



VI Jornada de Encuentros Doctorales LSI

Deep learning applied to driving environments

Student:

- **Carlos Guindel Gómez**

Supervisors:

- **Ph.D. José María Armingol**
- **Ph.D. David Martín**

Laboratorio de Sistemas Inteligentes | Intelligent Systems Lab
Universidad Carlos III de Madrid



-
- **Deep learning basics**
 - **Applications in driving environments**
 - **Future prospects**
 - **Conclusions**



-
- **Deep learning basics**
 - Applications in driving environments
 - Future prospects
 - Conclusions

- **Why deep learning?**
 - What is this?



- Why deep learning?
 - Query



Source: BVLC Caffe demo, <http://demo.caffe.berkeleyvision.org>

- Why deep learning?
 - Result (after ~0.06 seconds)



Maximally accurate

Maximally specific

car

motor vehicle

self-propelled vehicle

wheeled vehicle

vehicle

Source: BVLC Caffe demo, <http://demo.caffe.berkeleyvision.org>

- Why deep learning?
 - Query



Source: BVLC Caffe demo, <http://demo.caffe.berkeleyvision.org>

- **Why deep learning?**
 - Result (after 0.062 seconds)



Maximally accurate

Maximally specific

golfcart

motor vehicle

self-propelled vehicle

wheeled vehicle

golf equipment

Source: BVLC Caffe demo, <http://demo.caffe.berkeleyvision.org>

- Why deep learning?
 - Query



Source: Microsoft Captionbot: <https://www.captionbot.ai/>

- Why deep learning?
 - Result



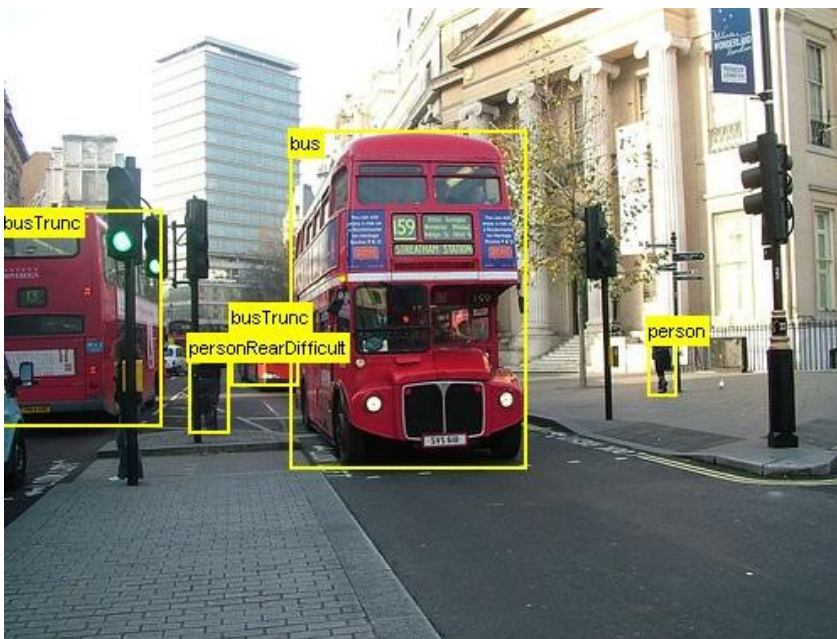
I think it's a group of people standing on the side of a road.



Source: Microsoft Captionbot: <https://www.captionbot.ai/>

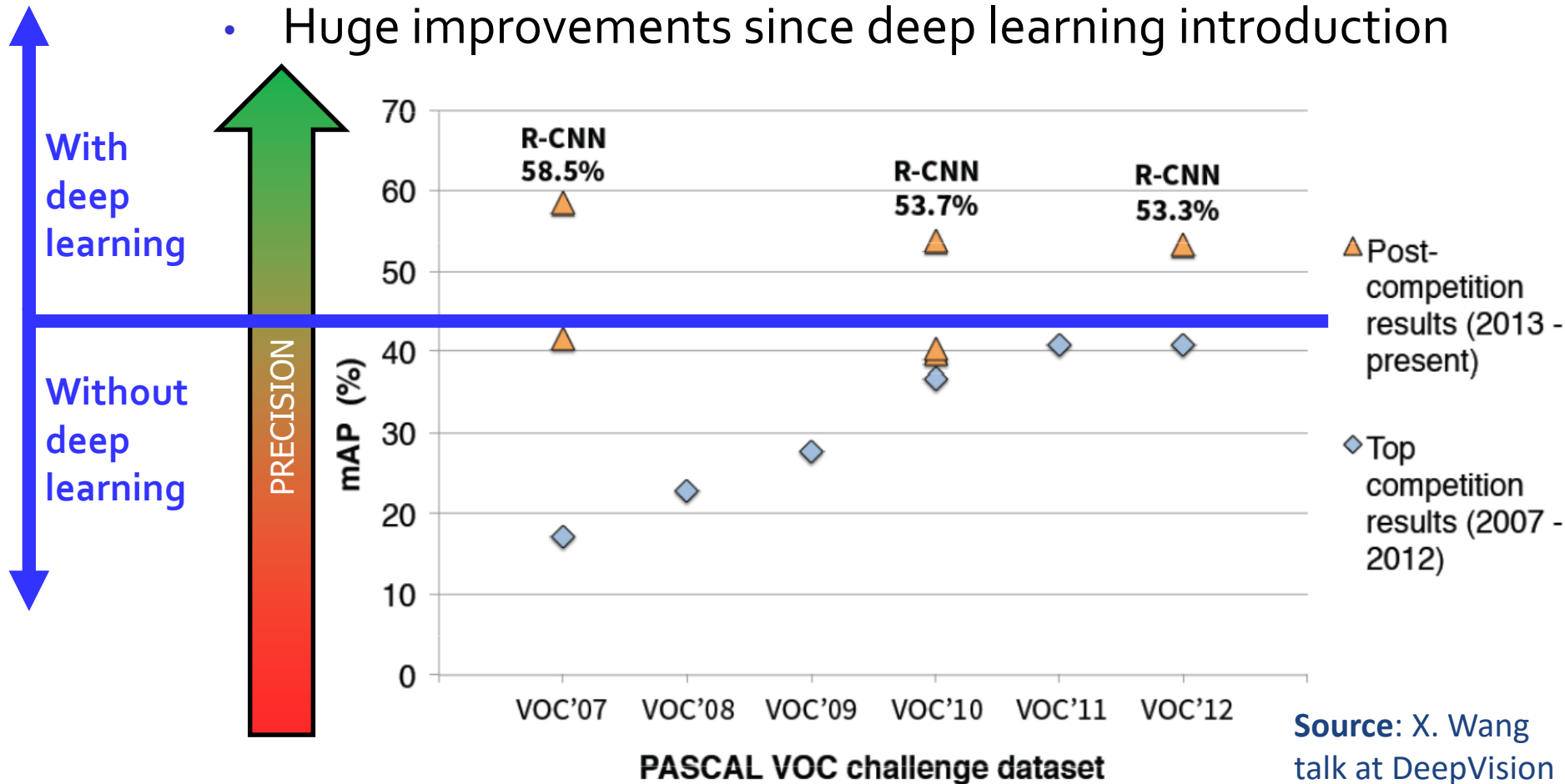
○ Why deep learning?

- Image classification and object detection challenges: comparable results across scientific works
- PASCAL Visual Object Classes Challenge: 20 classes



Why deep learning?

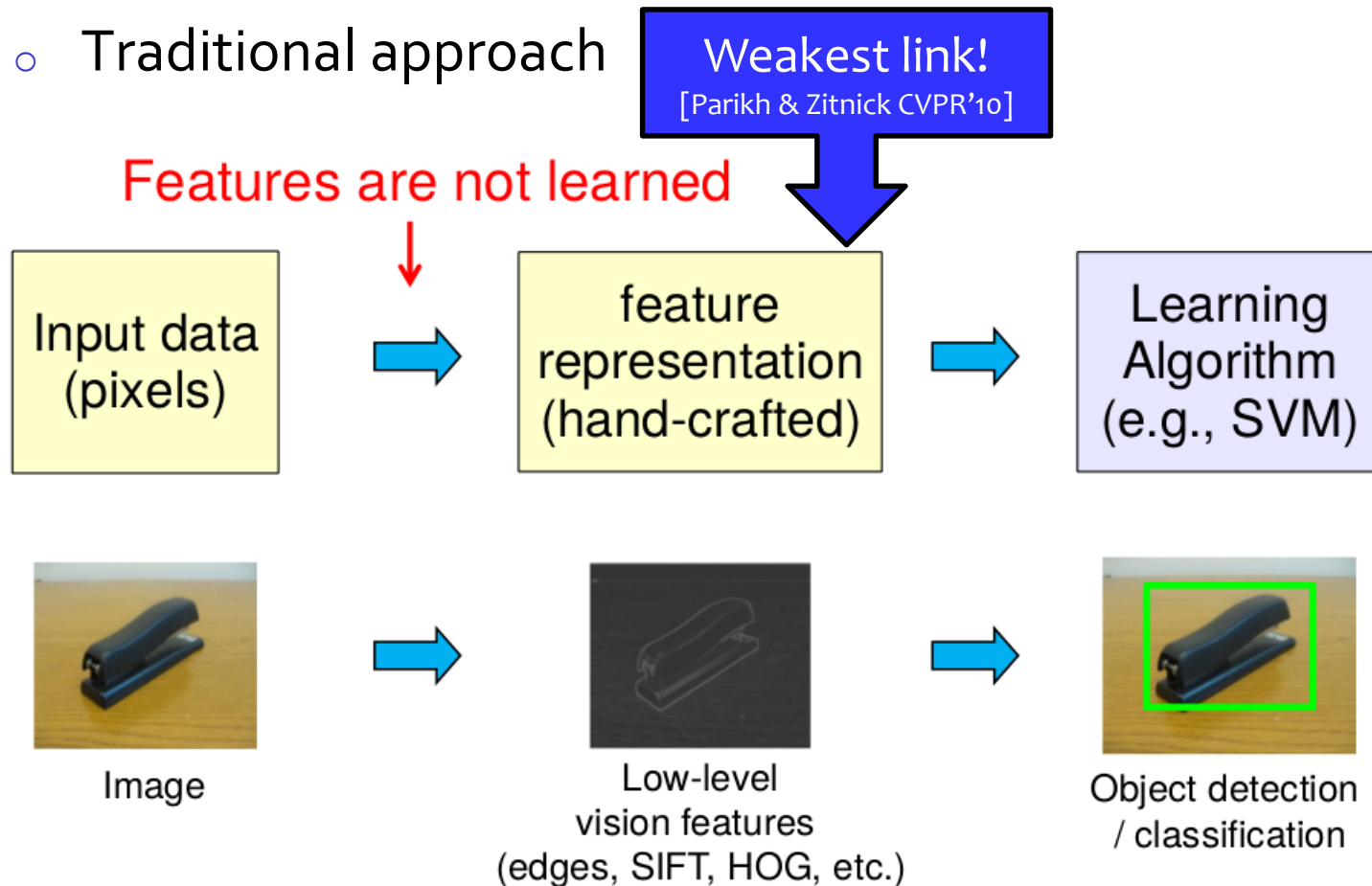
- Huge improvements since deep learning introduction



