Deep-Learning for Automated Vehicles

Pr. Fabien MOUTARDE
Center for Robotics
MINES ParisTech
PSL Université

Fabien.Moutarde@mines-paristech.fr
http://people.mines-paristech.fr/fabien.moutarde

What types of Intelligences are needed for Automated Vehicles?

- "Semantic" interpretation of vehicle's environment:
  – Detect and categorize/recognize objects (cars, pedestrians, bicycles, traffic signs, traffic lights, …)
  – Ego-localization
  – *Predict movements of other road users*
  – *Infer intentions of other drivers and pedestrians (or policeman!) from their movements/gestures/gazes*
- Planning of trajectories (including speed)
  In a dynamic and uncertain environment
- Coordinated/cooperative planning of multiple vehicles

- For Advanced Driving Assistance Systems (ADAS) and partial automated driving (level 3-4):
  – Analyze and understand attention and activities or gestures of the "driver-supervisor"

Outline

- **What can Deep-Learning perform with images?**
- Visual Object detection & Semantic Segmentation
- Image-based ego-localization
- Human posture and movement analysis

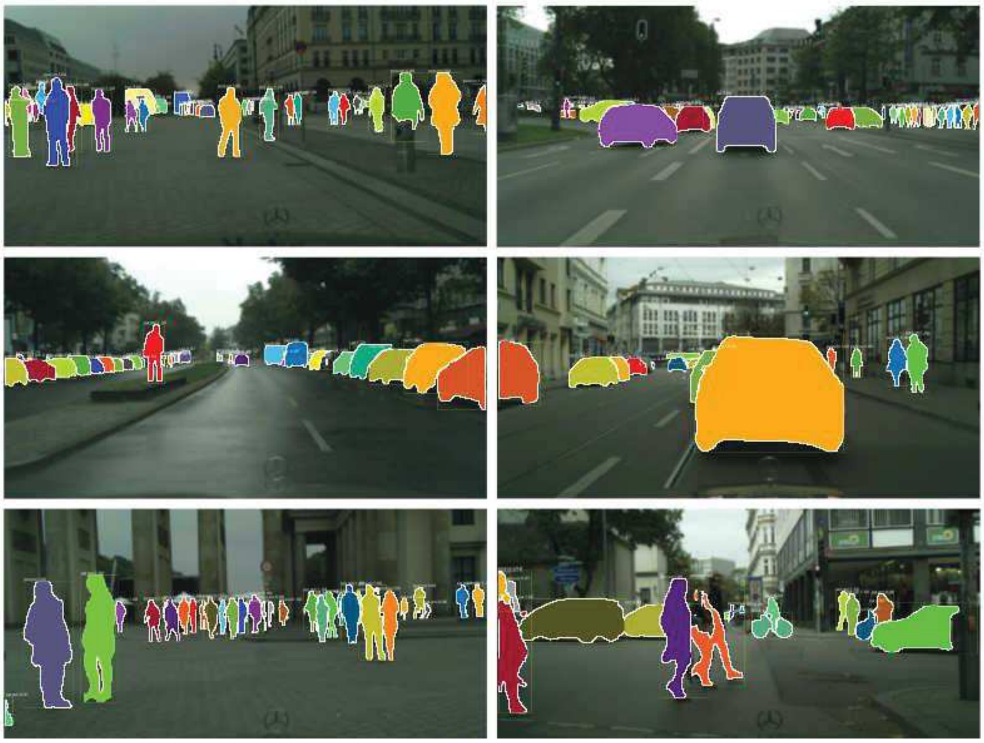
Image-based Deep-Learning

- Image classification
- Visual object detection and categorization
- Semantic segmentation of images
- Realistic image synthesis
- Image-based localization
- Estimation of Human pose
- Inference of 3D (depth) from monocular vision
- Learning image-based behaviors
  - End-to-end driving from front camera
  - Learning robot behavior from demonstration/imitation

Visually detect & categorize objects

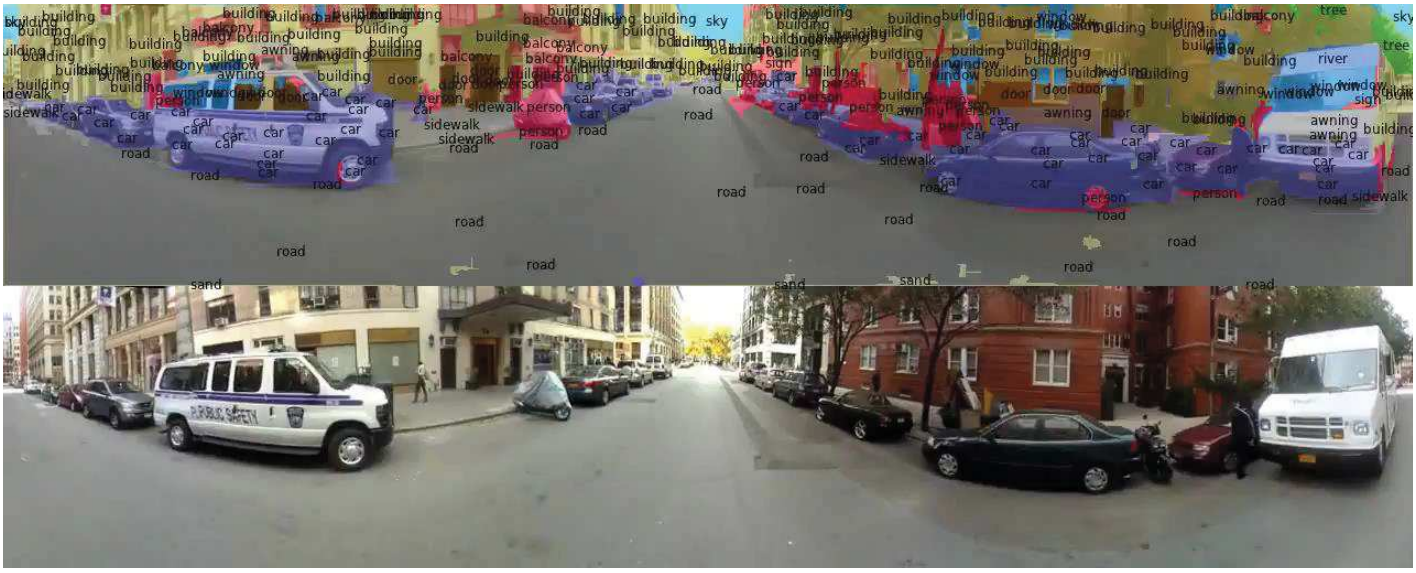
Visual objects Simultaneous Detection and Categorization with Faster_RCNN

Beyond bounding-boxes: getting *contours* of objects



Mask R-CNN extracts detailed contours and shapes of objects instead of just bounding-boxes

Example result of semantic segmentation by Deep-Learning



[C. Farabet, C. Couprie, L. Najman & Yann LeCun: Learning Hierarchical Features for Scene Labeling, IEEE Trans. PAMI, Aug.2013.

