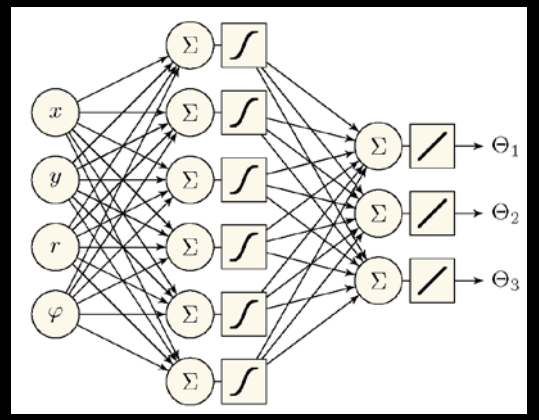
Potenzial von Deep Learning für das automatische Fahren

Uwe Franke !!et al.!!, Daimler R&D

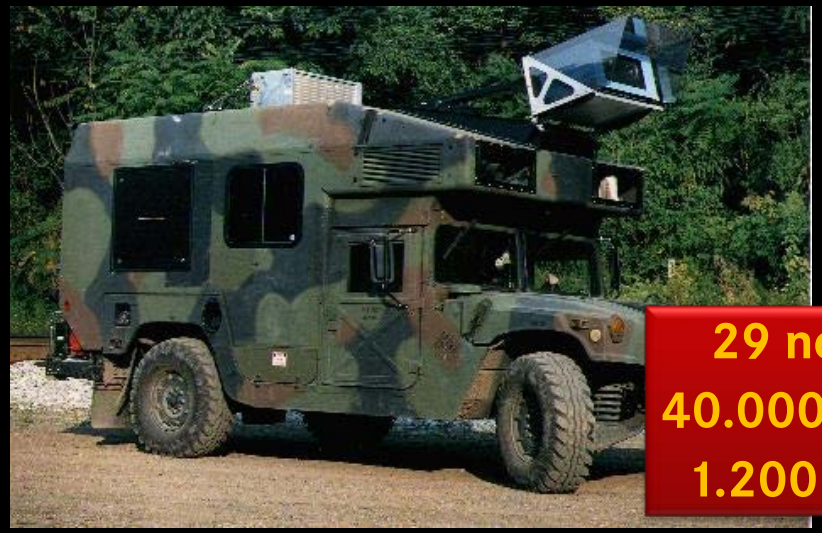
Mercedes-Benz
The best or nothing.

The Multi-Layer Perceptron (1986)

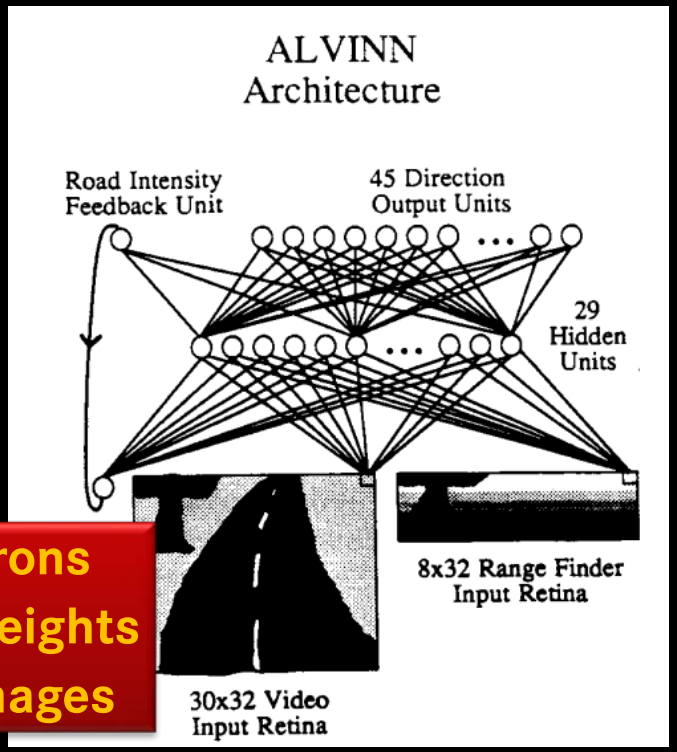


Replacing the binary activation function of the Perceptron by a non-linear function it became possible to have layer of hidden units AND to train their weights by means of error back-propagation (gradient descent).

Prominent Example: ALVINN (Dean Pommerleau, 1989)



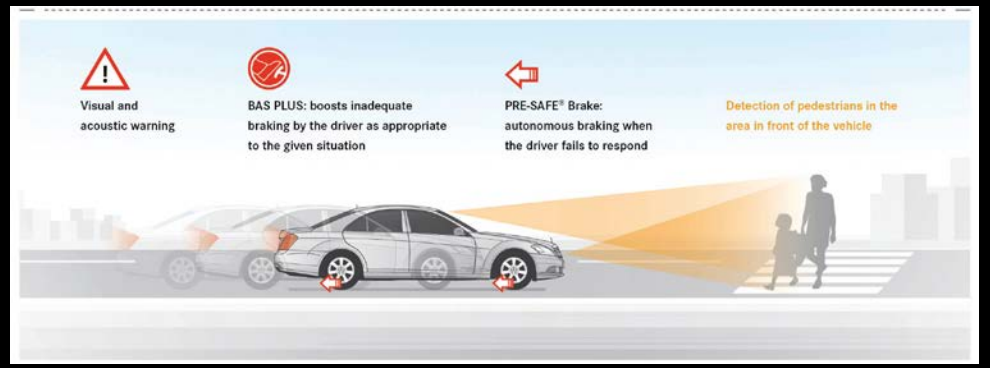
29 neurons
40.000 weights
1.200 images

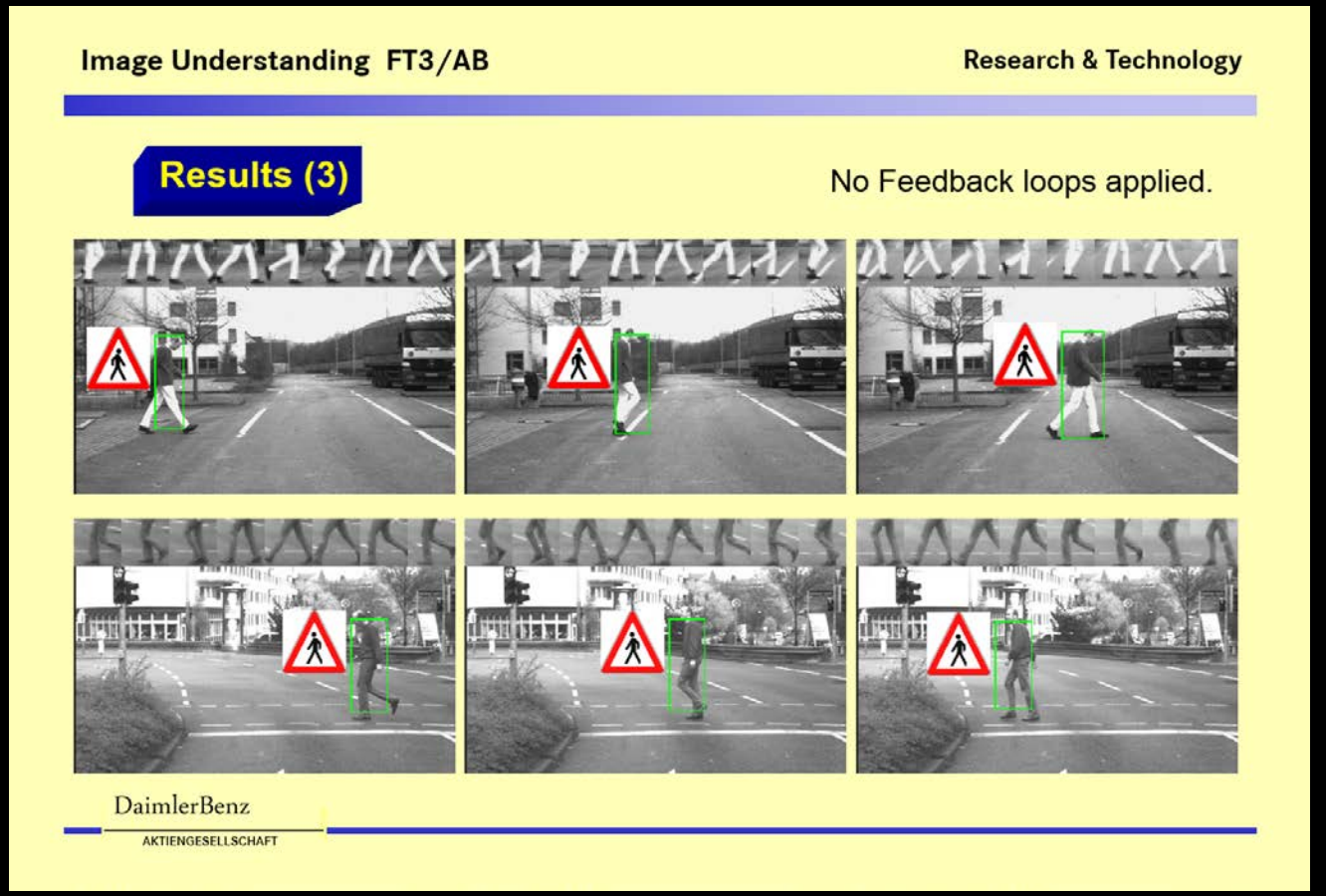


Own work: R.Neußer, J.Nijhuis, L.Spaanenburg, U.Franke, H.Fritz: „Neural Net based lateral vehicle control learned by human steering examples“, IEEE International Conference on Neural Networks, 1993