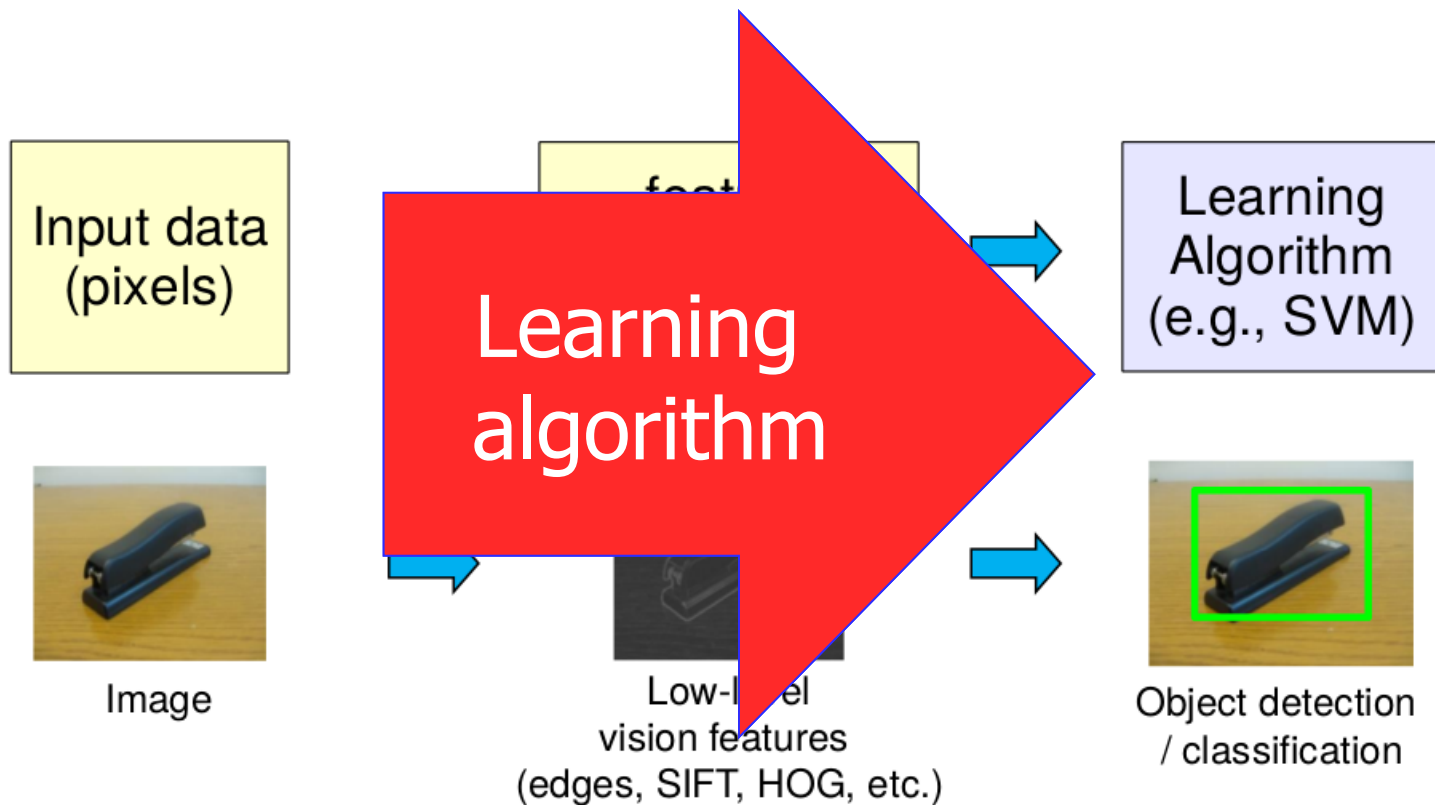
Source: X. Wang talk at DeepVision 2015 (CVPR 2015)

- **What is deep learning?**

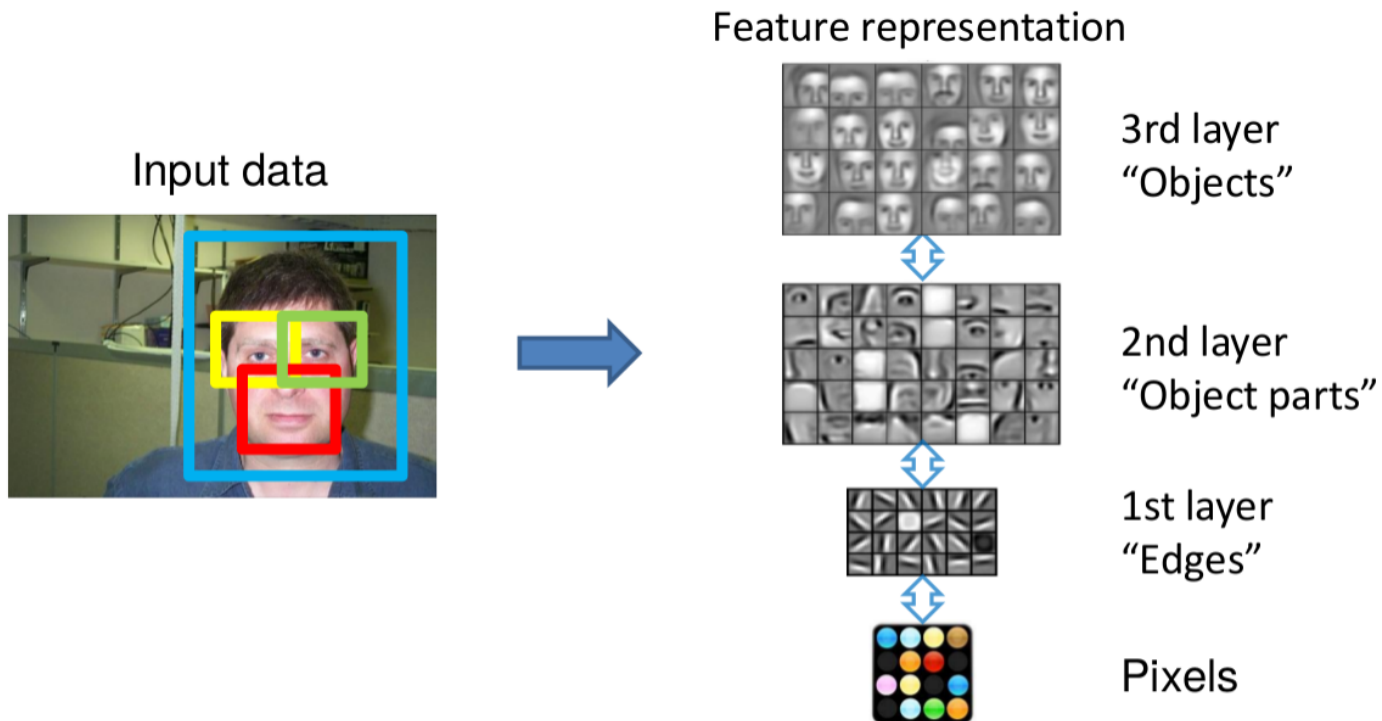
- Traditional approach



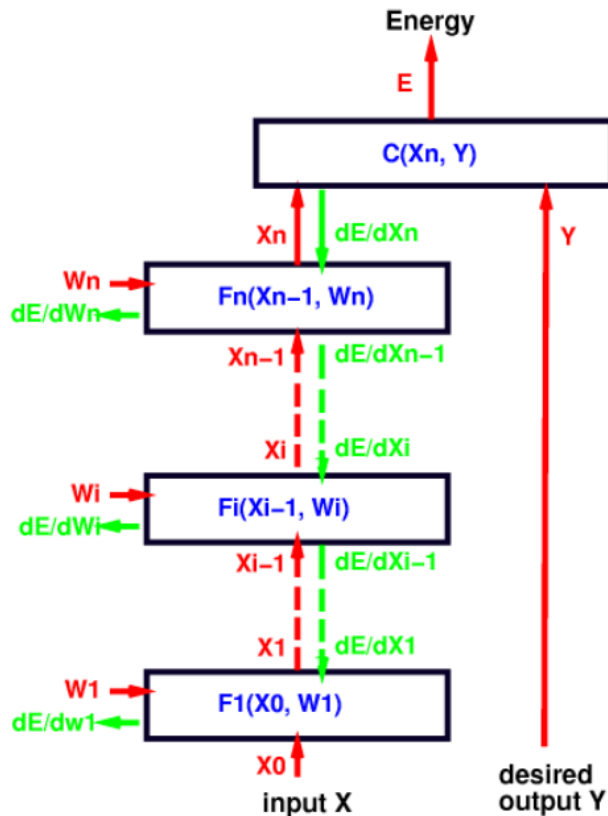
- **What is deep learning?**
 - New approach: deep learning



- **What is deep learning?**
 - **End-to-end learning** of deep architectures
 - They can learn a **hierarchy** of representations...

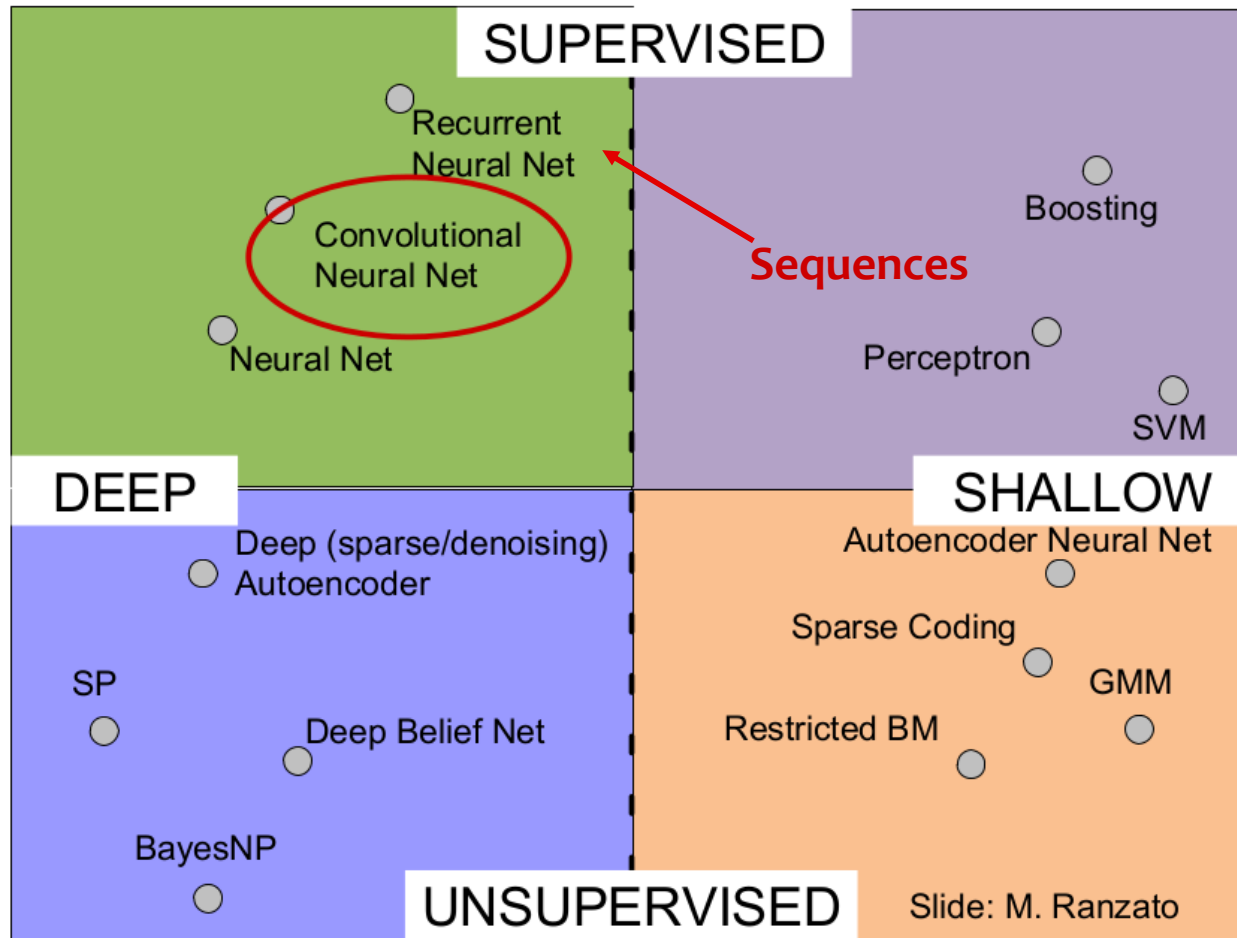


- **What is deep learning?**
 - ... and they are the best representations!