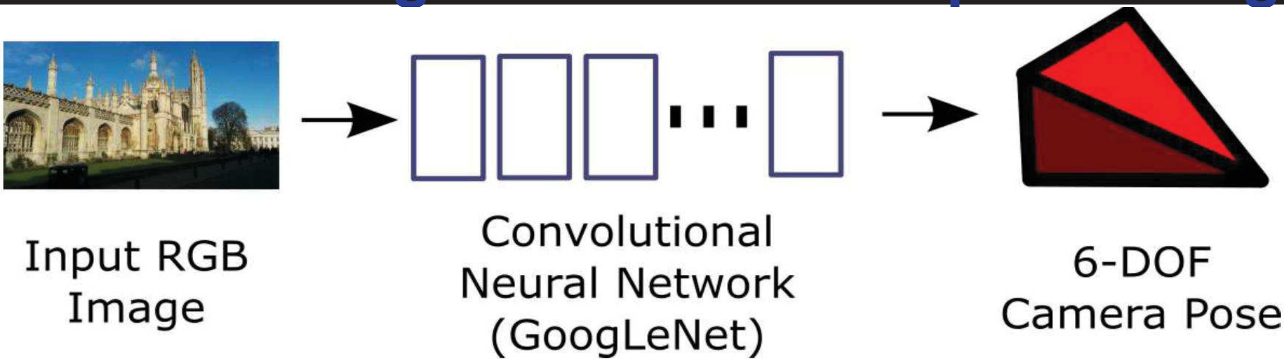
Semantic segmentation provides category information *also for large regions* (not only individualized « objects ») *such as « road », « building »*, etc…

DL for realistic Image synthesis

"Video-to-Video Synthesis", NeurIPS'2018 [Nvidia+MIT]

Using Generative Adversarial Network (GAN)

PoseNet: 6-DoF camera-pose regression with Deep-Learning



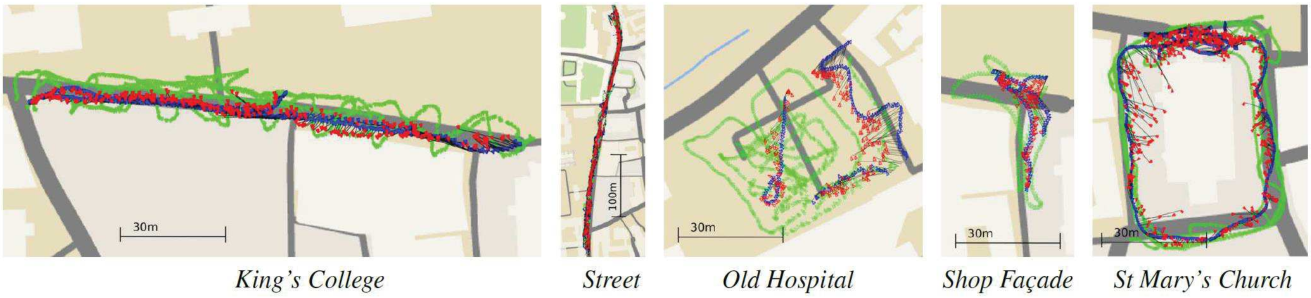
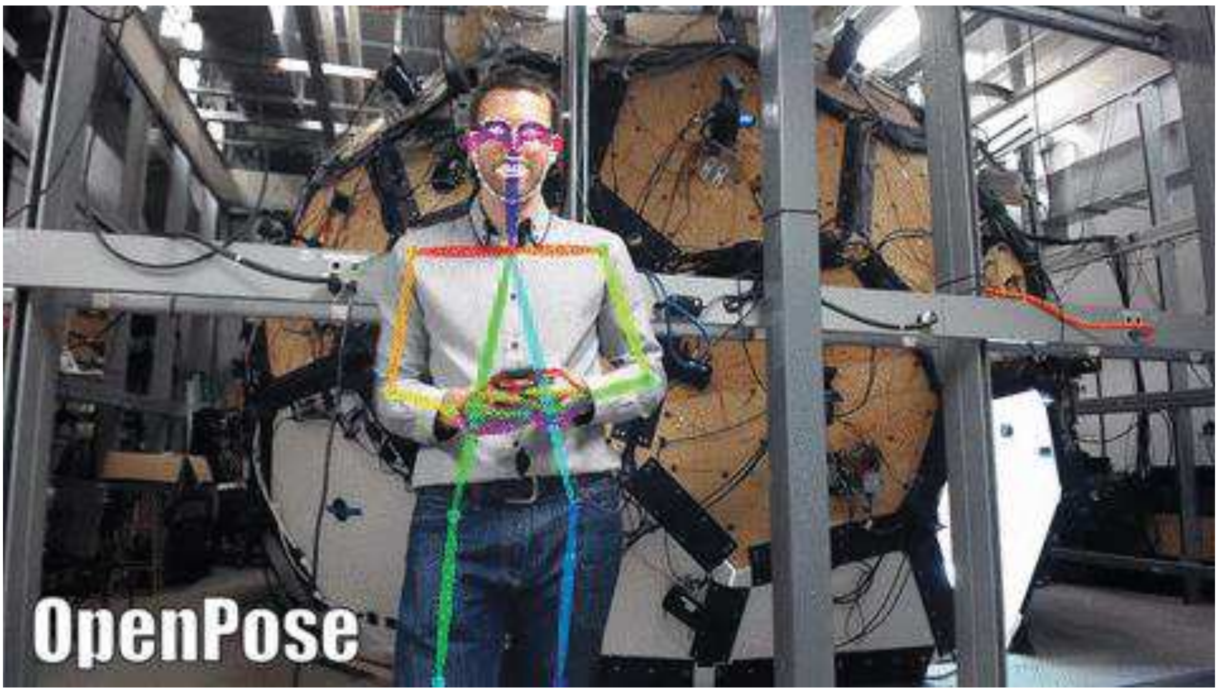


Figure 4: **Map of dataset** showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.

[A. Kendall, M. Grimes & R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization" , ICCV'2015, pp. 2938-2946]

Human pose estimation by Deep-Learning



Real-time estimation of Human poses on *RGB* video

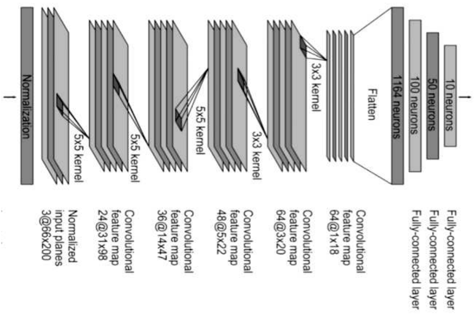
[Realtime Multi-Person 2D Pose Estimation using Part Affinity Field, Cao et al., CVPR'2017 [CMU]

Inference of 3D (depth) from monocular vision

[Unsupervised monocular depth estimation with left-right consistency]
[C Godard], [O Mac Aodha], [GJ Brostow] - *CVPR'2017 [UCL]*

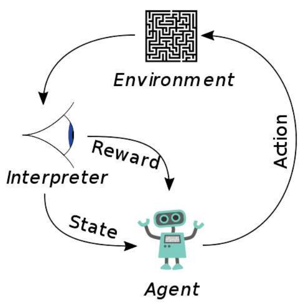
End-to-end driving from camera by Deep-Learning

