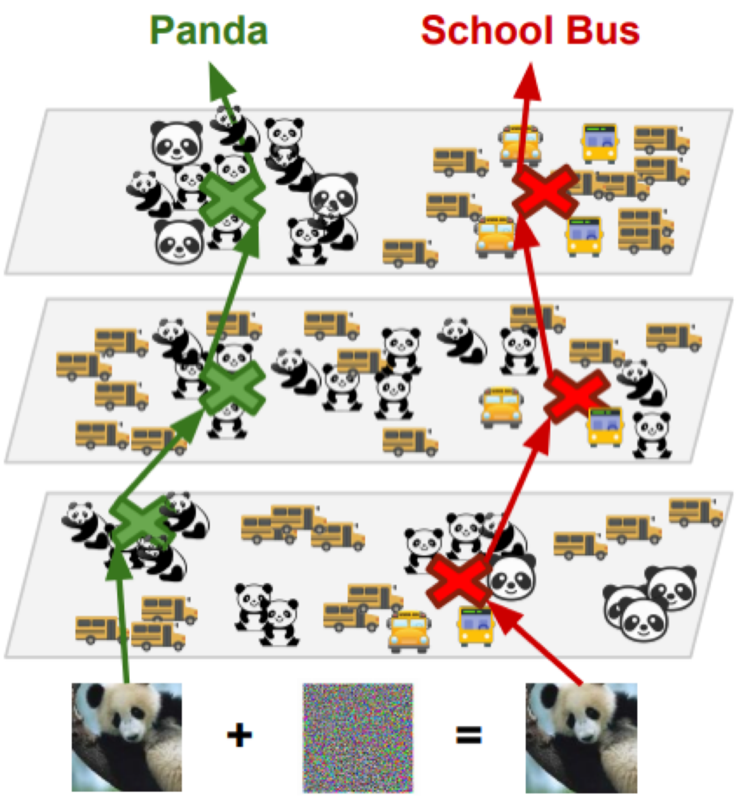
Bayesian Learning: Distribution over Weights



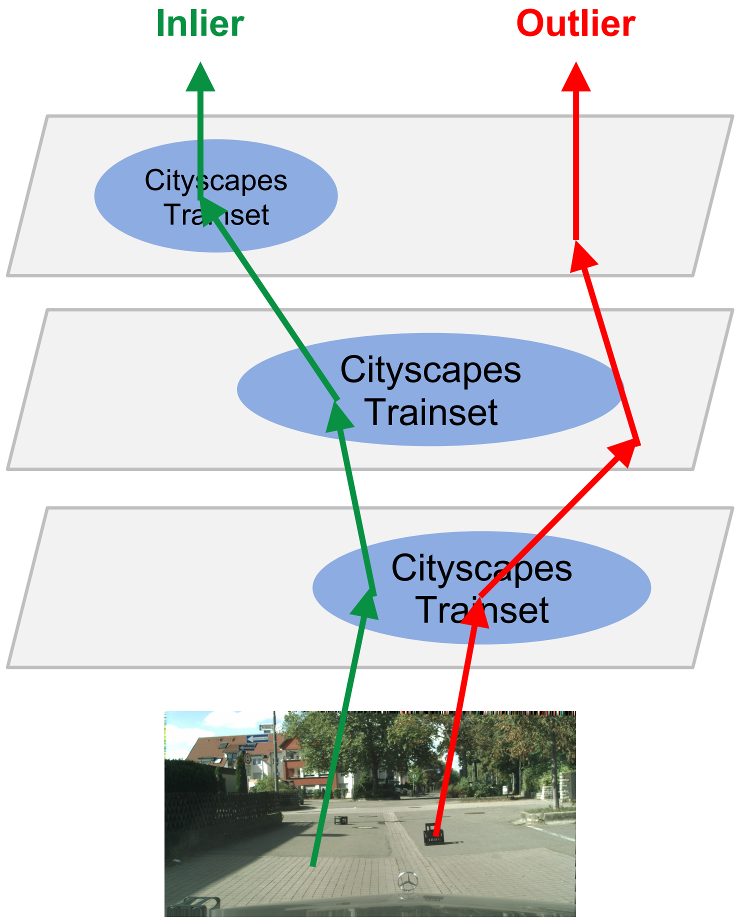
Epistemic Uncertainty

Gal, Y., & Ghahramani, Z. (ICML 2016). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.

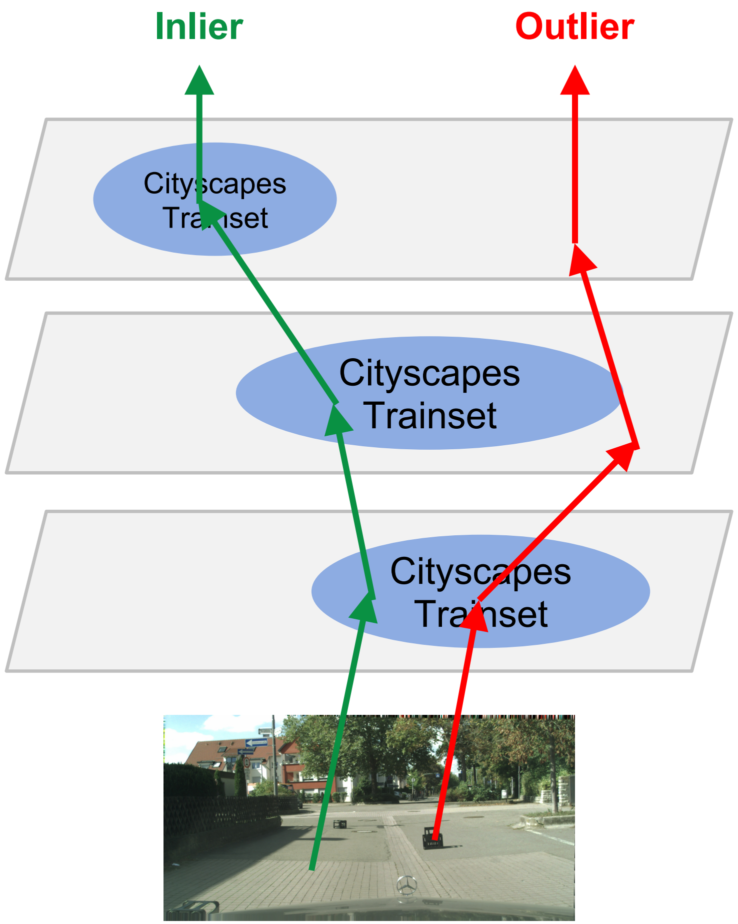
Embedding: Distribution in Layer Outputs