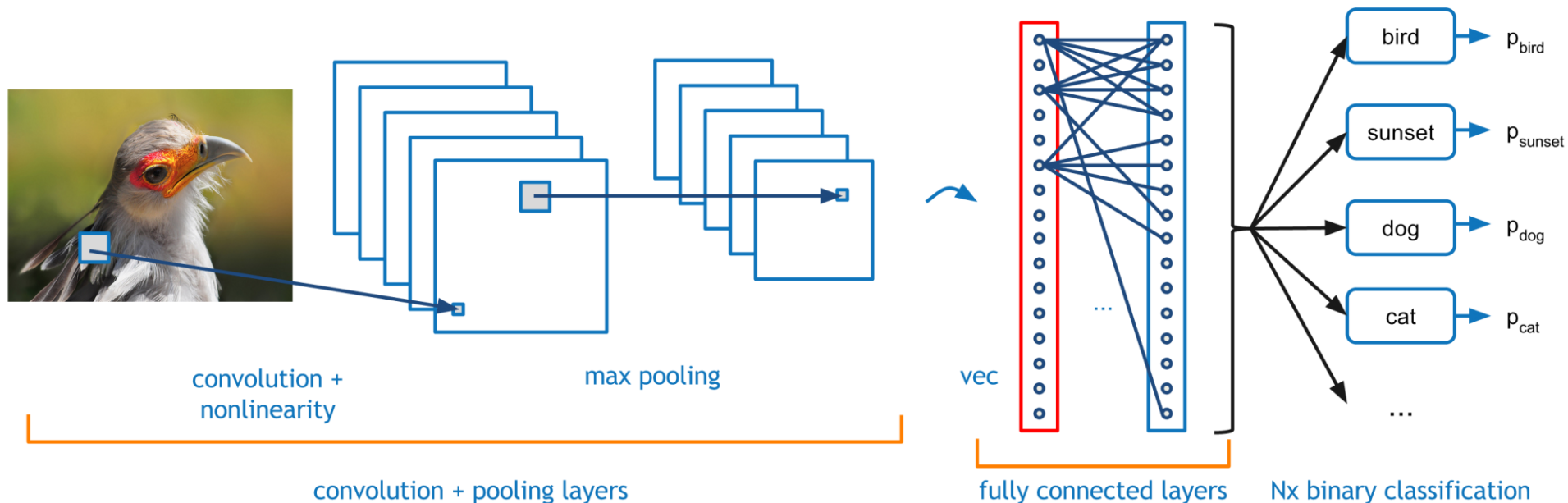
- Model parameters are jointly learnt through **back-propagation** to **optimize** the output for the task

- Deep learning in computer vision: CNNs



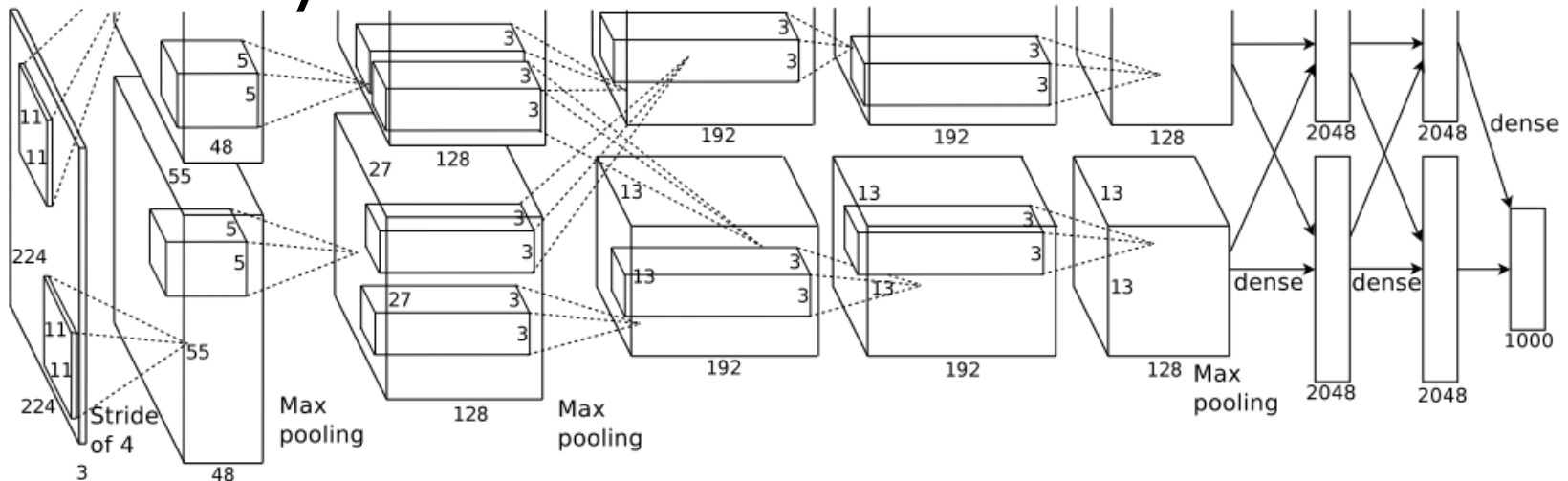
Source: P. Sermanet, Object Detection with Deep Learning, CVPR 2014 Tutorial

- Deep learning in computer vision: CNNs
 - Convolutional Neural Networks



Source: Flickr, Introducing: Flickr PARK or BIRD,
<http://code.flickr.net/2014/10/20/introducing-flickr-park-or-bird/>

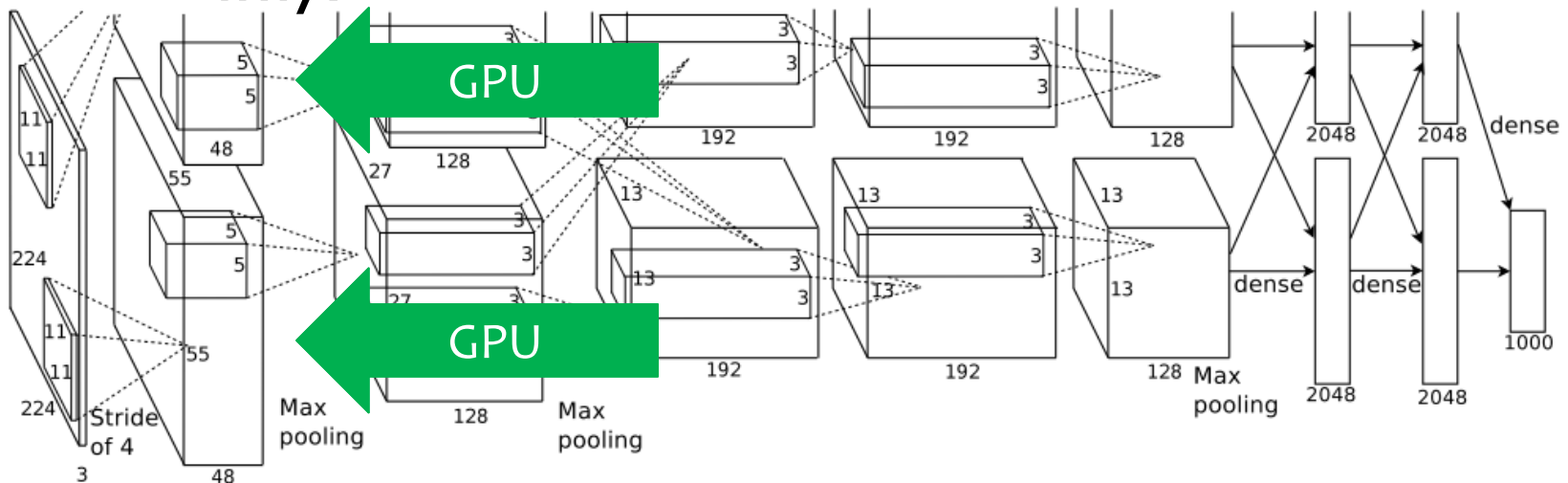
- **Deep learning in computer vision: CNNs**
 - They are basically huge neural networks (with improvements)...
 - ...but they could not be effectively trained until 2012, **why?**



A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in NIPS 2012, pp. 1097–1105.

- **Deep learning in computer vision: CNNs**
 - They are basically huge neural networks (with improvements)...
 - ...but they could not be effectively trained until 2012,

why?

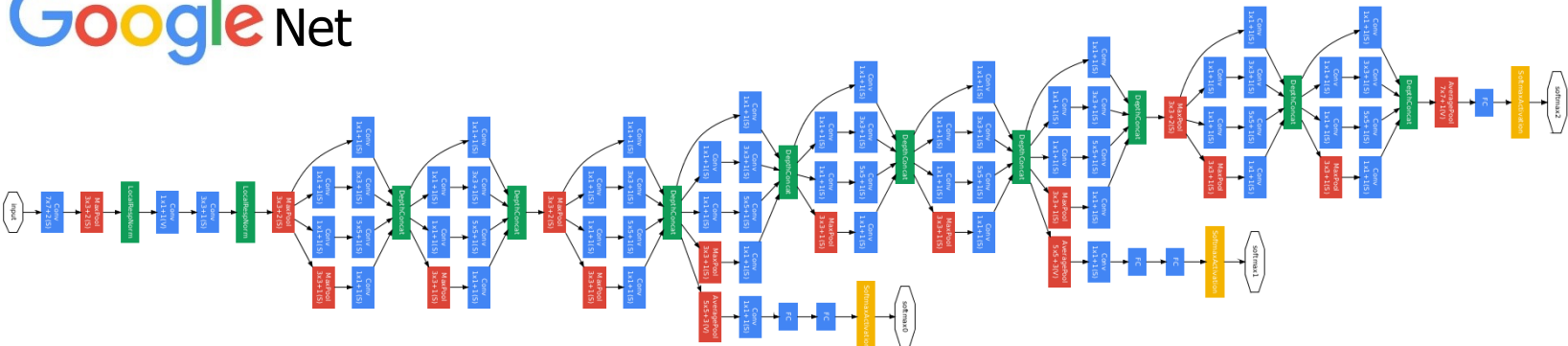


A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in NIPS 2012, pp. 1097–1105.

Deep learning basics

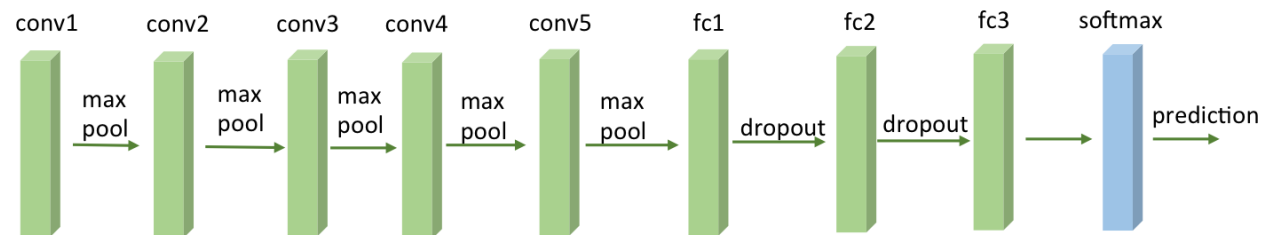
- **Deep learning in computer vision: CNNs**
 - A large number of architectures have been (and still are being) proposed

Google Net



Visual Geometry Group
(VGG Net)

each conv includes 3 convolutional layers



- **Deep learning in computer vision: CNNs**

“It can be concluded that from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task.”