ConvNet input: Cylindrical projection of fisheye camera



ConvNet output: steering angle

Imitation Learning from Human driving on real data



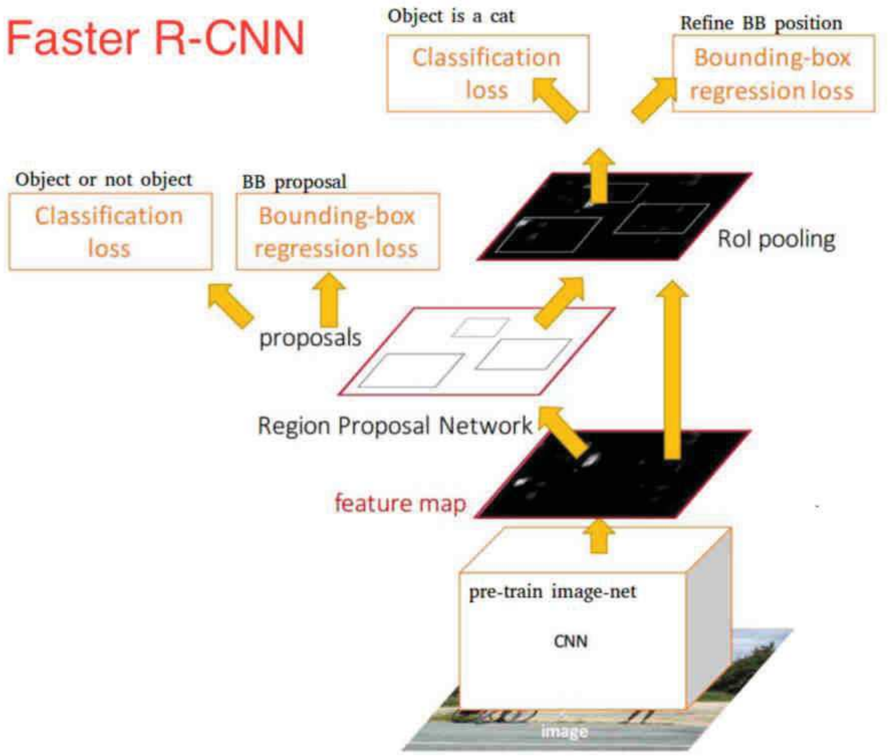
End-to-end driving via Deep *Reinforcement* Learning

[thèse CIFRE Valeo/MINES-ParisTech en cours]

Outline

- What can Deep-Learning perform with images?
- **Visual Object detection & Semantic Segmentation**
- Image-based ego-localization
- Human posture and movement analysis

Visual objects Detection and Categorization: Faster_RCNN



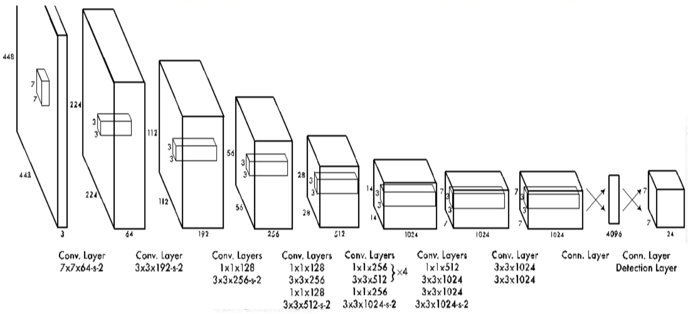
Region Proposal Network (RPN) on top of standard convNet.
End-to-end training with combination of 4 losses

Object visual detection *without proposal*

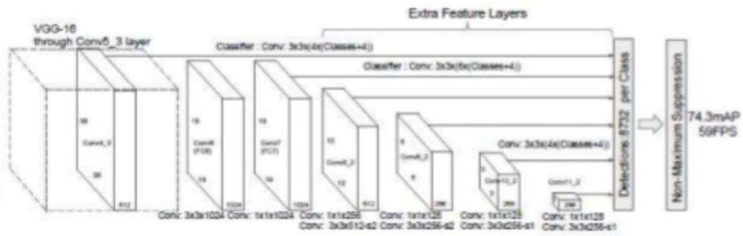
Solve detection as a regression problem (“single-shot” detection)

YOLO and SSD

YOU ONLY LOOK ONCE(YOLO)



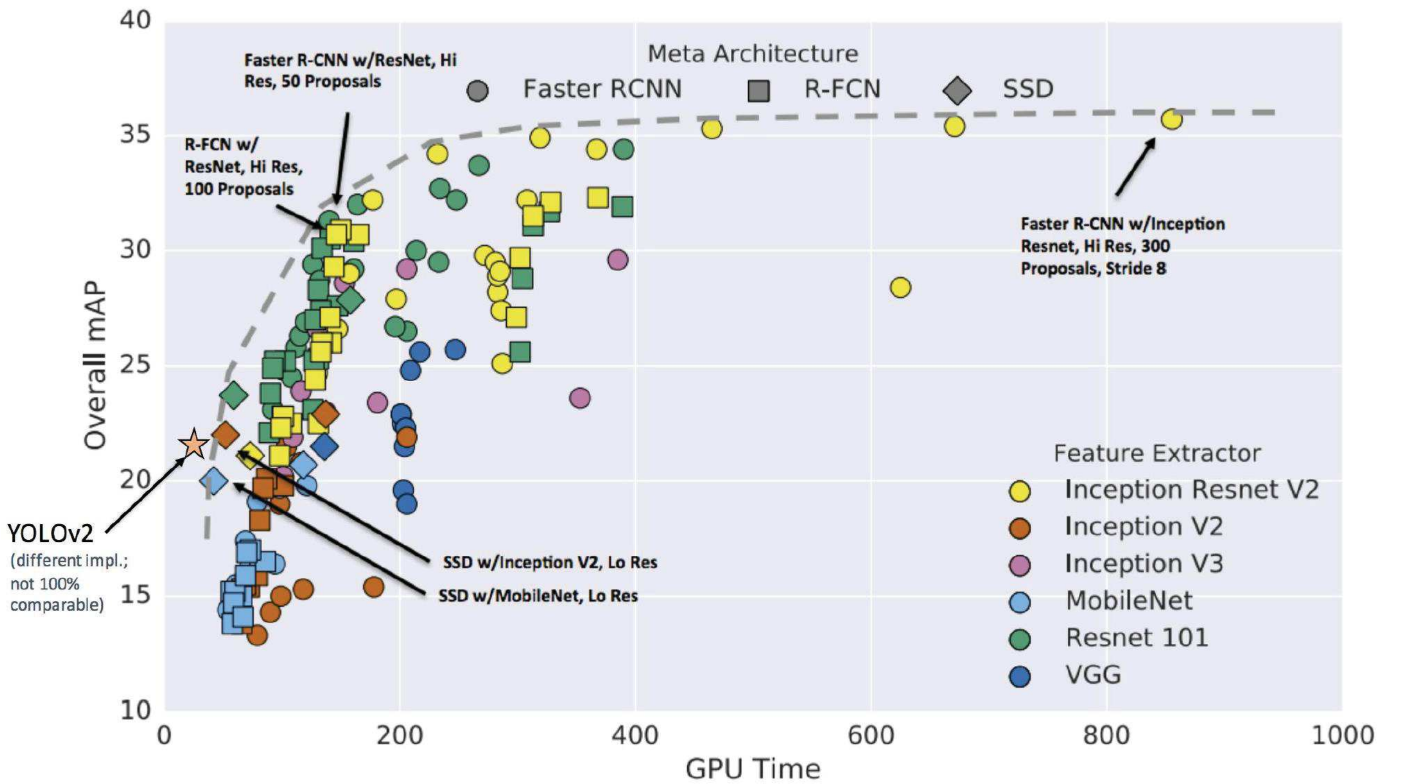
SINGLE SHOT MULTIBOX DETECTOR(SSD)