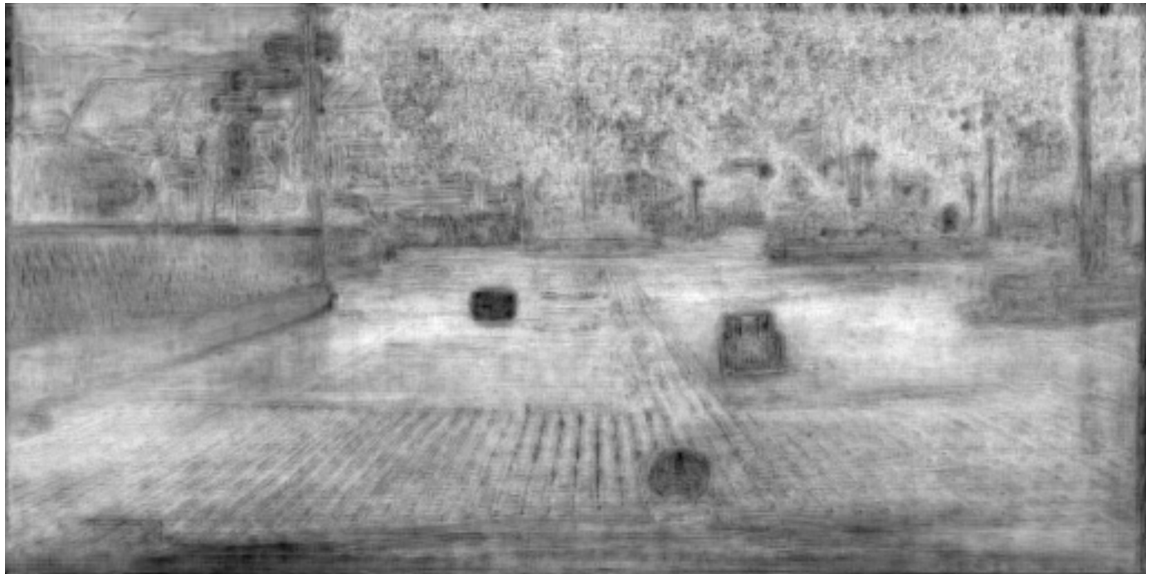
Papernot, N., & McDaniel, P. (2018). Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning.