A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN Features off-the-shelf: an Astounding Baseline for Recognition,” in CVPR 2014, pp. 512–519.



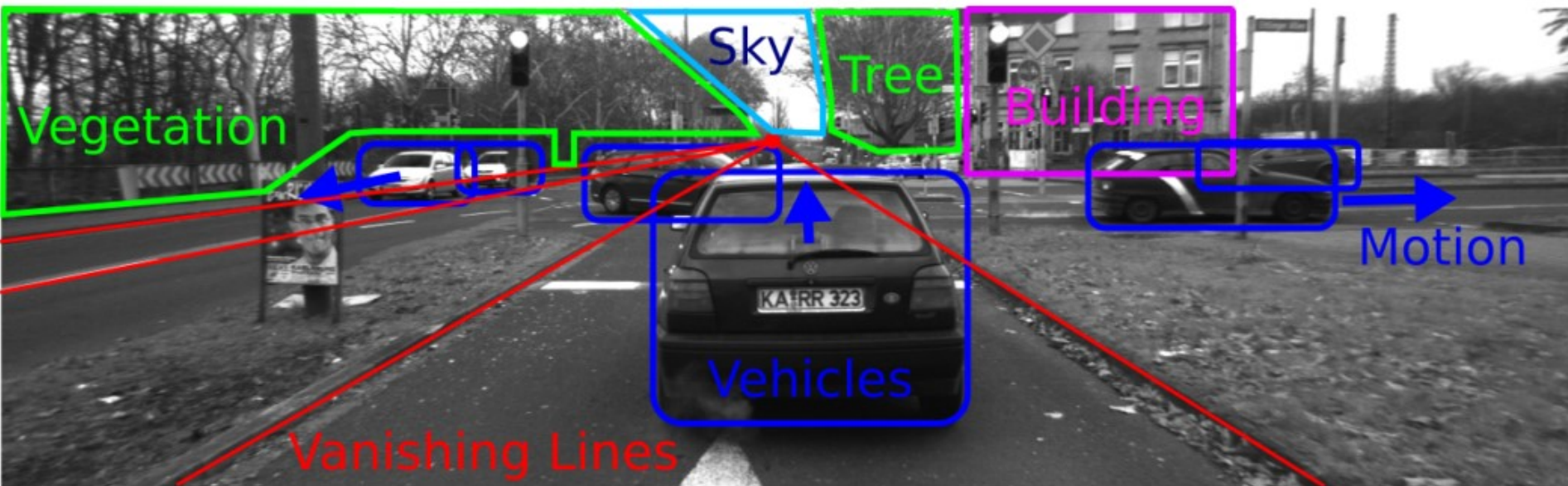
-
- Deep learning basics
 - **Applications in driving environments**
 - Future prospects
 - Conclusions

- What is this?



Source: A. Geiger, “Probabilistic Models for 3D Urban Scene Understanding from Movable Platforms,” Ph.D dissertation, Karlsruher Institut für Technologie, 2013.

- What is this?



Source: A. Geiger, "Probabilistic Models for 3D Urban Scene Understanding from Movable Platforms," Ph.D dissertation, Karlsruher Institut für Technologie, 2013.

- **Levels**
 - classification



Top 5:
pencil sharpener
pool table
hand blower
oil filter
packet

Groundtruth:
pencil sharpener

4.594C312_v4_00110000.JPG

- detection



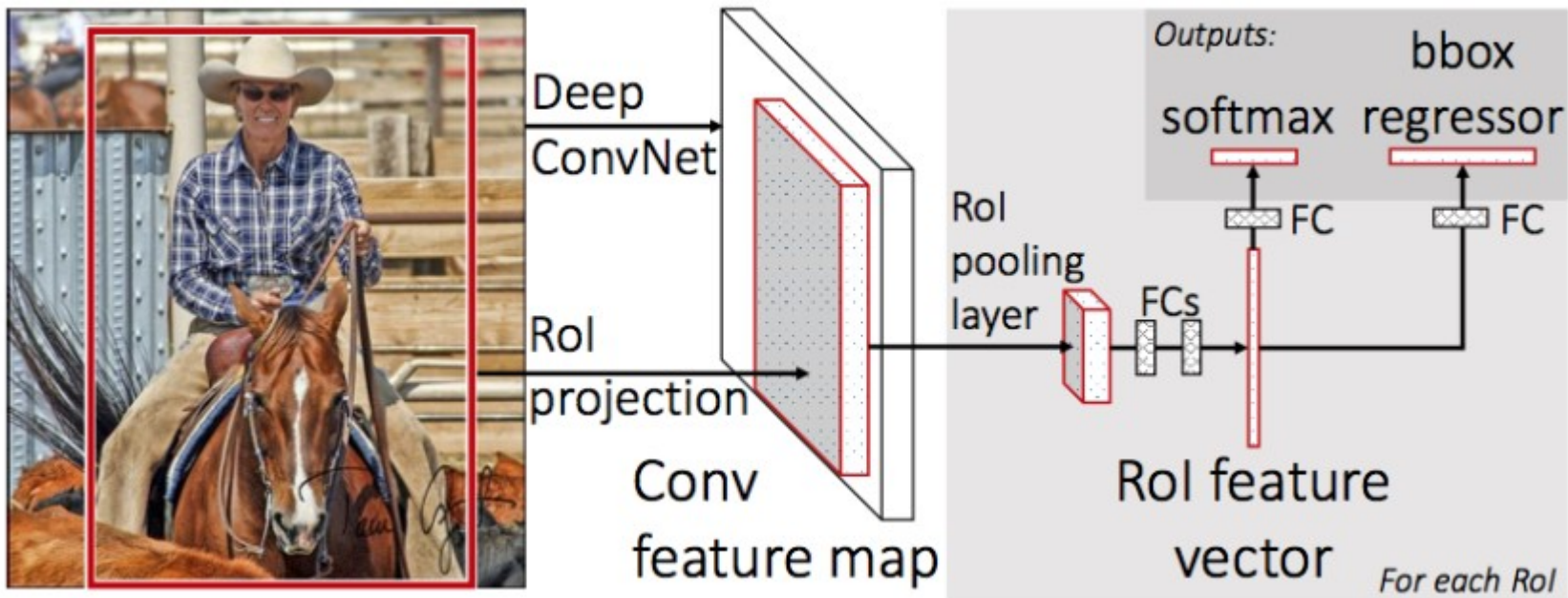
Groundtruth:
tv or monitor
tv or monitor (2)
tv or monitor (3)
person
remote control
remote control (2)

- segmentation



difficulty 27

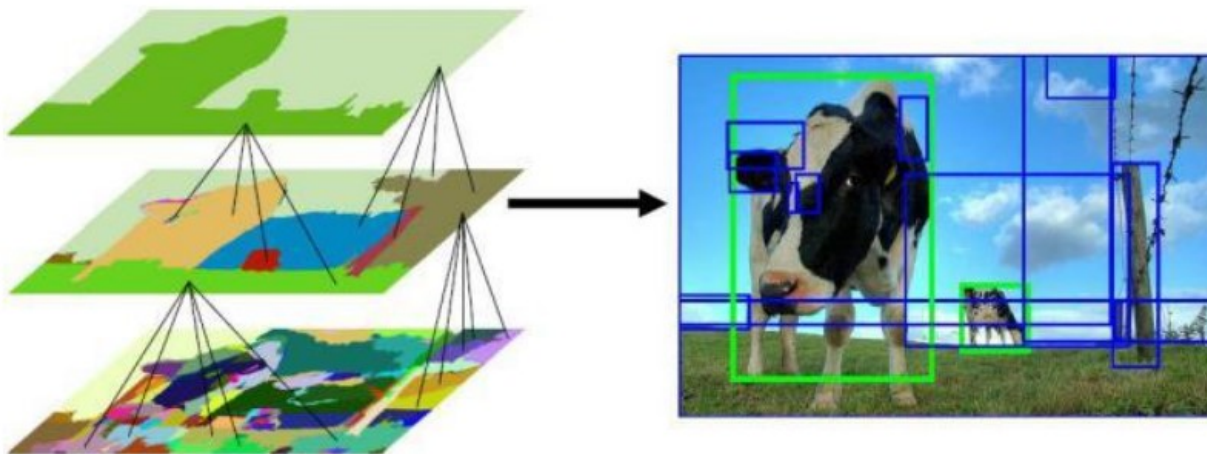
- **Object detection with CNNs: (Fast) R-CNN**
 - Fast R-CNN test time: < 0.5 second



R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in CVPR 2014, pp. 580–587.

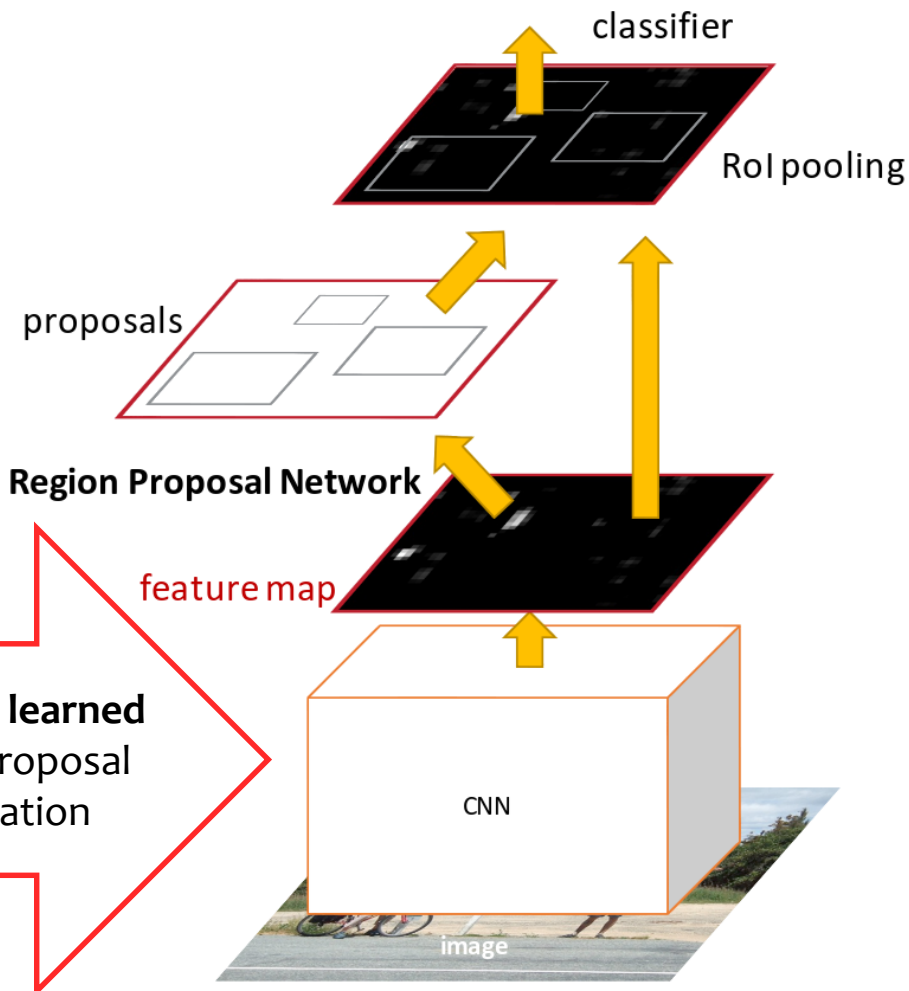
- **Where do Rols come from?**

- In the original work, they were proposed through Selective Search (a classical segmentation method) in a previous step.
- But they could be from anywhere (laser, stereo information, etc.)