Images from: https://www.slideshare.net/TaegyunJeon1/pr12-you-only-look-once-yolo-unified-realtime-object-detection

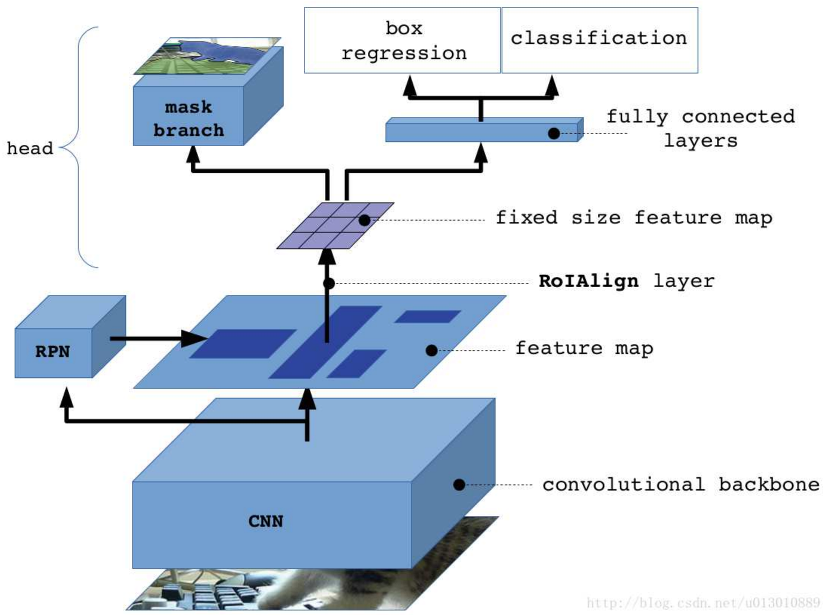
Both are faster, but less accurate, than Faster_R-CNN

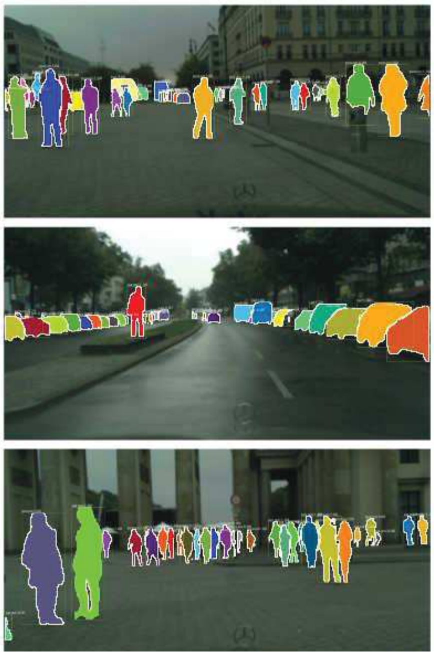
Recent comparison of object detection convNets



Slide from Ross Girshick’s CVPR 2017 Tutorial, Original Figure from Huang et al

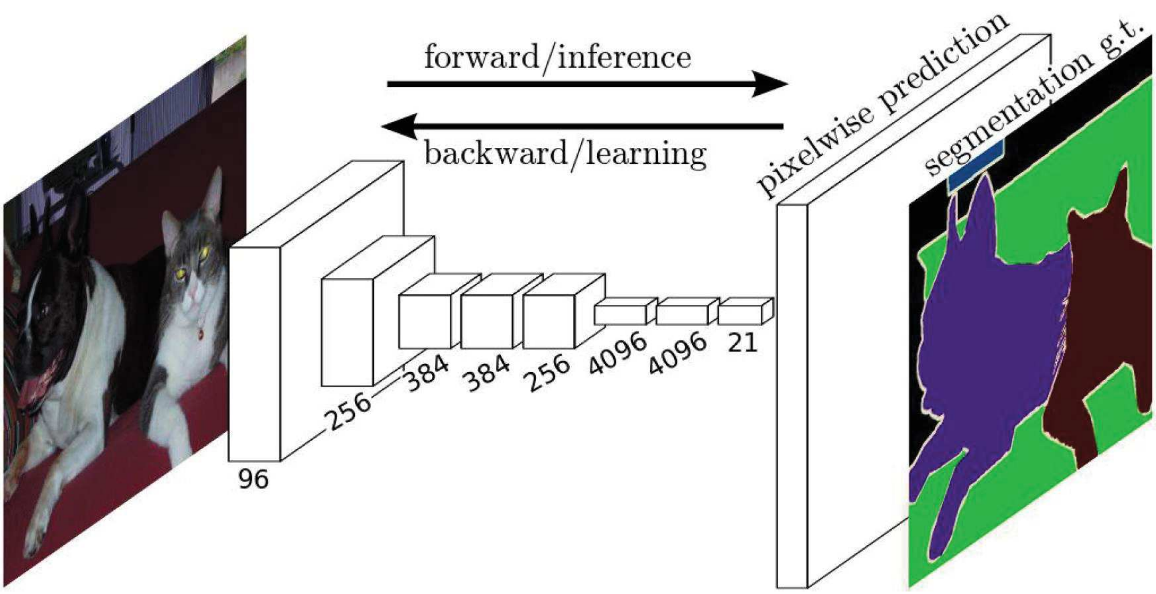
Mask_RCNN: Categorization and Localization with shape/*contours*