Local Receptive Fields (1998)

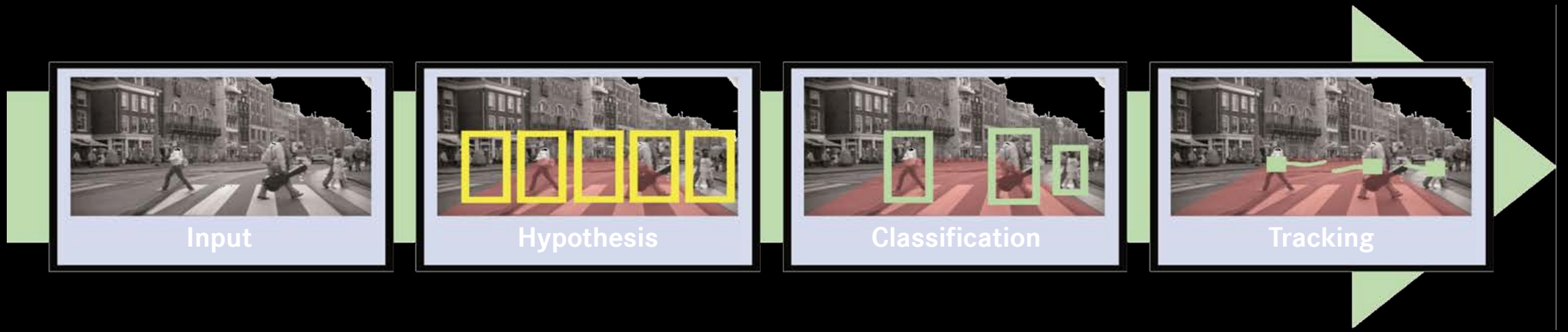
The human visual system applies the *same local filters* to all image pixel nearly independent of their position. Thus the number of weights to the hidden units can be reduced significantly and even 3D processing (temporal) becomes feasible.



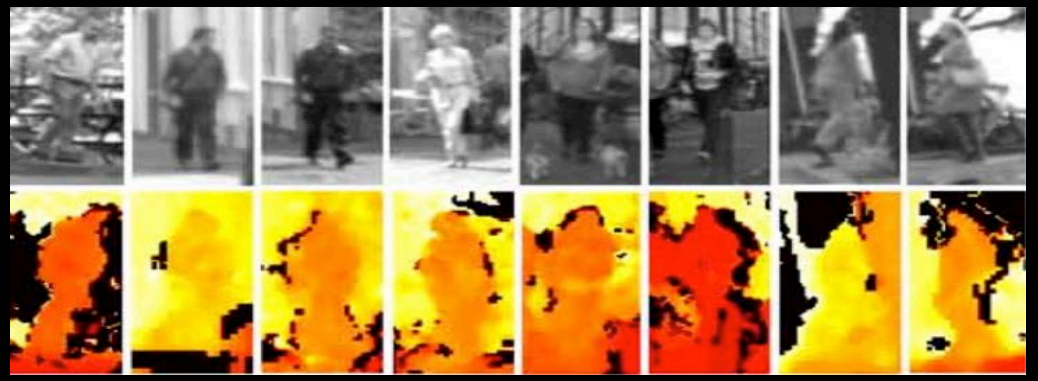


C.Wöhler, J.P.Anlauf, T.Pörtner, U.Franke: „A Time Delay Neural Network Algorithm for Real-Time Pedestrian Recognition“, IEEE Conference on Intelligent Vehicles ’98, Stuttgart

Pedestrian Recognition using Shallow Neural Nets



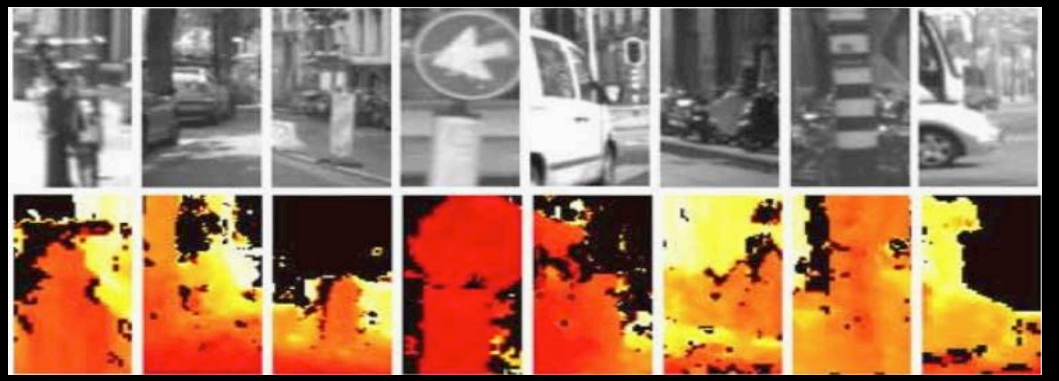
Pedestrian examples



Image

Depth

Non-Pedestrian examples



> 2 Million examples in data base

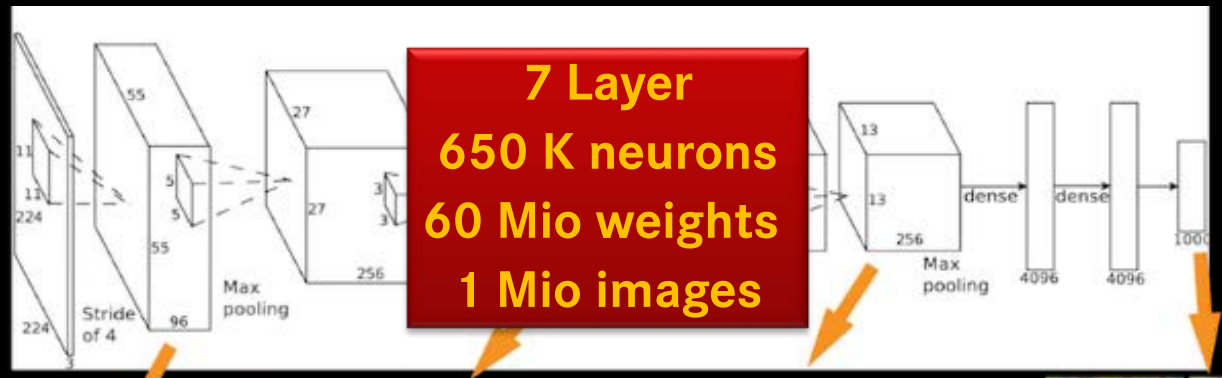
M.Enzweiler, D.M.Gavrila: "**A Multi-Level Mixture-of-Experts Framework for Pedestrian Classification**", IEEE Transactions on Image Processing, 2011

Pedestrian Recognition using Shallow Neural Nets

The AlexNet (2012)