Neighborhoods: Distribution in Layer Outputs



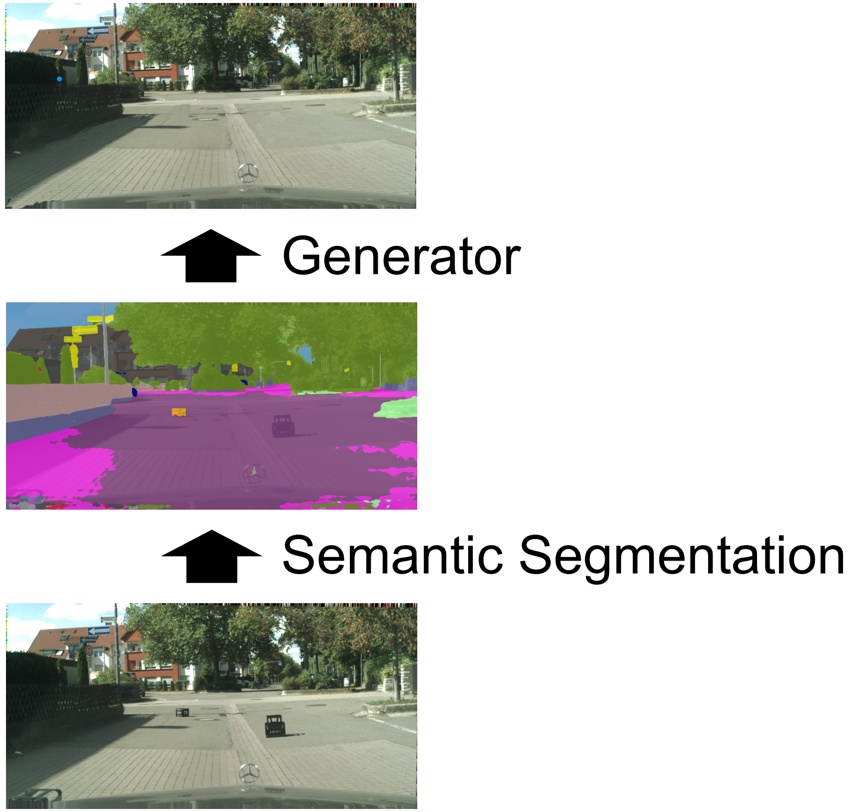
Embedding Density



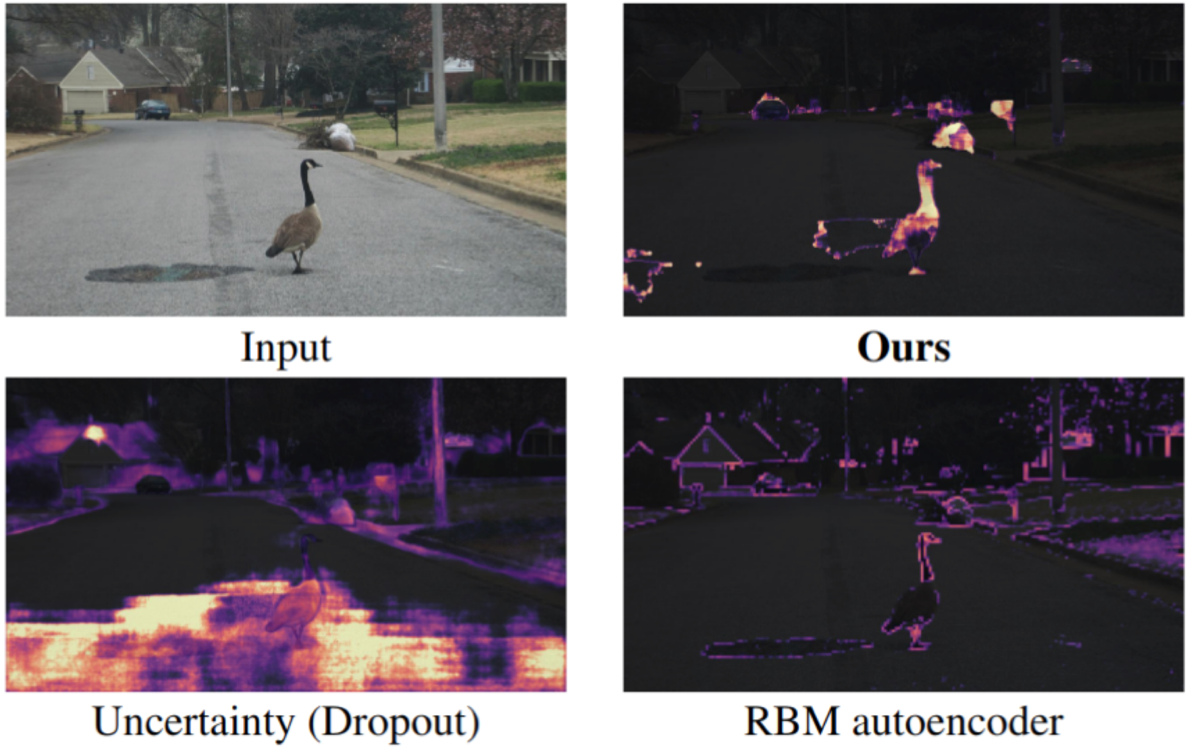
Mandelbaum, A., & Weinshall, D. (2017). Distance-based Confidence Score for Neural Network Classifiers.

Blum et al. (2019) The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. arXiv 2019

Reconstruct to measure discrepany

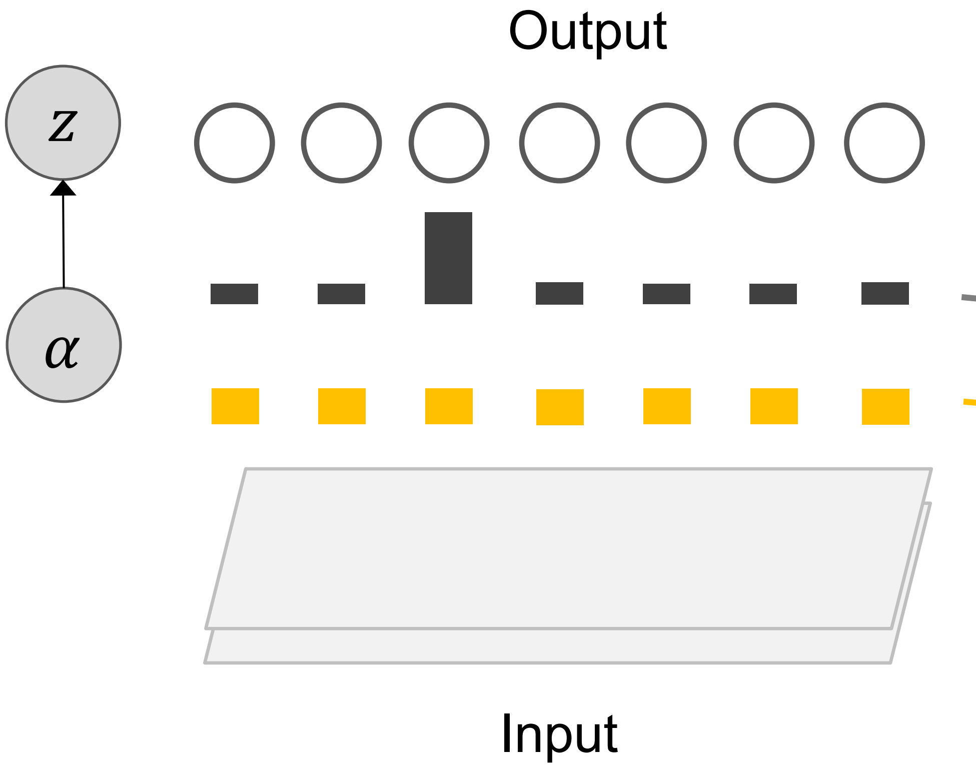


Haldimann, D., Blum, H., Siegwart, R., & Cadena, C. (2019). This is not what I imagined, arXiv 2019

