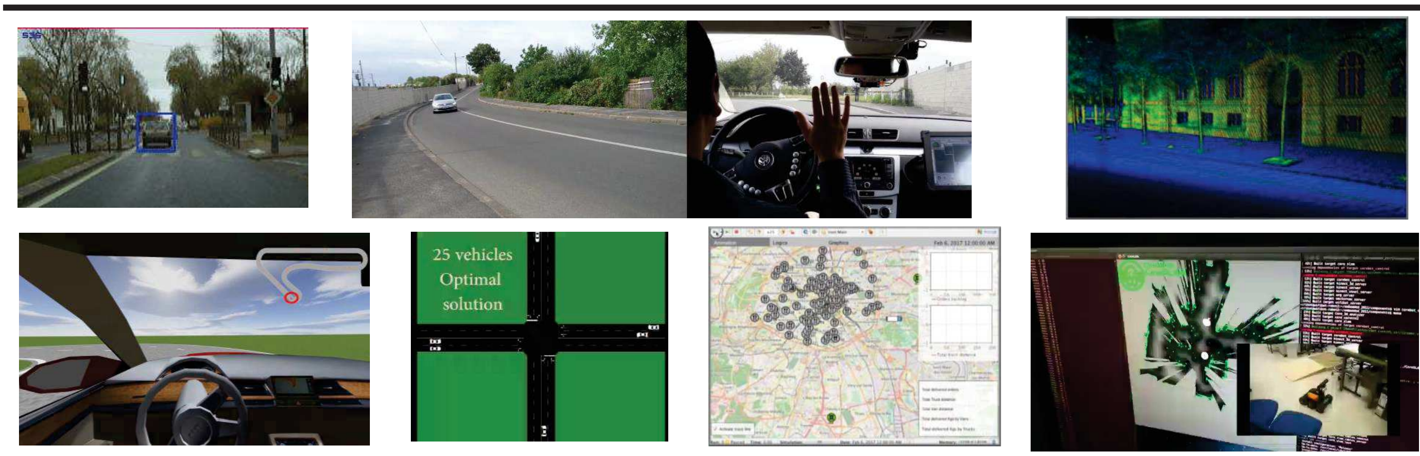
Deep-Learning for Intelligent Vehicles

Vehicle absolute *ego-localization from vision*, using only pre-existing geo-referenced panoramas

Pr. Fabien MOUTARDE
Center for Robotics
MINES ParisTech
PSL Université Paris

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

Some research results of the center for Robotics



Autonomous Vehicles, Intelligent Transport Systems, Mobile and/or Collaborative Robotics & Virtual Reality

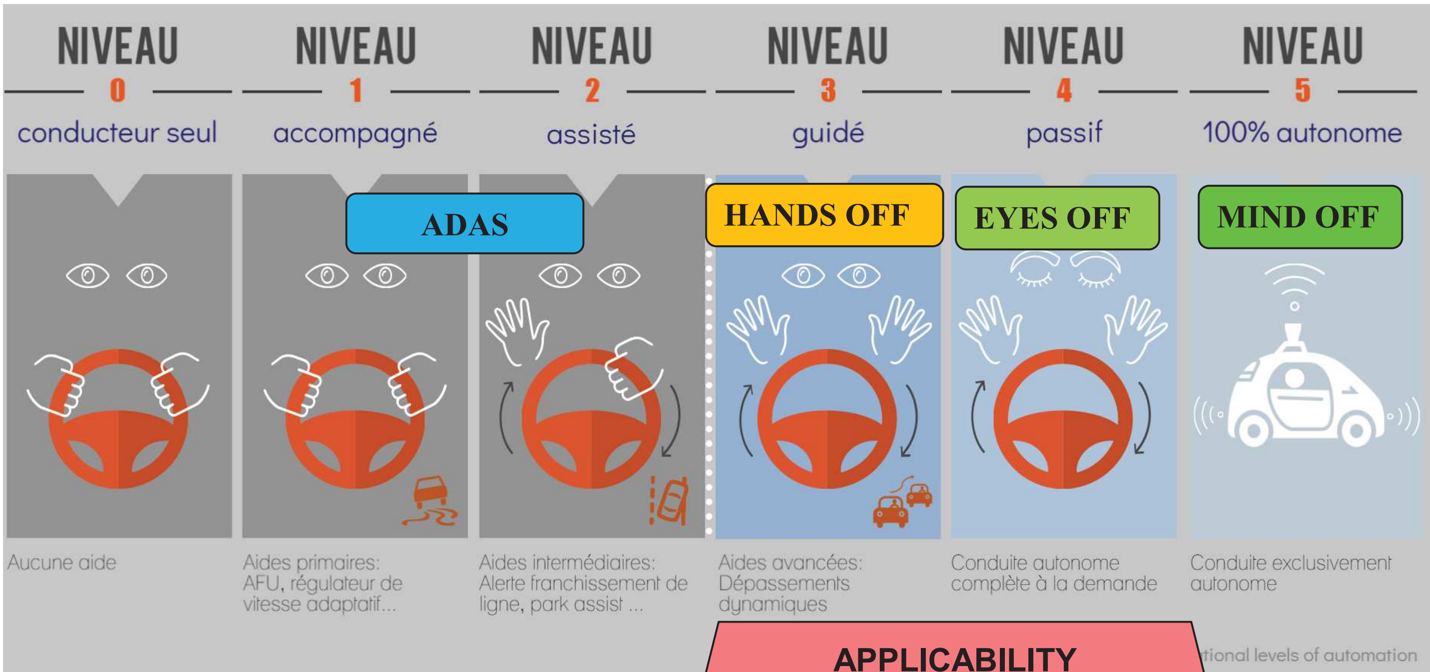
International collaborations with: Berkeley, EPFL, Shanghai JiaoTong,…
Industrial contracts with: Valeo, PSA, Safran, Thales, SoftbankRobotics, etc…

