# Deep-Learning for Robotics & Autonomous Vehicles

Pr. Fabien Moutarde
Centre de Robotique
MINES ParisTech
PSL Université

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

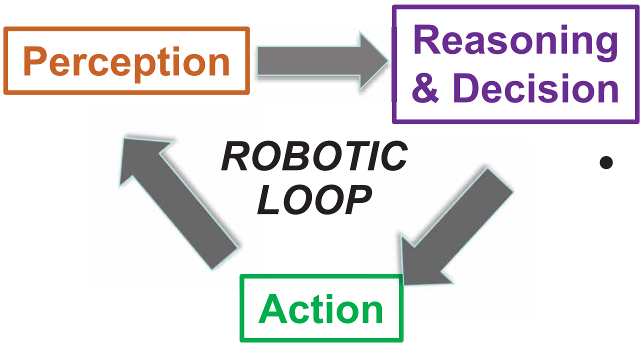


# Outline

- **Introduction: Artificial Intelligences & Machine-Learning**
- AIs for robotics & Autonomous Vehicles
- What can Deep-Learning perform with images?
- Recognition of Gestures/Actions for Human-Robot Collaboration
- Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior

# What is (human) intelligence??

- **Intelligence = reasoning?**
  **or Intelligence = adaptation?**
- **In fact, MANY DIFFERENT TYPES OF INTELLIGENCE**




- **<u>A possible typology:</u>**
  - ***Perception* Intelligence**
  - ***Prediction* Intelligence**
  - ***Reasoning* Intelligence**
  - ***Behavior* Intelligence**
  - ***Interaction* Intelligence**
  - ***Curiosity***

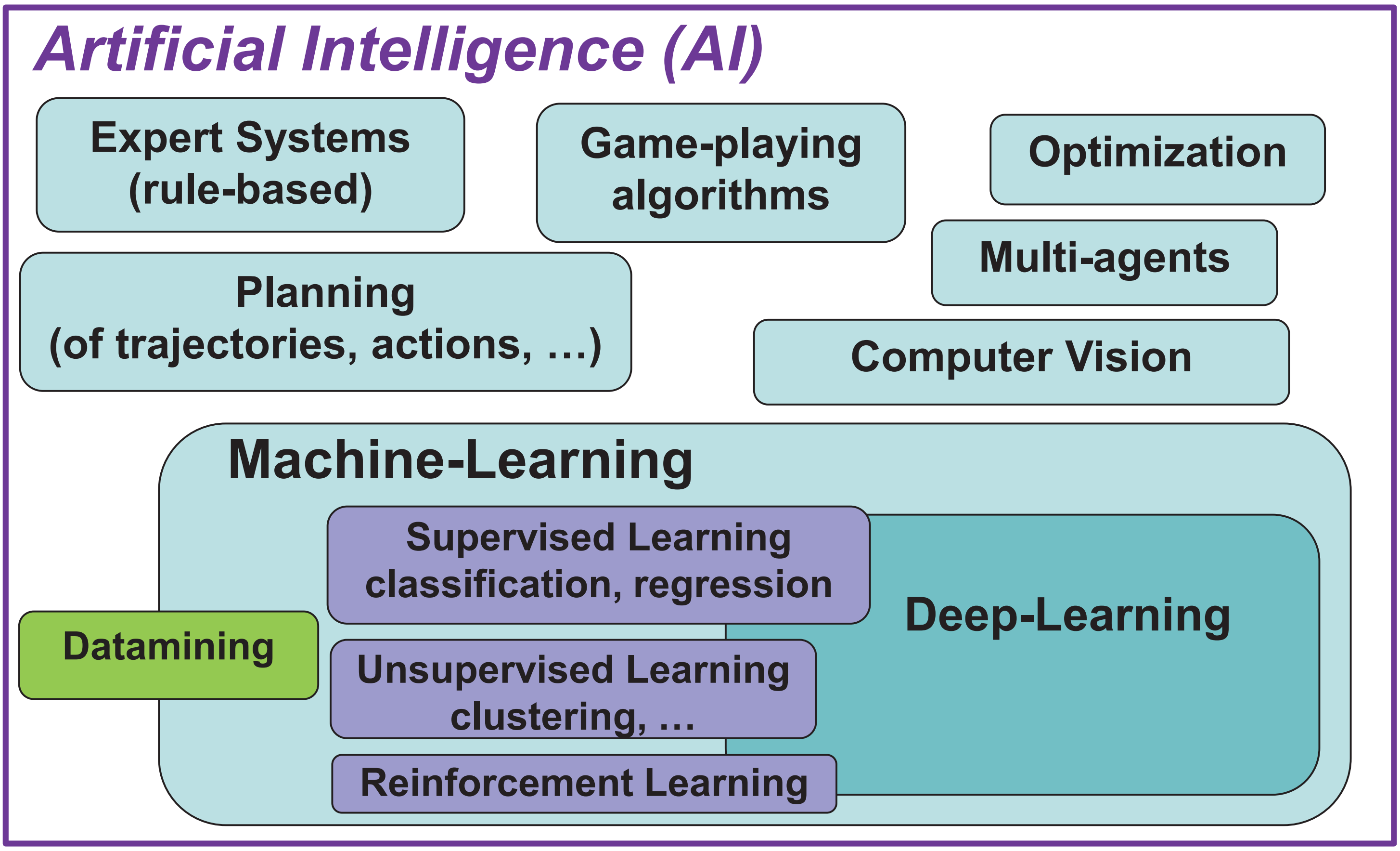


# What is AI?

**Artificial Intelligence, a vast and heterogeneous domain:**

- Rule-based reasoning, expert systems
- Algorithms for playing games (chess, Go, etc..)
- Multi-agents, emergence of collective behavior
- …
- Optimization, Operational Research,
  Dynamic Programming
- Planning (of trajectories, tasks, etc…)
- Computer vision, pattern recognition
- <u>Machine-Learning</u>
  = empirical data-driven modelling
  *(optimization, based on examples, of a parametric model)*

# Artificial IntelligenceS






# Outline

- Introduction: Artificial Intelligences & Machine-Learning
- **AIs for robotics & Autonomous Vehicles**
- What can Deep-Learning perform with images?
- Recognition of Gestures/Actions for Human-Robot Collaboration
- Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior

# "*Traditional*" (industrial) Robots

**Repetitive actions, fast, strong, …**
**BUT dangerous and NOT VERY ADAPTIVE**
**(simple "*automatons*")**

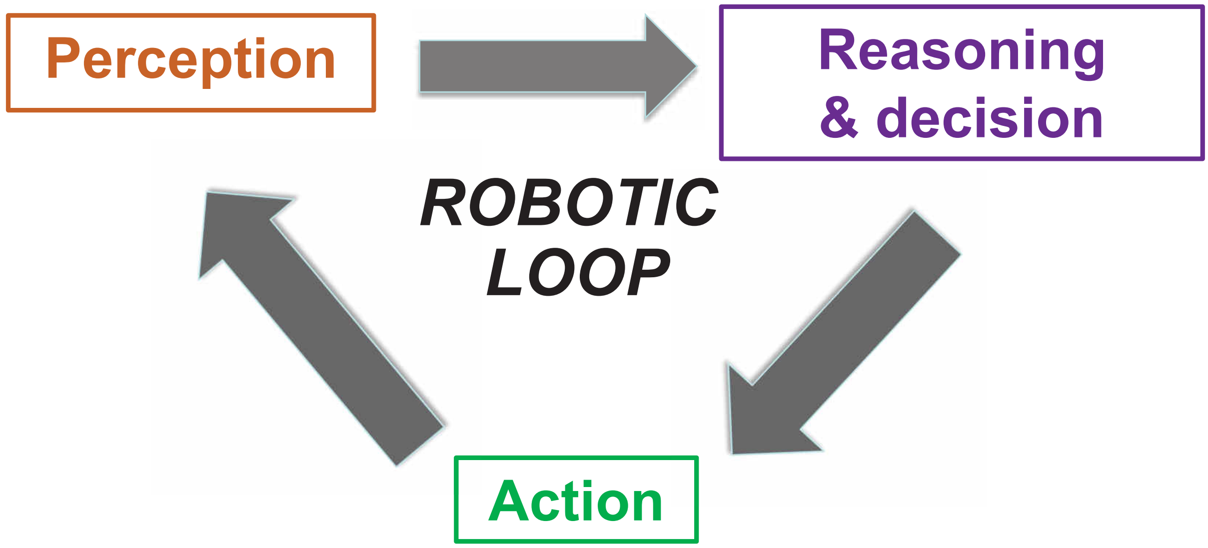


# "Intelligents" robots (≈ adaptive and/or interactive)


**React adaptively to environment…**

**… and/or interact with Humans**

# General principle of robots






# Autonomous Vehicles are mobile robots!

# What types of AIs are needed for Autonomous Vehicles?

- "Semantic" interprétation 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

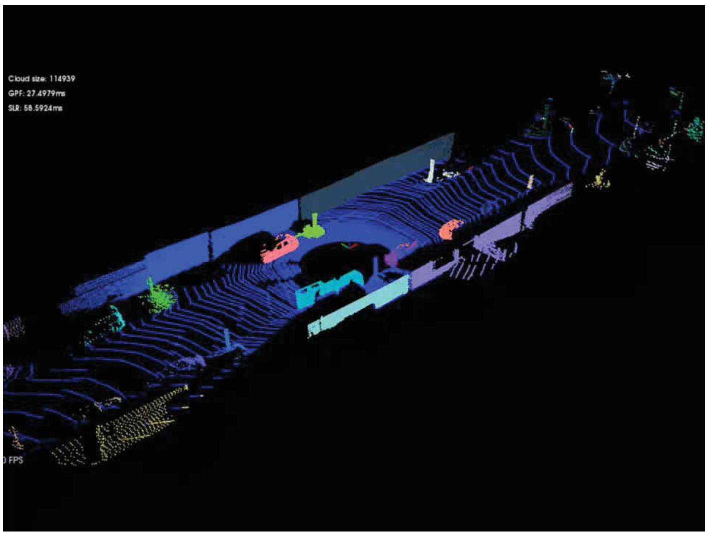
- For Advanced Driving Assistance Systems (ADAS) and partial automated driving (level 3-4):
  - Analyze and understand attention and activities or gestures of the "driver-supervisor"



# Intelligent Perception for Autonomous Vehicles

## Essential need: real-time "understanding" of surroundings


From camera



From LIDAR

**Strong real-time constraint: process ≥ 15 frames/seconde**

# What types of "intelligence" do ROBOTS need?

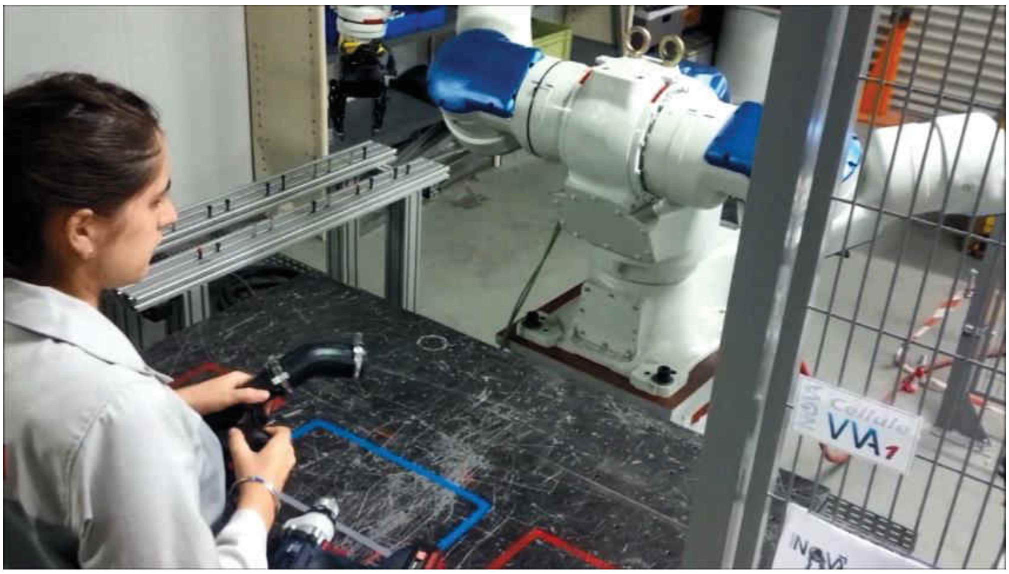
- Analyze & interpret a dynamic environment
  - Recognize a place & ego-localize
  - Detect/localize & categorize "objects"
  - Track & predict their movements
  - Guess "intentions"
- Choose action/movement to be performed
  - Decision logics
- Adapt/optimize chosen action/movement
  - Having a BEHAVIOR rather than rigid rules
- Interact with humans or other robots
  - Speech Recognition
  - Natural Language Processing, ability to dialog
  - Recognition of Gestures/Actions, of emotions?
  - Coordination/collaboration between robots