J. R. R. Uijlings, K. E. a Van De Sande, T. Gevers, and a. W. M. Smeulders, “Selective search for object recognition,” *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

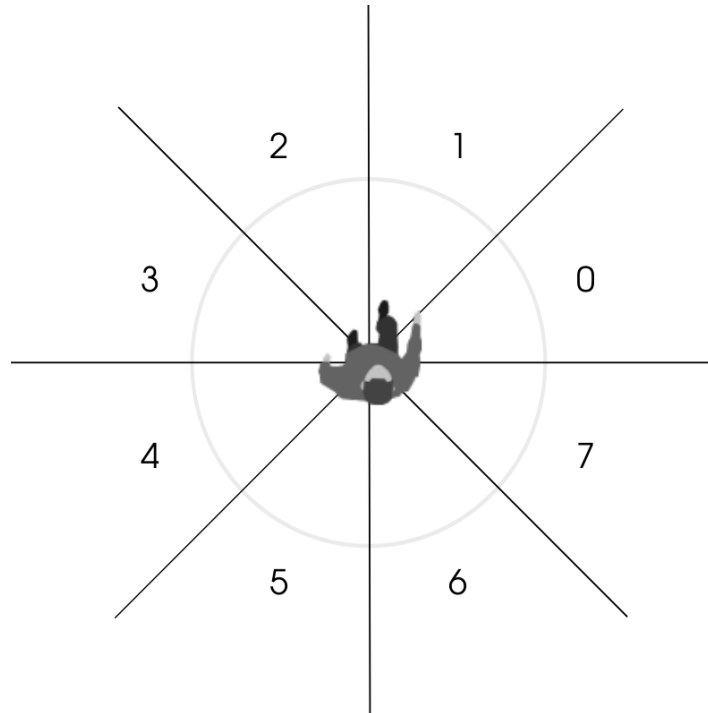
- **Skipping the RoI proposal step: Faster R-CNN**



- Real-time: up to 17 fps (with small resolutions)

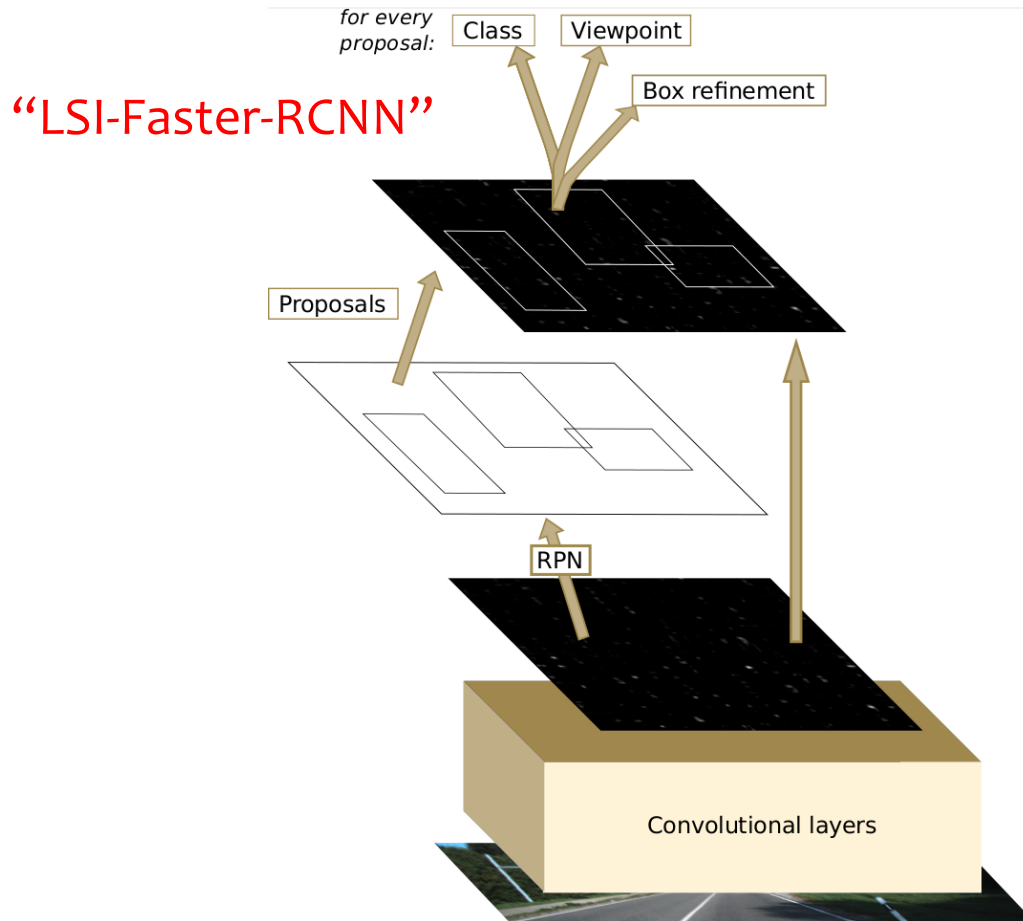
S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in NIPS 2015.

- **Why settle for detection?**
 - Obstacle viewpoint estimation
 - Important feature when understanding driving situations
 - Discrete viewpoint approach



- **Obstacle viewpoint estimation**

- Naturally integrated into the Faster R-CNN framework



- Based on the Zeiler-Fergus architecture

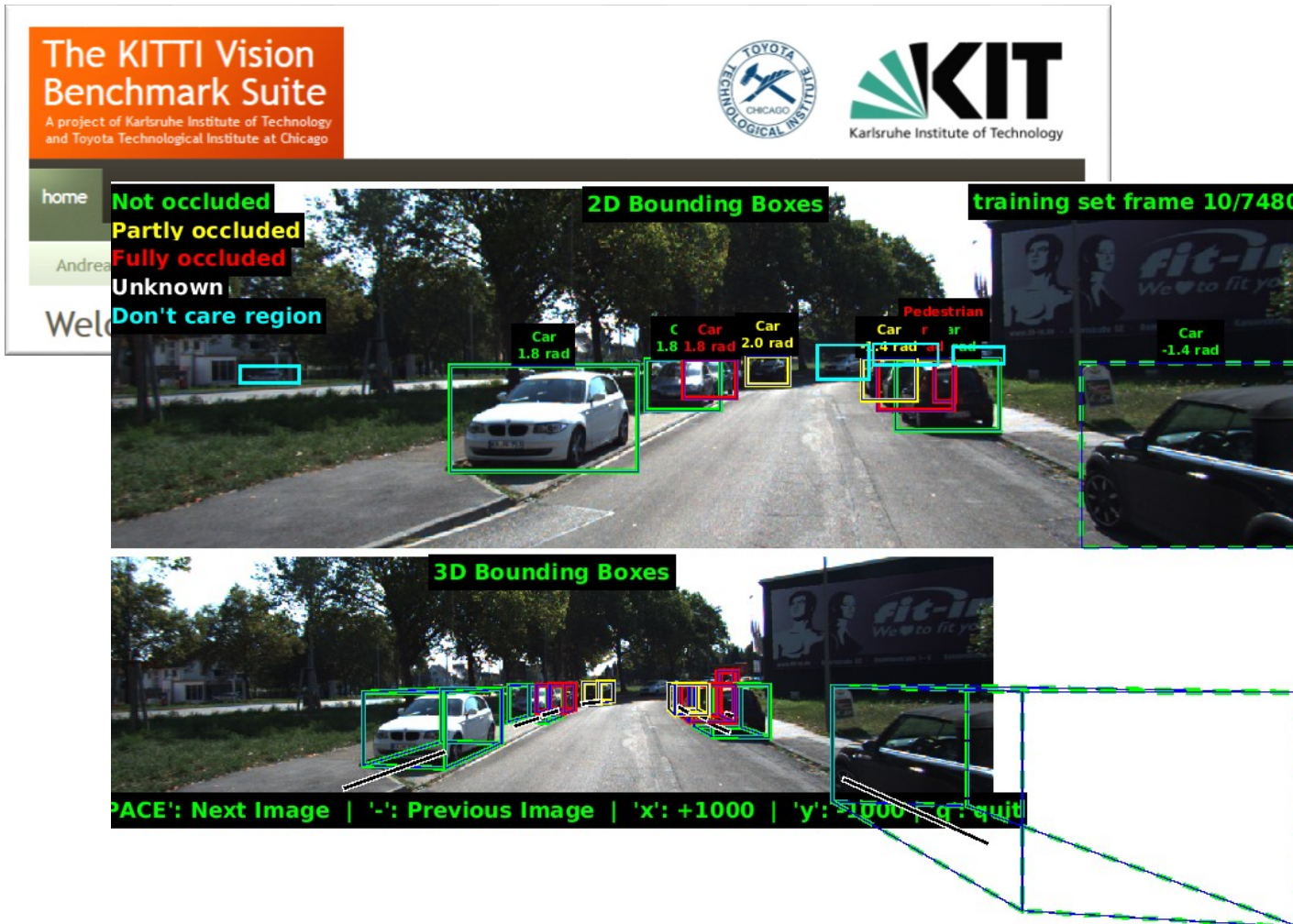
- Input:

- RGB Image (no additional features)

- Outputs

- Bounding boxes (Rols)
- Class (for every Rol)
- **Viewpoint** (for every Rol)

- Let's apply it to driving environments



Why do I love KITTI?

9 classes

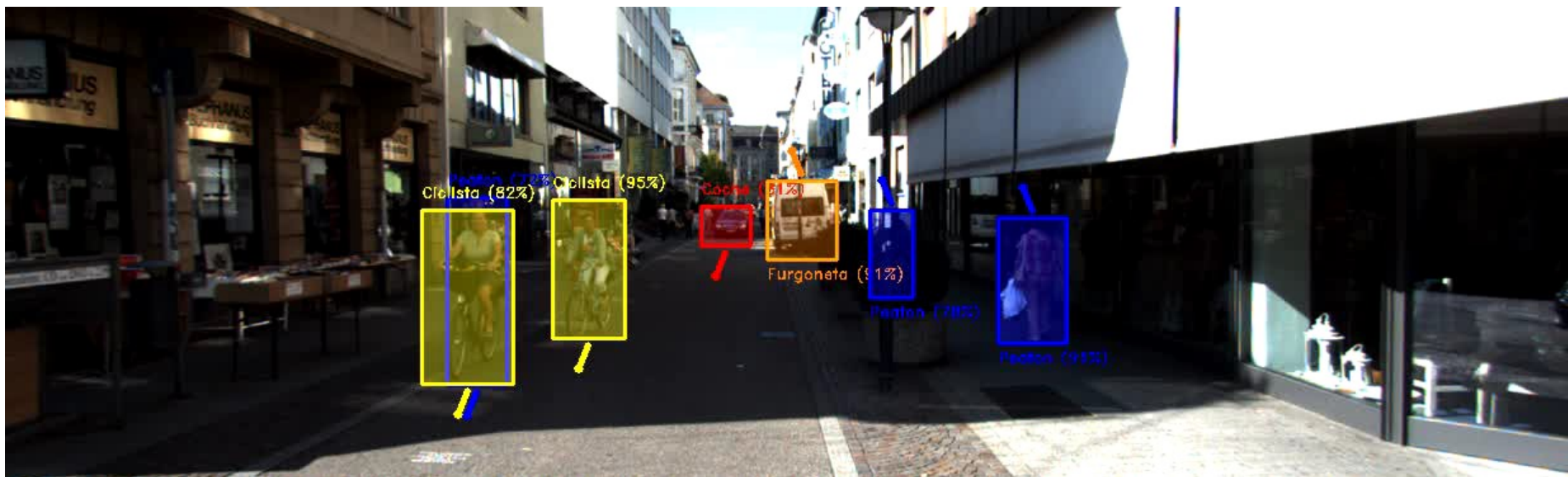
- Car
- Van
- Truck
- Pedestrian
- Person_sitting
- Cyclist
- Tram
- Misc
- DontCare