Outline

- **Introduction on AI for Intelligent Vehicles**
- Visual ego-localization from pre-existing geo-referenced panoramas: classic approach vs. *Deep-Learning* inference
- Wider outlook on Deep-Learning for Intelligent Vehicles

From Intelligent Vehicles to Autonomous Vehicles

The 5 « automation levels » for vehicles defined by SAE



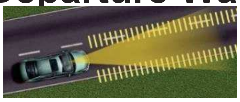
APPLICABILITY CAN BE CONDITIONAL (e.g. RESTRICTED TO ONLY MOTORWAYS, …)

What are ADAS?

Acronym of Advanced Driving Assistance Systems
= Intelligent functions for safer and/or easier driving

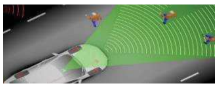
Warning or Information

- Lane Departure Warning (LDW)

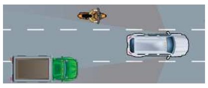


- Forward Collision Warning (FCW)

- Pedestrian Collision Warning



- Blind Spot Monitoring



- Speed Limit Assistant

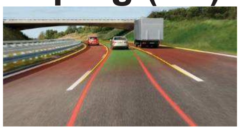
- etc…

Active ADAS

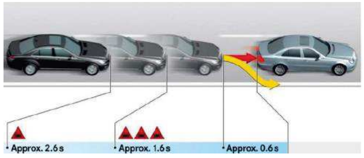
(*acting* on the vehicle)

- Adaptive Cruise Control (ACC)

- Lane Keeping (LK)



- Autonomous Emergency Braking



- Automated Parking
- etc…

Artificial Intelligence (AI) for Autonomous Vehicles

Autonomous Vehicles are mobile robots !

Perception

Reasoning & decision

ROBOTIC LOOP

Action

Real-time "understanding" of surroundings

Ego-Localization

Predicting movements/actions of others

Reactive behavior (obstacle avoidance)

Trajectory planning

Outline

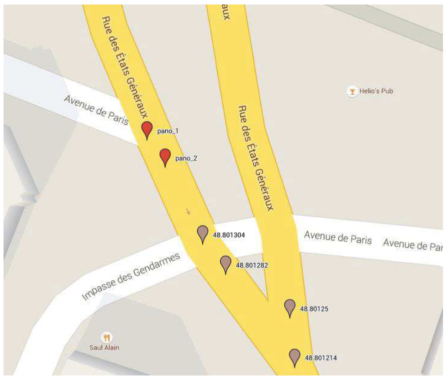
- Introduction on AI for Intelligent Vehicles
- *Visual* ego-localization *from pre-existing geo-referenced panoramas*:
  classic approach vs. *Deep-Learning* inference
  [work by former VeDeCom PhD student Li YU, co-supervised with G. Bresson and C. Joly]
- Wider outlook on Deep-Learning for Intelligent Vehicles