# Intelligent Perception for Collaborative Robots

**Strong need:**

**monitoring and interpreting movements, actions & activities of Humans around**




Action recognition for Human-Robot Collaboration

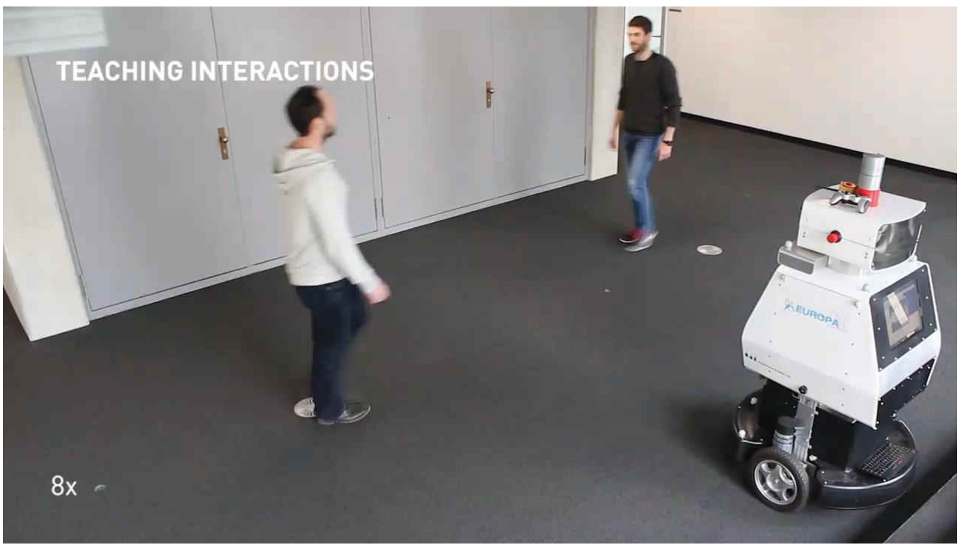
*[centre de Robotique de MINES ParisTech, Chaire PSA "Robotique et Réalité Virtuelle"]*

# Major challenges for Intelligent Robots & Autonomous Vehicles

- **Inference of INTENTIONS of Humans**
- **Human activity understanding**
- **Learning of adaptive BEHAVIOR**
  - Learning by demonstration/imitation
  - Learning by reinforcement
  - Abstraction of task rather than recording of trajectory
  - One/few shot(s) learning
- **Coordination/collaboration**
  - between robots (cooperative planning, etc…)
  - with Humans:
    - Non-verbal communication (gestures, movement, gaze)
    - Learning of implicit "social rules"



# Coordination with Humans: "human-aware" AI




**Challenge: learn implicit "social rules" of interaction**

# Outline

- Introduction: Artificial Intelligences & Machine-Learning
- AIs for robotics & Autonomous Vehicles
- **What can Deep-Learning perform with images?**
- Recognition of Gestures/Actions for Human-Robot Collaboration
- Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior



# Image-based Deep-Learning

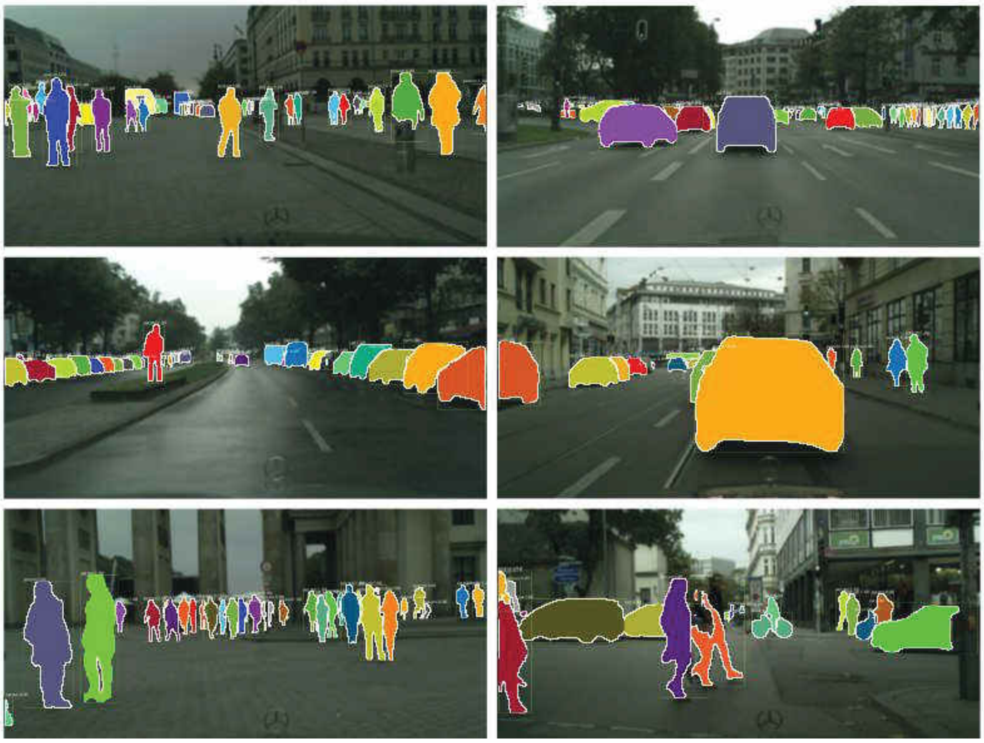
- Image classification
- Visual object detection and categorization
- Semantic segmentation of images
- Realistic image synthesis
- Image-based localization
- Estimation of Human pose
- Inference of 3D (depth) from monocular vision
- Learning image-based behaviors
  - End-to-end driving from front camera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

