

Visual scene real-time analysis for Intelligent Vehicles: **Visual ego-Localization with Deep-Learning using GIS images & on-board camera**

**Pr. Fabien Moutarde
Center for Robotics
MINES ParisTech
PSL Université Paris**

`Fabien.Moutarde@mines-paristech.fr`

`http://people.mines-paristech.fr/fabien.moutarde`

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 1

Acknowledgements

Content of several of these slides are borrowed from:

- Alex Kendall (University of Cambridge): slides on “Learning-based Visual Localization” from his CVPR’2017 tutorial

https://alexgkendall.com/media/presentations/lsvpr_2017_cvpr_tutorial_alex_kendall.pdf

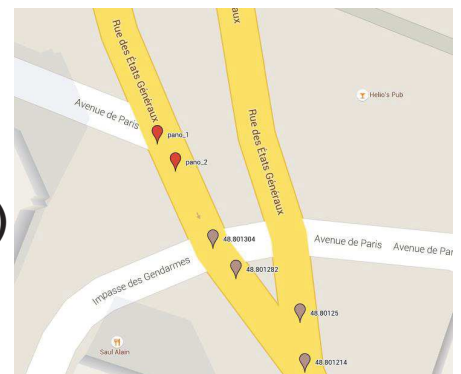
Several slides are also based on work and PhD thesis manuscript (<https://pastel.archives-ouvertes.fr/tel-01863297>) of my former PhD student Li YU.

- **GIS geo-tagged images**
- Visual localization from GIS images using BoVW+RANSAC
- Visual Localization with Deep-Learning
- Visual Localization from GIS images using Deep-Learning

Outdoor visual ego-localization



Where am I?
(position+bearing)



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)** expensive if precise, and subject to cumulative errors

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 5

Geographical Information System (GIS)

	GoogleMaps	HERE	Bing Maps	OpenStreetMap	BaiduMaps	TomTom	Mappy
Geo-data	+	+	+	+	+	+	+
Depth	+	+	-	-	+	+	-
2D Maps	+	+	-	-	+	+	-
HD Maps	-	+	-	-	-	+	-
3D Models	+	+	-	-	-	-	-
Live Maps	+	+	+	-	+	+	+
Street View	+	-	-	-	+	-	-
Public Access	+	+	+	+	-	-	-
Route Planer	+	+	+	+	+	+	+
Coverage	++	++	++	+	+	+	+
Accuracy	++	++	++	+	+	++	+

Several GIS now contain *millions of geo-tagged images*

Geo-tagged images

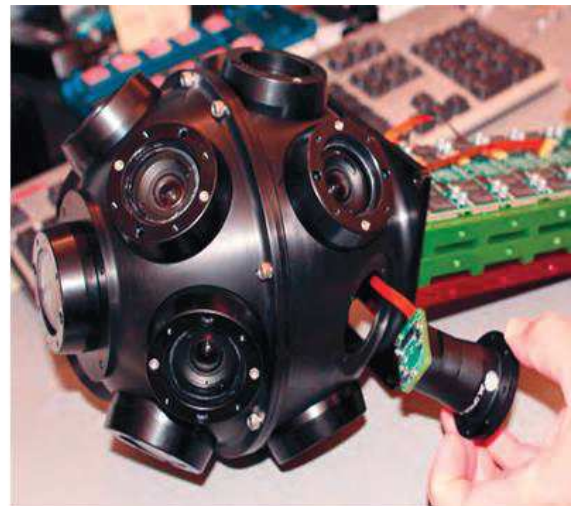


An extracted Street View of the Arc de Triomphe by setting parameters as 640×320 resolution, latitude= 48.8738, longitude= 2.2950, 0° heading, 0° pitch and 120° field of view.

Google StreetView sensors



(a)



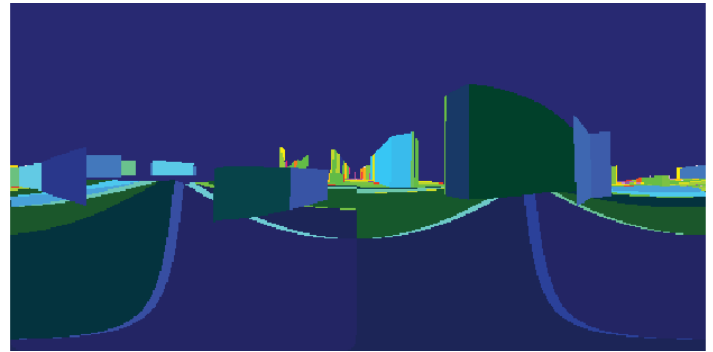
(b)

Collected from a car-mounted panoramic camera system + a LIDAR laser scanner.

R7 panoramic camera system = rosette of 15 identical and synchronized cameras with 5-megapixel CMOS image sensors and low-flare, controlled-distortion lenses.



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



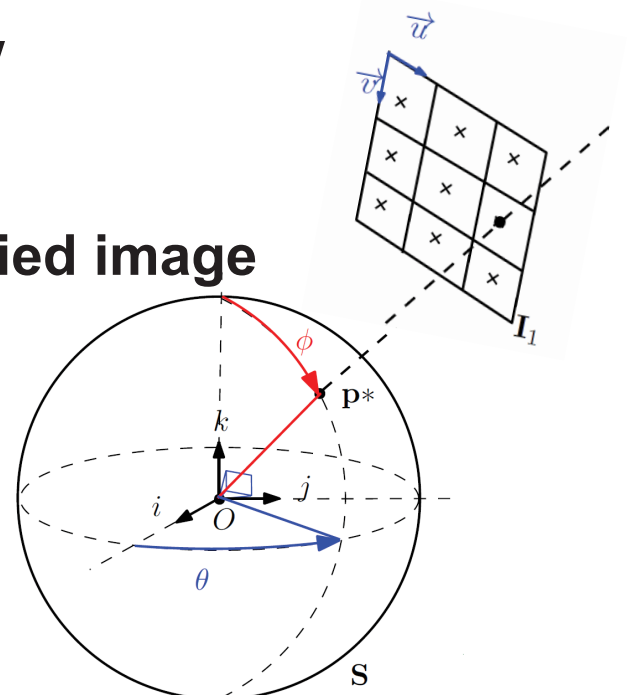
Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 9