IMAGENET Benchmark set up in 2010

- 1.000.000 images
- 1.000 categories

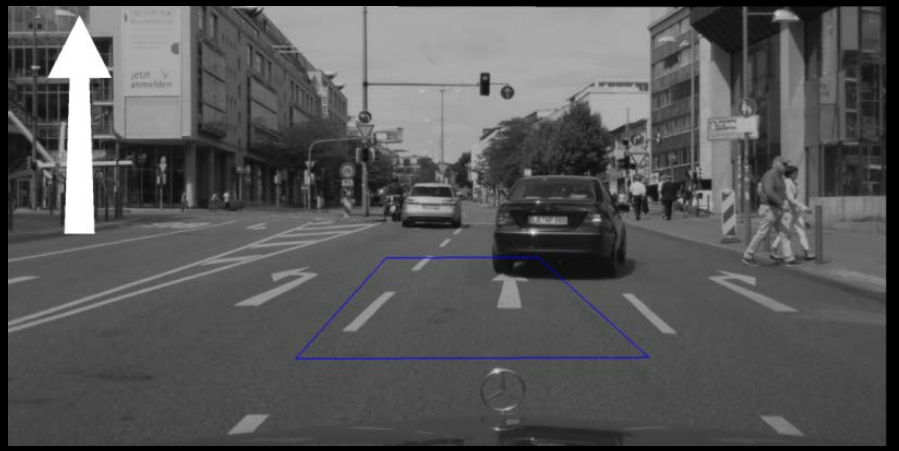


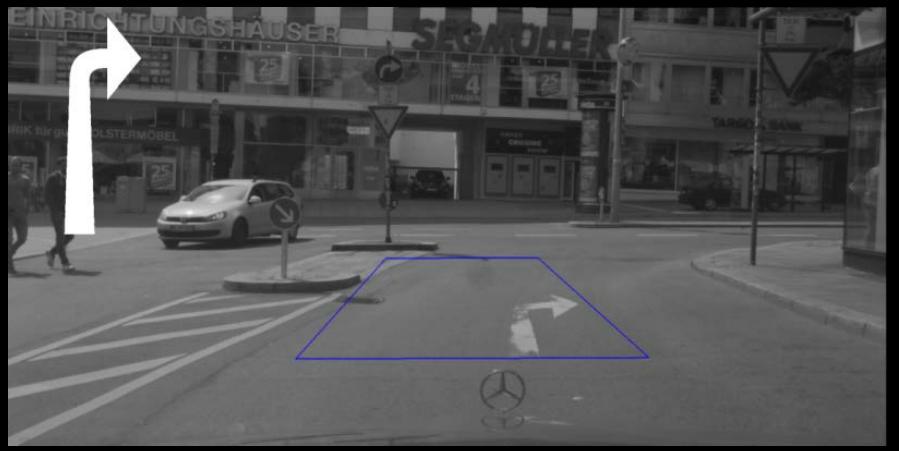
Alex Krizhevsky et al.: "**ImageNet Classification with Deep Convolutional Neural Networks**", NIPS 2012

AlexNet: Versatile and Easy to Use

Since 2014 it is common knowledge that it is easy to adapt AlexNet to any categorization problem:

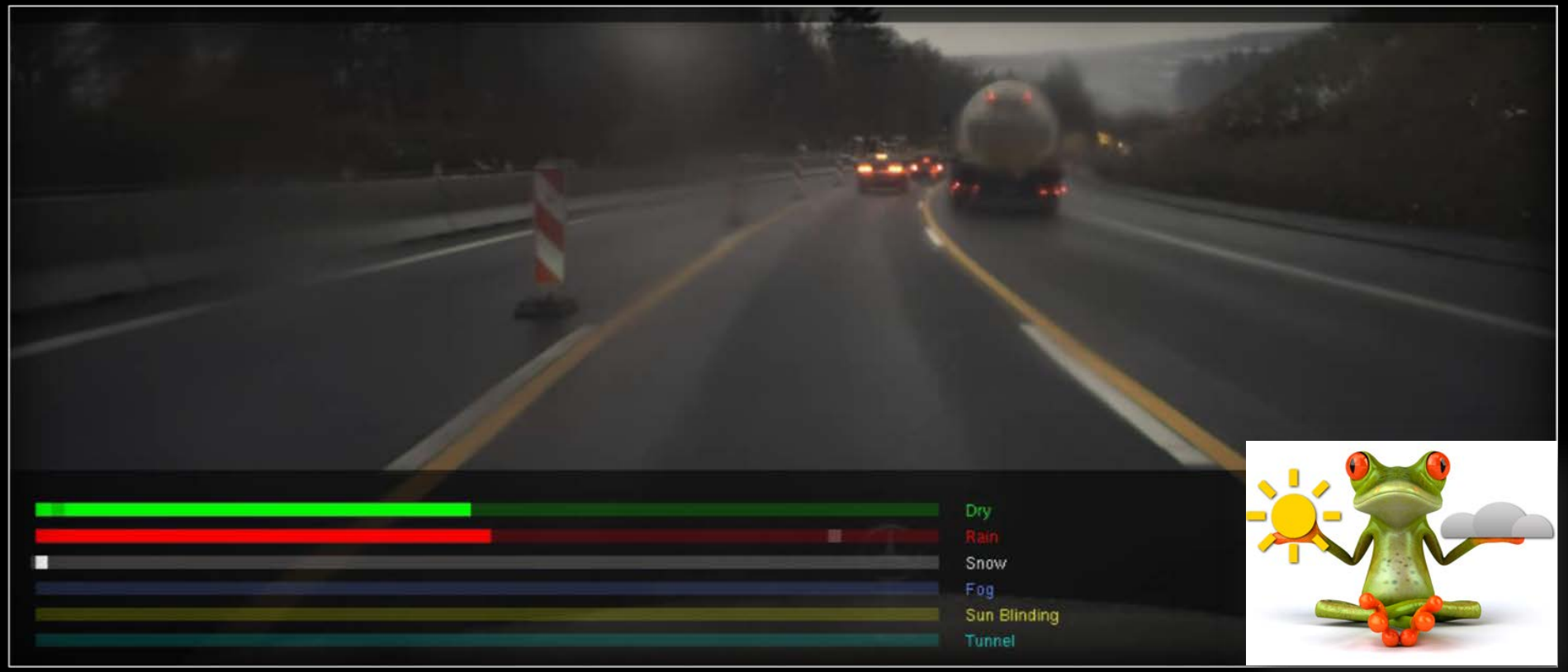
- Resize your image to 224x224 pixel
- Retrain the fully connected layers





Weather Frog – Vision Based Weather Recognition

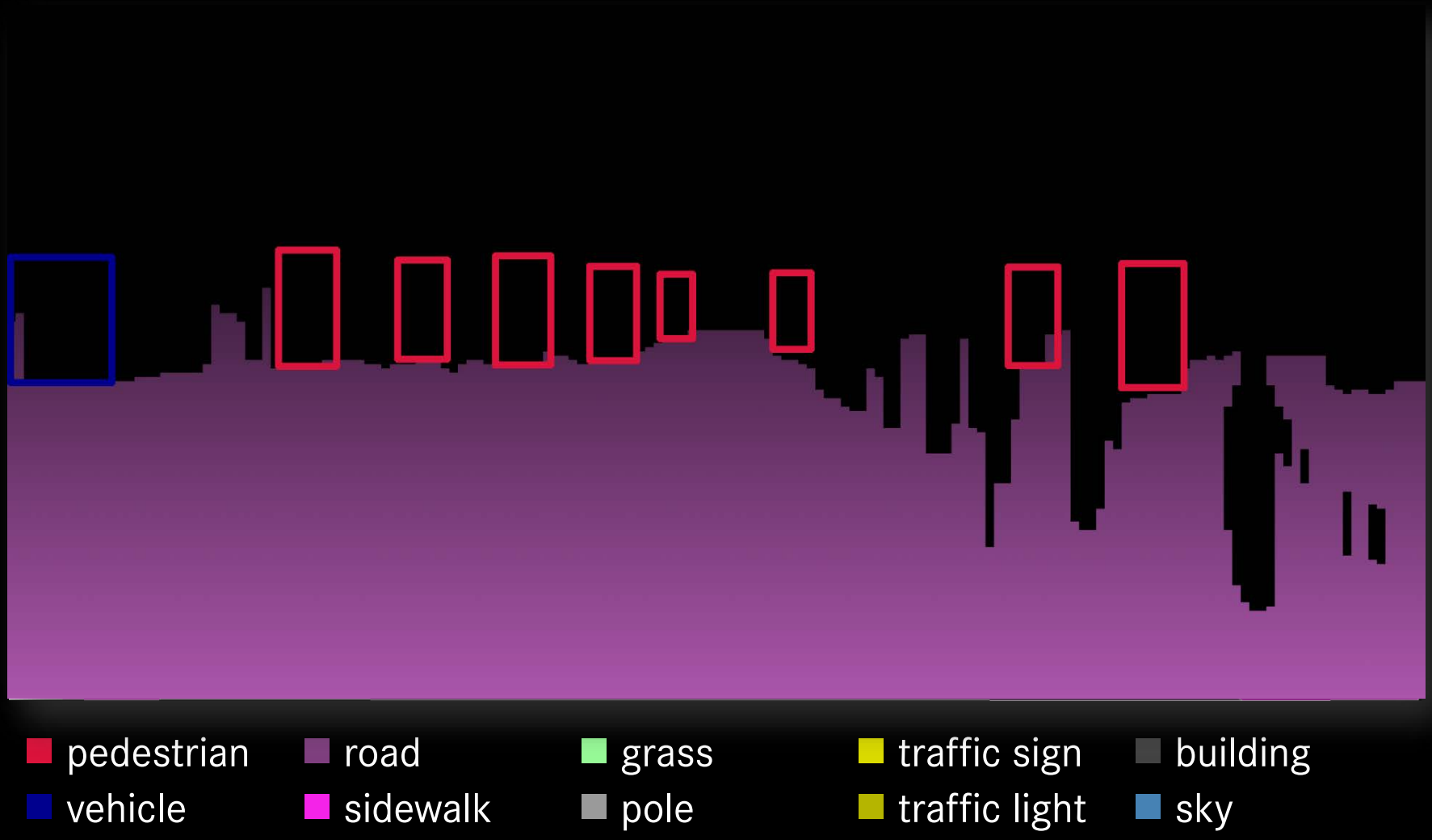
The performance of vision systems will always depend on weather and lighting conditions. Switching from a classifier using hand-crafted features to an AlexNet improves the performance and results in more classes – at the same time.