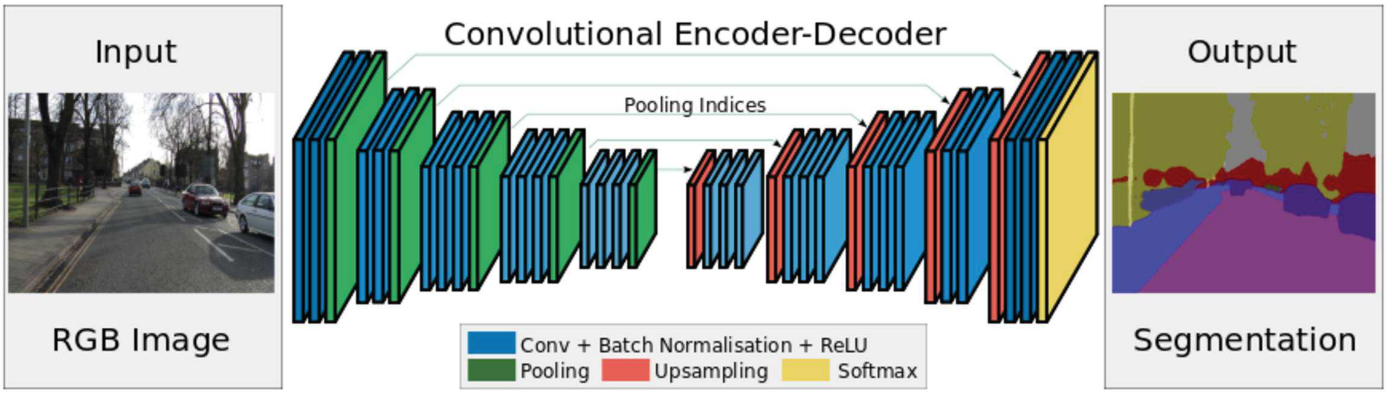


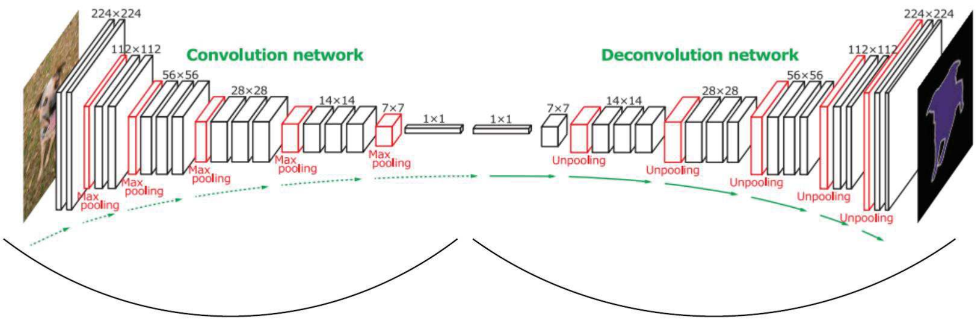
Mask R-CNN architecture (left) extracts detailed contours and shape of objects instead of just bounding-boxes

Deep-Learning approach for semantic segmentation



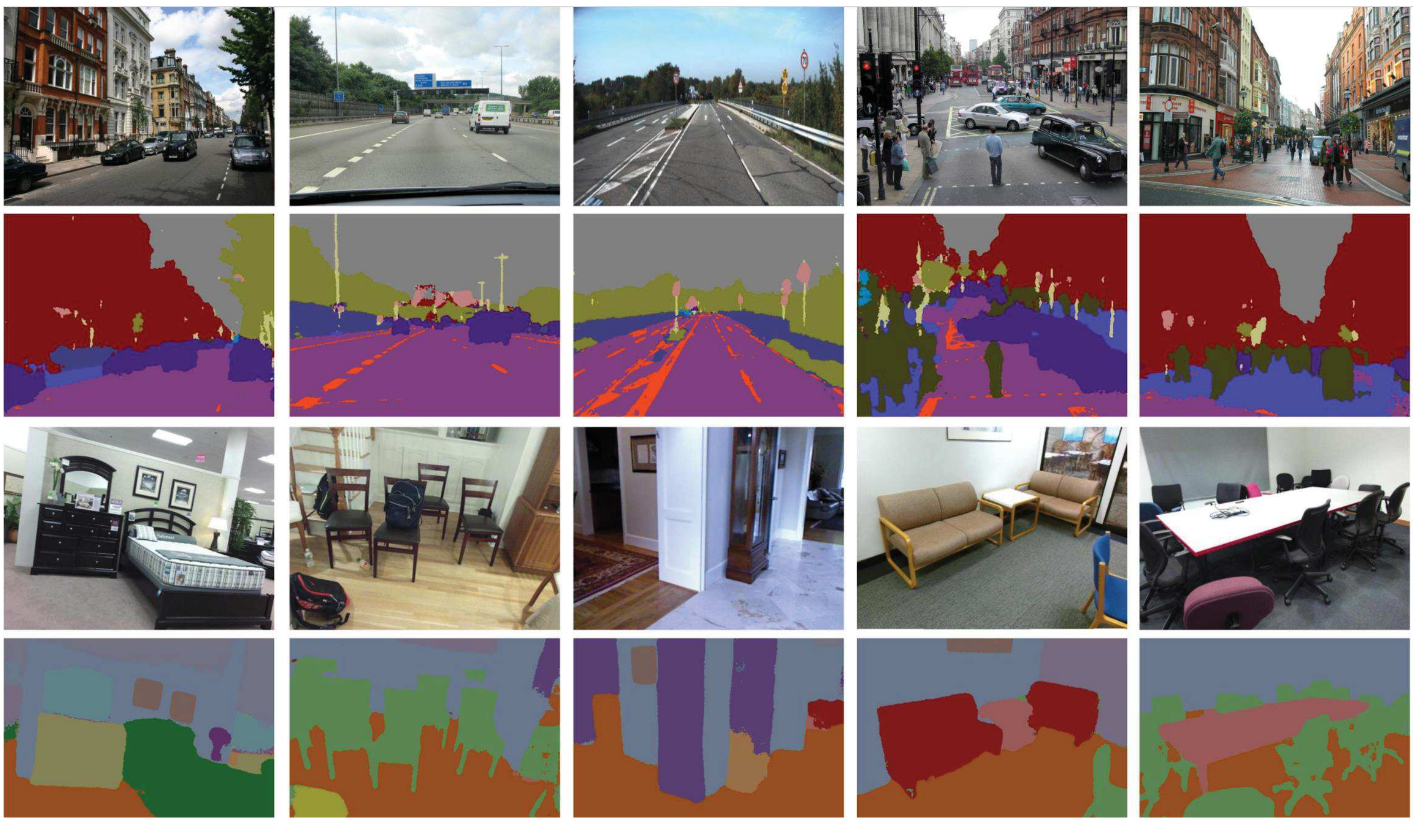
Convolutional Encoder-Decoder





Feature extractor **Shape generator**

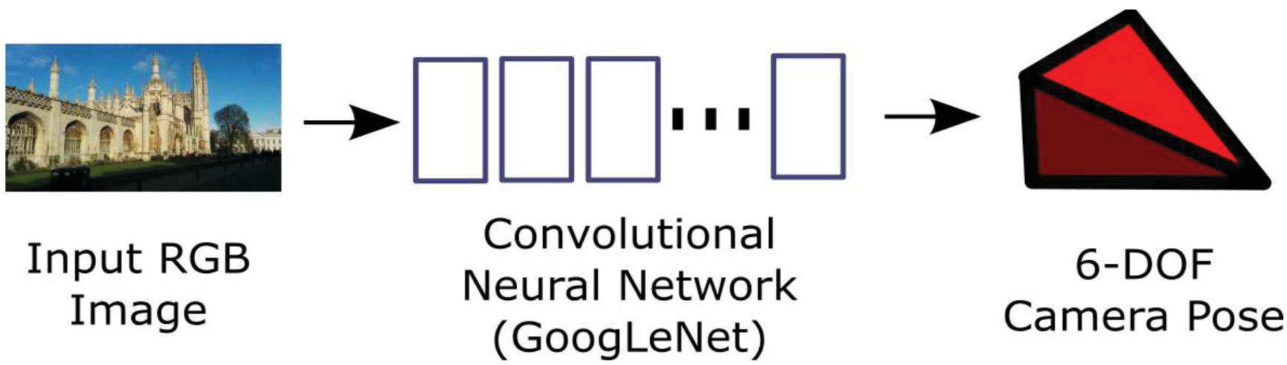
SegNet example results



Outline

- What can Deep-Learning perform with images?
- Visual Object detection & Semantic Segmentation
- **Image-based ego-localization**
- Human posture and movement analysis