Synthesis of rectified views from panoramic image

Specify:

- Orientation θ, Φ
- Focal length ~ Field of View
- Resolution

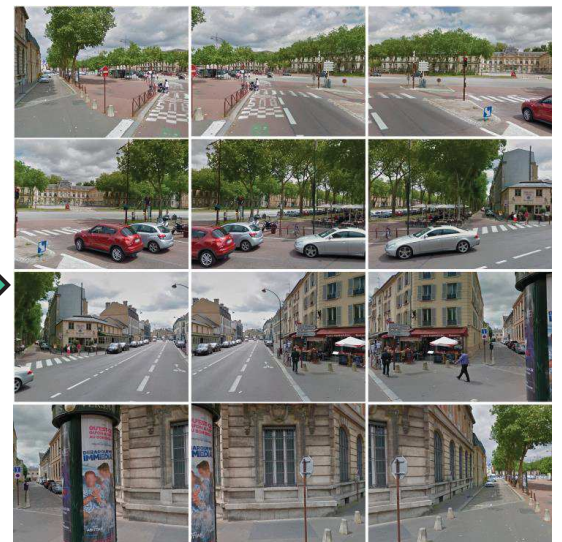
→ Compute a synthetic rectified image



- GIS geo-tagged images
- **Visual localization from GIS images using BoVW+RANSAC**
- Visual Localization with Deep-Learning
- Visual Localization from GIS images using Deep-Learning

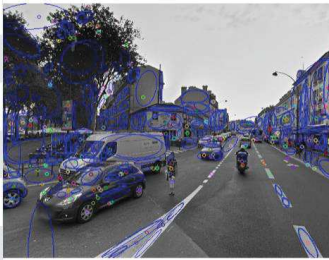
Using StreetView panorama for visual ego-localization

- Distorsion of 360° images
+ unknown query viewpoint
- ➔ **Generate *synthetic* views (with same focal length as on-board camera) in several orientations**



Visual place recognition with GIS images

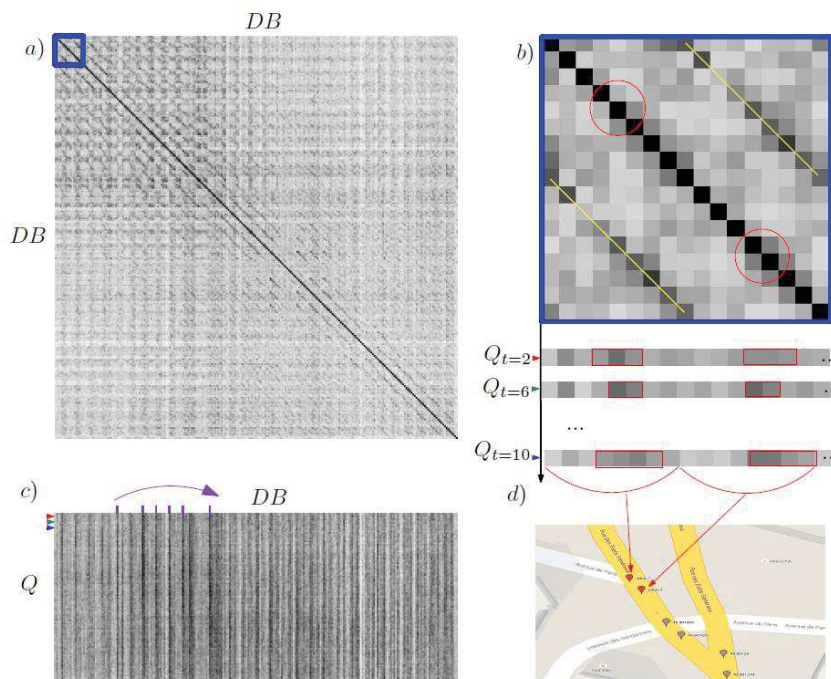
With enough (~8-12) rectified synthetic images generated with several viewpoints, coarse visual place recognition by standard Bag of VisualWords (BoVW) is possible