CNN-based Lane Recognition: more robust than ever

End-to-End allows to exploit all visual hints to estimate the parameters required for LaneKeeping.

We need a deeper Understanding of the World



We need a deeper Understanding of the World

pedestrian road grass traffic sign building
vehicle sidewalk pole traffic light sky

What is in the Scene?

all objects in the scene

The Cityscapes Dataset

50 major German Cities

5000 precisely labeled frames (2MPx stereo)

3 seasons

Visit: www.cityscapes-dataset.net

M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele,
"The Cityscapes Dataset for Semantic Urban Scene Understanding", CVPR 2016

The Cityscapes Dataset

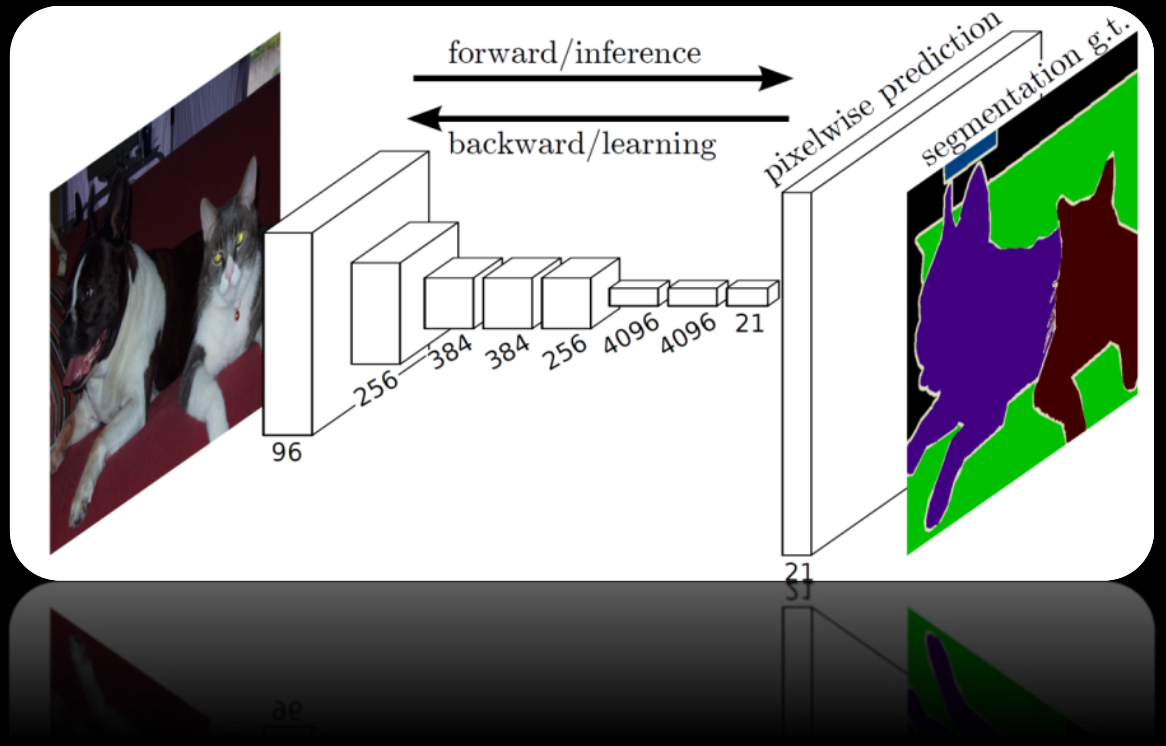
M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele,
"The Cityscapes Dataset for Semantic Urban Scene Understanding", CVPR 2016

Fully Connected Neural Networks (2015)

At CVPR 2015 Long and Shelhammer (UC Berkeley) showed how to classify EVERY pixel with reasonable computational effort.

At the same conference, Szegedy et al. published a new architecture named “GoogLeNet”, which won the 2014 ImageNet competition. This network performs 10 times faster than the often used very deep “VGG16” architecture.

16 Layer
24 Mio neurons (224x224)
128 Mio parameter



C. Szegedy et al.: “**Going deeper with convolutions**”, CVPR 2015
J. Long, E. Shelhamer, and T. Darrell: “**Fully convolutional networks for semantic segmentation**,” CVPR 2015

Progress in Scene Labeling

Our current real-time performance

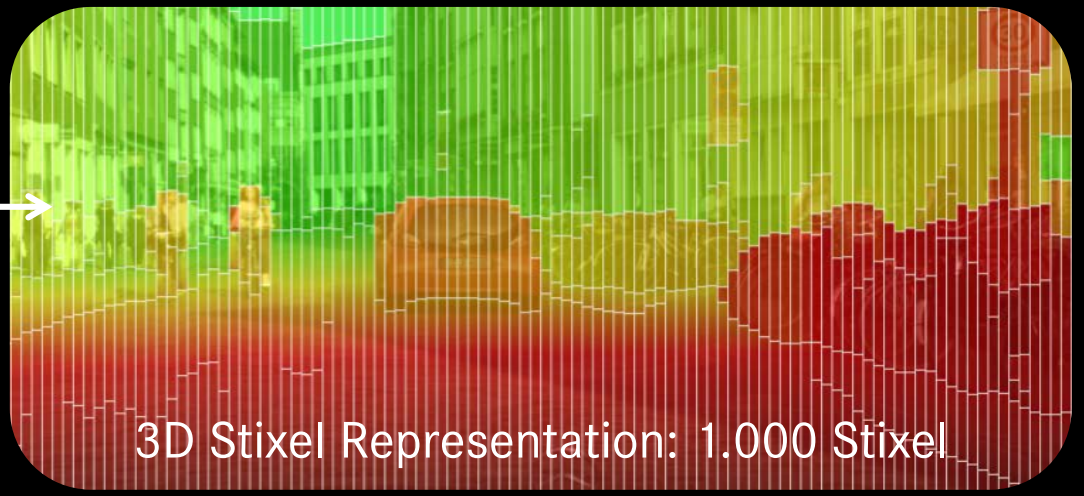
Pixel-level	IoU
Adelaide Context [Lin et al. CVPR‘16]	**71.6**
DeepLab v2 [Chen et al. arXiv’16]	70.4
Dilation 10 [Yu and Koltun ICLR’16]	67.1
DPN [Liu et al. ICCV’15]	66.8
FCN 8s [Long et al. CVPR‘15]	65.3
Pixel-level Encoding [Uhrig et al. GCPR‘16]	64.3
CRFasRNN [Zheng et al. ICCV‘15]	62.5
ENet [Paszke et al. arXiv’16]	58.3
Segnet [Badrinarayanan et al. arXiv‘15]	57

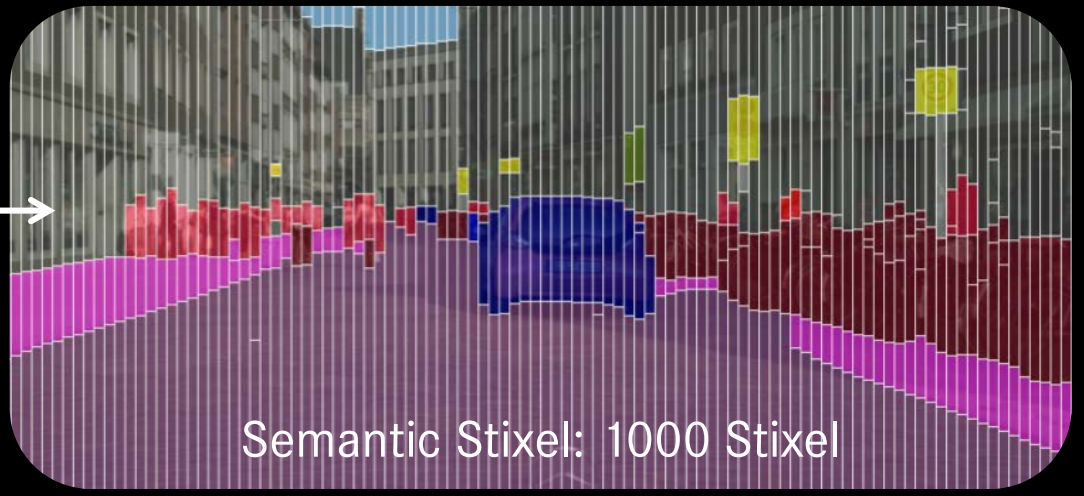
Adelaide Context [Lin et al. ‘16]