Visual ego-localization motivations

- GPS not always available (indoor, tunnels, underground parkings, *"urban canyons"*)
- GPS precision quite low (up to 10m error ! [except for differential GPS]
- GPS directly provides position but *NOT the orientation* (only the local orientation of TRAJECTORY can be estimated over time)
- Odometry is quite imprecise (cf. wheel slip!), and subject to large rapid cumulative errors
- Inertial Measurement Unit (IMU) expansive if precise, and subject to cumulative errors

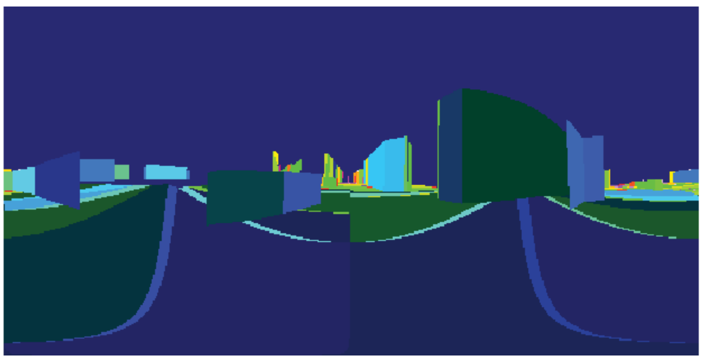
Outdoor visual ego-localization

Where am I?
(position+bearing)



Google StreetView data

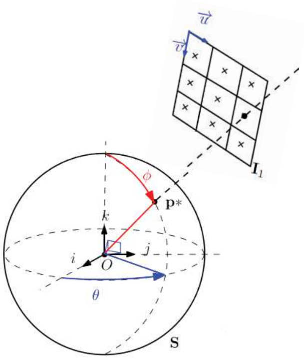
360° panoramas (RGB in UHD 13,312x6,656 pixels
+ *coarse 360° depthMap*
<u>~ every 10-50 m</u> in ~3000 city centers worldwide

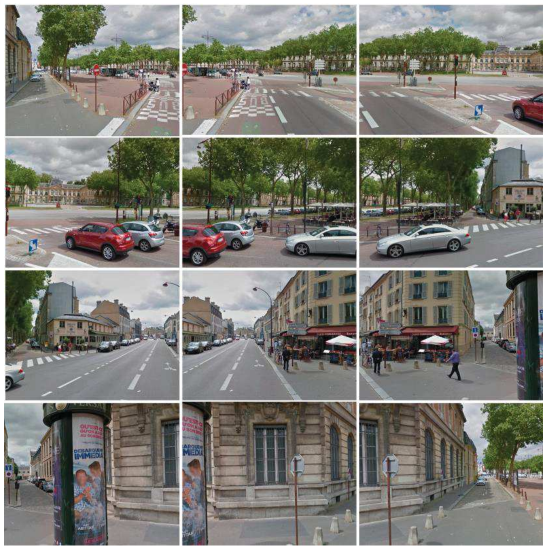


Using StreetView panorama for synthesizing rectified images

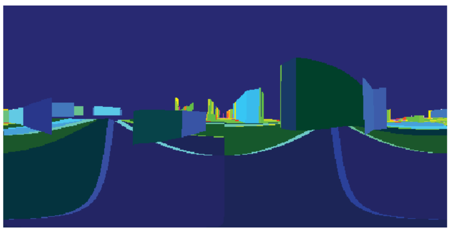
- **Distorsion of 360° images**
  + unknown query viewpoint

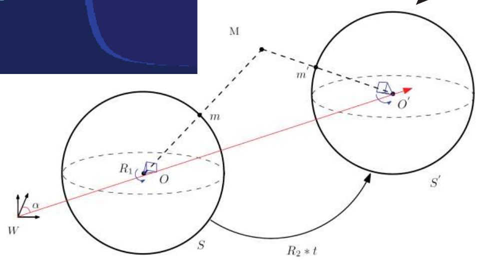
➔ Generate *synthetic* rectifiedviews (with <u>same focal length as on-board camera</u>) in several orientations