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

Co-similarity between GIS images

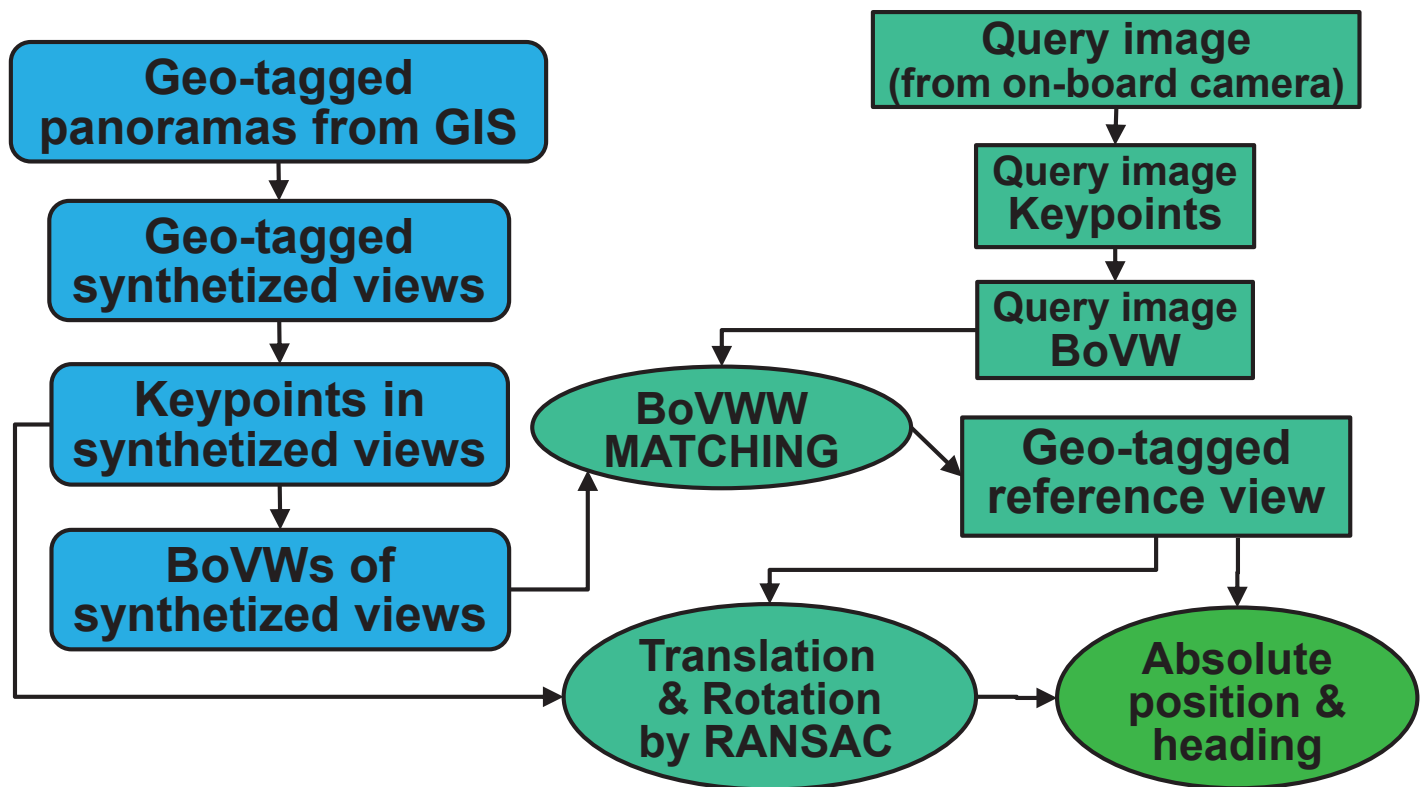


Pre-compute co-similarity matrix between all synthesized rectified views + filter by topologic proximity to help finding several pertinent best matches

Visual metric localization from GIS images

OFFLINE

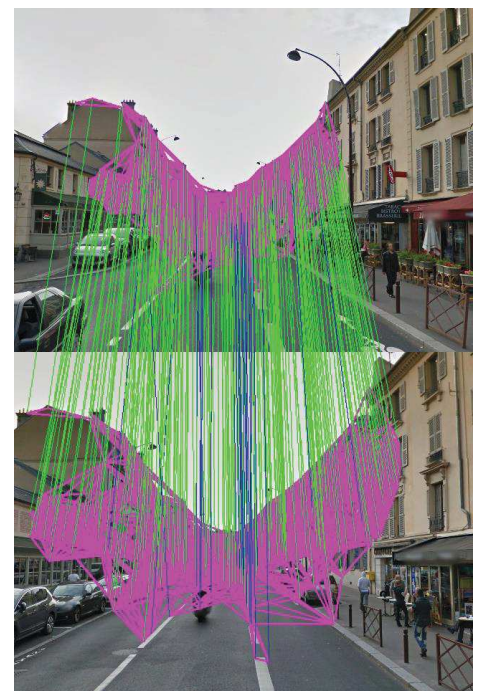
ONLINE (onboard)



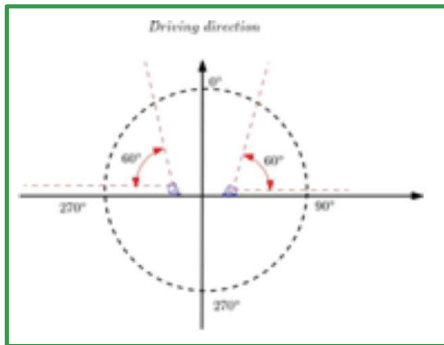
Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 15

Visual metric localization from geo-tagged reference view

- Estimation of translation+rotation from reference view to query image by multiple matches of keypoint descriptors (with outliers filtering by RANSAC)
- Use geo-tag of reference view + estimated translation&rotation to estimate current absolute position and heading



Experiment: set-up



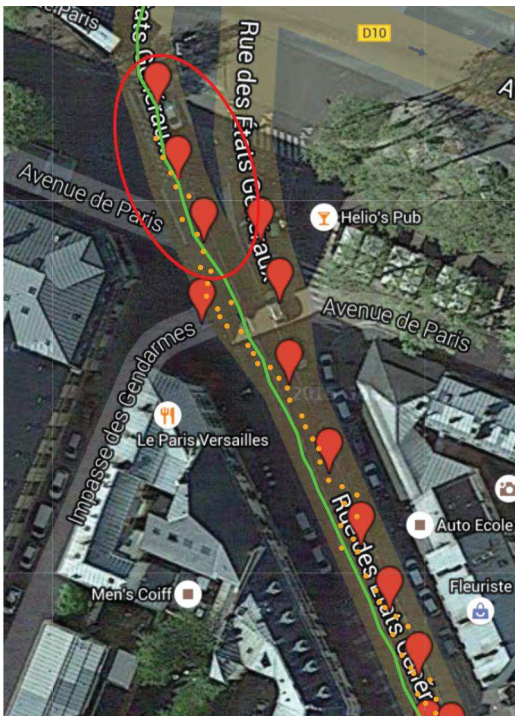
Techniques:

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

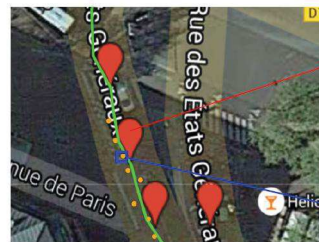
Work by my former PhD student Li YU

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 17

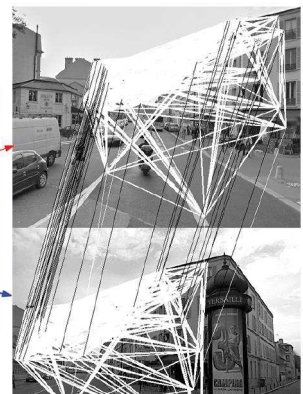
Experiment: results

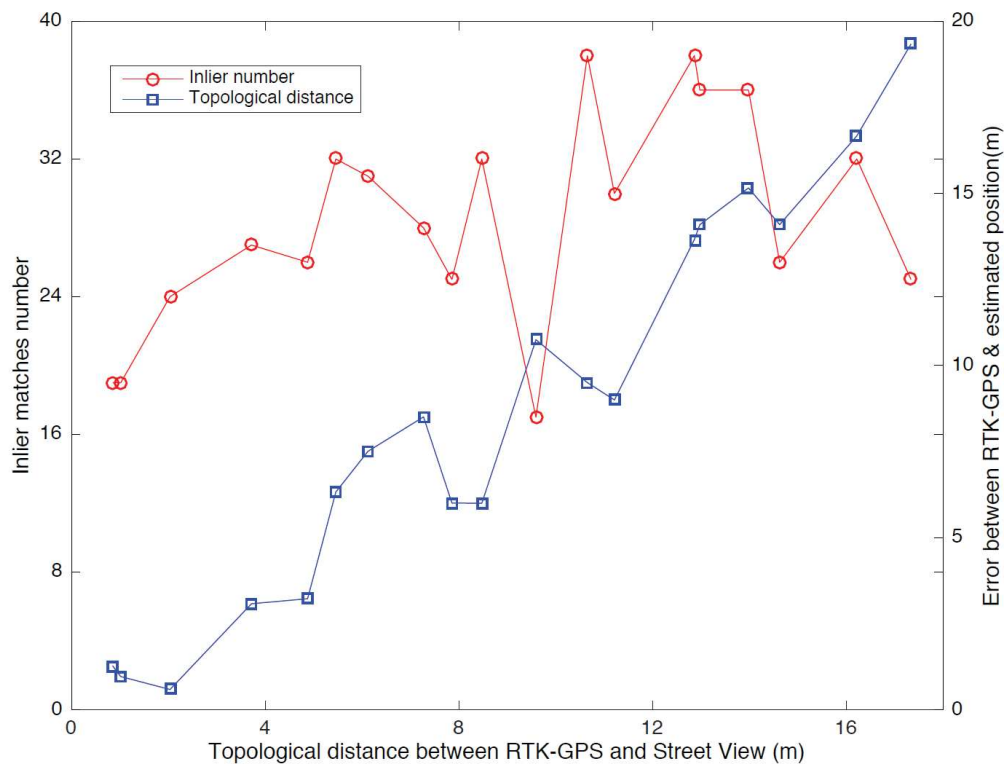


- 13 panoramas in a 287m street
- Ground truth in green
- 58/423 images localized
- Average error <6.5m, 58.6% <2m
- Standard GP <8m



Initial Matches = 36
Topological Distance = 2.03m

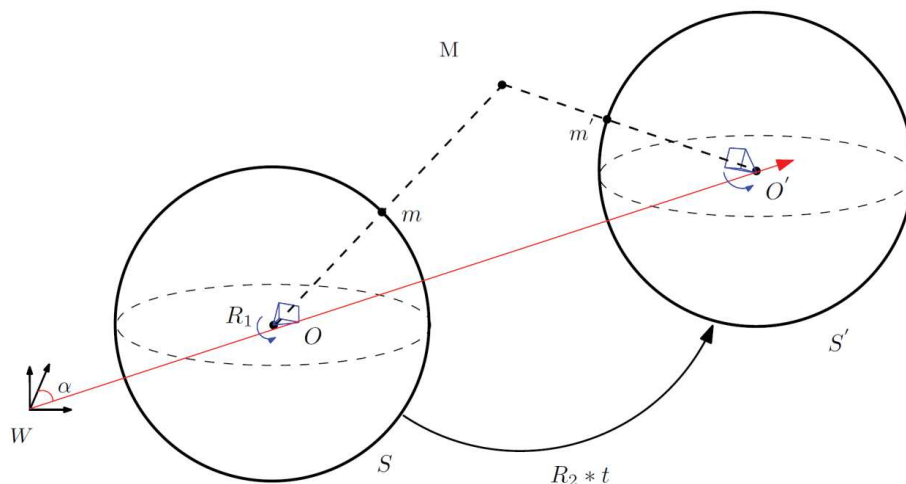




Generating virtual views BETWEEN StreetView panoramas

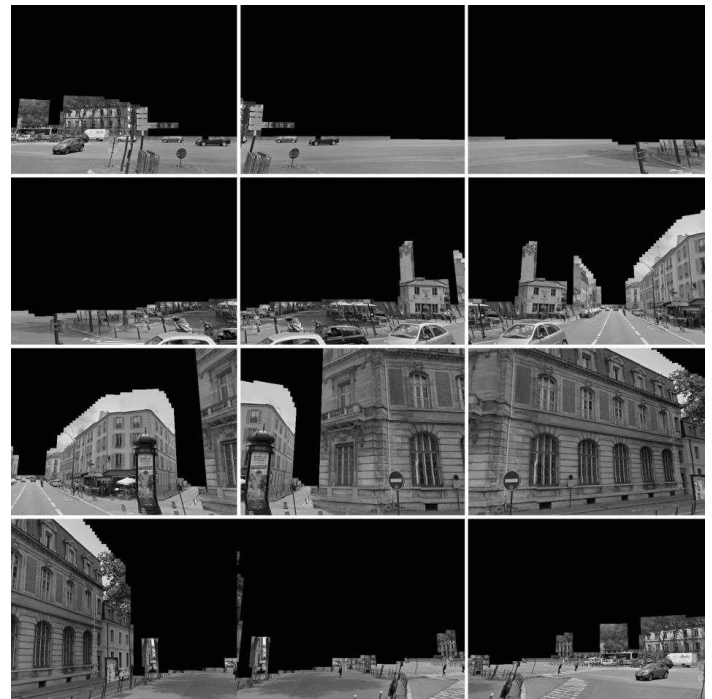
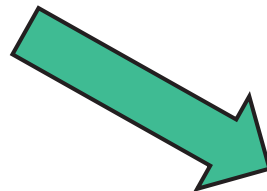
Too long distance between 2 panoramas !