The Stixel-Representation

H.Badino, U.Franke, D.Pfeiffer: **"The Stixel World – A Compact Medium Level Representation of the 3D-World"**, DAGM Symposium 2009
D. Pfeiffer and U. Franke: **„Towards a Global Optimal Multi-Layer Stixel Representation of Dense 3D Data"**, BMVC 2011

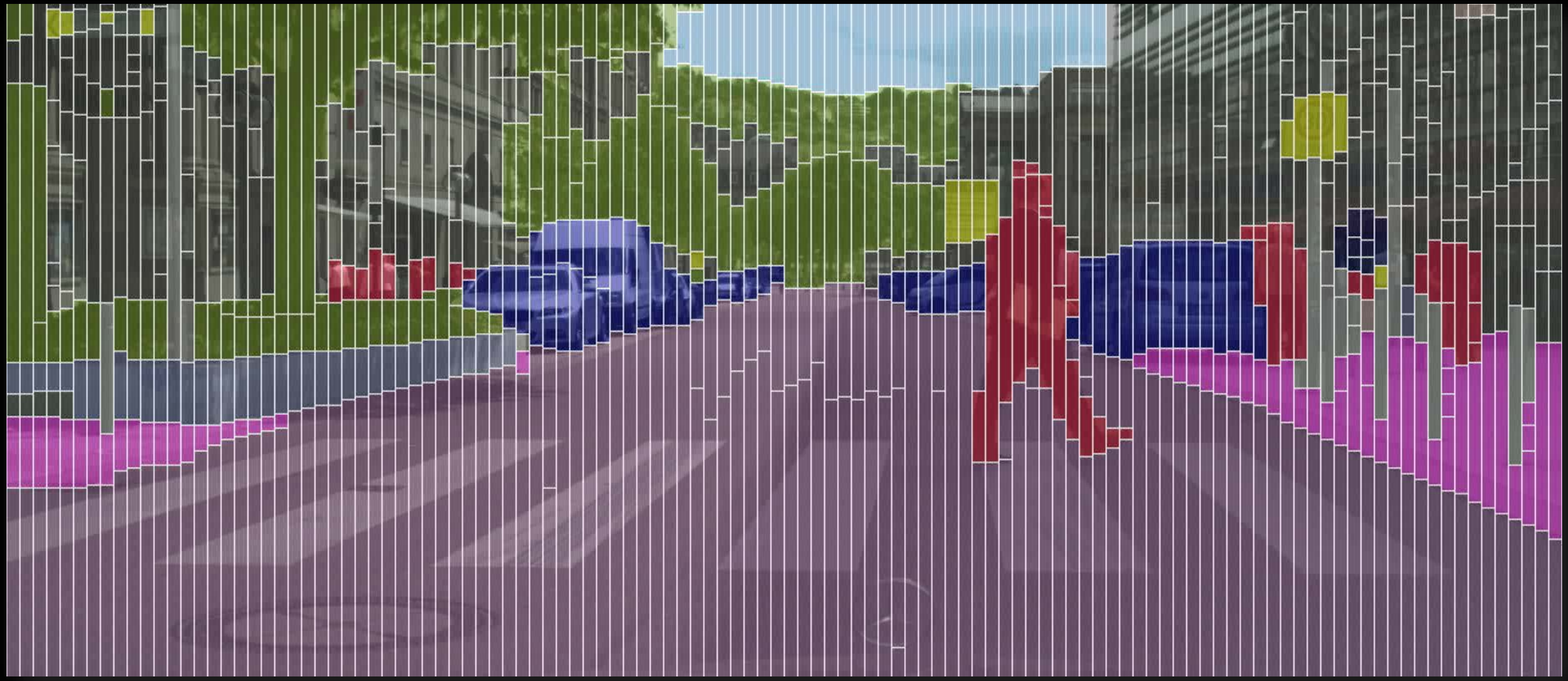
Stixel and Semantics





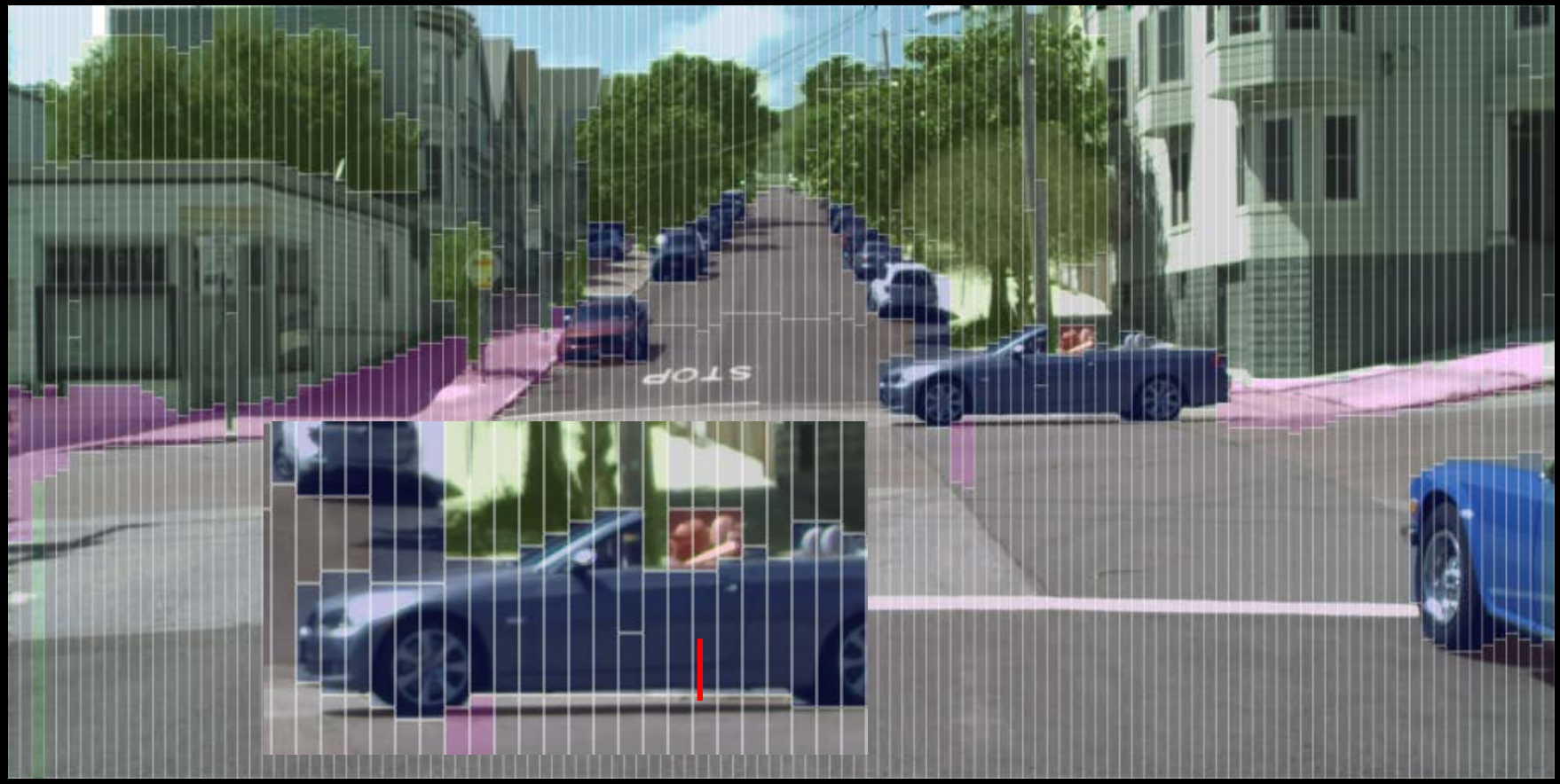
L. Schneider et al.: **"Semantic Stixel: Depth is not Enough"**, Int. Vehicles Symposium 2016 (Best Paper Award)

Semantic Stixel World in Downtown Stuttgart



Most Serious Problem... Solved!

The common Stixel-World assumes a good 3D road estimation, otherwise ghost objects may appear.



Lost! ... and Found by CNN

Bertha was blind for small objects on the road. An autonomous vehicle has to avoid any hazard on the road.