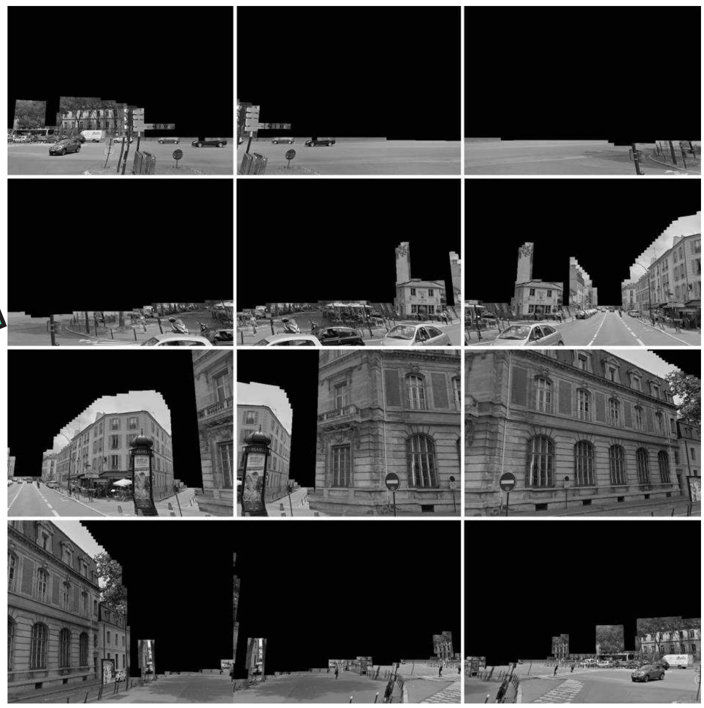




Generating virtual views BETWEEN StreetView panoramas







Possible thanks to availability of (coarse) panoramic depth map in StreetView

Visual place recognition using geo-localized images

With enough (~8-12) rectified synthetic images generated for several viewpoints, coarse visual place recognition by standard Bag of VisualWords (BoVW) is possible

t=0

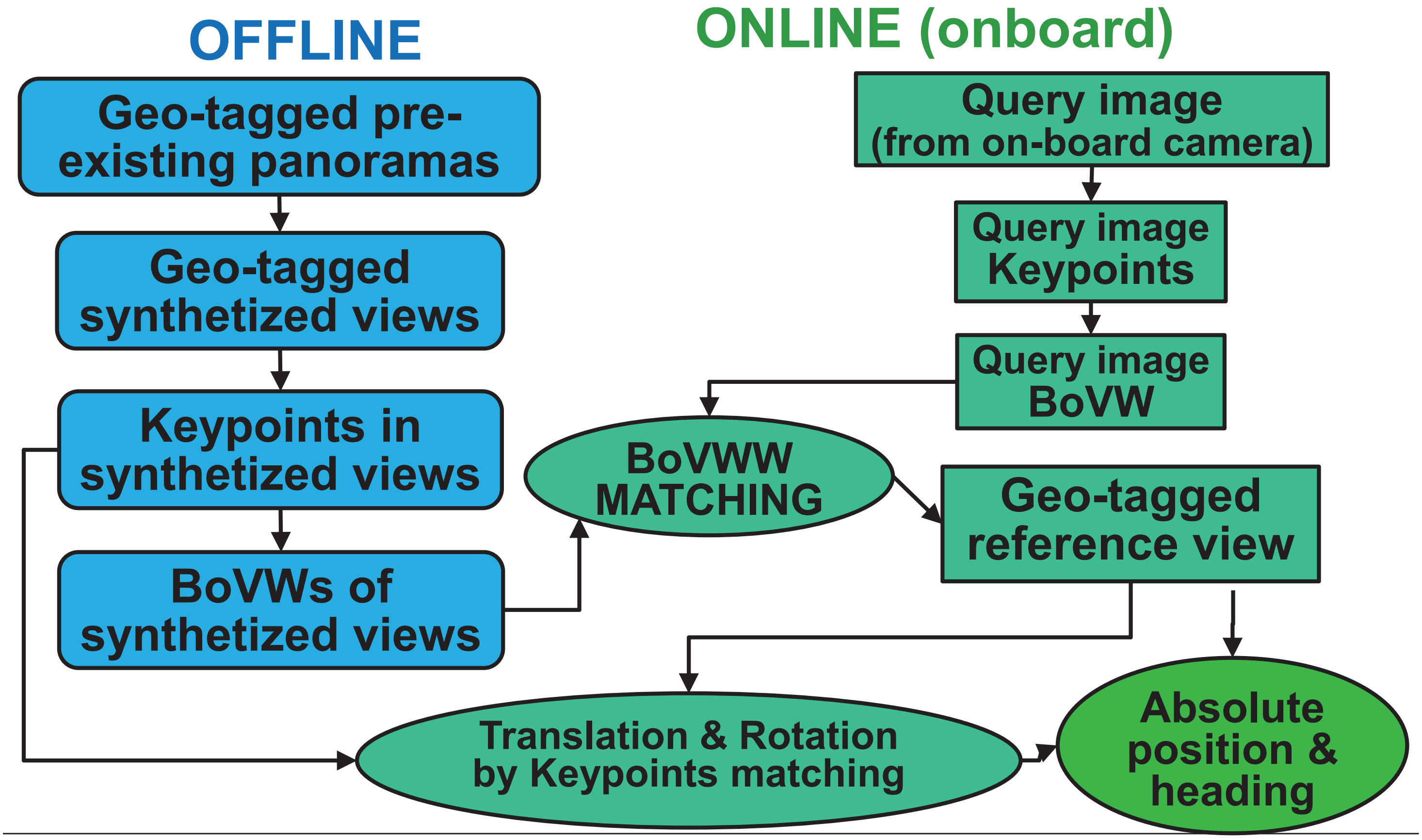
....