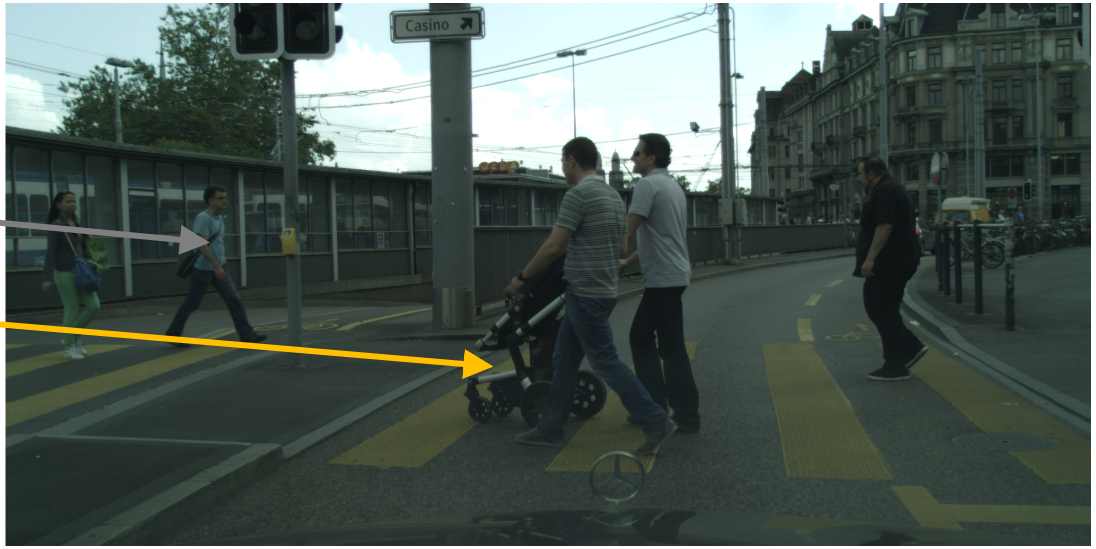


Lis, K., Nakka, K., Fua, P., & Salzmann, M. Detecting the Unexpected via Image Resynthesis. ICCV 2019