S.Ramos: **„Lost and Found: Detecting Small Road Hazards for Self-Driving Cars"**, IROS 2016

Lost! ... and Found by CNN

Bertha was blind for small objects on the road. An autonomous vehicle has to avoid any hazard on the road.

S.Ramos: **„Lost and Found: Detecting Small Road Hazards for Self-Driving Cars"**, IROS 2016

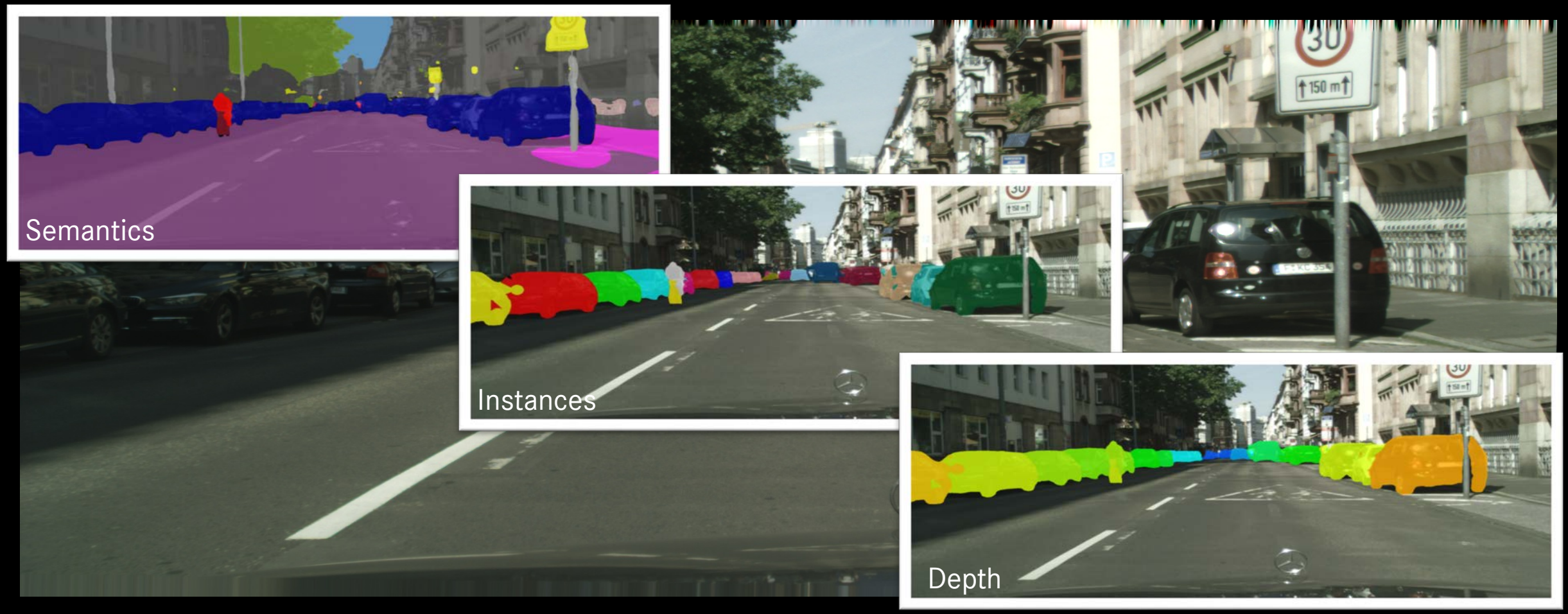
Lost! ... and Found by CNN

Bertha was blind for small objects on the road. An autonomous vehicle has to avoid any hazard on the road.

S.Ramos: „**Lost and Found: Detecting Small Road Hazards for Self-Driving Cars**”, IROS 2016

Instance Segmentation and Depth from a Single Image

What can we conclude from a single image?

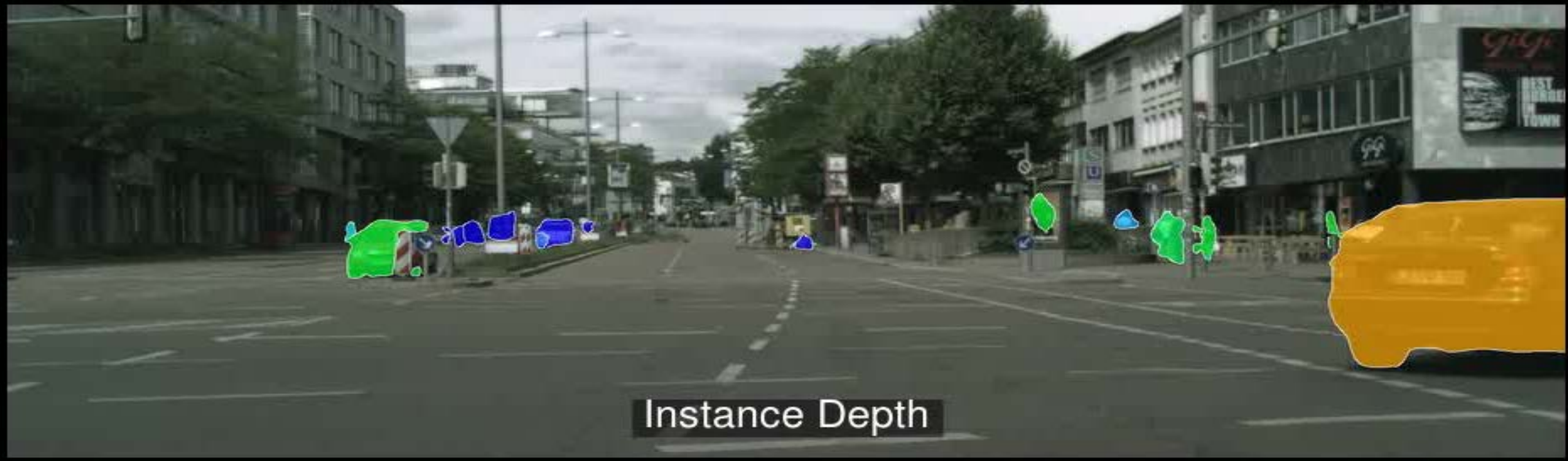


Simultaneous Estimation of Semantics, Depth and Instances

J. Uhrig et al.: **"Pixel-level Encoding and Depth Layering for Instance-level Semantic Labeling"**, GCPR 2016 (2nd best paper award)

Simultaneous Estimation of Semantics, Depth and Instances

Note: each frame is processed independently. The low temporal noise proofs the stability of the estimation.