Construct 2 independent bags of words		
No.	1	2
Detectors	SIFT	MSER
Descriptors	484202	91026
Parameterization		
Size of bags	5000	2000
IF-ITF weighing		
Combination of two bags		
Search by cosine similarity		

SIFT - local point
MSER - local region

➔ Pre-compute 1 BoVW x ~10 views for each geo-tagged panorama

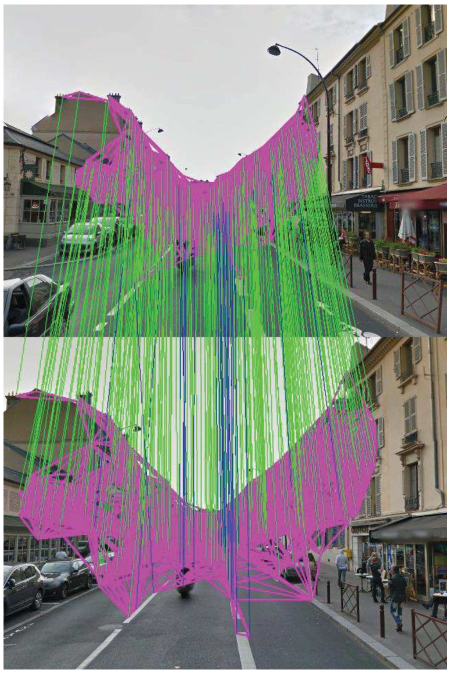
Visual metric localization from geo-localized images



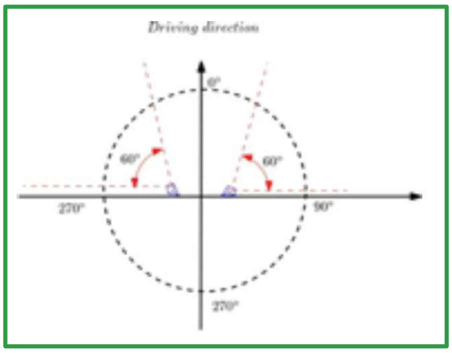
Visual metric localization from geo-tagged reference view

- Estimation of translation+rotation from reference view to query image by ***Bundle Adjustment of keypoint descriptors matches (with outliers filtered by RANSAC)***

- Use geo-tag of reference view + estimated translation&rotation to estimate current absolute position and heading



Experiment: set-up



Techniques:
- MIPSee Cameras 57.6° Fov / 20 fps
- 640*480 resolution
- Real Time Kinematic(RTK) GPS as ground truth (<20cm)

Results of experiment with "augmented" StreetViews

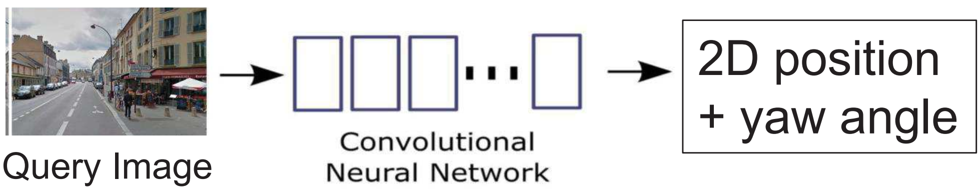
	Original Street View	Augmented Street View
Continuity	137/1046	281/1046
Average Error	3.82m	3.19m
Ratio in $[0m, 1m[$	21.89%	41.28%
Ratio in $[1m, 2m[$	28.47%	27.40%
Ratio in $[2m, 3m[$	44.53%	19.22%
Ratio in $[3m, 4m[$	5.11%	12.10%