Supervised Anomaly Learning: Dirichlet Prior

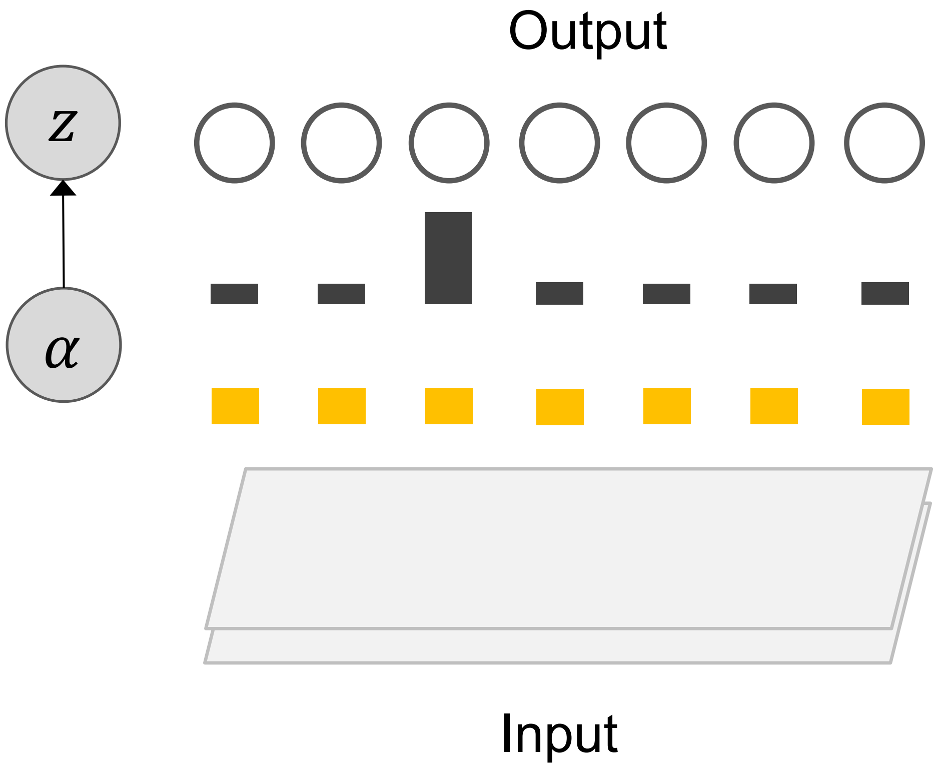


Cityscapes Training Image

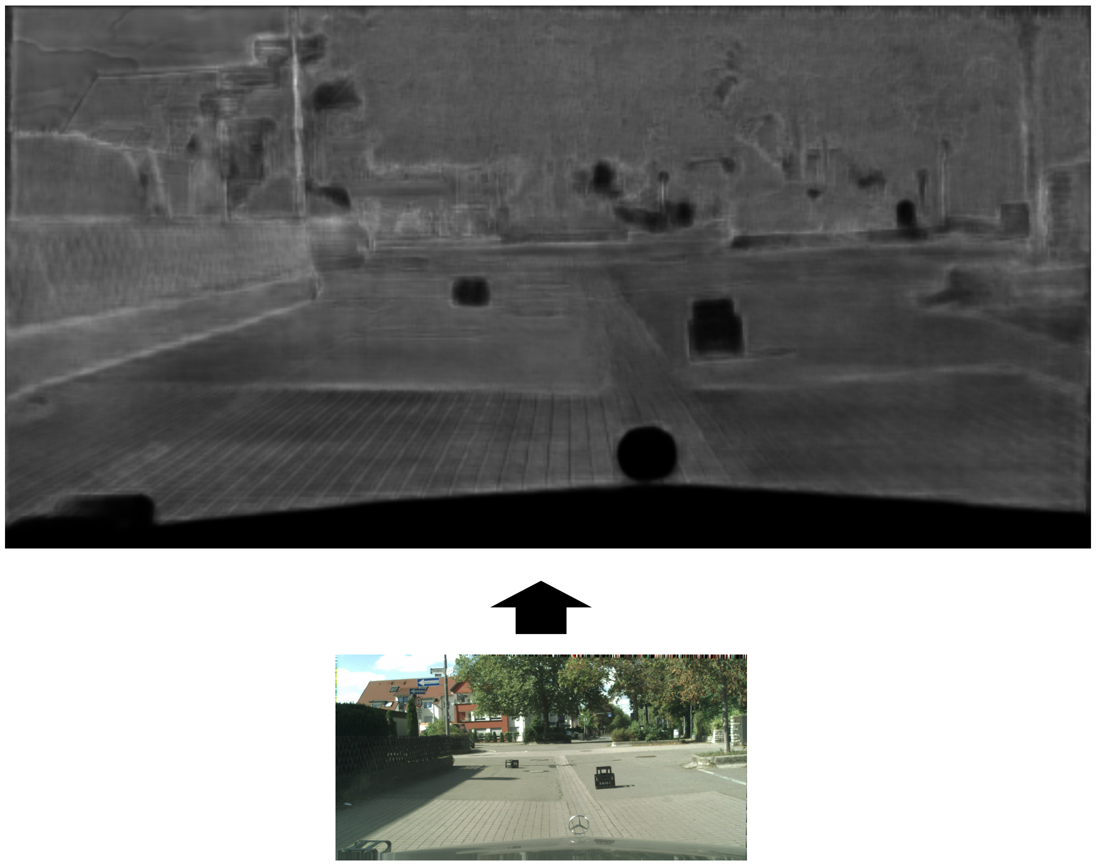


Malinin, A., & Gales, M. (2018). Predictive Uncertainty Estimation via Prior Networks.

Supervised Anomaly Learning: Dirichlet Prior



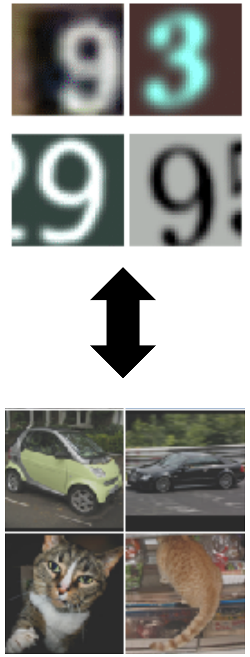
Dirichlet Entropy



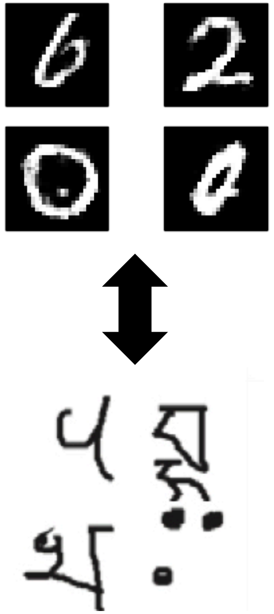
Malinin, A., & Gales, M. (2018). Predictive Uncertainty Estimation via Prior Networks.

ETH zürich

Which method works best for anomaly detection?

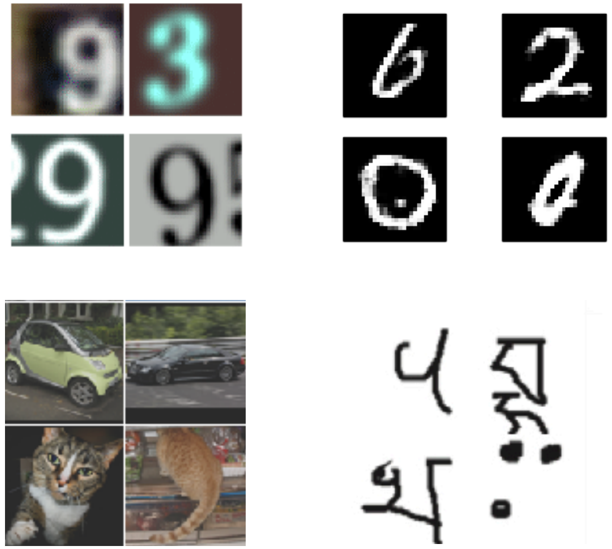


Method	SVHN vs STL-10 AUROC	MNIST vs OMNIGLOT AUROC
Dirichlet Prior		100%
Dropout		99%
Embedding	90%	
Softmax	87%	99%



Which method works best for anomaly detection?

Existing ML Research



Real World

ETH zürich

The Fishyscapes Benchmark

Does it work
on real data?

Does it work
on unkown objects?

The Fishyscapes Benchmark

What are the training requirements?

Does it work on real-world data?

Score	Method Requirements		Cityscapes	FS Lost & Found		FS Web Sept. 2019		FS Web June 2019		FS Web March 2019		FS Static	
	retraining	OoD Data	mIoU	AP	FPR_{95}	AP	FPR_{95}	AP	FPR_{95}	AP	FPR_{95}	AP	FPR_{95}
Dirichlet DeepLab *Malinin & Gales, 'Predictive Uncertainty Estimation via Prior Networks'*													
prior entropy	✓	✓	70.5	34.28	47.43	43.44	59.18	43.58	78.16	27.7	93.6	31.3	84.6

How does it perform on the original task?

Does the method generalize to diverse objects in an open world?

Method	FS Lost & Found		FS Web Sept.	Cityscapes
	AP	FPR95	AP	mIoU
Dirichlet Prior	34%	47%	43%	(70%)
Dropout	10%	38%	53%	(74%)
Embedding	5%	24%	40%	(80%)
Softmax	2%	45%	19%	(80%)

Trade-off between segmentation and anomaly detection

domain shift makes real-world dataset harder

different metrics have inverse ranking

no good method yet

Pascal VOC 2012: 97% AP
ImageNet DET 2017: 73% AP

fishyscapes.com is open for submissions!

Test and submit your method!