PoseNet: 6-DoF camera pose regression with Deep-Learning



Trained with a naïve end-to-end loss function to

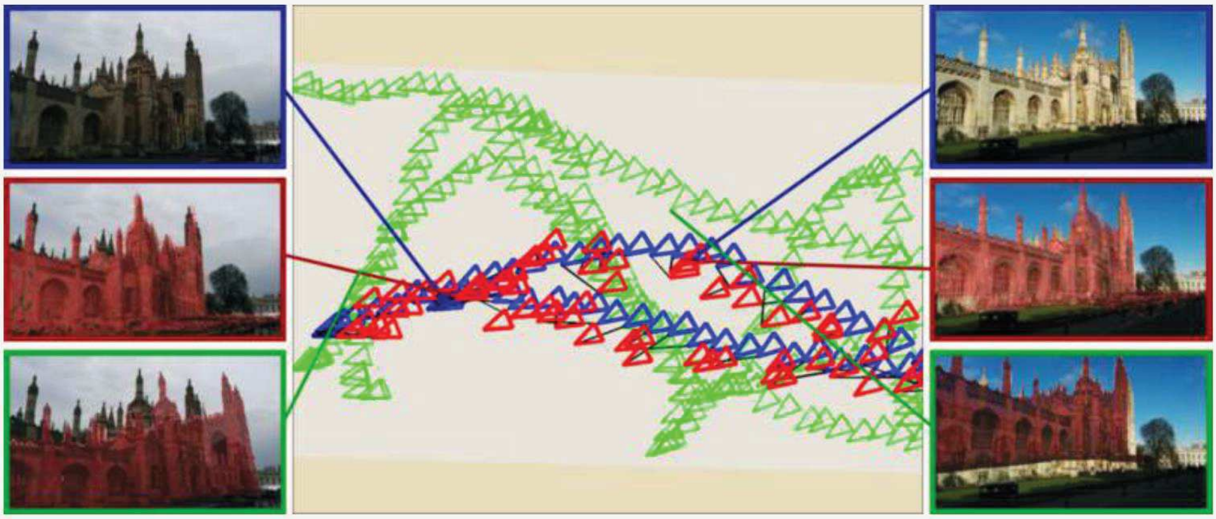
regress camera position, **x**, and orientation, **q**

$$\mathrm{loss}(I) = \|x - \hat{x}\|_2 + \beta \left\| q - \frac{\hat{q}}{\|\hat{q}\|} \right\|_2$$

[A. Kendall, M. Grimes & R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization« , ICCV'2015, pp. 2938-2946]

PoseNet training data and test results

training data in green, test data in blue, PoseNet results in red



Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

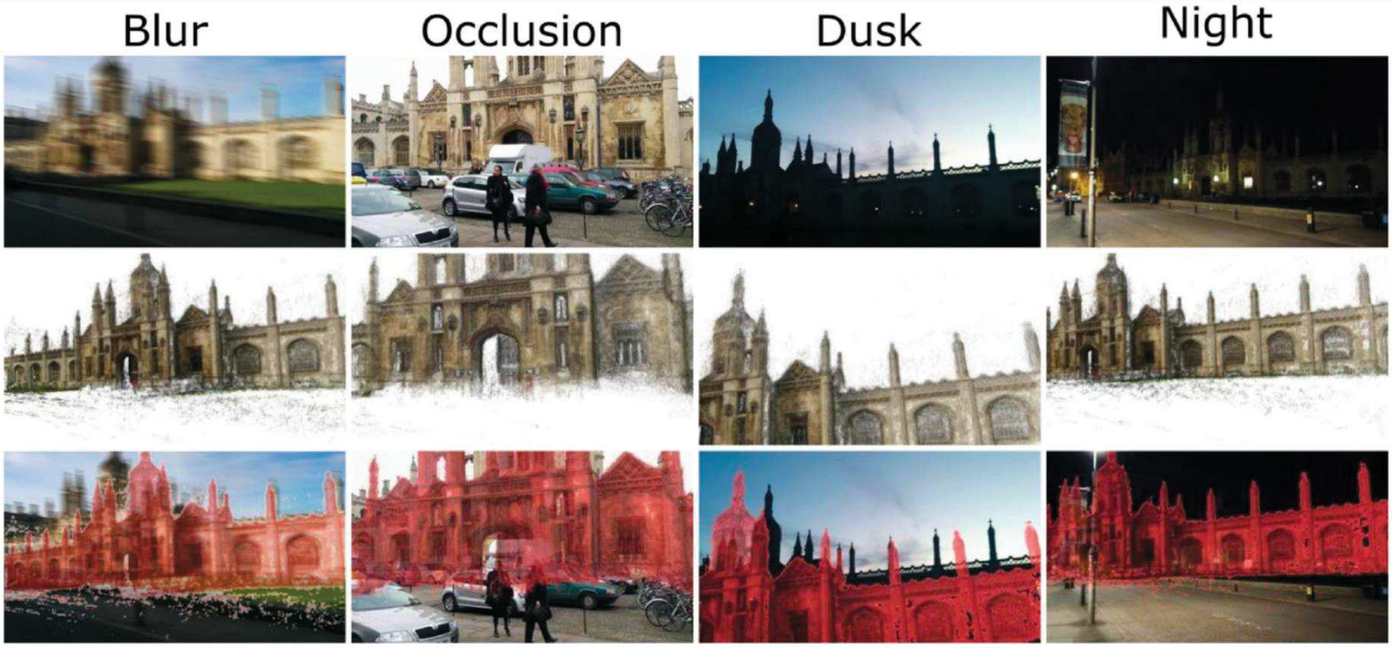
PoseNet results on other tests

King's College *Street* *Old Hospital* *Shop Façade* *St Mary's Church*

Figure 4: **Map of dataset** showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.

PoseNet robustness

Tolerance to environment, unknown intrinsics, weather, etc.



Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.

PoseNet summary: robust to scene change + very fast

✓Robust to lighting, weather, dynamic objects

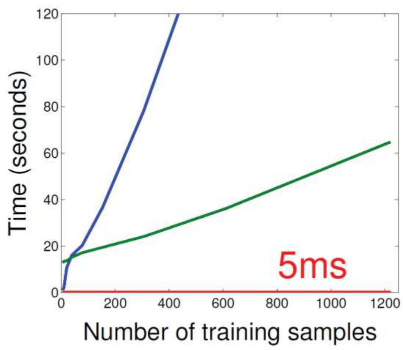
✓Fast inference, <2ms per image on Titan GPU

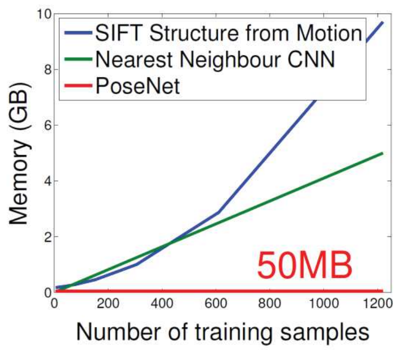
✓Scale not dependent on number of training images

✗Coarse accuracy

✗Difficult to learn both position vs orientation

Alex Kendall, Matthew Grimes and Roberto Cipolla. **PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization.** ICCV, 2015.