- 1046 query images
- 498m trajectory
- 28 existing panoramas
- 53 virtual panoramas synthesized

with augmented Street View:
More query images are localized

68.7% of estimated positions with error <2m

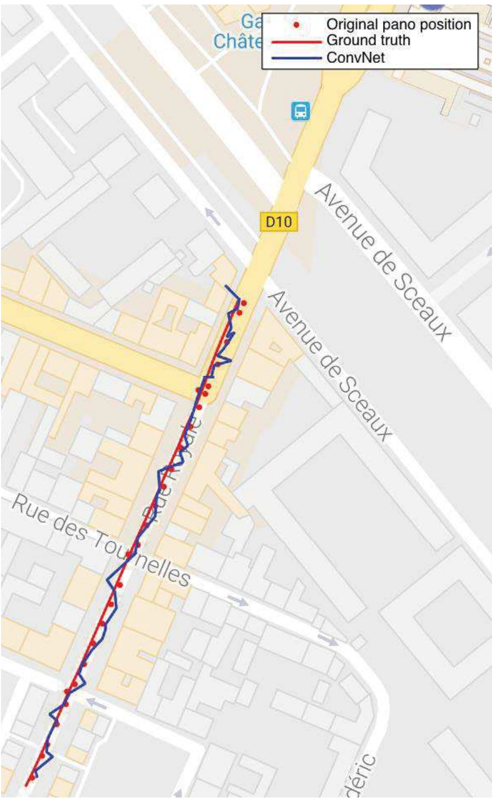
Deep-Learning of pose regression using only pre-existing images



Adapt PoseNet approach *[Kendall et al. 2015]*:

- **Learn an only 3-DoF pose (x,y,q) [instead of 6DoF]**
- **Start *transfer learning from ResNet50* model [instead of InceptionV3] 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**
- **Train using ONLY images generated from PRE-EXISTING geo-referenced Google-StreetView panoramas [instead of many images from prior 1st pass]**

Results of DL visual localization trained on pre-existing geo-referenced images



SeqID (length)	Nb of images	Nb of StView panoramas (nb of virtual ones)	Average localization errors	
			Image features + geometry	Pose regression CNN
1 (234 m)	897	29 (1160)	**2.85 m**	**7.62 m**
2 (271 m)	898	29 (1160)	**2.63 m**	**7.93 m**
3 (222 m)	895	29 (1160)	**Fail**	**Fail**
4 (216 m)	901	34 (1360)	**2.82 m**	**7.55 m**
F (265 m)	554	29 (1160)	**Fail**	**7.87 m**

DL ego-localization errors (~ 7m) larger than with BoVW+geometry

BUT

Error comparable to GPS, and much faster inference (~75ms) than using BoVW+geometry (~3s !)

Conclusions on DL for Visual ego-localization of Intelligent Vehicles

- **<u>Vision-based ego-localization</u> using BoW place recognition + keypoints matching, even if <u>using as only references pre-existing geo-tagged panoramas</u> from Google-StView, can provide in city centers <u>positioning acuracy of ~3m</u>, comparable to plain GPS**
- **<u>Deep-Learning pose regression</u> = very interesting alternative to standard visual localization methods: currently still <u>~ 2 times less precise (positioning error~7m), BUT but much more real-time at inference (*75ms/image*</u> vs. 3s/image)**

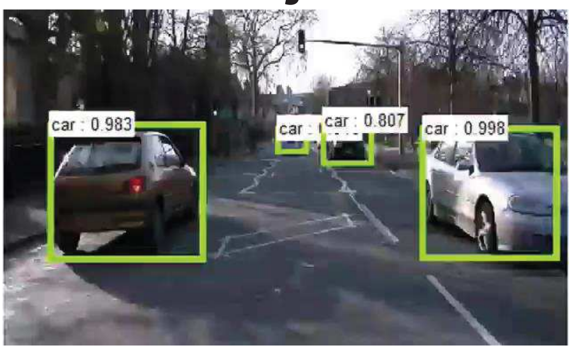
➔ Potential use as a fallback for mitigating GPS outages (urban canyons, tunnels, etc…)

Outline

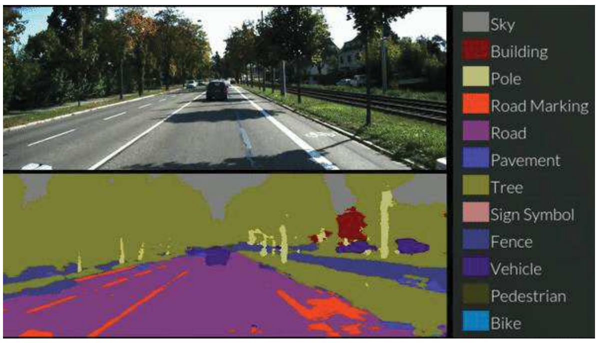
- Introduction on Intelligent Vehicles
- Visual ego-localization from pre-existing geo-referenced panoramas: classic approach vs. *Deep-Learning* inference
- **Wider outlook on Deep-Learning for Intelligent Vehicles**