→ Also generate *virtual views at positions between 2 successive panoramas*



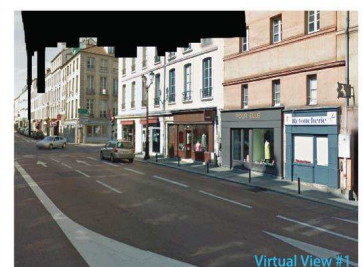
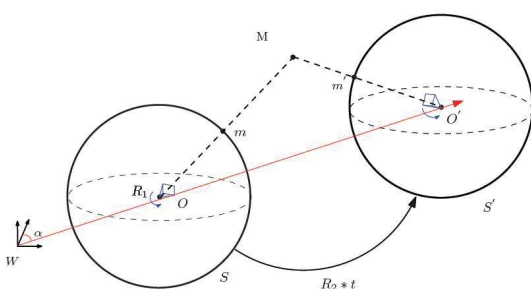
Possible thanks to availability of (coarse) panoramic depth map in StreetView

Typical virtual views BETWEEN StreetView panoramas



Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 21

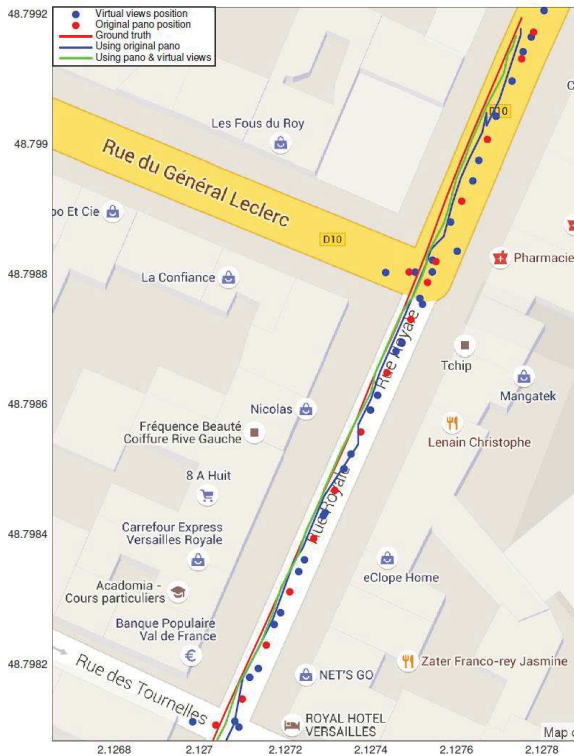
Choice of translation offset for virtual views



Translation distance	2m	4m	6m	8m
Invalid camera position	0	3	11	27
Uniform distribution	N	Y	Y	N
Ratio of virtual views with null pixels	0	0.125	0.5	1

4-meter forward/backward virtual panoramas are constructed from the original panorama.

Results of experiment with « augmented » StreetView



	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**

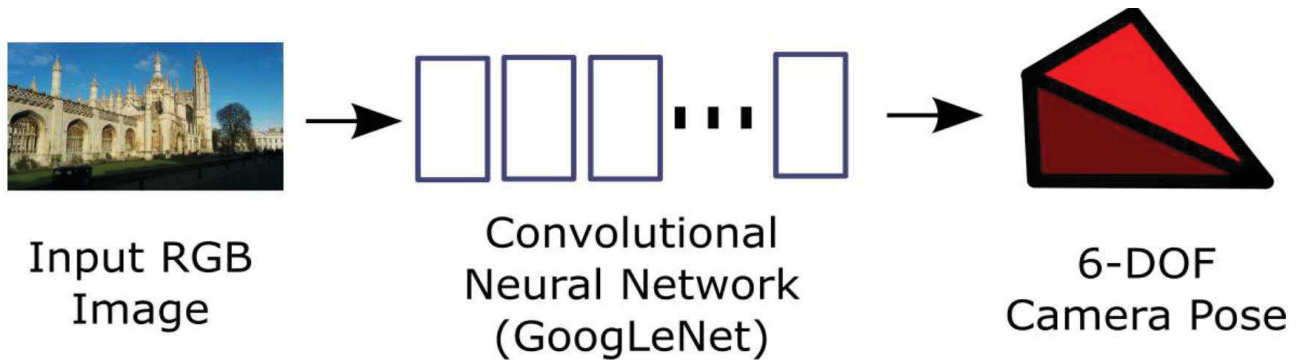
Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 23

Outline

- GIS geo-tagged images
- Visual localization from GIS images using BoVW+RANSAC
- **Visual localization with Deep-Learning**
- Visual Localization from GIS images using Deep-Learning

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 24

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

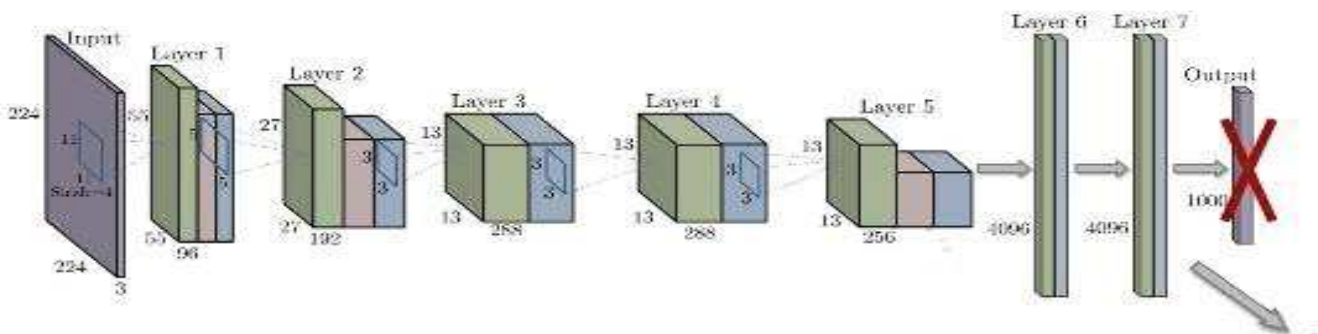


Trained with a naïve end-to-end loss function to regress camera position, \mathbf{x} , and orientation, \mathbf{q}