Part of BDL Benchmarks

Medical Diagnosis

+

Urban Driving (fishyscapes)

+

[Galaxy Zoo Challenge]

+

…

Angelos Filos, Sebastian Farquhar, Aidan N. Gomez, Tim G. J. Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon & Yarin Gal. Benchmarking Bayesian Deep Learning with Diabetic Retinopathy Diagnosis, 2018

WildDash

Zendel et al, ECCV 2018

Algorithm	Meta AVG IoU Class	Classic IoU Class	Classic iIoU Class	Classic IoU Cat.	Classic iIoU Cat.	Negative IoU Class	Impact (IoU class) Blur	Coverage	Distortion	Hood	Occl.	Overexp.	Particle	Screen	Underexp.	Variation
LDN_OE	42.7%	43.3%	31.9%	60.7%	50.3%	52.8%	-11%	-13%	-7%	-10%	-5%	-24%	0%	-6%	**-30%**	-7%
LDN_BIN	41.8%	43.8%	37.3%	58.6%	53.3%	54.3%	-14%	-14%	-22%	-14%	-3%	**-35%**	-3%	-9%	-25%	-8%
DN169_CAT_DUAL	41.0%	41.7%	34.4%	57.7%	49.7%	52.6%	-4%	-7%	-11%	-10%	-5%	-24%	-7%	-4%	**-26%**	-9%
AHiSS_ROB	39.0%	41.0%	32.2%	53.9%	39.3%	43.6%	-11%	-12%	-2%	-24%	0%	-27%	-13%	-13%	**-28%**	-16%
MapillaryAI_ROB	38.9%	41.3%	38.0%	60.5%	57.6%	25.0%	-15%	-5%	-4%	-23%	0%	-23%	-12%	-21%	**-25%**	-6%
PSP-IBN-SA_ROB	38.5%	39.4%	33.6%	60.6%	51.0%	65.3%	-18%	-3%	-5%	-18%	-3%	**-27%**	-17%	-13%	-27%	-12%
DN_2_4_CWVI_BIN_SEG	36.6%	37.9%	30.9%	52.5%	43.7%	63.5%	-16%	-7%	0%	-15%	-2%	-30%	-9%	-10%	**-41%**	-14%
IBN-PSP-SA_ROB	33.6%	34.7%	30.8%	55.1%	38.9%	68.5%	-8%	0%	0%	-22%	0%	-27%	-23%	-23%	**-36%**	-8%
IBN-PSA-SA_ROB	32.5%	33.6%	30.1%	53.8%	39.3%	69.5%	-9%	-1%	0%	-25%	0%	-28%	-25%	-20%	**-32%**	-11%
LDN2_ROB	32.1%	34.4%	30.7%	56.6%	47.6%	29.9%	-7%	-0%	-11%	-36%	0%	-37%	-16%	-24%	**-42%**	-6%
BatMAN_ROB	31.7%	31.4%	17.4%	51.9%	37.3%	36.3%	-9%	-8%	-11%	-20%	-11%	-29%	-5%	-10%	**-37%**	-6%
HiSS_ROB	31.3%	31.0%	16.3%	50.3%	34.6%	44.1%	-11%	-10%	-11%	-25%	-10%	-32%	-2%	-10%	**-44%**	-0%
DeepLabv3+_CS	30.6%	34.2%	24.6%	49.0%	38.6%	15.7%	-13%	-15%	-15%	-34%	0%	**-55%**	-17%	-23%	-53%	-6%
AdapNetv2_ROB	29.5%	28.7%	16.5%	51.5%	38.0%	43.6%	-15%	-10%	-20%	-24%	-14%	-21%	-8%	-7%	**-37%**	-7%

Softmax is a good indicator for misclassifications.

Method	WildDash + FoggyZurich + Mapillary		Cityscapes
	max J	mIoU	mIoU
Dropout	42%	(30%)	(74%)
Embedding	41%	(46%)	(80%)
Softmax	44%	(46%)	(80%)

no big difference between methods

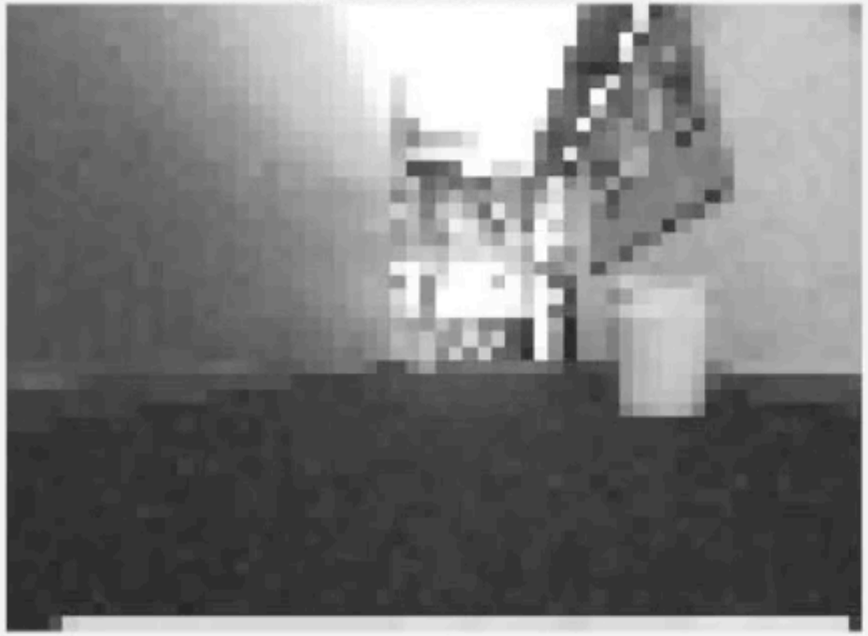
benchmarking challenge: decreasing segmentation performance can make detection easier

misclassification mixes many effects

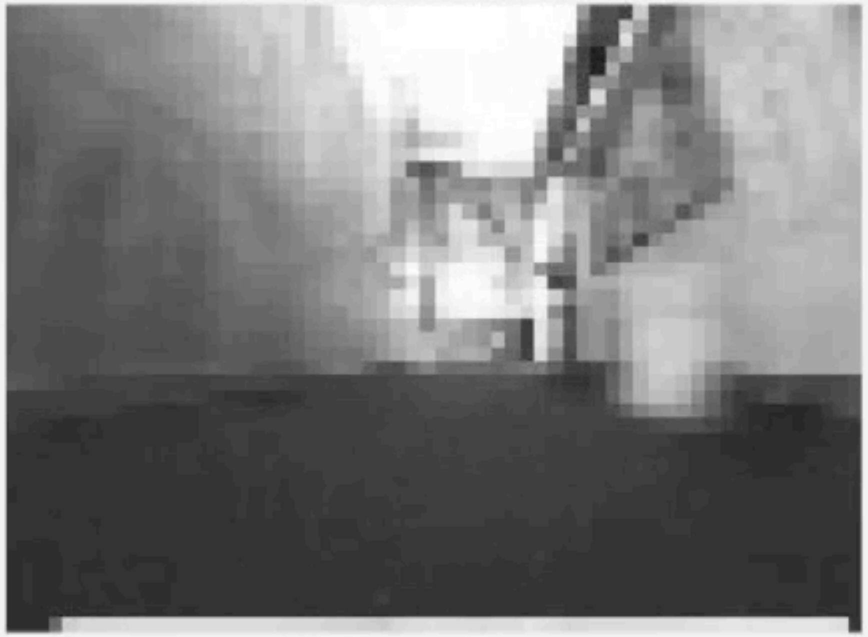
no method is much better than softmax entropy