# Visually detect & categorize objects


**Visual objects Simultaneous Detection and Categorization with Faster_RCNN**

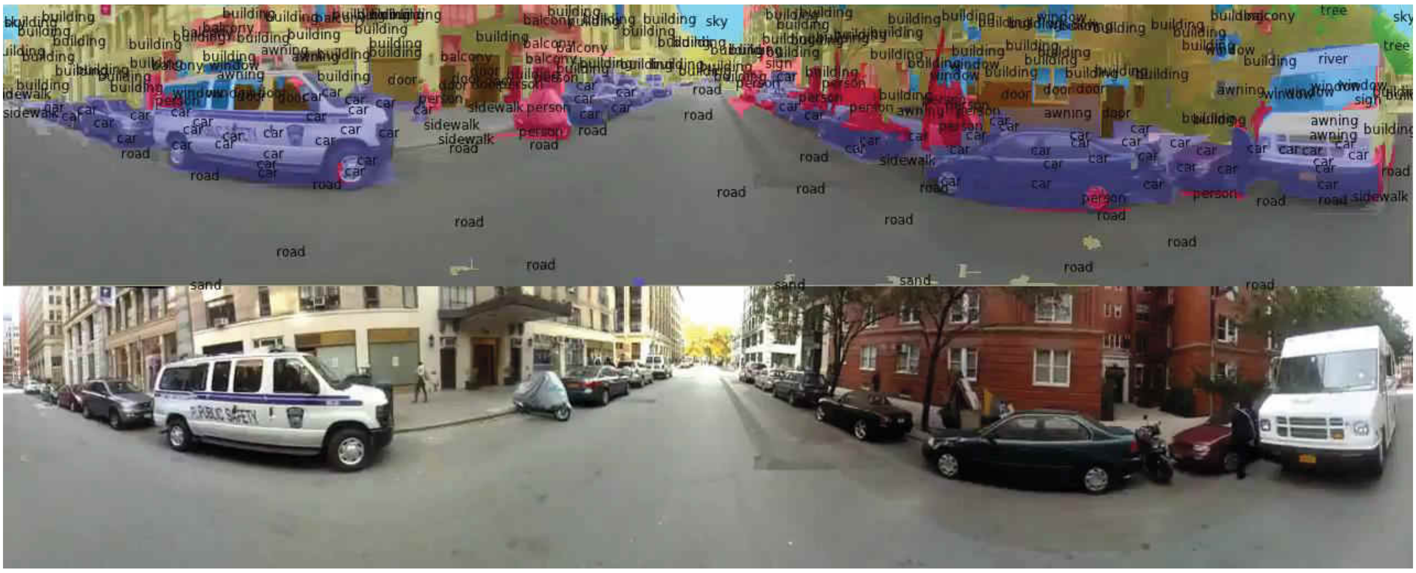


# Beyond bounding-boxes: getting *contours* of objects



**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

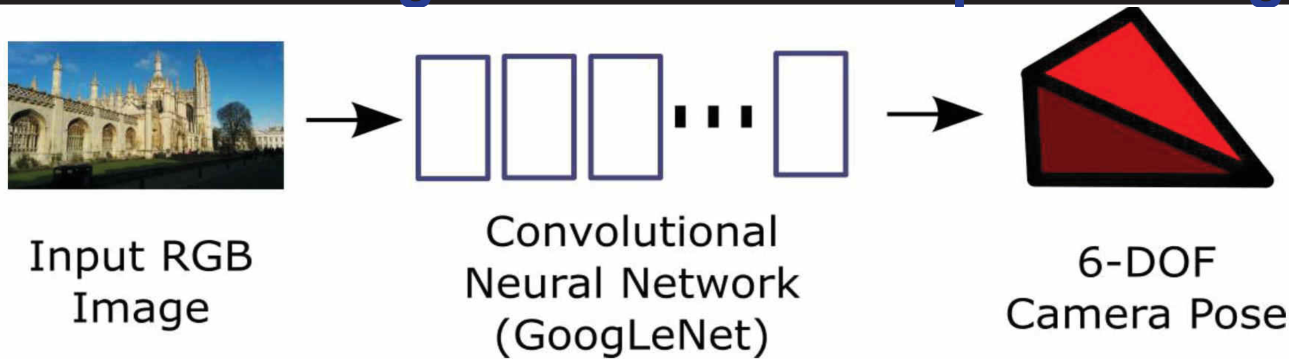


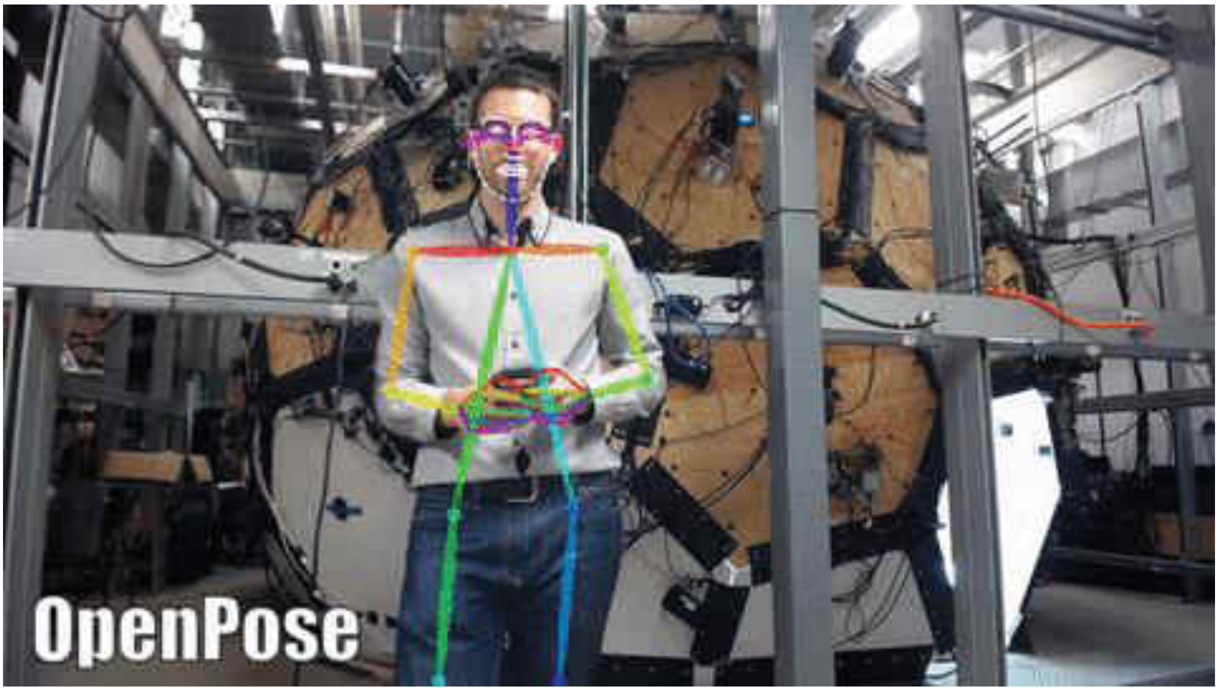


Figure 4: **Map of dataset** showing training frames (green), testing frames (blue) and their predicted camera pose (red). The testing sequences are distinct trajectories from the training sequences and each scene covers a very large spatial extent.

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



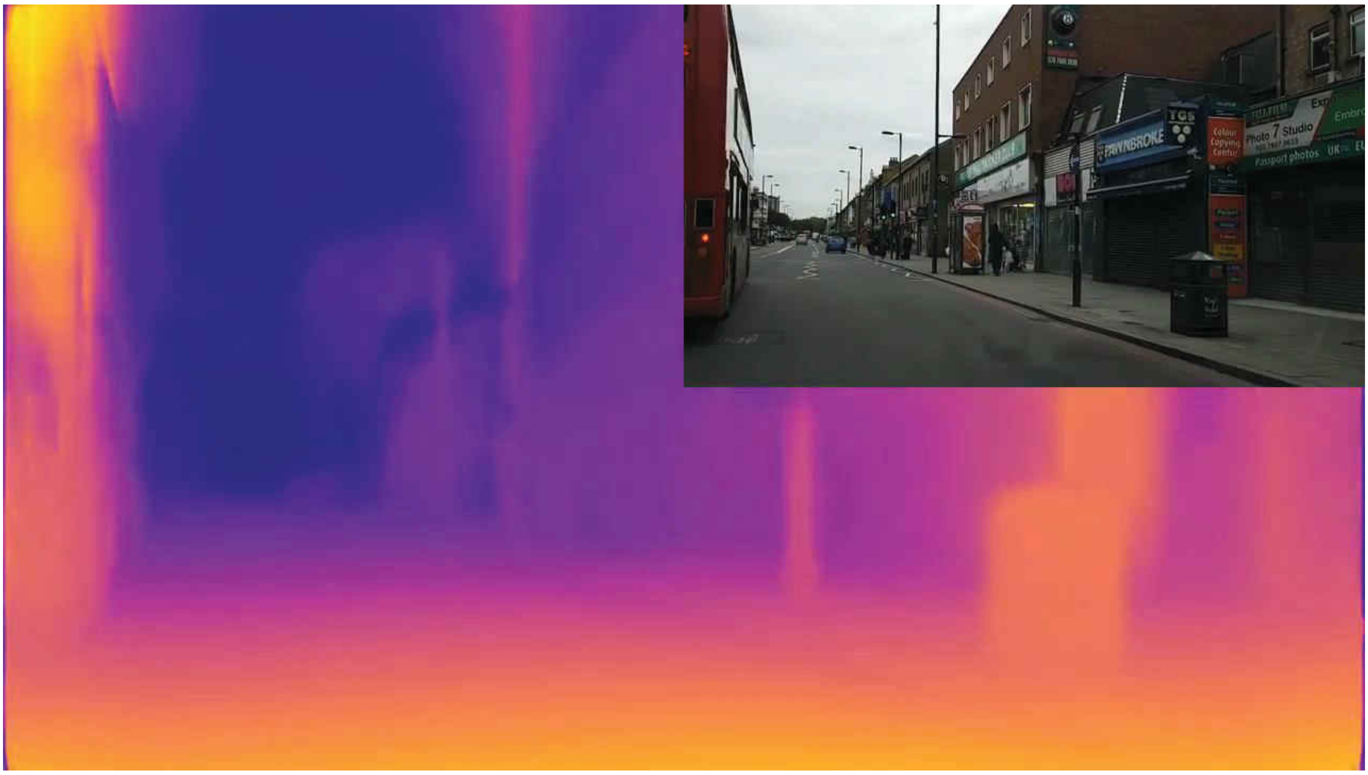
# Human pose estimation by Deep-Learning




**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]*

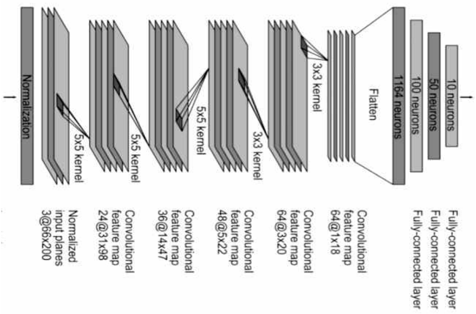


# Autonomous learning of a task by a robot


**Supervised learning, but with success/failure easily estimated automatically, for a bin-picking task**

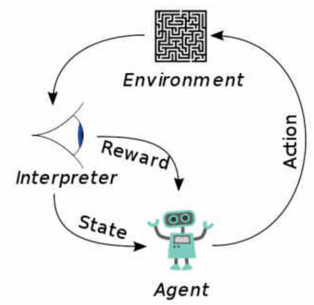
# End-to-end driving from camera by Deep-Learning

ConvNet input:
Cylindrical projection of fisheye camera




ConvNet output:
steering angle

## Imitation Learning from Human driving on real data

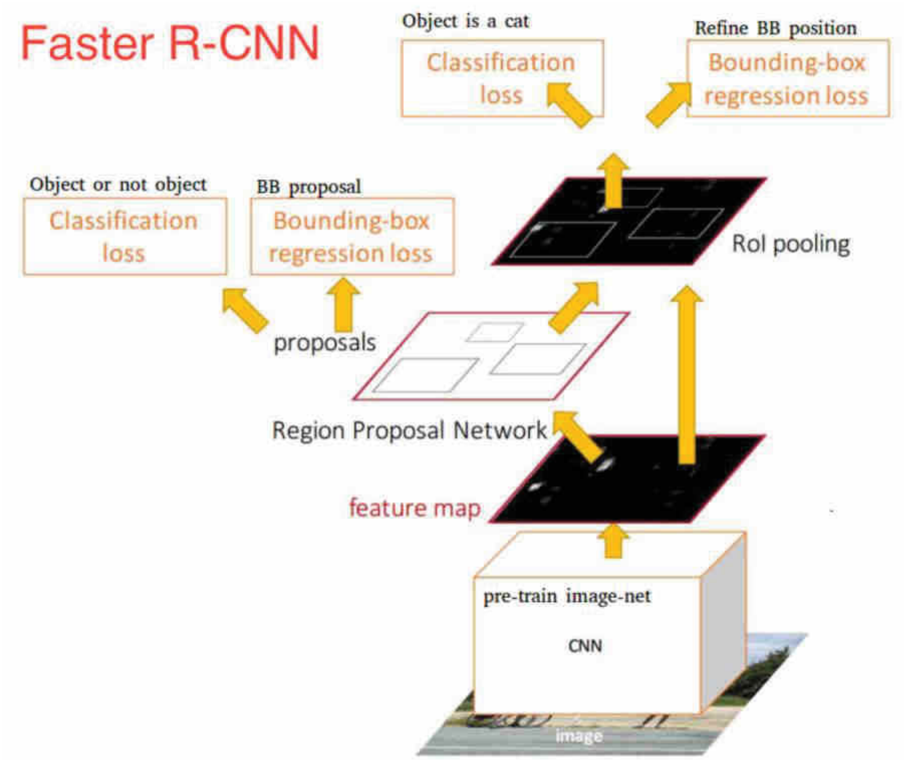



## End-to-end driving via Deep *Reinforcement* Learning

*[thèse CIFRE Valeo/MINES-ParisTech en cours]*



# Visual objects Detection and Categorization: Faster_RCNN



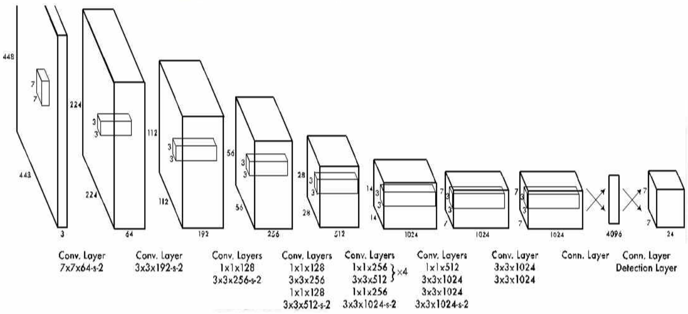

**Region Proposal Network (RPN) on top of standard convNet.**
**End-to-end training with combination of 4 losses**

# Object visual detection *without proposal*

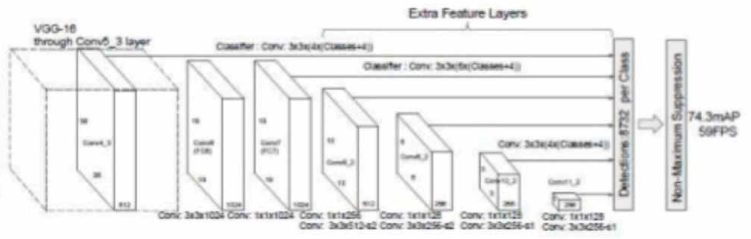
## Solve detection as a regression problem ("single-shot" detection)

## YOLO and SSD

YOU ONLY LOOK ONCE(YOLO)




SINGLE SHOT MULTIBOX DETECTOR(SSD)