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

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

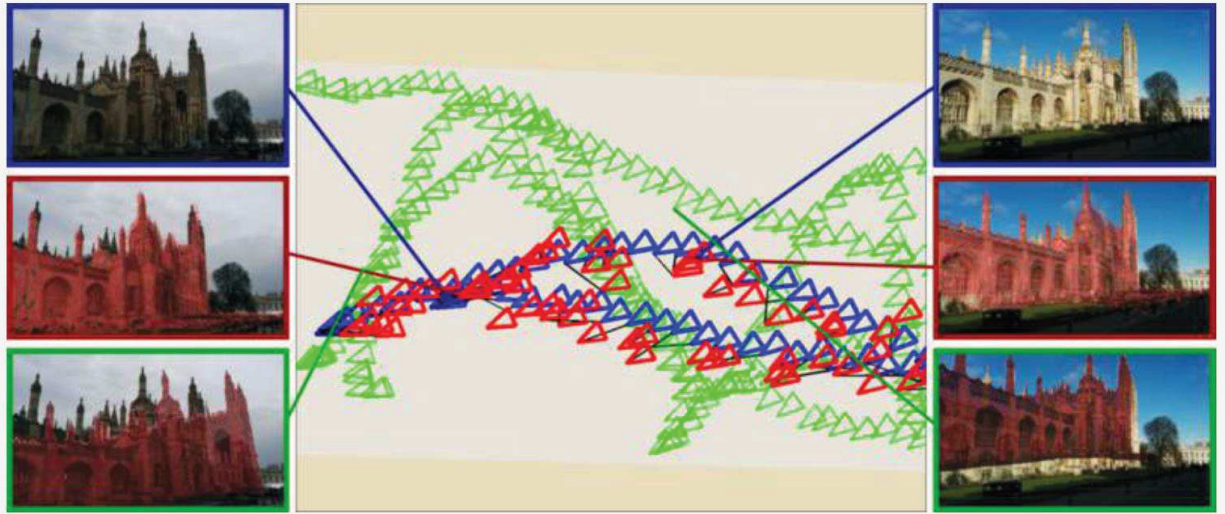
PoseNet applies Transfer learning for a task totally different from classification!



By removing last layer(s) (those for classification) of a convNet trained on ImageNet, one obtains a transformation of any input image into a semi-abstract representation, which can be used for learning SOMETHING ELSE (« transfer learning ») by creating new convNet output and perform learning of new output layers + fine-tuning of re-used layers

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.

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 27

PoseNet results on other tests



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.

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 28

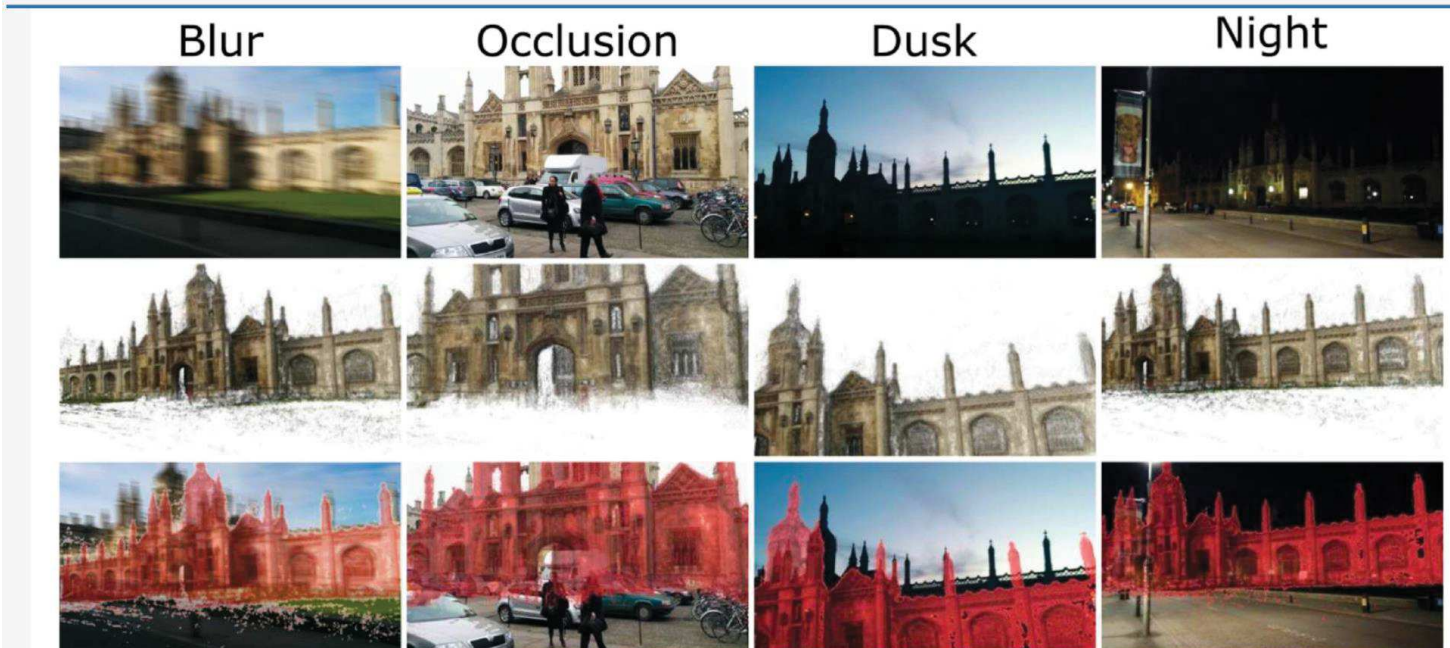
PoseNet results summary

Scene	# Frames		Spatial Extent (m)	SCoRe Forest (Uses RGB-D)	Dist. to Conv. Nearest Neighbour	PoseNet	Dense PoseNet
	Train	Test					
King's College	1220	343	140 x 40m	N/A	3.34m, 2.96°	1.92m, 2.70°	1.66m, 2.43°
Street	3015	2923	500 x 100m	N/A	1.95m, 4.51°	3.67m, 3.25°	2.96m, 3.00°
Old Hospital	895	182	50 x 40m	N/A	5.38m, 4.51°	2.31m, 2.69°	2.62m, 2.45°
Shop Façade	231	103	35 x 25m	N/A	2.10m, 5.20°	1.46m, 4.04°	1.41m, 3.59°
St Mary's Church	1487	530	80 x 60m	N/A	4.48m, 5.65°	2.65m, 4.24°	2.45m, 3.98°
Chess	4000	2000	3 x 2 x 1m	0.03m, 0.66°	0.41m, 5.60°	0.32m, 4.06°	0.32m, 3.30°
Fire	2000	2000	2.5 x 1 x 1m	0.05m, 1.50°	0.54m, 7.77°	0.47m, 7.33°	0.47m, 7.02°
Heads	1000	1000	2 x 0.5 x 1m	0.06m, 5.50°	0.28m, 7.00°	0.29m, 6.00°	0.30m, 6.09°
Office	6000	4000	2.5 x 2 x 1.5m	0.04m, 0.78°	0.49m, 6.02°	0.48m, 3.84°	0.48m, 3.62°
Pumpkin	4000	2000	2.5 x 2 x 1m	0.04m, 0.68°	0.58m, 6.08°	0.47m, 4.21°	0.49m, 4.06°
Red Kitchen	7000	5000	4 x 3 x 1.5m	0.04m, 0.76°	0.58m, 5.65°	0.59m, 4.32°	0.58m, 4.17°
Stairs	2000	1000	2.5 x 2 x 1.5m	0.32m, 1.32°	0.56m, 7.71°	0.47m, 6.93°	0.48m, 6.54°

Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 29

PoseNet robustness

Tolerance to environment, unknown intrinsics, weather, etc.