PoseNet vs traditional methods

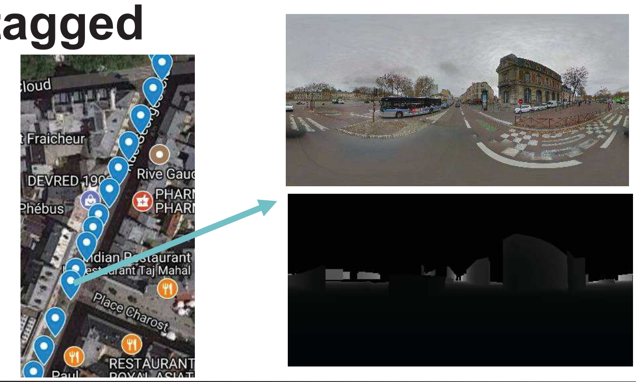
Dataset	PoseNet with Geometry [1]	Active Search (SIFT + Geometry) [2]
King's College	0.88m, 1.04°	**0.42m, 0.55°**
Resolution	**256 x 256 px**	1920 × 1080 px
Inference Time	**2 ms**	78 ms

PoseNet less precise, but much faster and can work with much smaller images

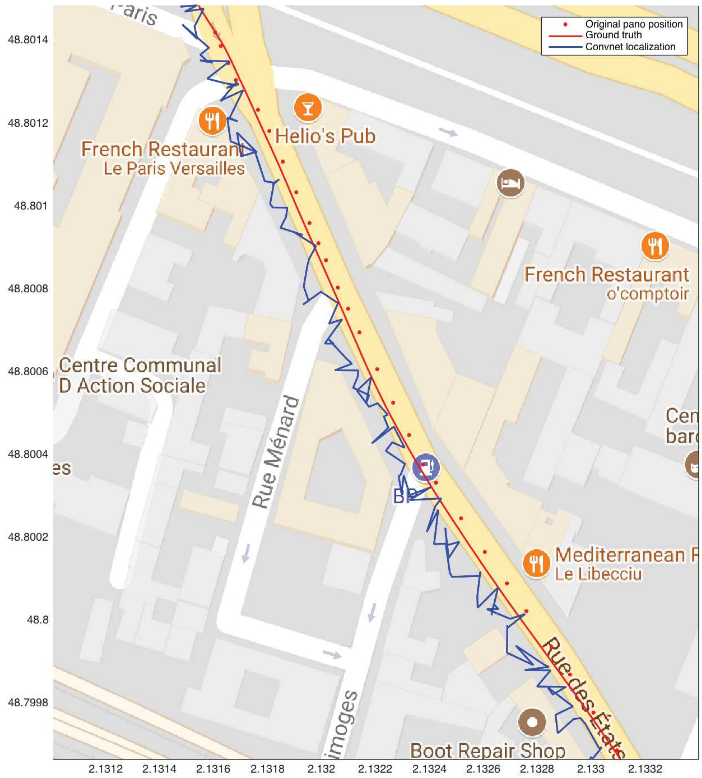
Deep-Learning pose regression from geo-tagged images

- **Learn an only 3-DoF pose (x,y,θ)**
- **Start *transfer learning from <u>InceptionV3</u>* model modified as follows:**
  - **final classifier replaced by a dropout layer**
  - **fully connected layer with 256 neurons added and connected to final 3-dimension pose regressor**
- **Use StreetView "augmented" with virtual views added 4m after each geo-tagged panorama**

Work by Dr Li YU during his PhD thesis @ VeDeCom-MINES_ParisTech (defended in Apr.2018)



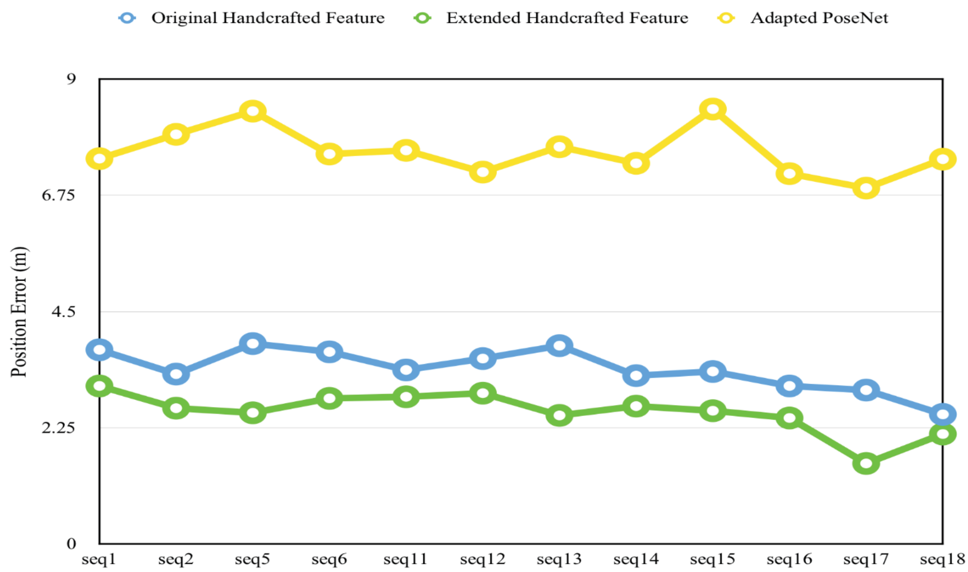
Results of Deep-Learning visual localization trained on GIS images



Results:

- Average error of 7.62m
- 54.2% within a 4m error

GIS-trained adapted PoseNet vs. Coarse-to-fine image matching



- Handcrafted feature method (2x) more accurate + smooth positions
- BUT convNet based method much faster to compute, and reaches accuracy of a standard GPS.

Outline

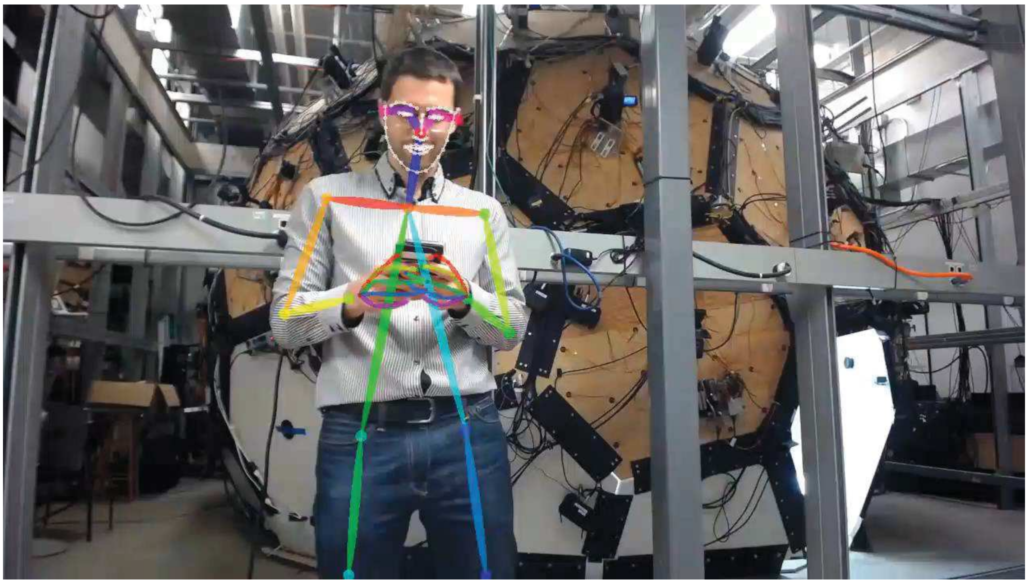
- What can Deep-Learning perform with images?
- Visual Object detection & Semantic Segmentation
- Image-based ego-localization
- **Human posture and movement analysis**