Deep-Learning for *Perception* by Intelligent Vehicles

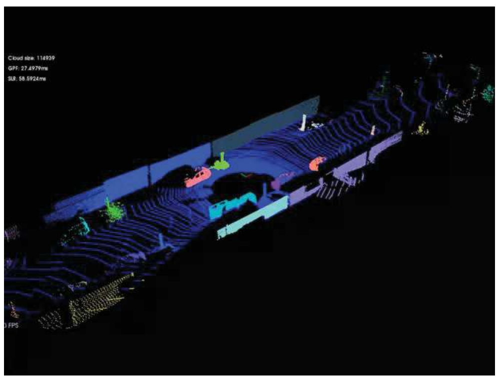
Visual objects detection, semantic segmentation



From camera



Real-time estimation of Human poses *[OpenPose, 2017]*



From LIDAR (3D points cloud)

Image-based Deep-Learning

- Image classification
- Visual <u>object detection and categorization</u>
- <u>Semantic segmentation</u> of images

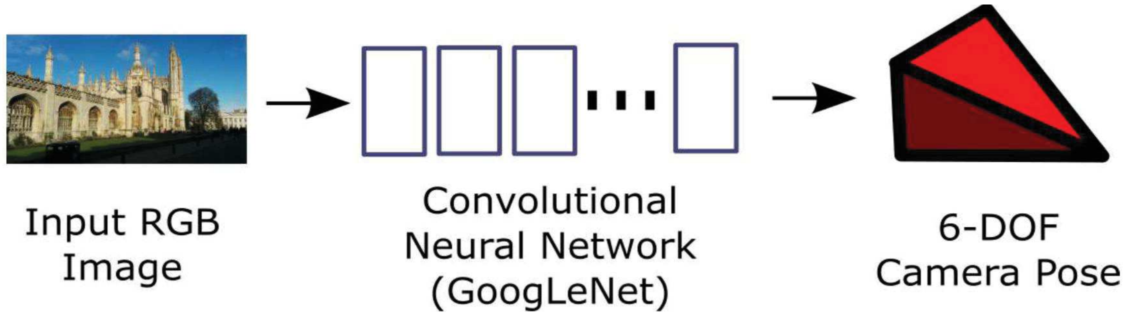
- Estimation of <u>Human pose</u>
- Inference of <u>3D (depth) from monocular vision</u>
- Image-based <u>ego-localization</u>
- <u>Realistic image synthesis</u>

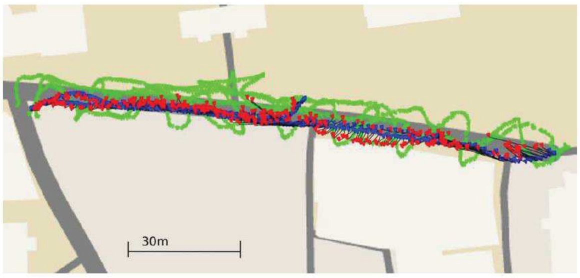
- Learning <u>image-based behaviors</u>
  - End-to-end driving from front camera
  - Learning behavior by <u>*Imitation of Humans*</u>, or with <u>*Reinforcement Learning*</u>