Open-Set Learned Control

Input Image



Reconstructed Image



Reconstruction Error: 2.22e-03
Classification: Familiar

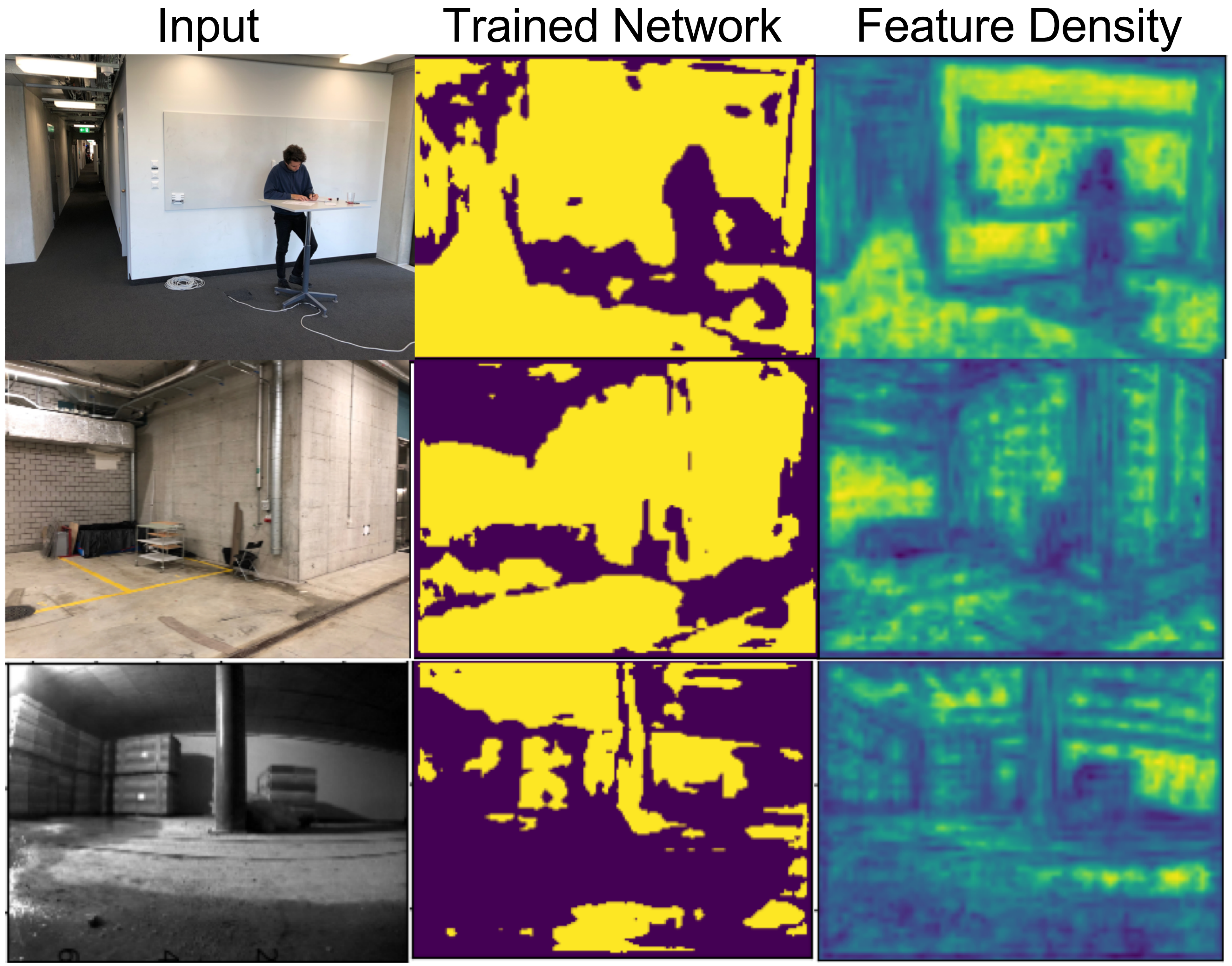
Richter & Roy, RSS'17

Open-Set Detections

Miller et al., ICRA 2018

Open-Set Segmentation

Training: Learn how background and foreground look like.



Marchal et al., arXiv, 2019

How well does uncertainty estimation actually work?