Automated Vehicles interactions with Humans

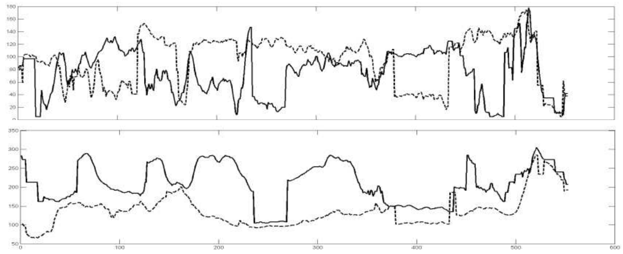
Need to monitor and interpret Human movements, actions & activities:

- Inference of Human intentions (pedestrians and drivers) for Automated Vehicles
- Gestual communication with Humans

Pose estimation now possible from RGB camera (openPose)



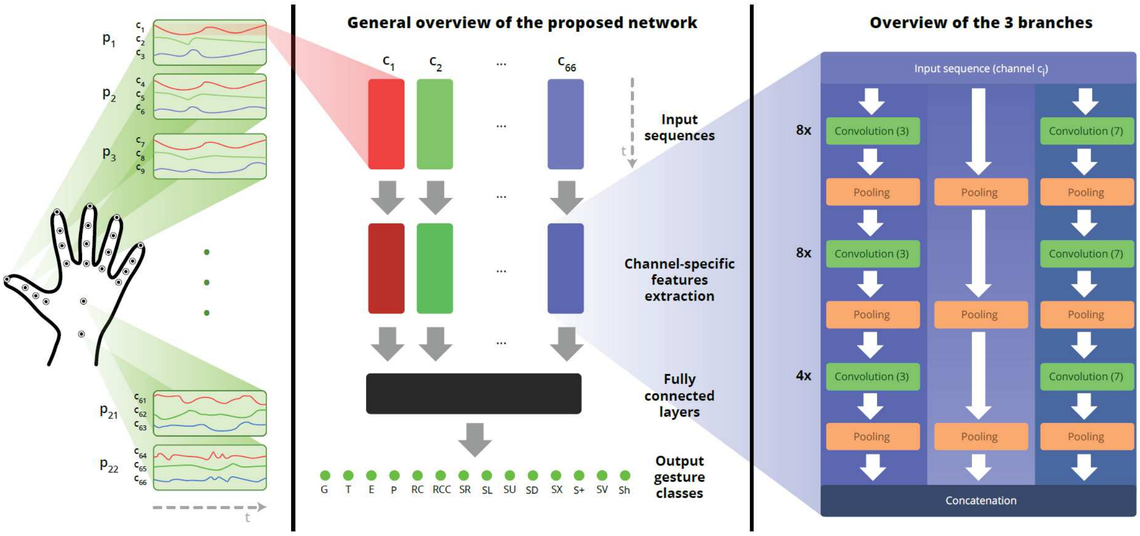
Trajectories of joints



Deep-Learning for time-series

Two main approaches:

- Deep Recurrent Neural Network (RNN) e.g. LSTM or GRU
- Temporal Convolutions



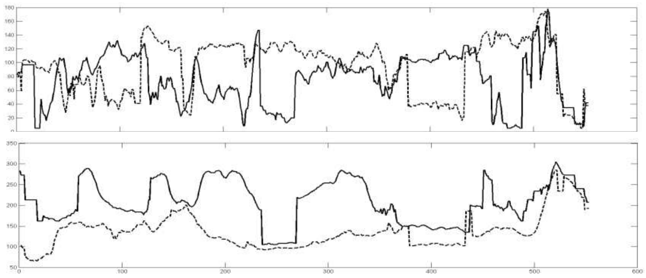
["Convolutional Neural Networks for Multivariate Time Series Classification using both Inter- and Intra- Channel Parallel Convolutions"](), *G. Devineau, W. Xi, F. Moutarde and J. Yang, RFIAP'2018.*
["Deep Learning for Hand Gesture Recognition on Skeletal Data"](), *G. Devineau, W. Xi, F. Moutarde and J. Yang, FG'2018.*

[PhD thesis of Guillaume Devineau @ MINES_ParisTech, supervised by me]

Combining DL pose estimation + Deep Temporal Convolution (or/and RNN)?

Camera

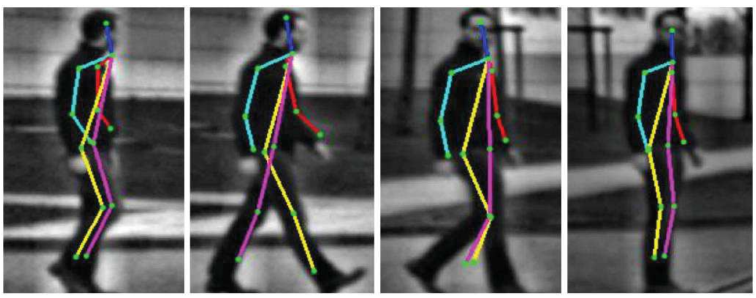
DL pose estimation (openPose/alphaPose)

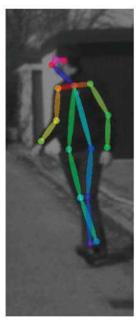


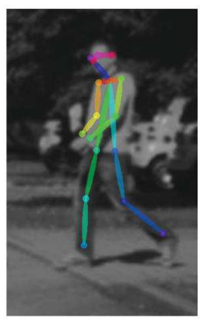
Deep Temporal Convolution (or/and Deep RNN?) for Multivariate Time-Series

Recognized action/gesture

Inferring pedestrian intention from posture?







New PhD thesis started at VeDeCom by Joseph GESNOUIN (supervised by Bogdan Stanciulescu and me)

Conclusions

- **Deep Convolutional Neural Networks already can perform many more things than just image classification:** semantic segmentation, localization from vision, estimation of Human pose, inference of depth from monovision, generation of realistic synthetic images, and learning complex image-based adaptive behaviors

- **The above can be leveraged for many AI challenges for Automated Vehicles:**
  - image-based ego-localization by convNet
  - for Human movements or intents analysis, combining human-pose estimation by DL with Deep Temporal Convolution of time-series seems promising
  - *adaptive behavior learning as an image-based end-to-end driving task [NEXT DECK OF SLIDES]*

Questions ?