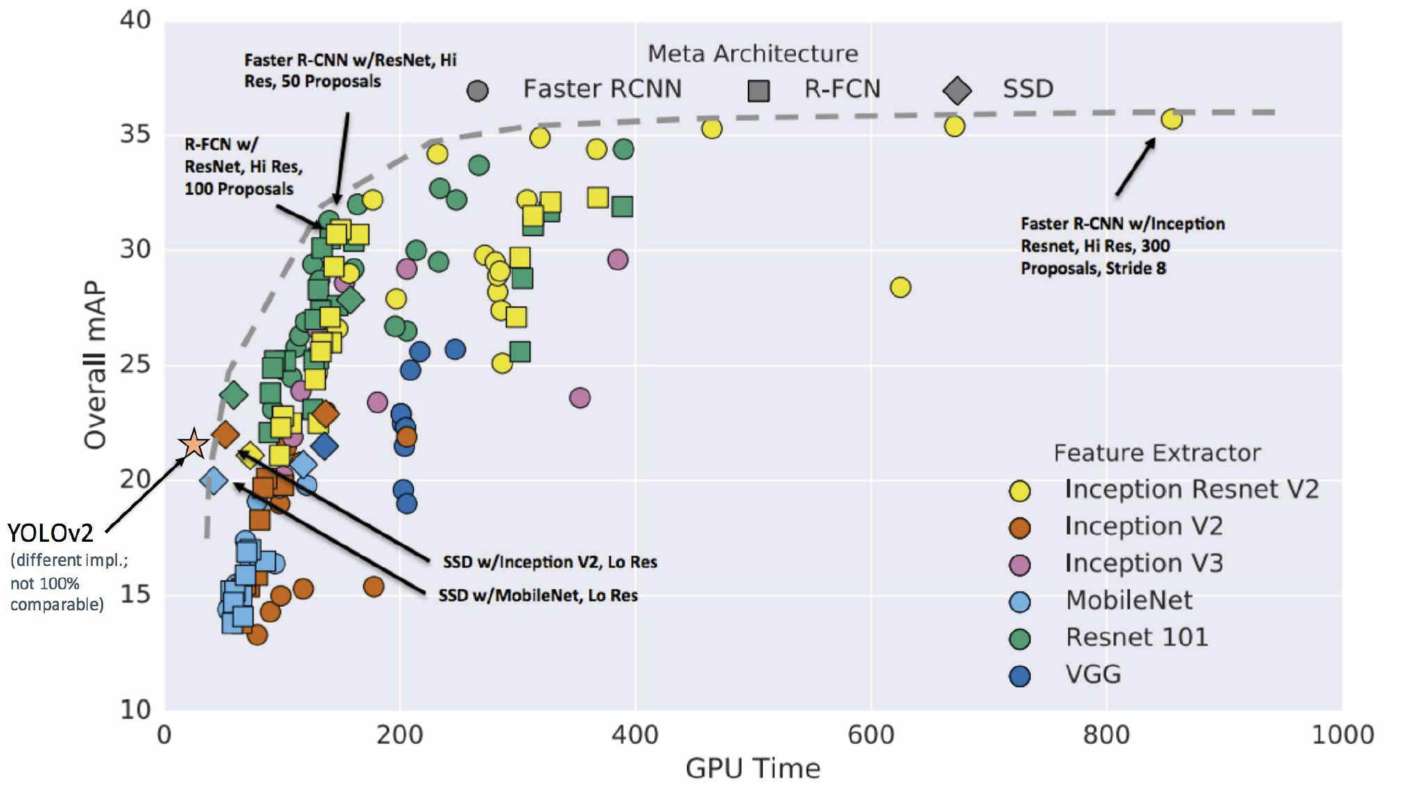



Images from: https://www.slideshare.net/TaegyunJeon1/pr12-you-only-look-once-yolo-unified-realtime-object-detection

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

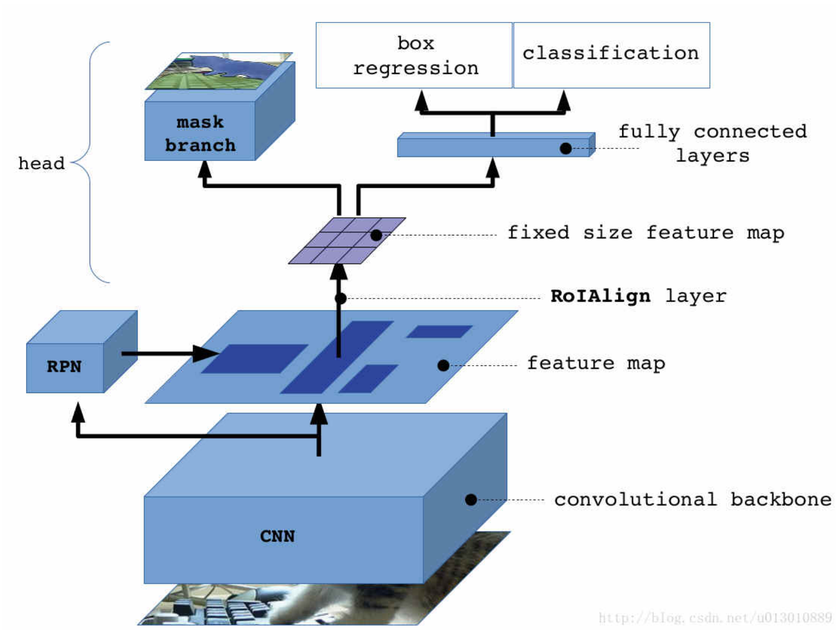


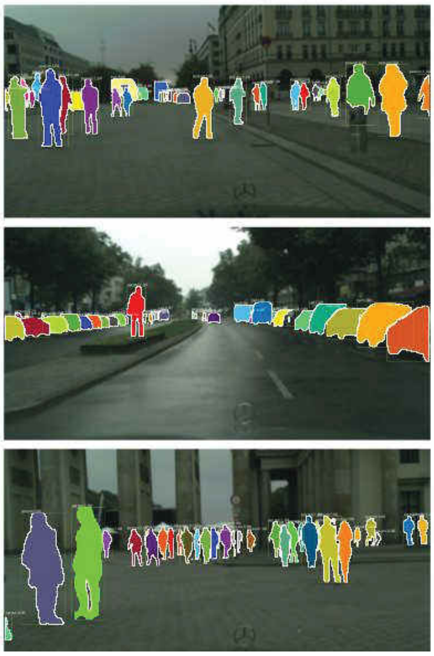
# Recent comparison of object detection convNets




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

# Mask_RCNN: Categorization and Localization with shape/*contours*

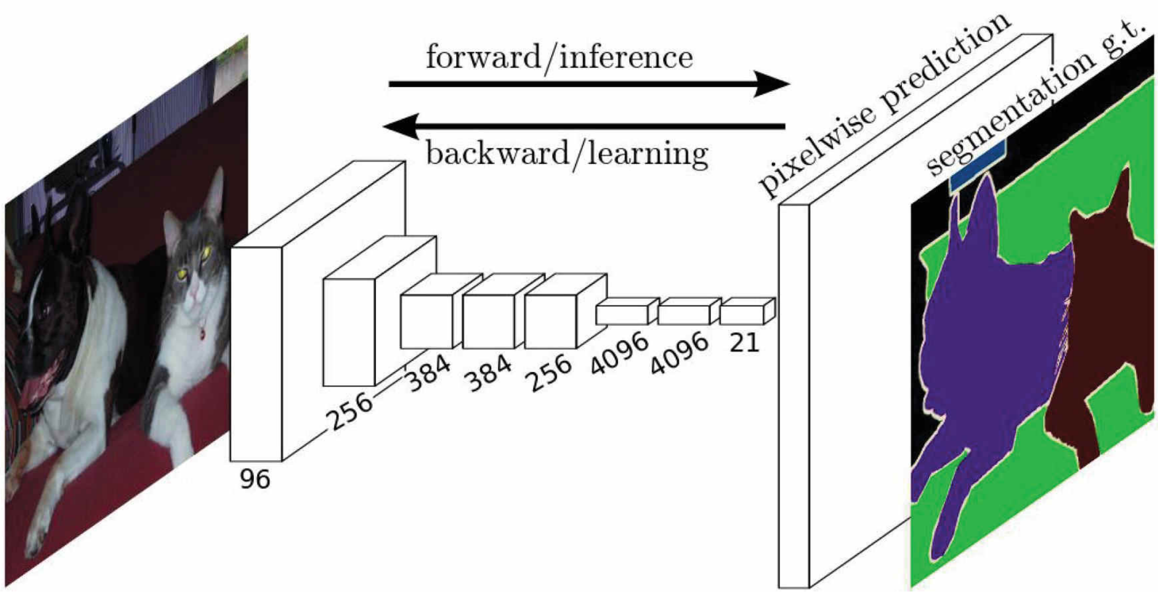


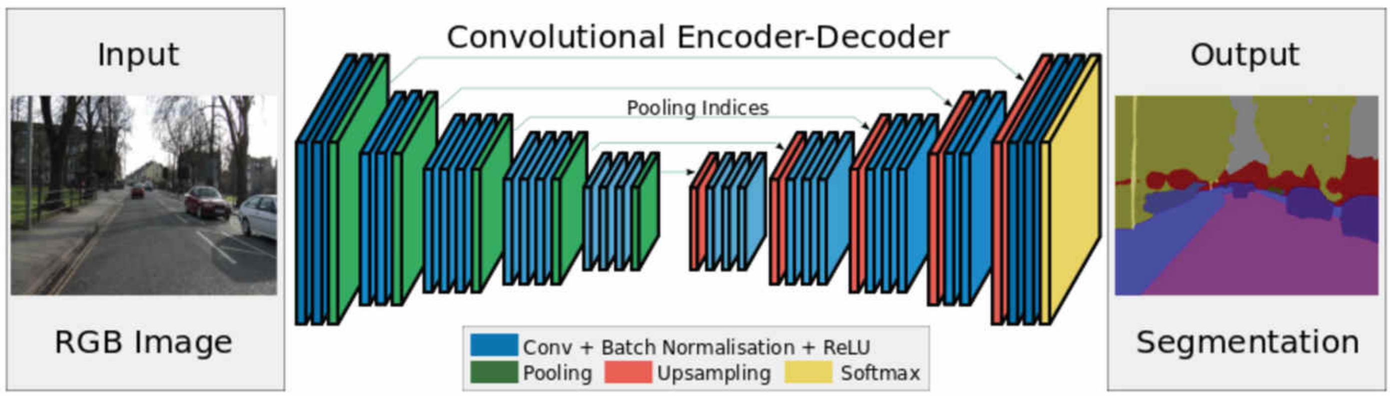
**Mask R-CNN architecture (left) extracts detailed contours and shape of objects instead of just bounding-boxes**

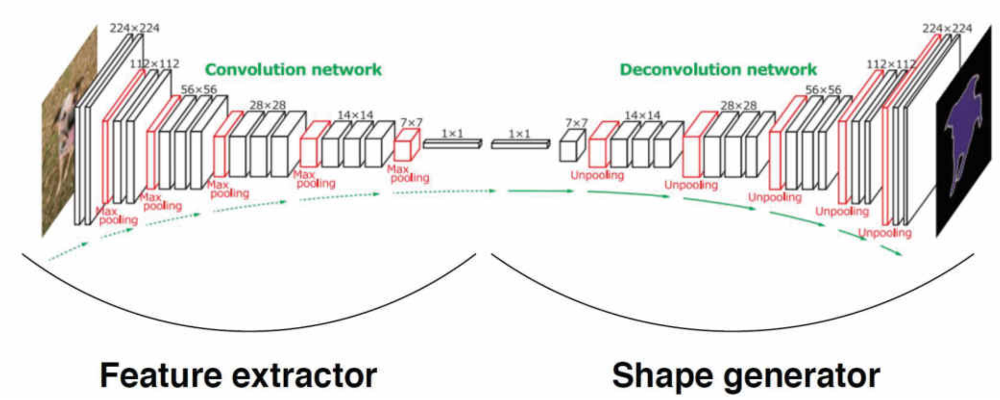


# Deep-Learning approach for semantic segmentation



# Convolutional Encoder-Decoder






# Outline

- Introduction: Artificial Intelligences & Machine-Learning
- AIs for robotics & Autonomous Vehicles
- What can Deep-Learning perform with images?
- **Recognition of Gestures/Actions for Human-Robot Collaboration**
- Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior

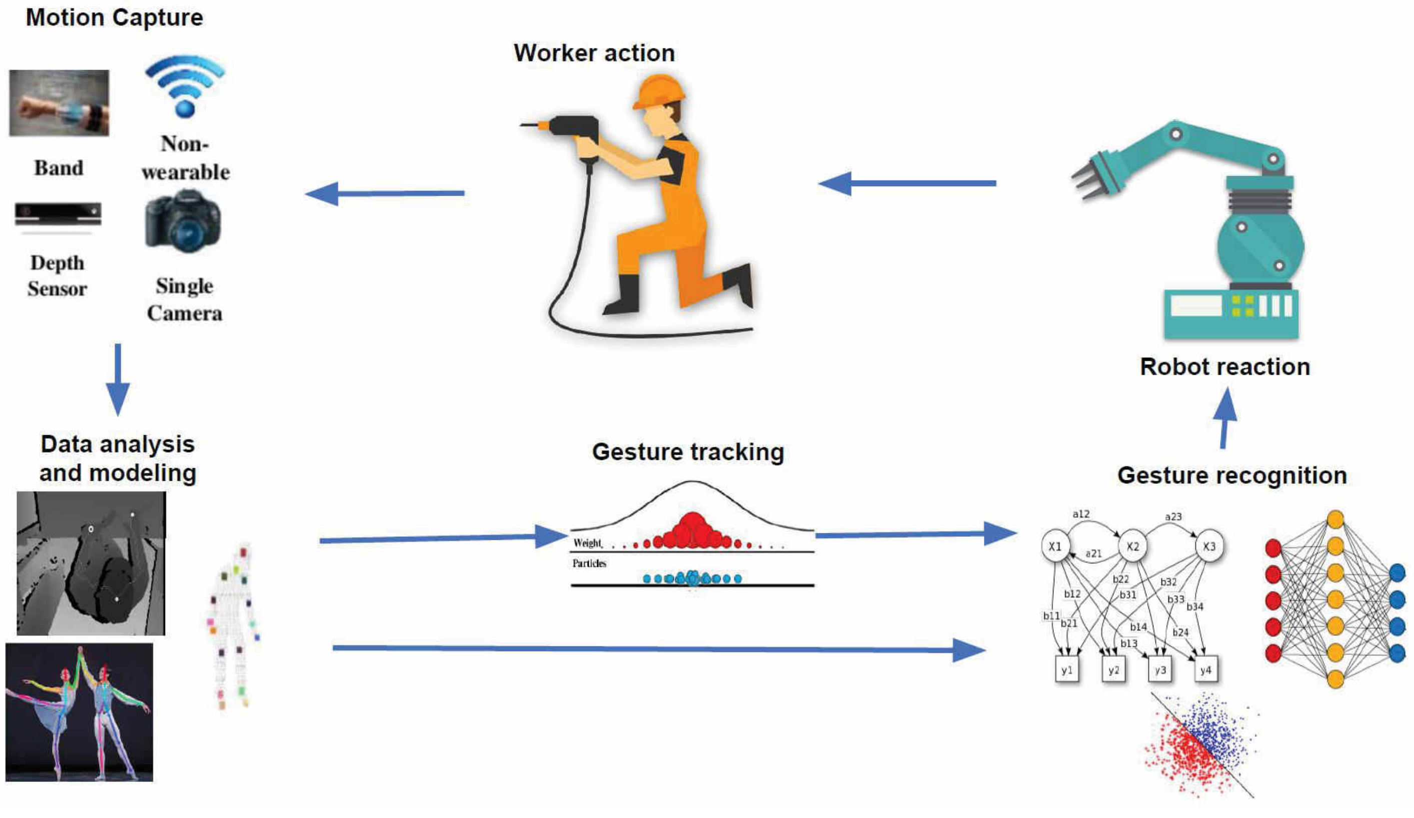
# Human-Robot/AutonomousVehicle interactions

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

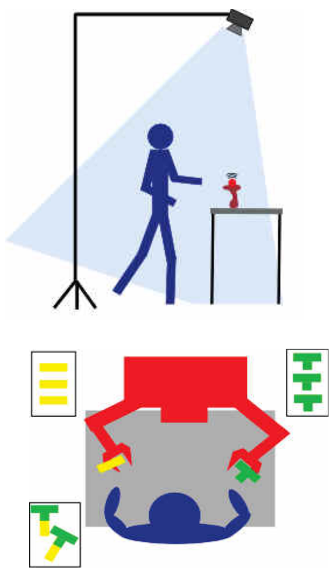
- Action recognition for collaborative robots
- Inference of Human intentions (pedestrians and drivers) for Autonomous Vehicles
- Gestual communication with Humans for both

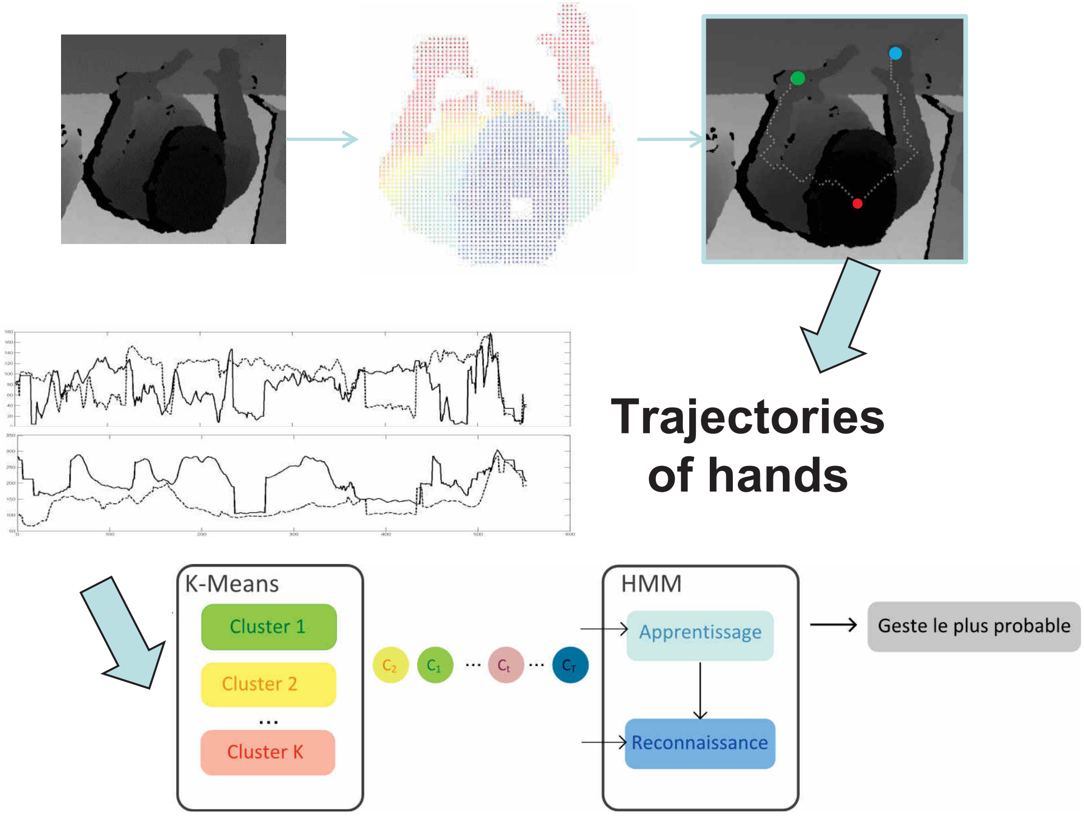


# Human-Robot Collaboration (HRC)



# Example of Action Recognition for HRC

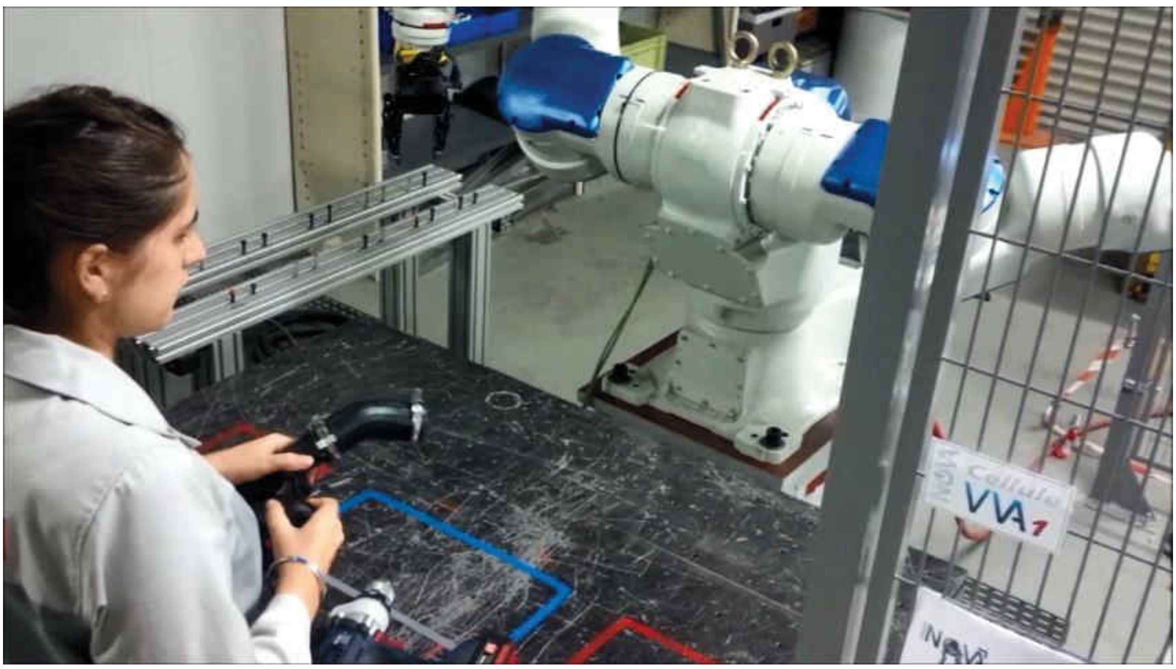





***PhD thesis of Eva Coupeté at MINES_Paris (defended in 2016), sponsored by Chaire PSA "Robotique et Réalité Virtuelle"***



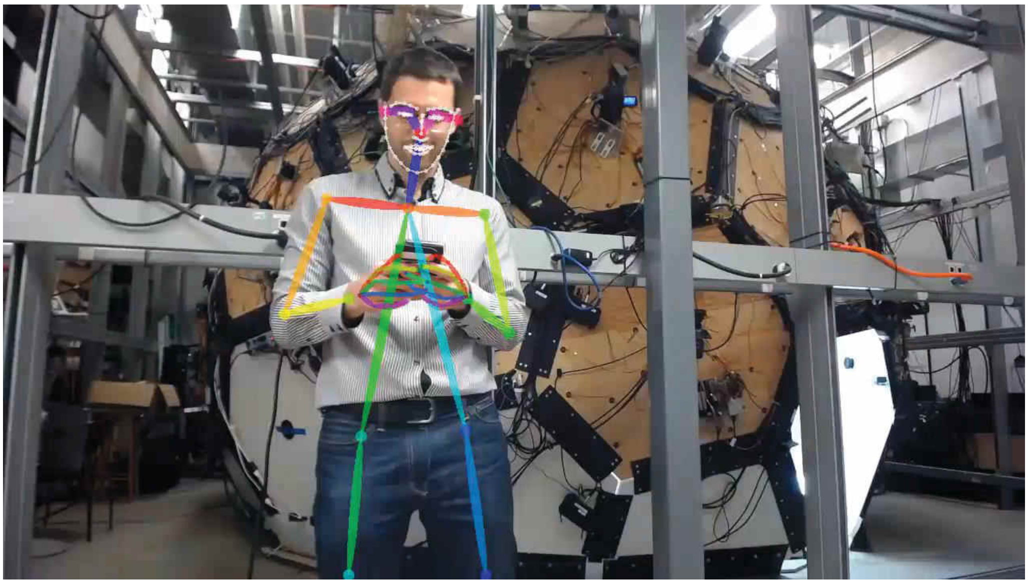
# Example of Action Recognition result

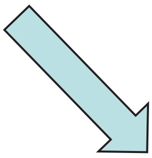


**Action recognition for Human-Robot Collaboration**

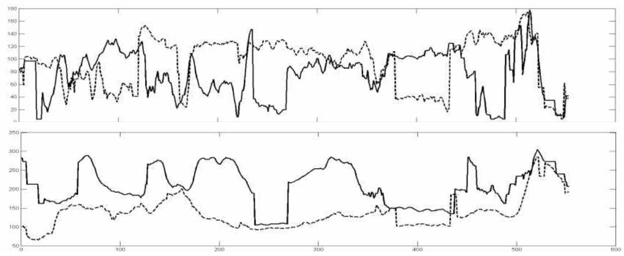
*[centre de Robotique de MINES ParisTech, Chaire PSA "Robotique et Réalité Virtuelle"]*