Additional info:

- Truncated
- Occluded
- Viewpoint
- Dimensions
- Location
- Yaw angle

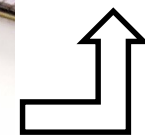
+ Table of results

- Now available in IVVI 2.0



- Detection time per image: ~100 ms (depending on the selected image scale)
- Work in progress

- **New GPU equipment: NVIDIA Tesla K40c**
 - Kindly donated by NVIDIA to the LSI
 - Deep learning training & test hugely accelerated



BEFORE

448 CUDA cores

6 GB GDDR5

cuDNN not supported



NOW

2880 CUDA cores

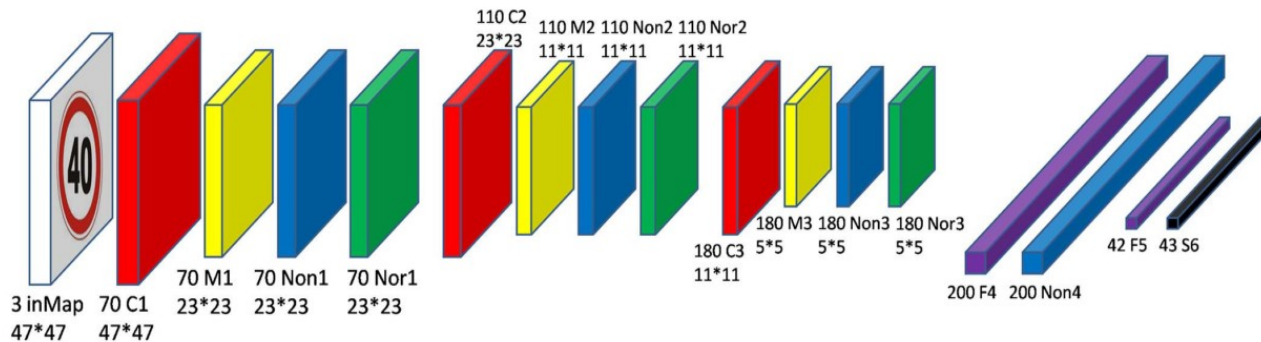
12 GB GDDR5

cuDNN supported



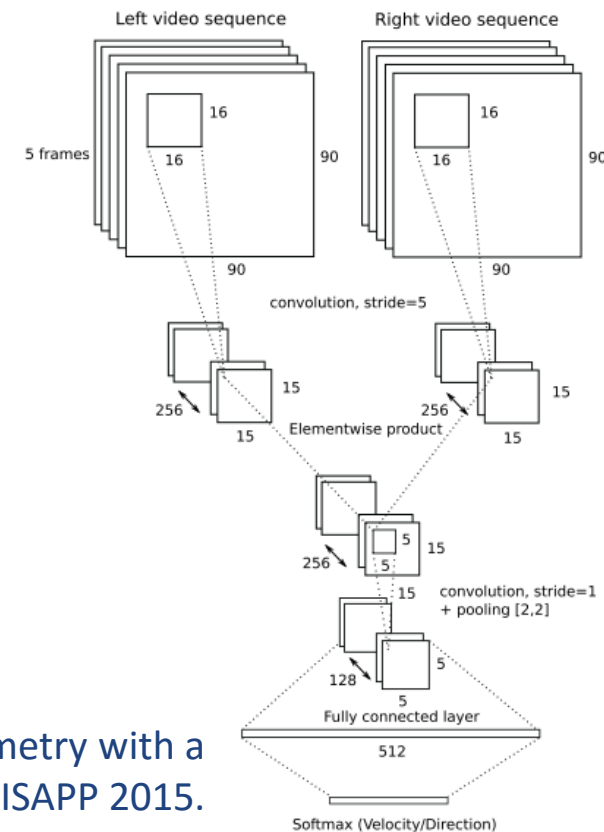
-
- Deep learning basics
 - Applications on driving environments
 - **Future prospects**
 - Conclusions

- Deep learning for traffic signs, vehicle color... even visual odometry!

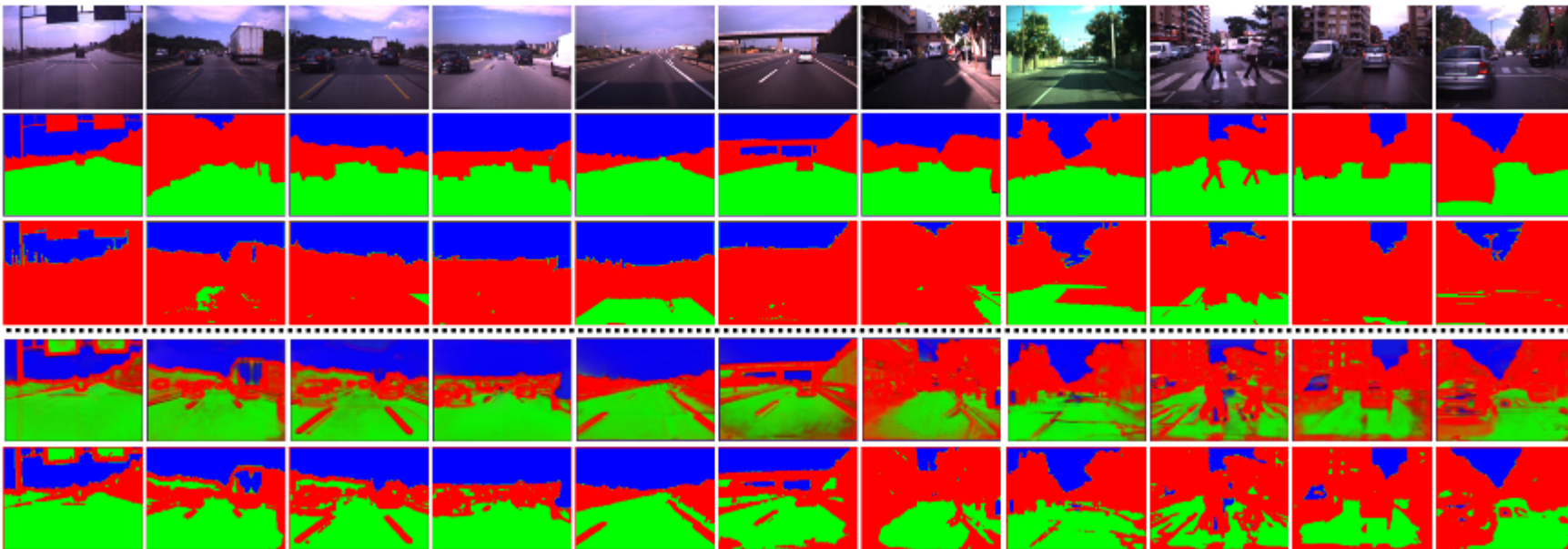


J. Jin, K. Fu, and C. Zhang, "Traffic sign recognition with hinge loss trained convolutional neural networks," IEEE TITS, vol. 15, no. 5, pp. 1991–2000, 2014.

K. Konda and R. Memisevic, "Learning Visual Odometry with a Convolutional Network," in VISAPP 2015.



- Per-pixel scene labeling using ConvNets



J. M. Alvarez, T. Gevers, Y. Lecun, and A. M. Lopez, “Road Scene Segmentation from a Single Image,” in ECCV 2012, pp. 376–389.

- Per-pixel scene labeling using ConvNets



C. Farabet, C. Couprie, L. Najman, and Y. Lecun, "Learning hierarchical features for scene labeling," IEEE Trans. PAMI, vol. 35, no. 8, pp. 1915–1929, 2013.

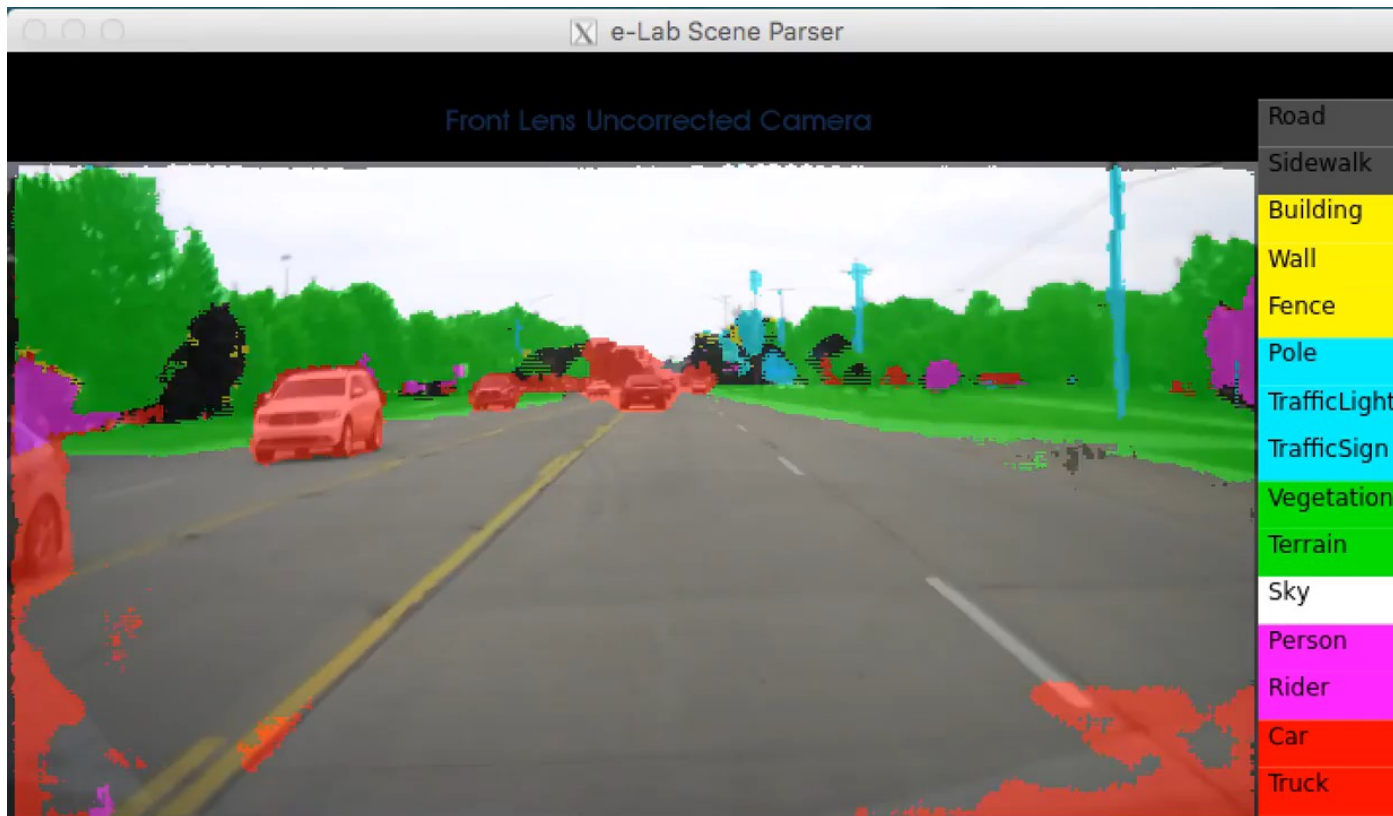


-
- **Scene understanding in driving environments remains an enormous challenge**