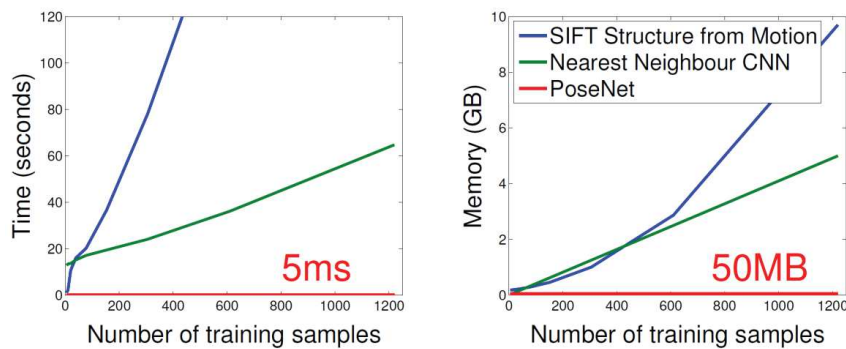


Alex Kendall, Matthew Grimes and Roberto Cipolla. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. ICCV, 2015.

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.



Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 31

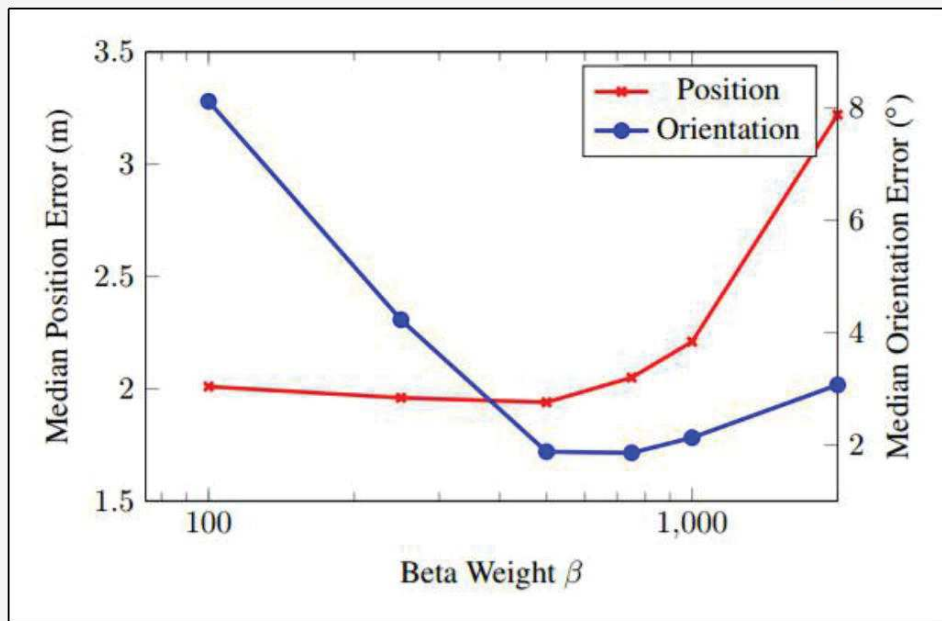
PoseNet vs traditional methods

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 x 1080 px
Inference Time	2 ms	78 ms

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

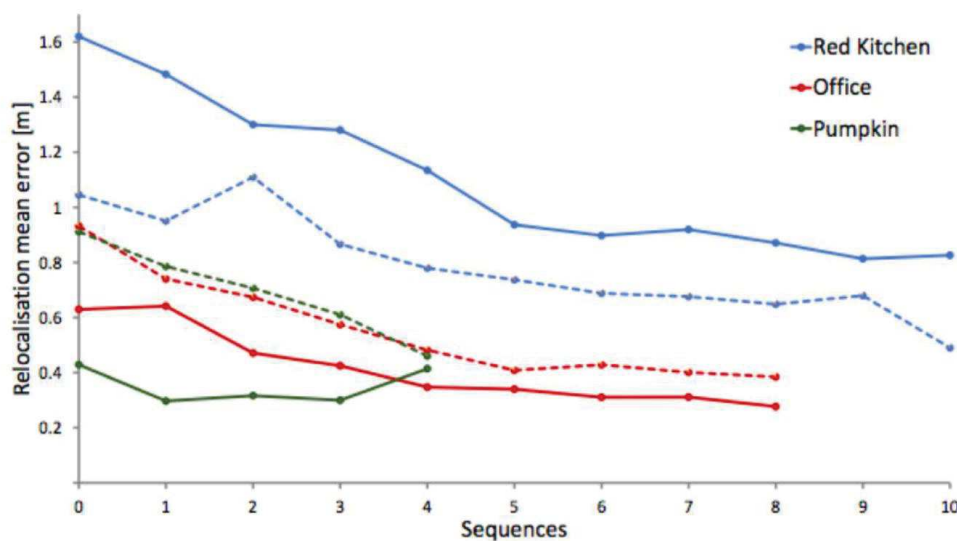
PoseNet: importance of relative weighting of position-orientation errors