# Pose estimation now possible from RGB camera (openPose)





Trajectories of joints

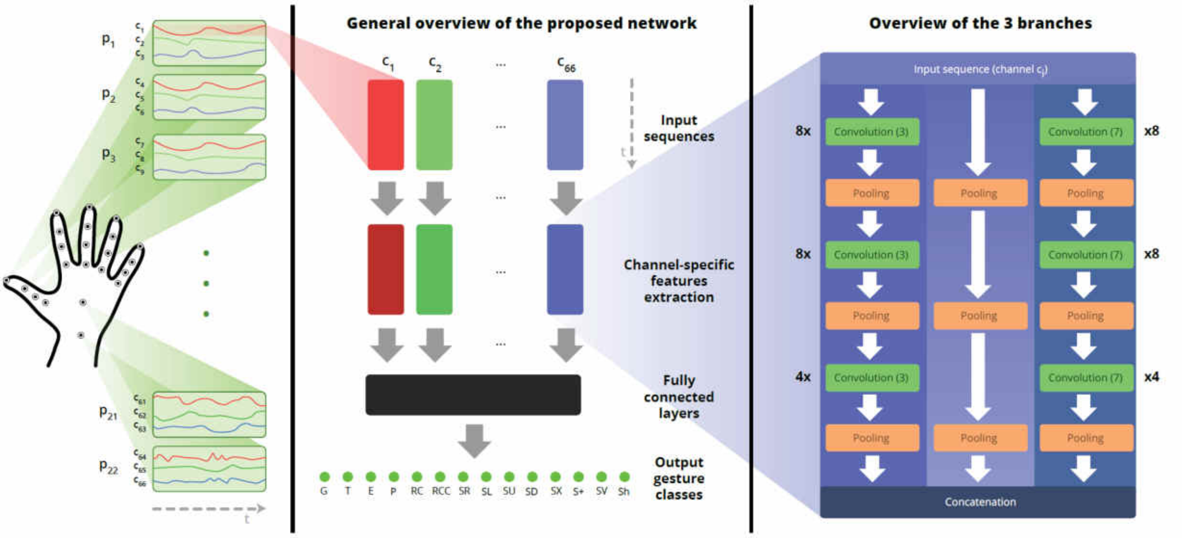




# Deep-Learning for time-series

**<u>Two main approaches:</u>**

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



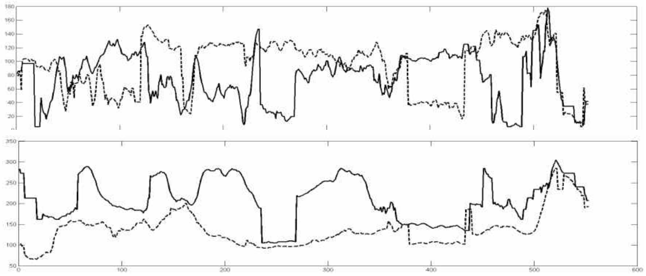

*"[Convolutional Neural Networks for Multivariate Time Series Classification using both Inter- and Intra- Channel Parallel Convolutions]"*, *G. Devineau, W. Xi, F. Moutarde and J. Yang, RFIAP'2018.*
*"[Deep Learning for Hand Gesture Recognition on Skeletal Data]"*, *G. Devineau, W. Xi, F. Moutarde and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

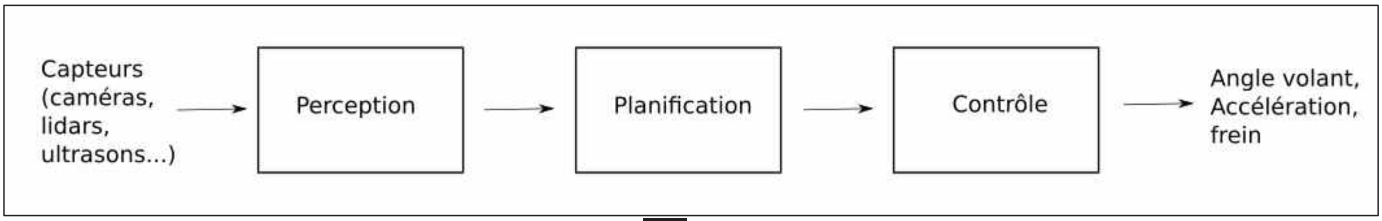
***Work in Progress (PhD thesis of Salwa El Kaddaoui at MINES_ParisTech, within H2020 European project COLLABORATE***

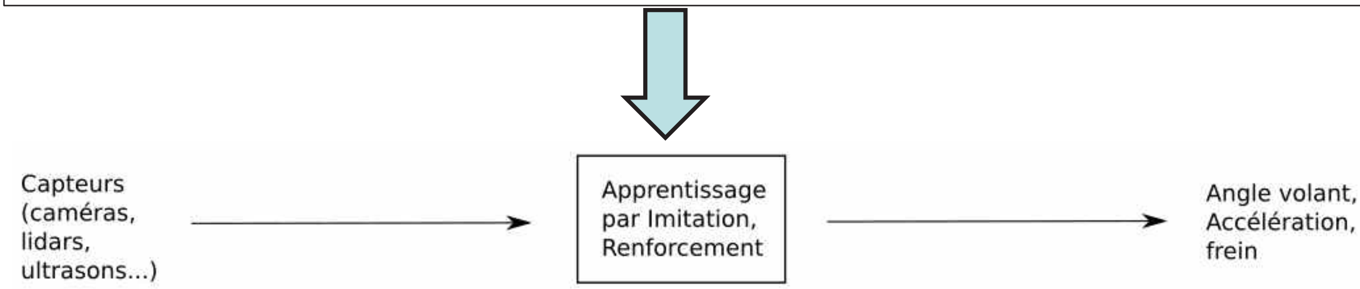


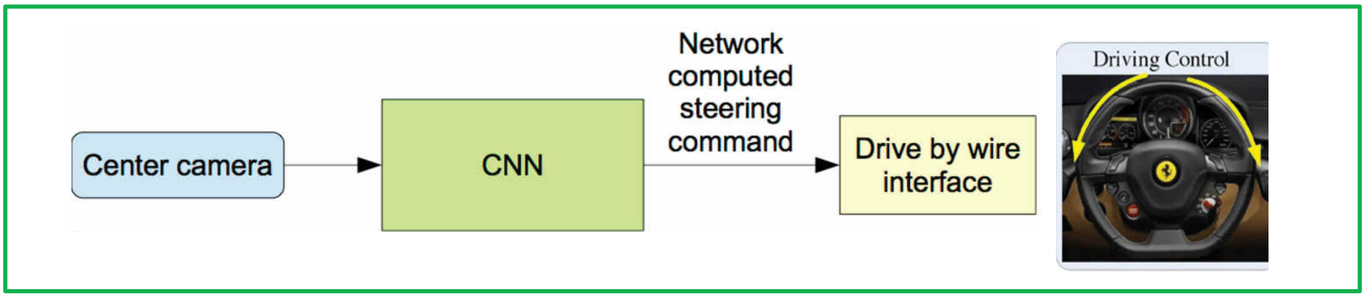
# Outline

- Introduction: Artificial Intelligences & Machine-Learning
- AIs for robotics & Autonomous Vehicles
- What can Deep-Learning perform with images?
- Recognition of Gestures/Actions for Human-Robot Collaboration
- **Imitation Learning & Deep Reinforcement Learning for Autonomous Driving and design of Robots behavior**

# Idea of end-to-end driving

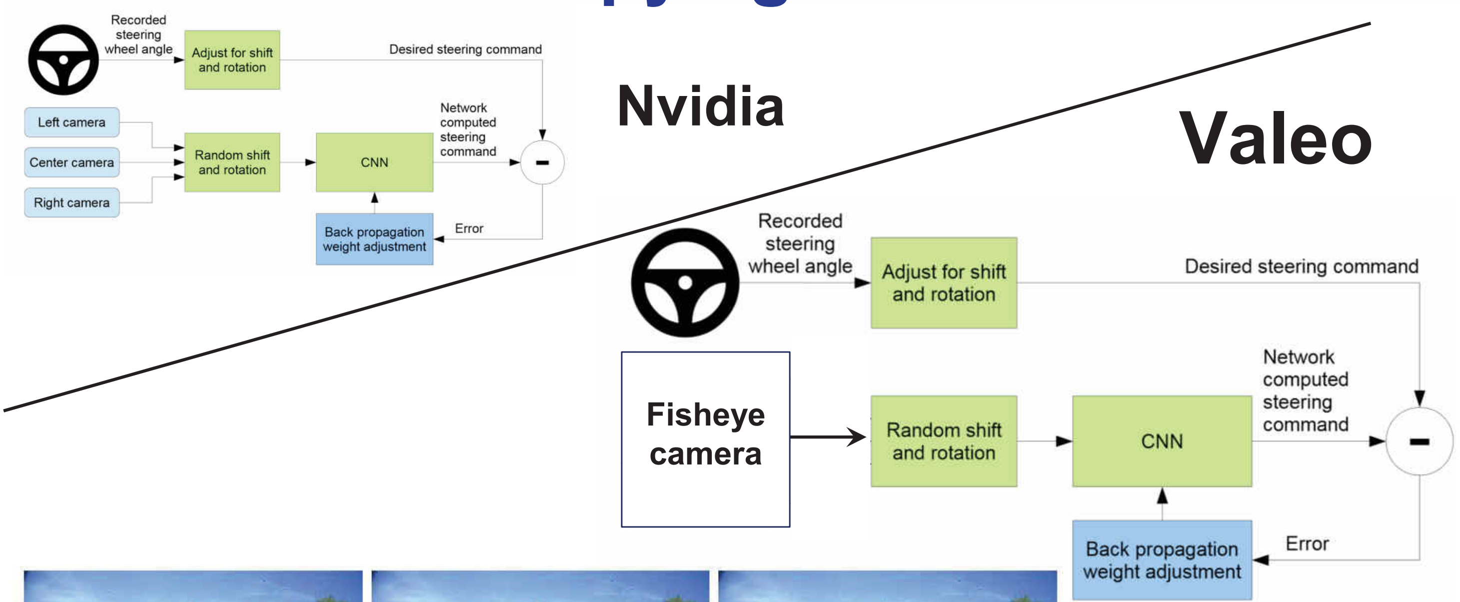





# Imitation Learning: "copying" human driver

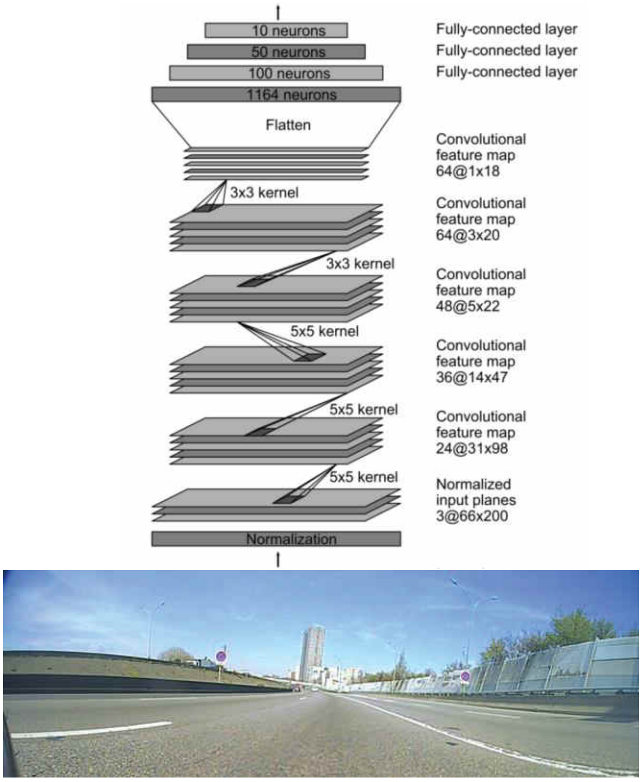



(a) Translation d'un mètre vers la gauche (b) Image originale (c) Translation d'un mètre vers la droite

"End to End Vehicle Lateral Control Using a Single Fisheye Camera", Marin Toromanoff, Emilie Wirbel, Frédéric Wilhelm, Camilo Vejarano, Xavier Perrotton et Fabien Moutarde, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018), Madrid, Spain, 1-5 oct. 2018.