Inference of 3D <u>(depth)</u> from <u>*monocular*</u> vision

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

PoseNet: camera-pose regression with Deep-Learning



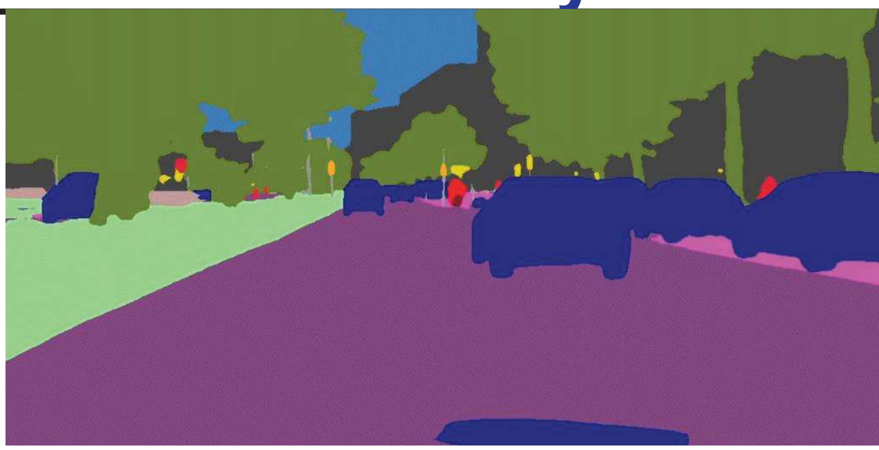


King's College

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

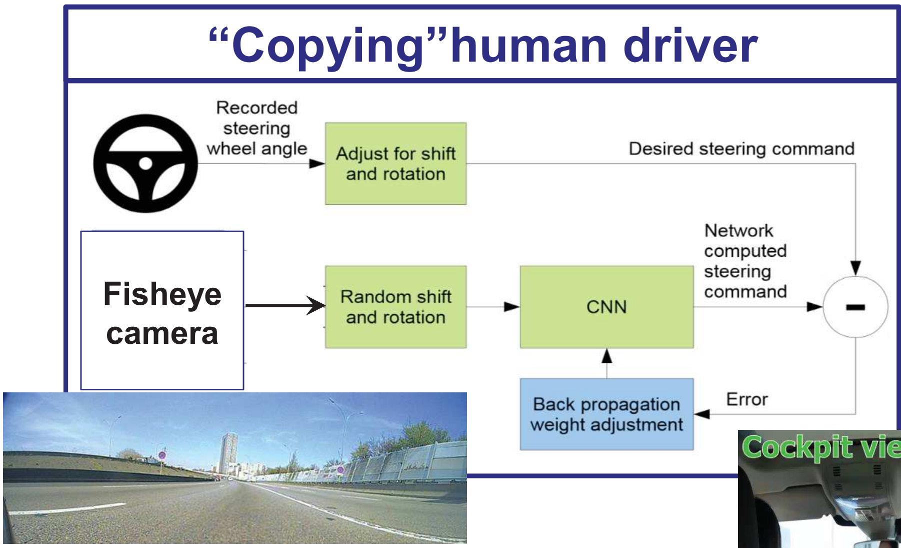
[A. Kendall, M. Grimes & R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization" , ICCV'2015, pp. 2938-2946]

DL for realistic Image synthesis



"Video-to-Video Synthesis", NeurIPS'2018 [Nvidia+MIT]
Using Generative Adversarial Network (GAN)

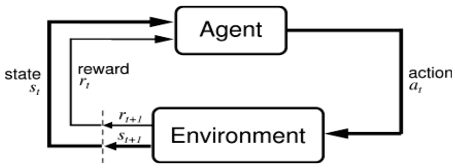
Imitation Deep Learning for Automated Driving by Visual servoing

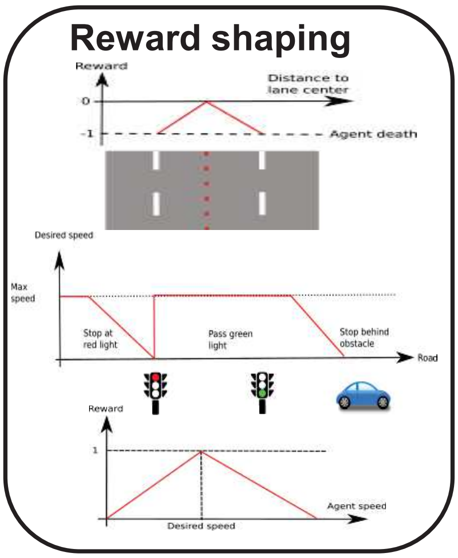


"End to End Vehicle Lateral Control Using a Single Fisheye Camera", Marin Toromanoff, Emilie Wirbel, Frédéric Wilhelm, Camilo Vejarano, Xavier Perrotton et Fabien Moutarde, 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018), Madrid, Spain, 1-5 oct. 2018.

Valeo demo @CES'2018

Deep *Reinforcement* Learning for vision-based Autonomous Driving





Work by my Valeo/MINES_ParisTech PhD student Marin Toromanoff

1st prize at « CARLA Autonomous Driving challenge » !!

Brief summary on DL for IV

- **Deep-Learning (DL) is now able to provide much more than just image analysis for visual objects detection**
- **For Intelligent Vehicles, DL is now even investigated *beyond perception*, for machine-learning of *reactive behavior or even trajectory planning***