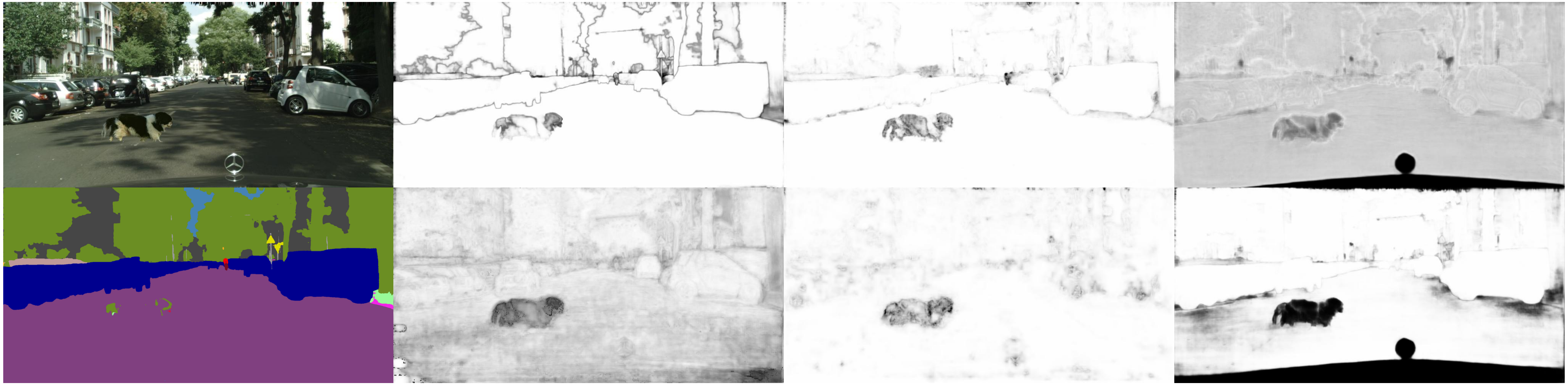
We can measure it, and measurements are clear: More work to be done!

Challenges

Match method to problem

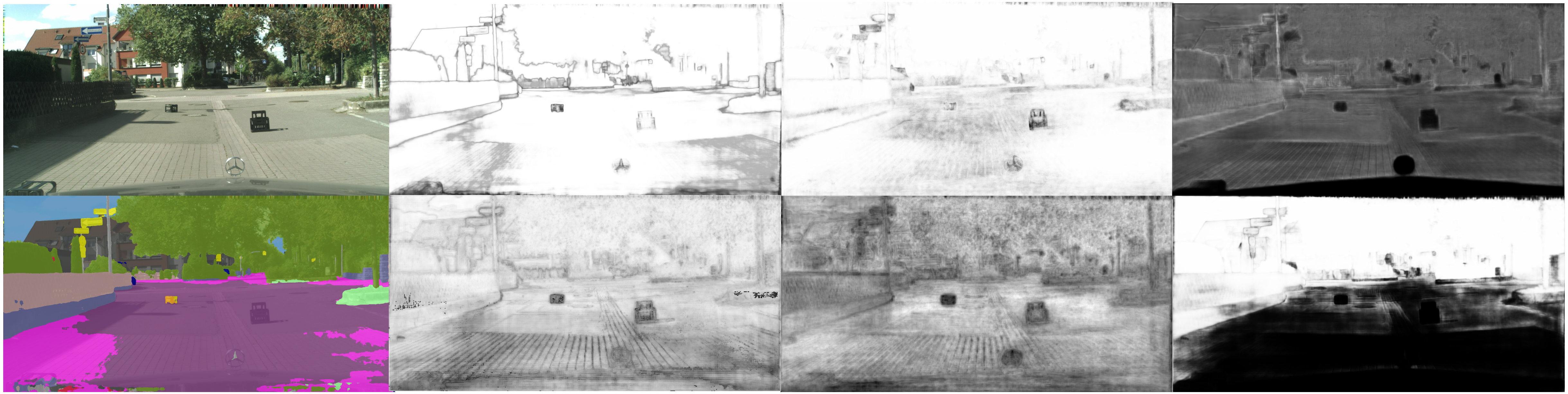
too much noise for safety

unsupervised methods lack behind



Fishyscapes Web March 2019

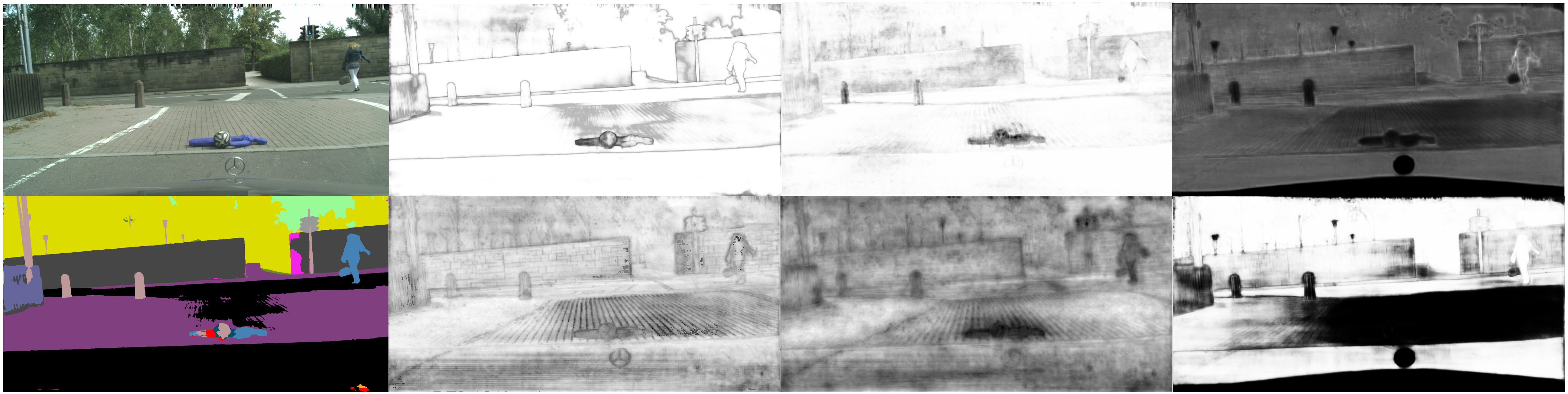
Input | Softmax Entropy | Monte-Carlo Dropout | Dirichlet Prior Network



DeepLabv3+ | kNN in Embedding | Density in Embedding | Learning a Void Class

Fishyscapes Lost & Found

Input | Softmax Entropy | Monte-Carlo Dropout | Dirichlet Prior Network



DeepLabv3+ | kNN in Embedding | Density in Embedding | Learning a Void Class

Fishyscapes Lost & Found