# End-to-end driving convNet

## ConvNet output: steering angle



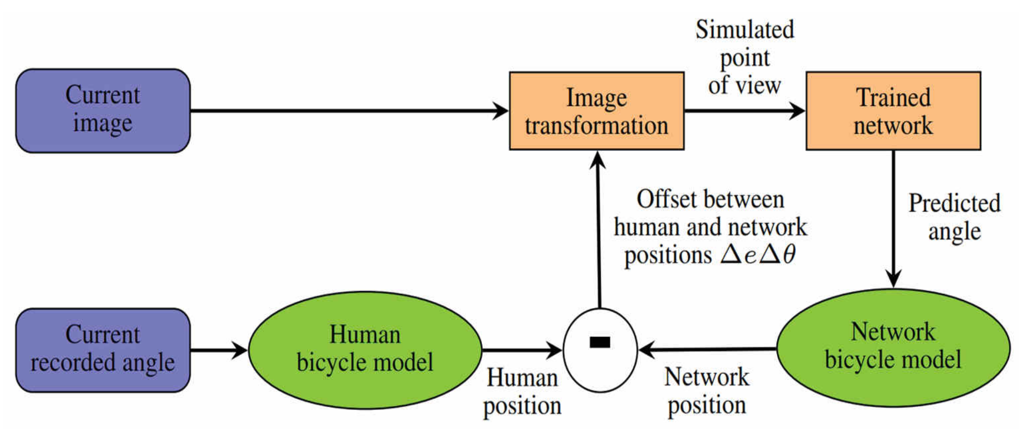

**ConvNet input: Cylindrical projection of fisheye camera**



# Real data + "simulator" with real images

**Training+testing dataset = <u>10000 km</u> and <u>200 hours</u> of human driving in openroad** (highways, urban streets, country roads, etc…) under various weather conditions

**TrainSet = 10 million images, TestSet = 3 million images.**




**"ConvNet in-the-loop" simulator with <u>real</u> images**

# End-to-end driving: closed loop evaluation

TABLE V: Autonomy (%) and mean absolute distance (MAD, in cm) according to data distribution and validation scenario, the baseline is just going straight.

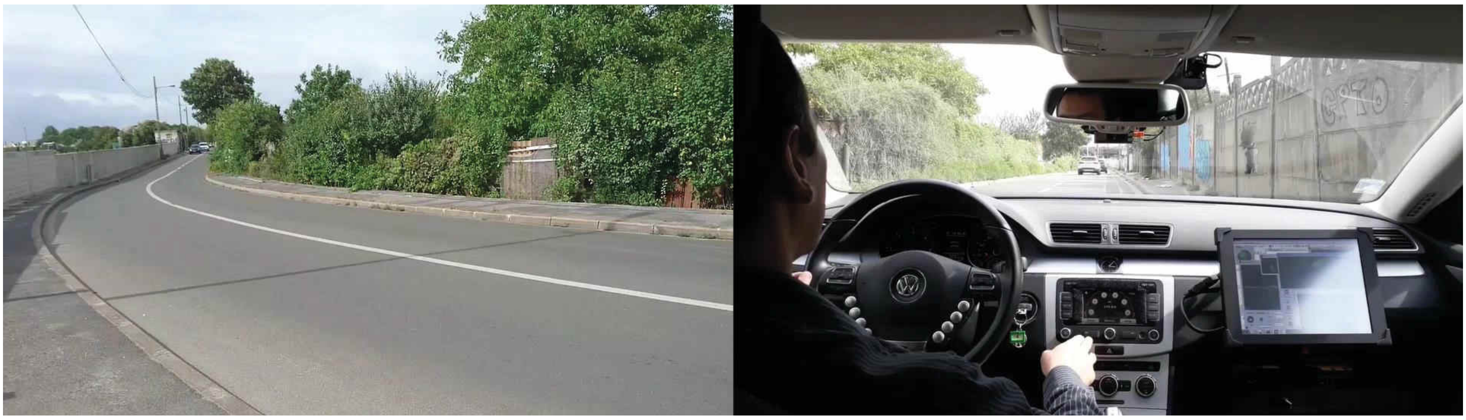
| Scenario | Urban | | Highways | | Sharp turns | |
|---|---|---|---|---|---|---|
| Metric | Aut. (%) | MAD (cm) | Aut. (%) | MAD (cm) | Aut. (%) | MAD (cm) |
| Original | 99.3 | 16 | **98.7** | **19** | **73.7** | **30** |
| Sel. #1 | 98.9 | **15** | 97.7 | 25 | 83.7 | **27** |
| Sel. #2 | **99.5** | 16 | 97.2 | 24 | **87.5** | 28 |
| Oversamp. | **98** | **18** | **91.8** | **29** | 82.5 | 29 |
| Baseline | 8 | 36 | 14 | 41 | 0 | 35 |

TABLE VI: Comparison of performance between individual networks and bagging

| Scenario | Urban | | Highways | | Sharp turn | |
|---|---|---|---|---|---|---|
| Metric | Aut. (%) | MAD (cm) | Aut. (%) | MAD (cm) | Aut. (%) | MAD (cm) |
| Weights #1 | **99.5** | 16 | 97.2 | 24 | **87.5** | 28 |
| Weights #2 | 98.9 | 15 | 97.7 | 25 | 83.7 | **27** |
| Weights #3 | 99.3 | 16 | **98.7** | **19** | **73.7** | **30** |
| Weights #4 | 98.6 | **18** | **92** | **26** | 85 | 29 |
| Weights #5 | **98.4** | 15 | 96.4 | 21 | 83.7 | 28 |
| Bagging | **99.5** | **13** | **98.7** | **19** | **87.5** | **27** |

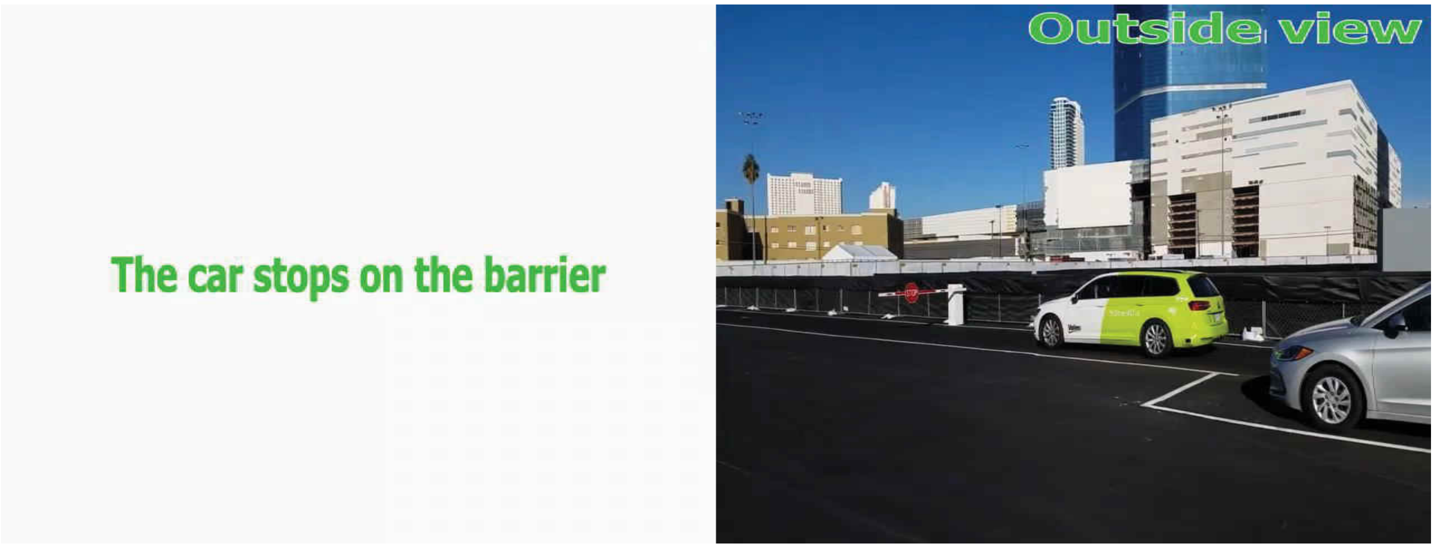


# Real vehicle end-to-end driving (learnt by imitation)



*[Work by Valeo using ConvNet trained by my CIFRE PhD student Marin Toromanoff]*

# Demonstration at CES'2018

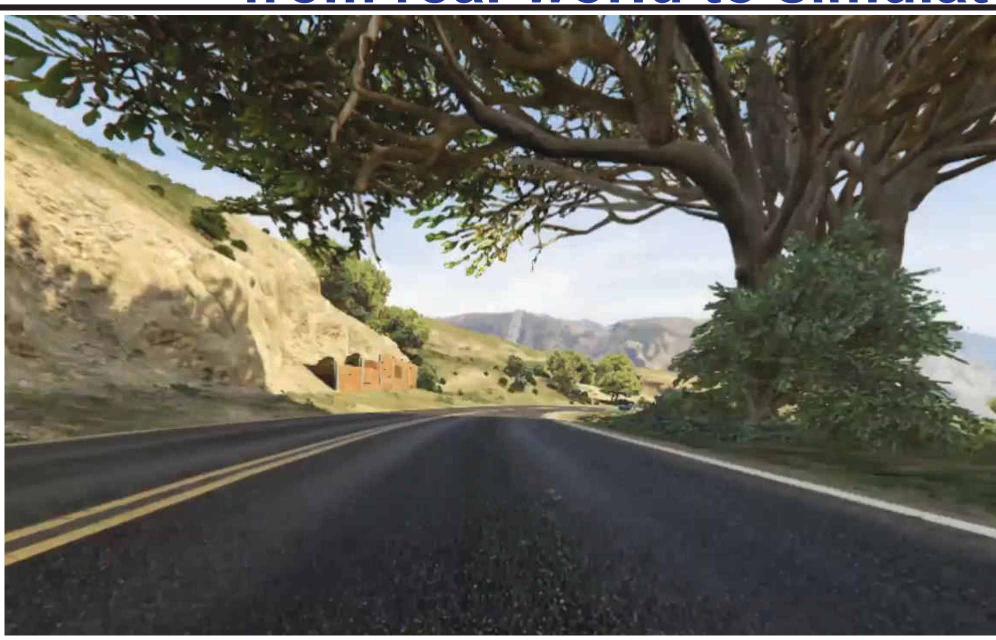



*[Work by Valeo using ConvNet trained by my CIFRE PhD student Marin Toromanoff]*


"End to End Vehicle Lateral Control Using a Single Fisheye Camera", Marin Toromanoff, Emilie Wirbel, Frédéric Wilhelm, Camilo Vejarano, Xavier Perrotton et Fabien Moutarde, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018), Madrid, Spain, 1-5 oct. 2018.