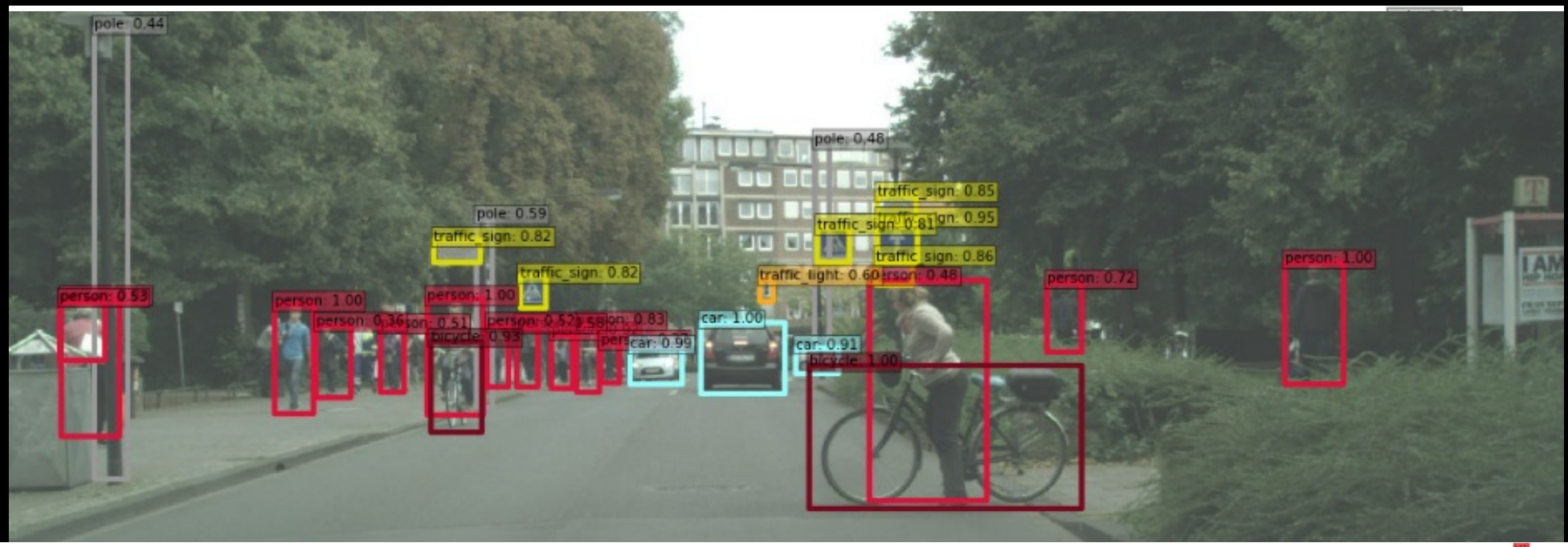


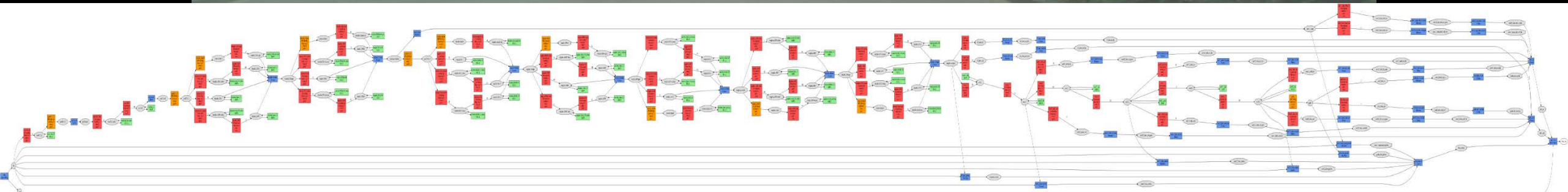
Instances of Interest: vehicles and pedestrians

J. Uhrig et al.: **"Pixel-level Encoding and Depth Layering for Instance-level Semantic Labeling"**, GCPR 2016 (2nd best paper award)

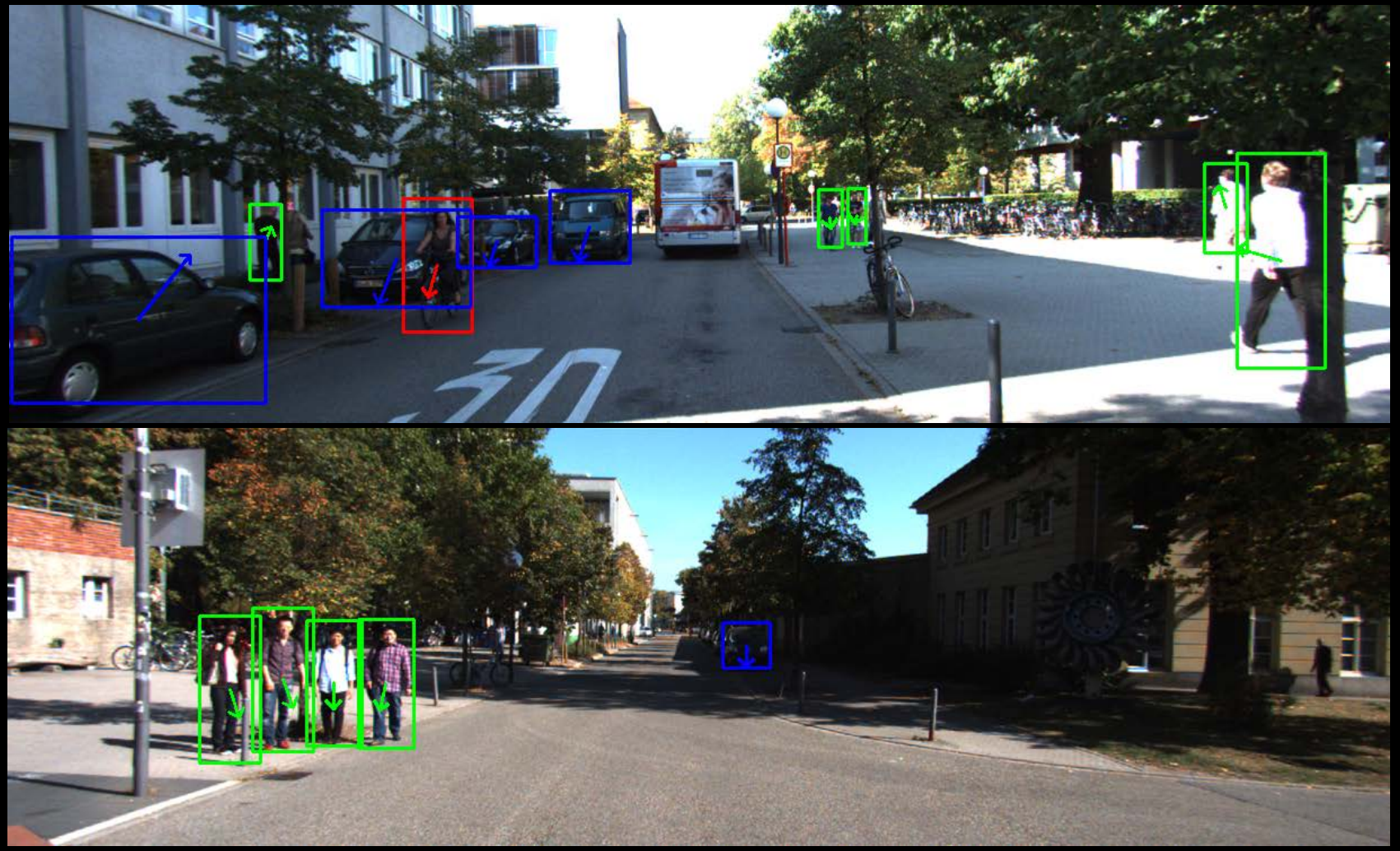
Recent Progress in Box based Object Detection

We aim at a network that does all tasks at once, including box based detection that we need for special objects like school buses, emergency vehicles etc.



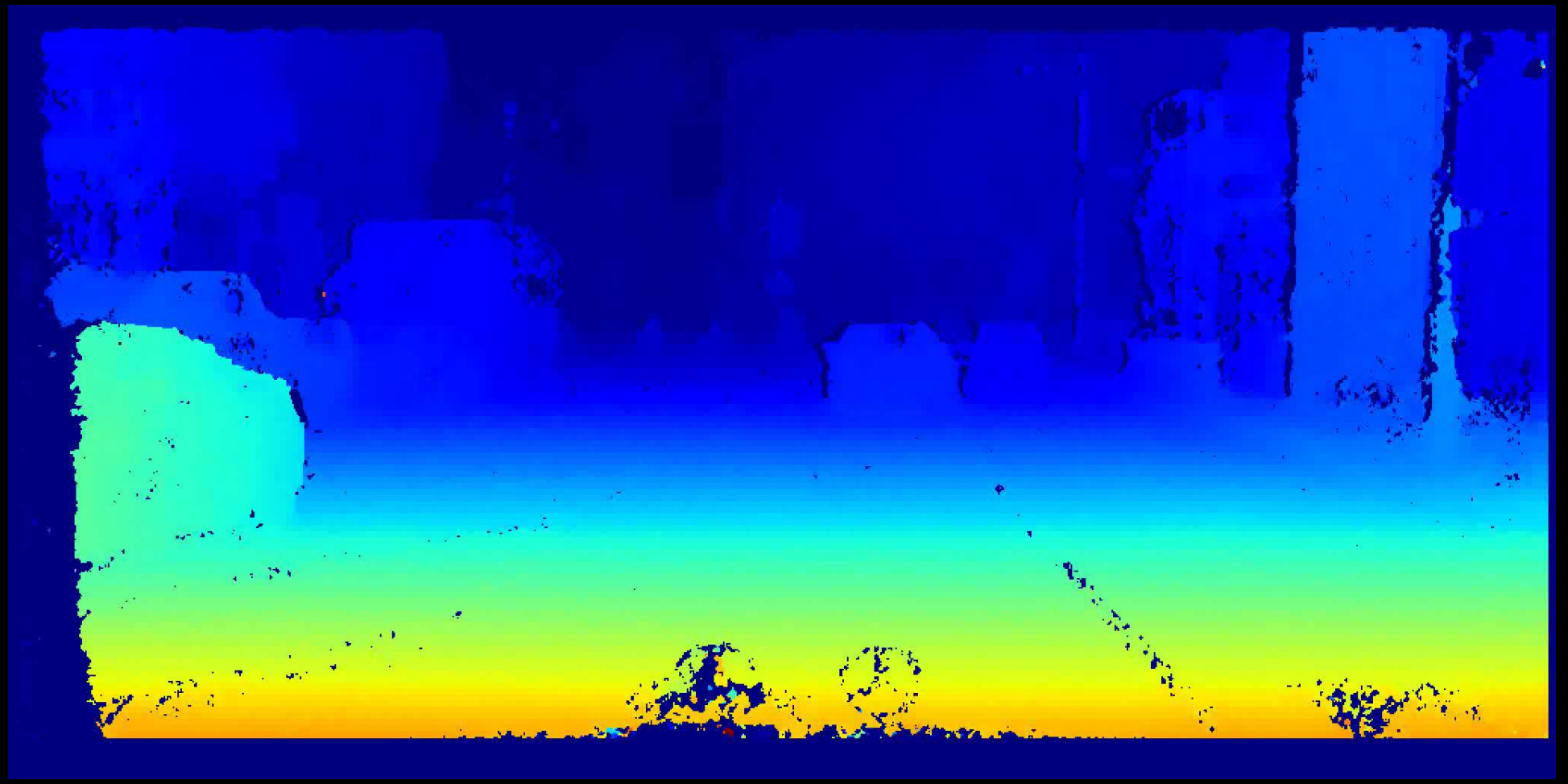


Pose and Orientation Estimation



M. Braun et al.: “**Pose-RCNN: Joint Object Detection and Pose Estimation Using 3D Object Proposals**”, ITSC 2016

Adding Depth to Scene Labeling – Does it helps?