-
- **Scene understanding in driving environments remains an enormous challenge**
 - **First of all, what is scene understanding?**

- Scene understanding meaning #1: video labeling (NNs)



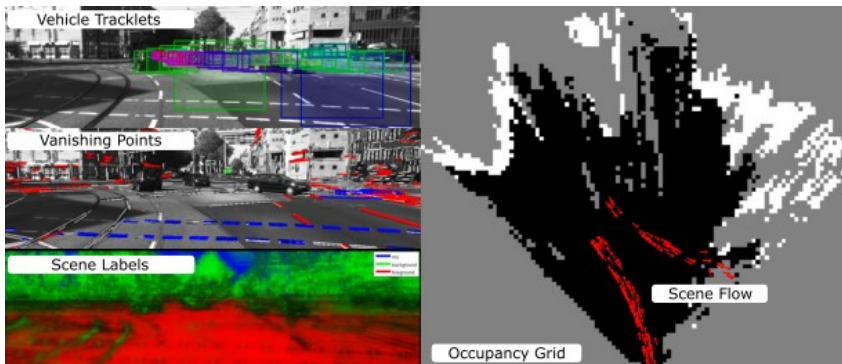
Source: Eugenio Culurciello, <https://www.youtube.com/watch?v=3jq4FnO5Nco>

- **Scene understanding meaning #2: video parsing (Recurrent ConvNets)**

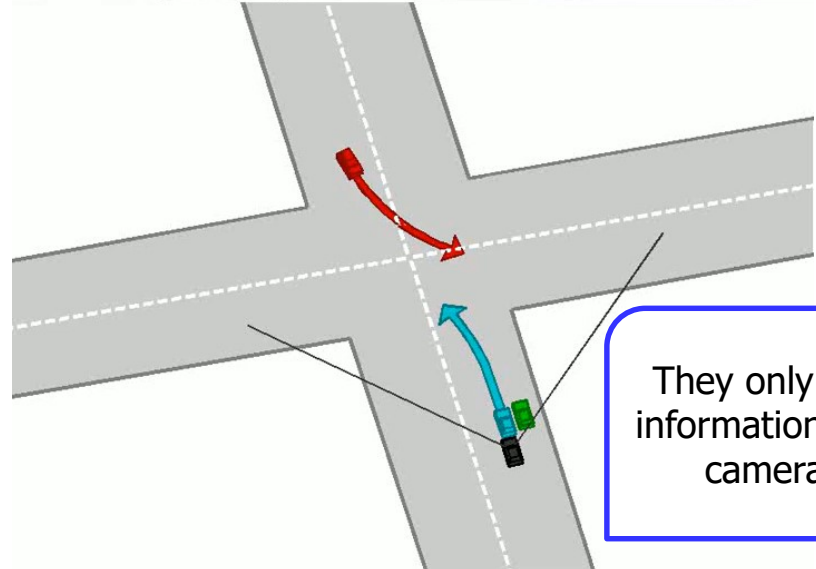
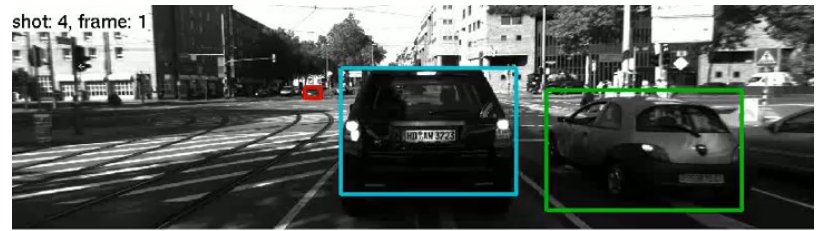
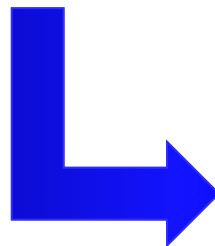


Source: <http://jeffdonahue.com/lrcn/>

- **Scene understanding meaning #3: scene models (probabilistic methods)**

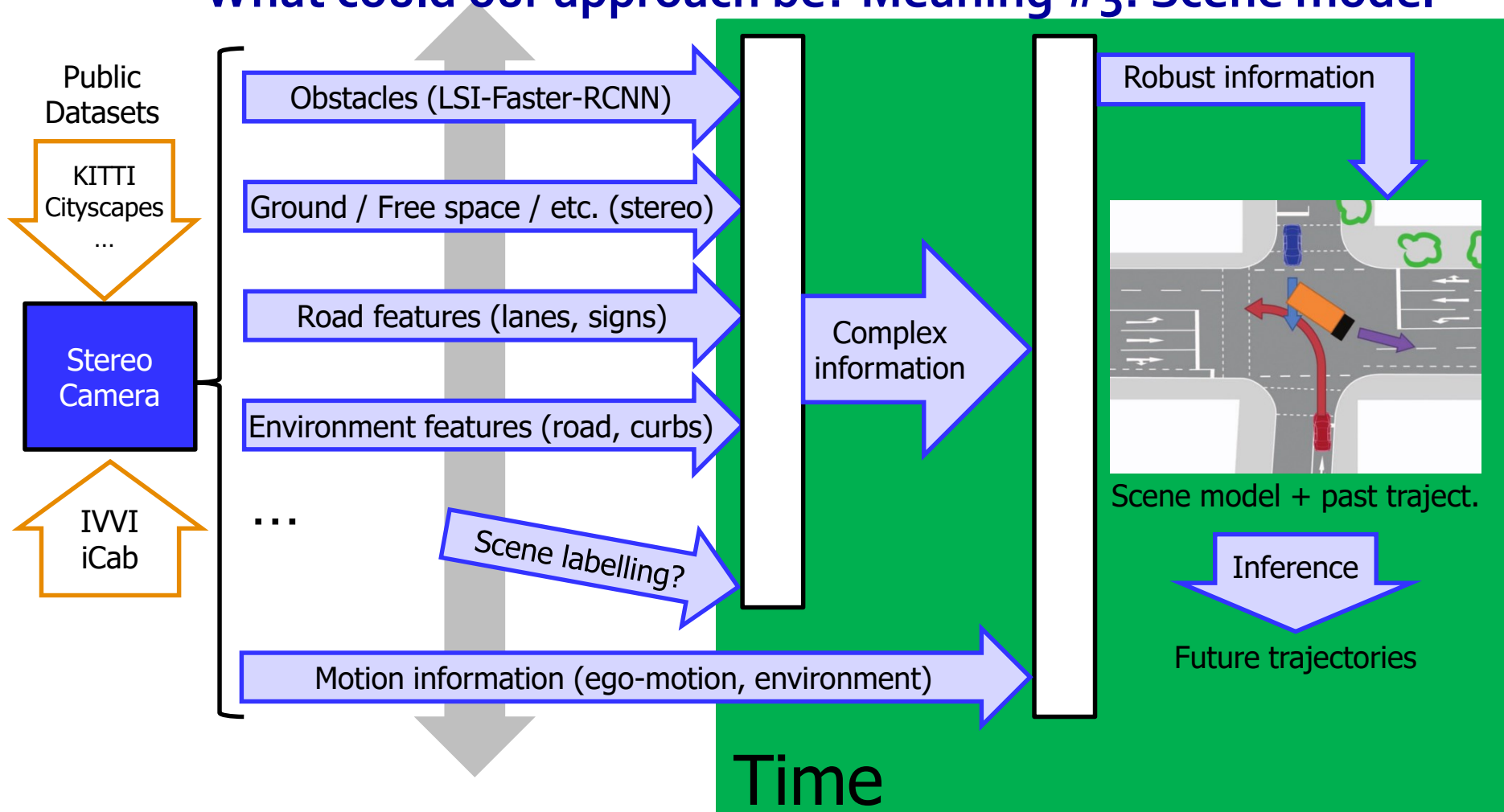


A large number of stochastic (non-perfect) sources of information

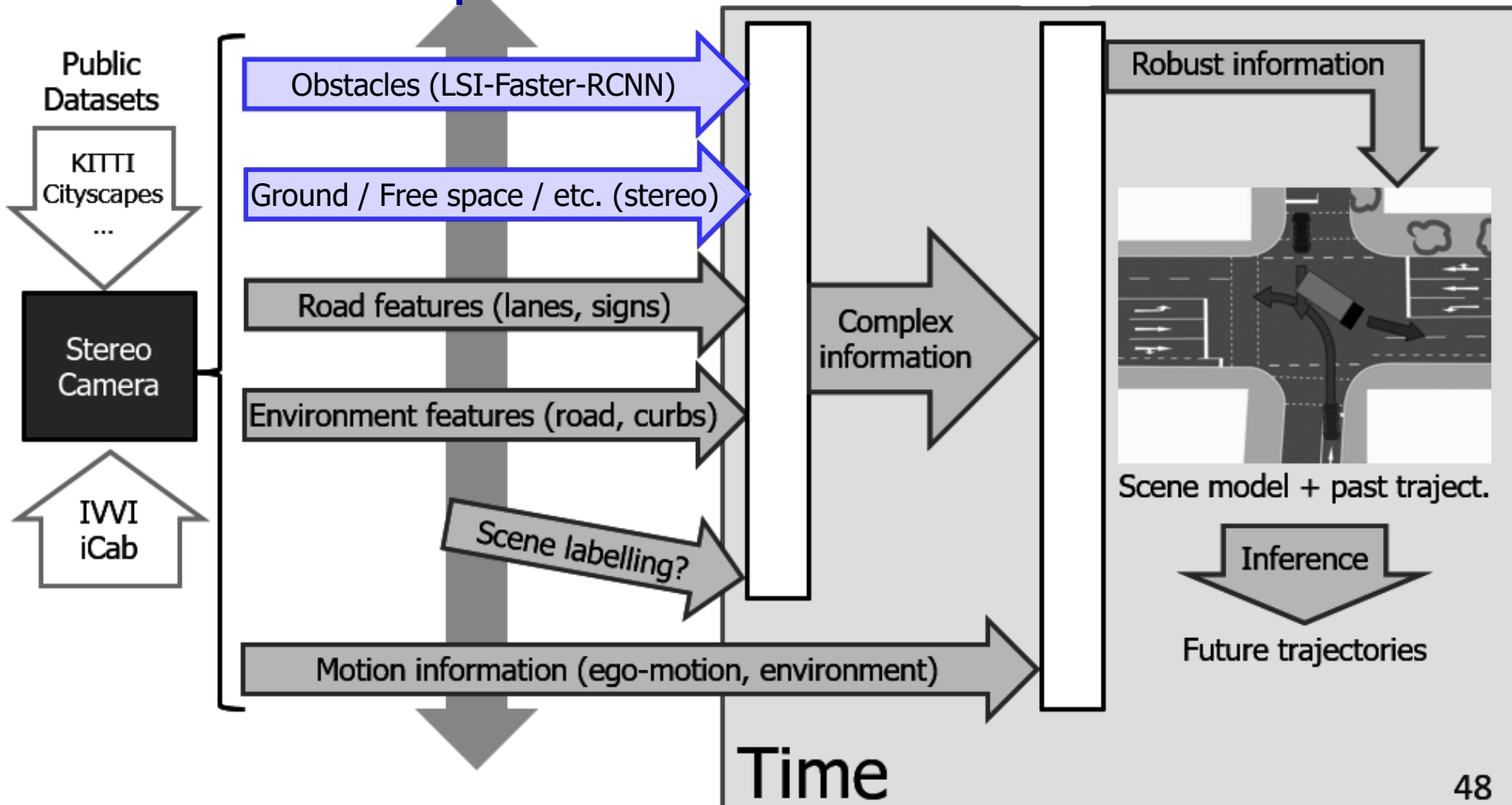


A. Geiger, "Probabilistic Models for 3D Urban Scene Understanding from Movable Platforms," Ph.D dissertation, Karlsruher Institut für Technologie, 2013