# Transferability from real-world to simulator

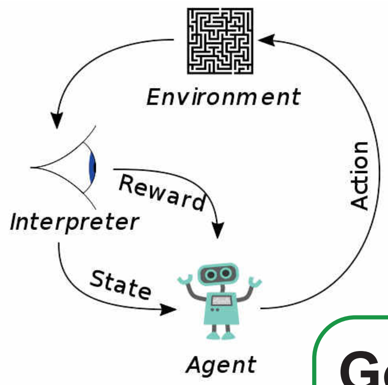


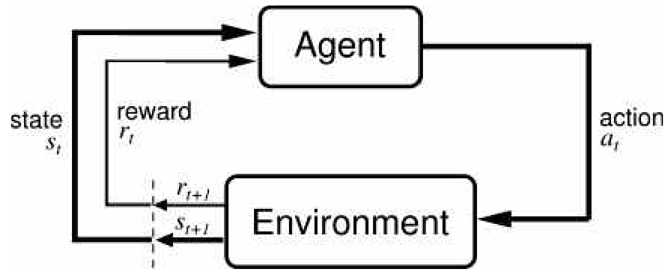
Test of driving convNet in GTA simulator

**Note that learning was done *only on real-world data* (by human driving imitation)**

*[Work by my Valeo CIFRE PhD student Marin Toromanoff]*

# Reinforcement Learning (RL)




**Goal: find a "policy" $a_t=\pi(s_t)$ that**

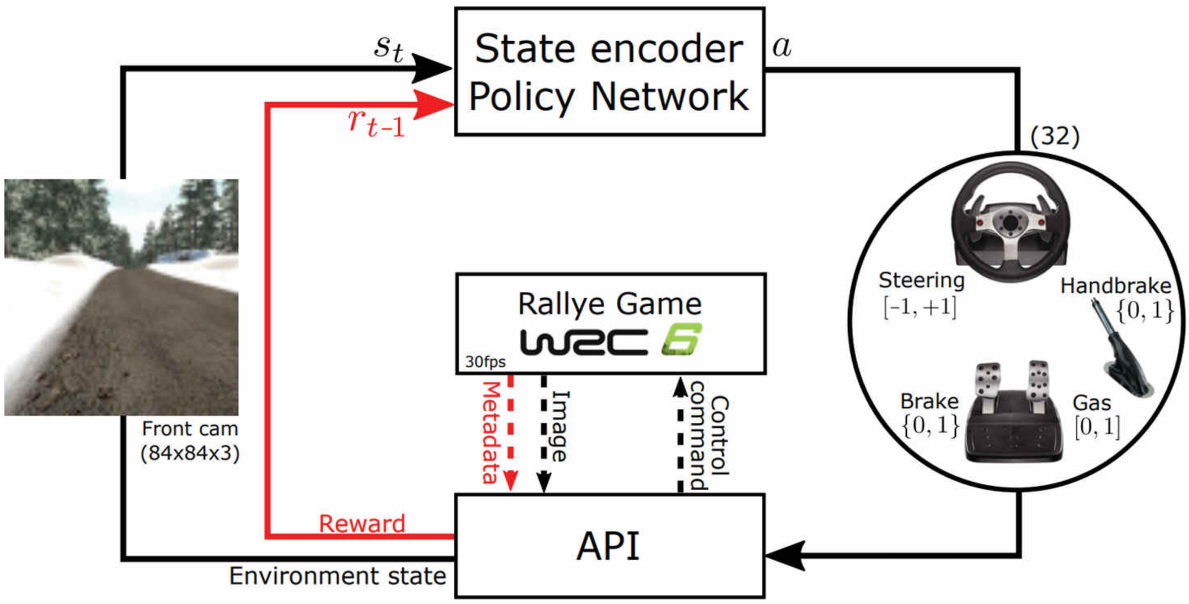
**Maximizes** $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}, \gamma \in [0,1[$

**Deep Reinforcement Learning (DRL) if Deep NeuralNet used as model (for policy and/or its "value"): DQN, Actor-Critic A3C**

**End-to-end driving: policy $\pi$ searched as ConvNet(front-image)**

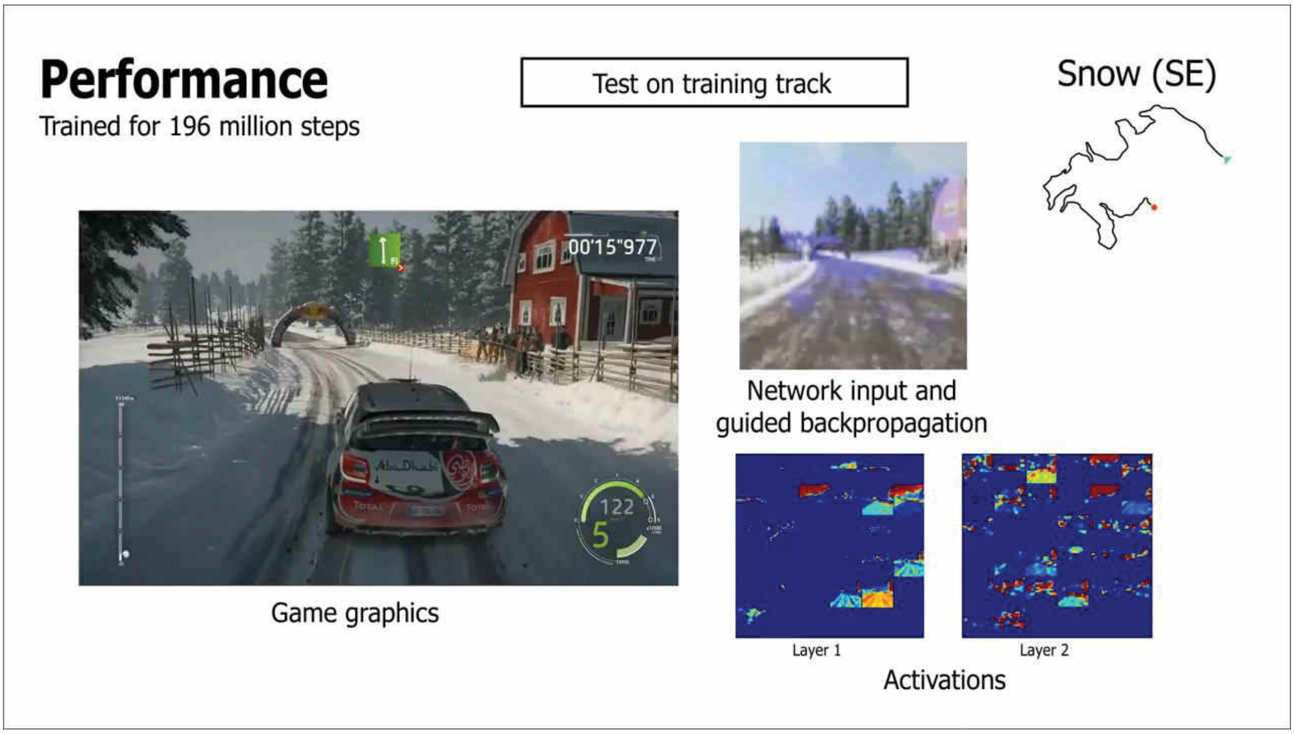


# End-to-end driving learning by RL in racing-car simulator




*Etienne Perot, Maximilian Jaritz, Marin Toromanoff, Raoul De Charette. End-to-End Driving in a Realistic Racing Game with Deep Reinforcement Learning, International conference on Computer Vision and Pattern Recognition - Workshop, Honolulu, United States, Jul. 2017.*

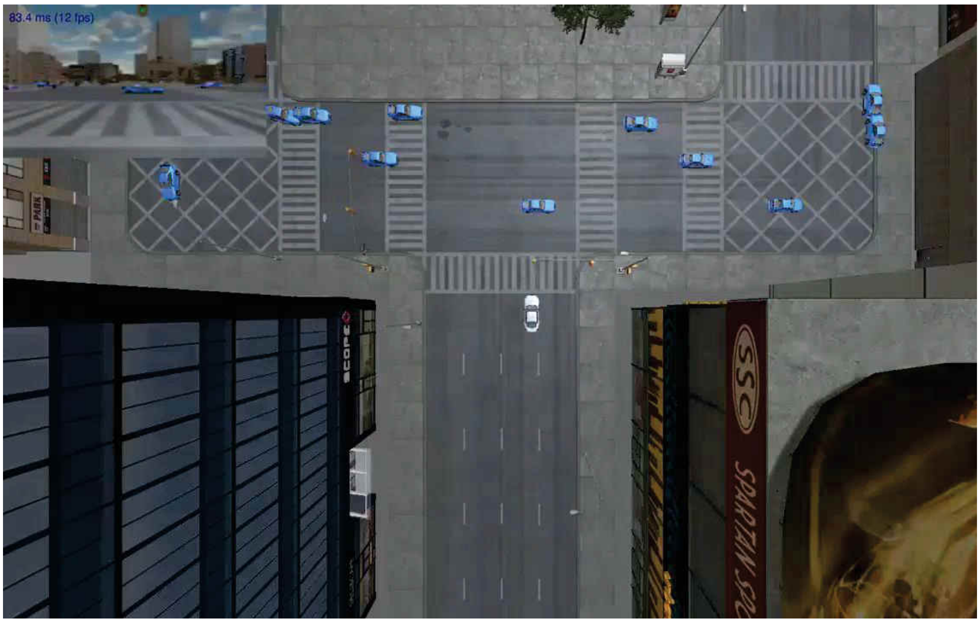
# End-to-end driving learnt by RL (in a racing-car simulator)

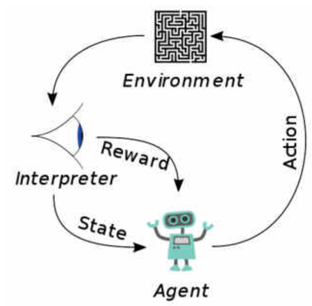



*[End-to-End Race Driving with Deep Reinforcement Learning](), Maximilian Jaritz, Raoul De Charette, Marin Toromanoff, Etienne Perot, Fawzi Nashashibi, ICRA 2018 - IEEE International Conference on Robotics and Automation, Brisbane, Australia, May 2018.*



# First RL experiment for end-to-end driving in urban environment



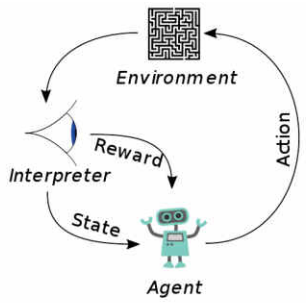



**End-to-end driving via Deep Reinforcement Learning**

***[thèse CIFRE Valeo/MINES-ParisTech en cours]***

***WORK IN PROGRESS…***

# Robot task learning using Reinforcement Learning




# Conclusions

- Most current AI challenges for Robotics and Autonomous Vehicles are related either to: Human-Robot Interaction, understanding of Human actions or behaviors, inference of Human intents, or learning of complex adaptive behaviors
- 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
- For Human movements or intents analysis, combining human-pose estimation by DL with Deep Temporal Convolution of time-series seems promising
- For behavior learning, Deep Reinforcement Learning from images already produces interesting results

# Questions ?