Depth from Stereo improves the classification performance by 2…3%.

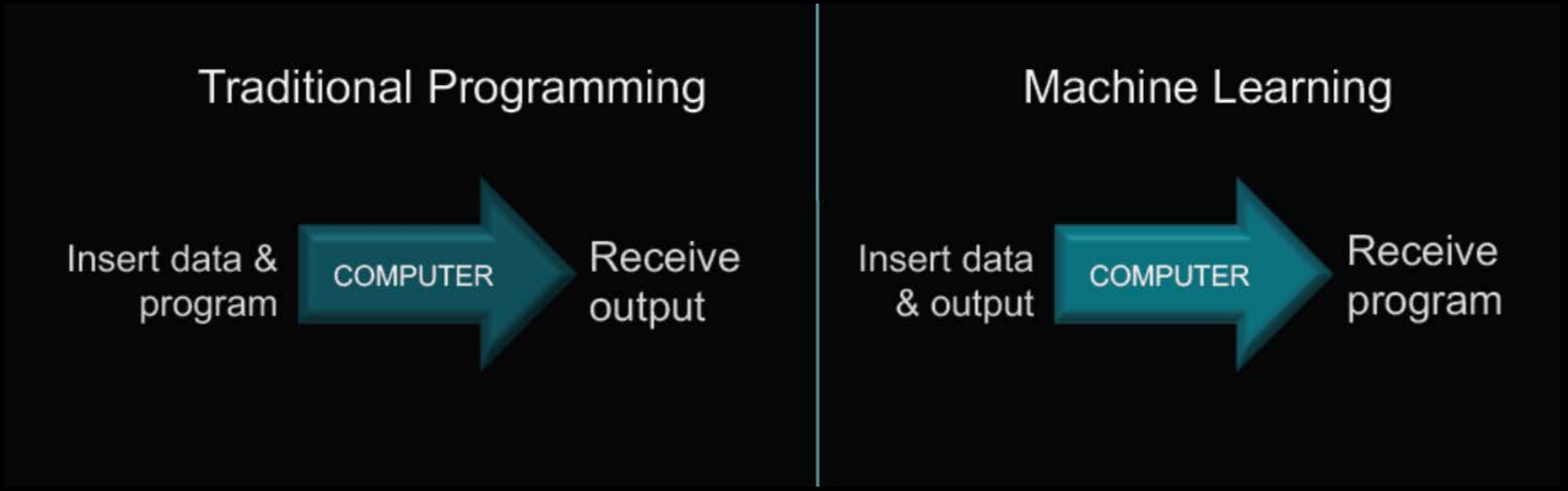
M. Jasch, L. Schneider, T. Weber, M. Rätsch: „**Fast and Robust RGB-D Scene Labeling for Autonomous Driving**“, International Conference on Systems, Control and Communications (ICSCC 2016)

Adversarial Images: Do you see the Difference?

Why had we to wait for 30 years?

Geoffrey Hinton (Univ Toronto & Google):

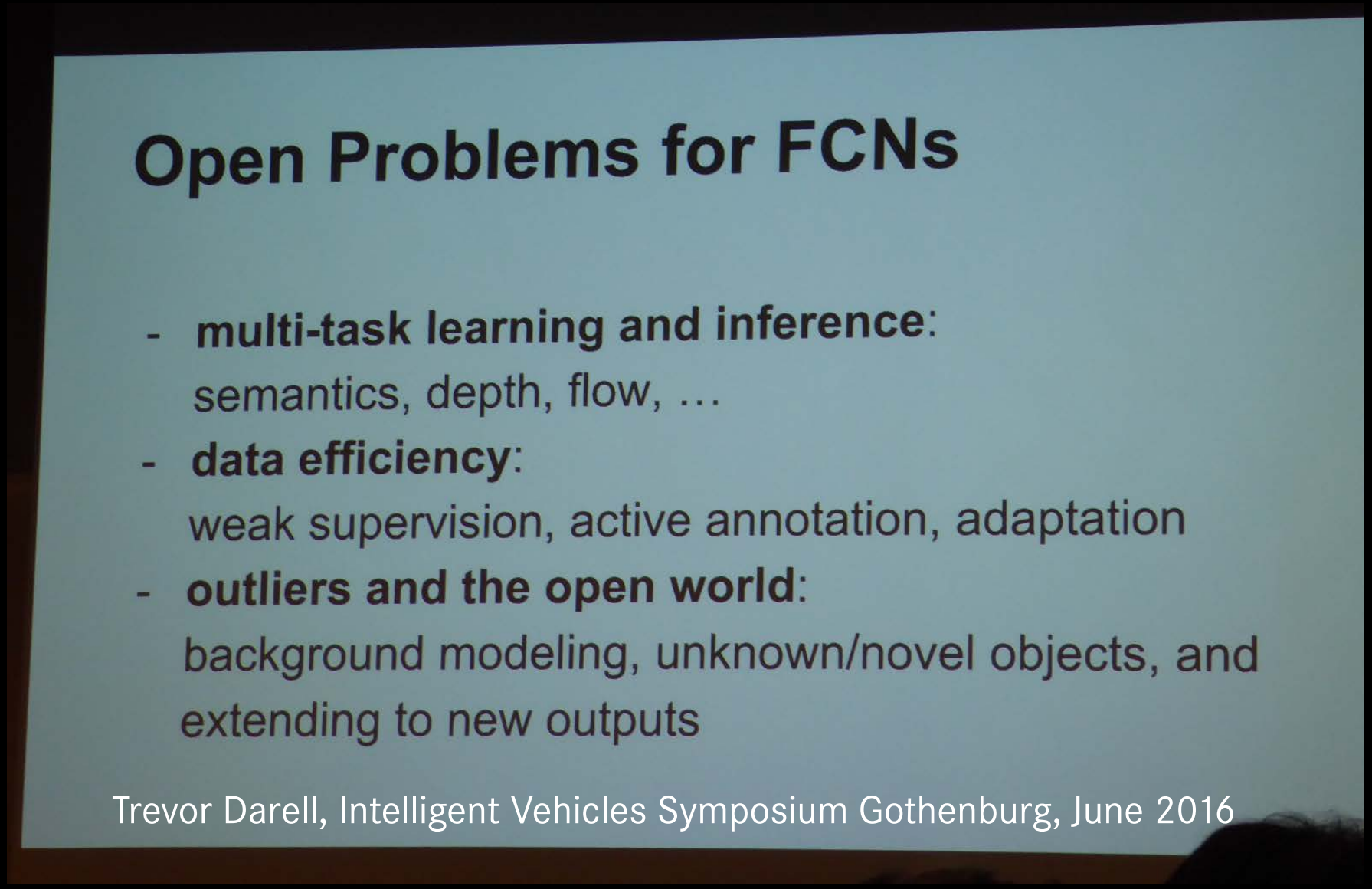
- Our labeled datasets were thousands of times too small.
- Our computers were millions of times too slow.
- We initialized the weights in a stupid way.
- We used the wrong type of non-linearity.



Lessons Learned

- Deep Neural Networks outperform every classical recognition scheme w.r.t performance and robustness.
- New HW components will reduce the power consumption of today's GPUs from kWs to acceptable numbers.
- Deep Neural Networks will drive our autonomous vehicles of the future.

What is next?


Open Problems for FCNs
- multi-task learning and inference:
semantics, depth, flow, …
- data efficiency:
weak supervision, active annotation, adaptation
- outliers and the open world:
background modeling, unknown/novel objects, and
extending to new outputs
Trevor Darell, Intelligent Vehicles Symposium Gothenburg, June 2016

The Questian asked by Thorsten Fleischer...