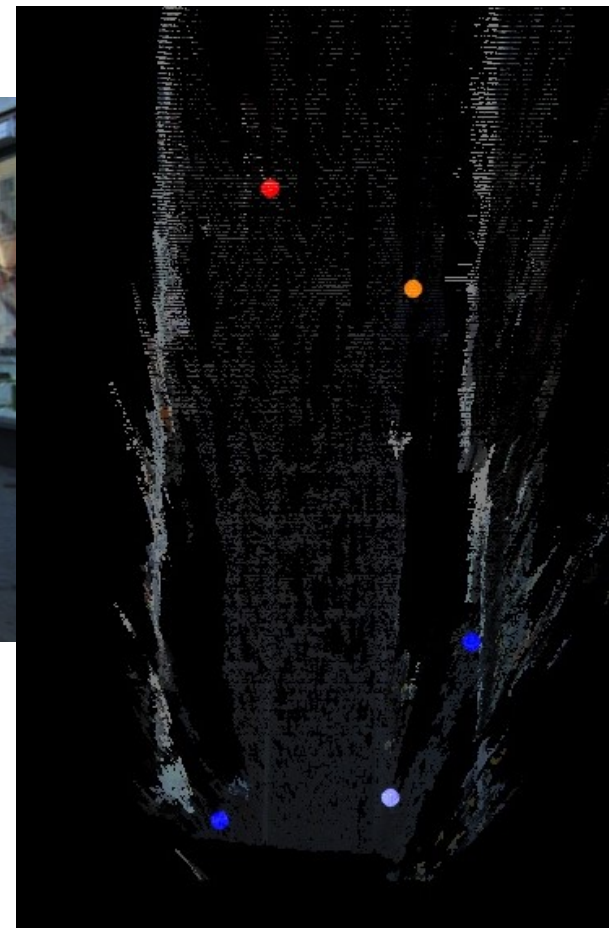
- **What could our approach be? Meaning #3: Scene model**



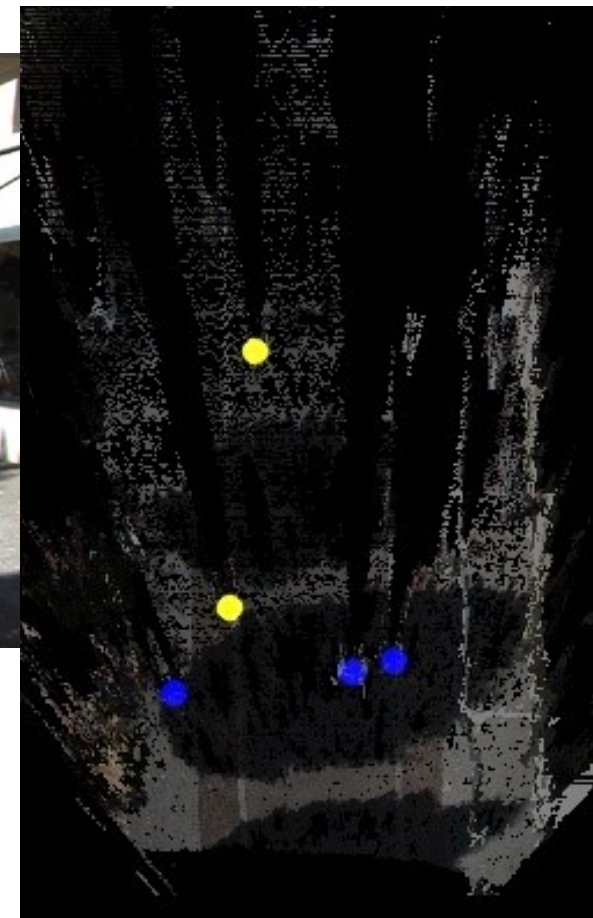
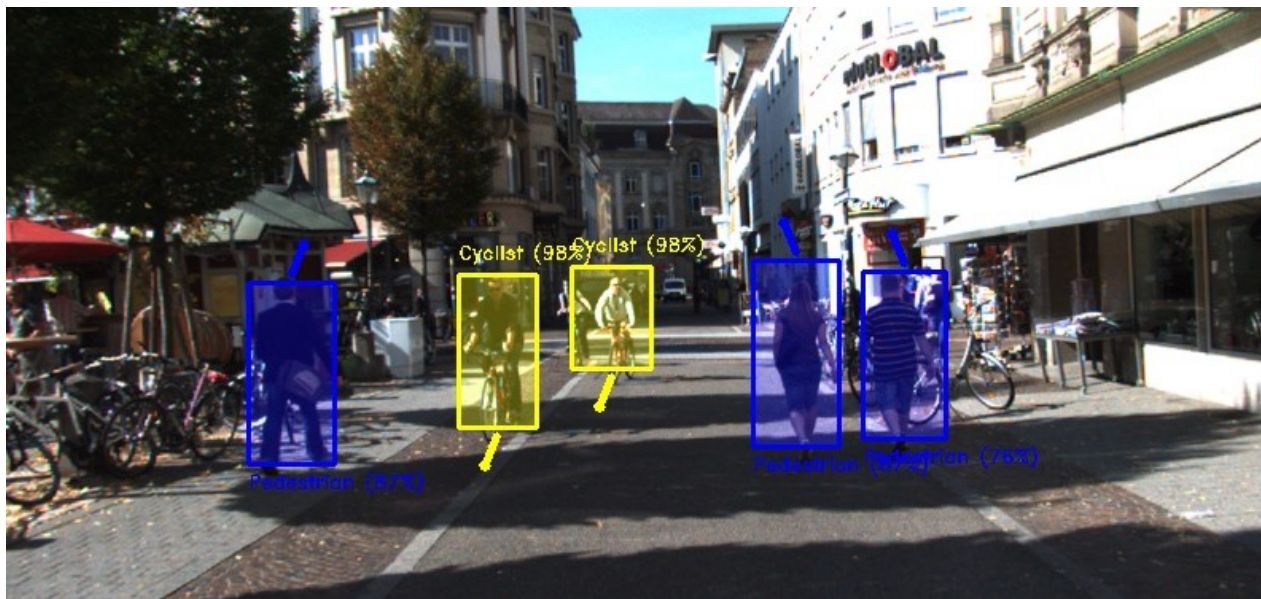
- **First attempts:**



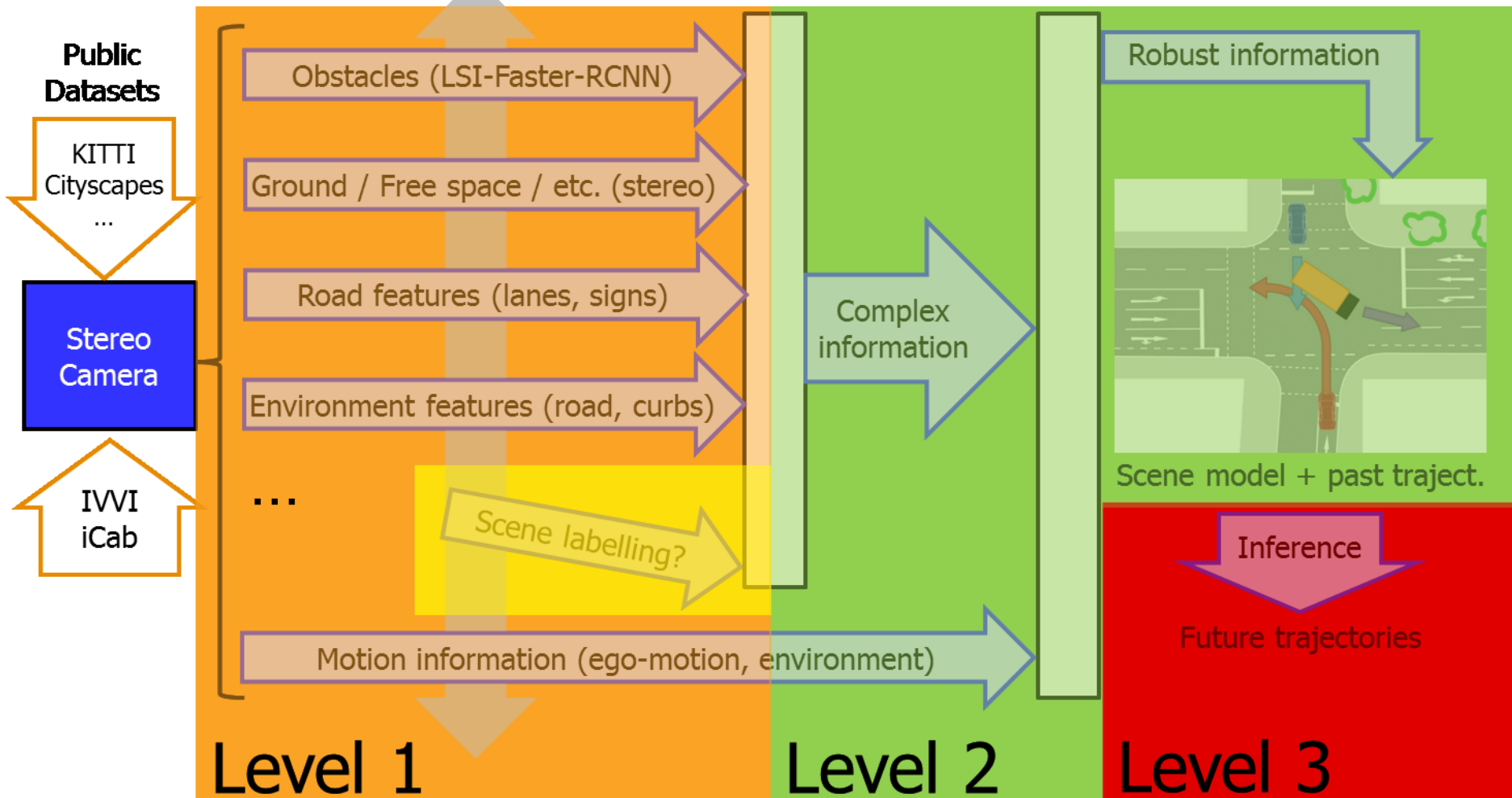
- **First attempts (now available in IVVI 2.0)**
 - A sparse 3D model or a loosely labeled pointcloud?



- **First attempts (now available in IVVI 2.0)**
 - A sparse 3D model or a loosely labeled pointcloud?



- A large number of information sources still to integrate...



-
- ...and obstacle detection can be largely improved
 - Thousands of *different* options:
 - **More sensors?**
 - Cons: lots of time-consuming engineering tasks (and also labeling tasks) before getting meaningful information
 - **Less sensors?**
 - E.g.: no stereo information. Cons: depth information is not straightforwardly available, structure-from-motion is a whole research field
 - **Less information sources?**
 - Cons: obstacles are not enough to understand the current driving scenario, inference would not be based on strong evidences
 - **Focus on only one thing?** Which one?
 - **Shortcuts to the last level?** “Scene understanding meaning #2”



-
- Deep learning basics
 - Applications on driving environments
 - Future prospects
 - **Conclusions**

-
1. Deep learning is becoming the new baseline in computer vision
 2. Intelligent transportation systems offer many opportunities to take advantage of it
 3. Scene understanding is a very high-level goal where multiple approaches can be adopted



VI Jornada de Encuentros Doctorales LSI

Deep learning applied to driving environments

Thank you for your attention

The End

- **Bonus track:**



IN CS, IT CAN BE HARD TO EXPLAIN
THE DIFFERENCE BETWEEN THE EASY
AND THE VIRTUALLY IMPOSSIBLE.