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



Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 33

PoseNet performance improves with more data

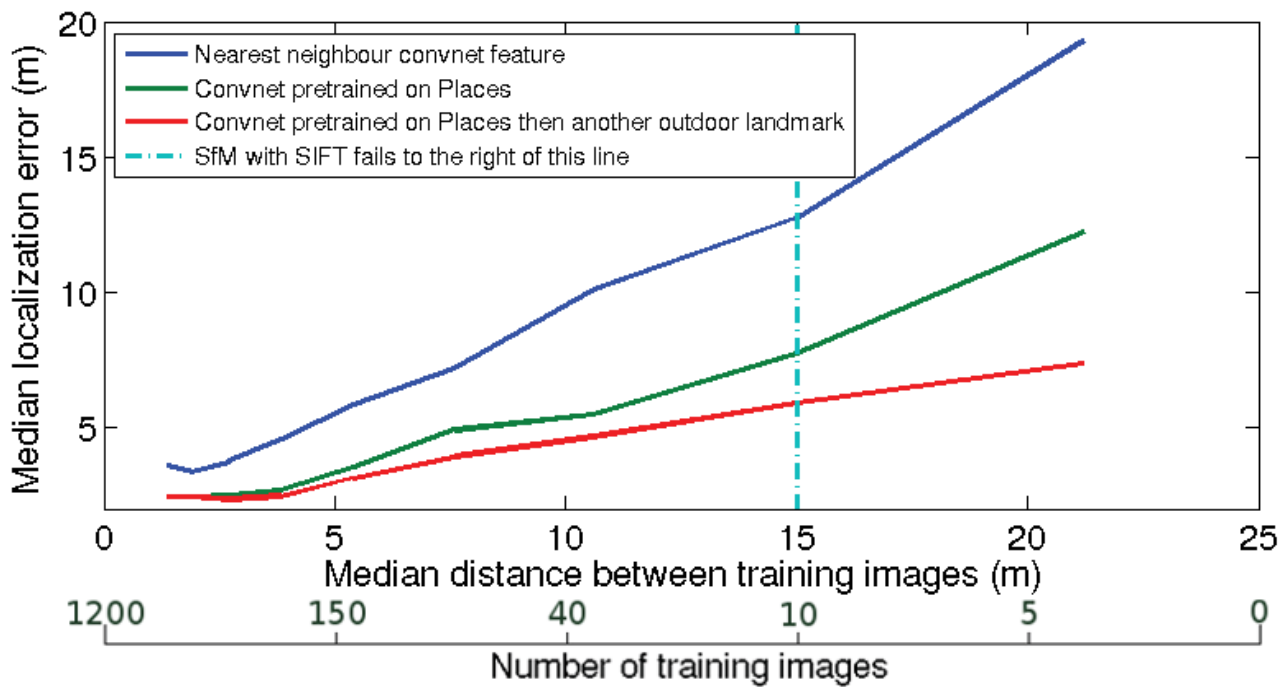


Scales very well:

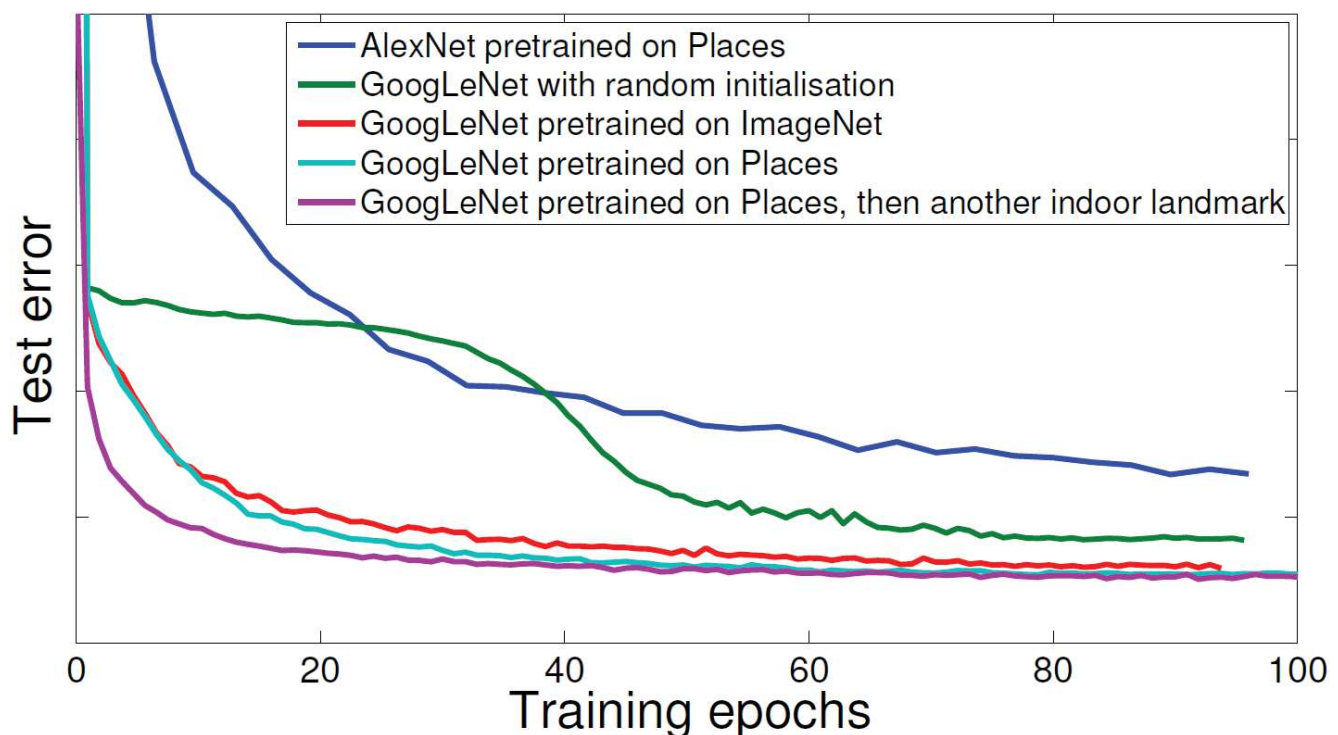
- Constant inference time (single forward pass of the network)
- Constant memory (~5 MB of neural network weights)

Contreras, Luis, and Walterio Mayol-Cuevas. Towards CNN Map Compression for camera relocalisation. arXiv:1703.00845, 2017.

PoseNet: graceful degradation with increased spacing of training images



PoseNet: importance of transfer learning



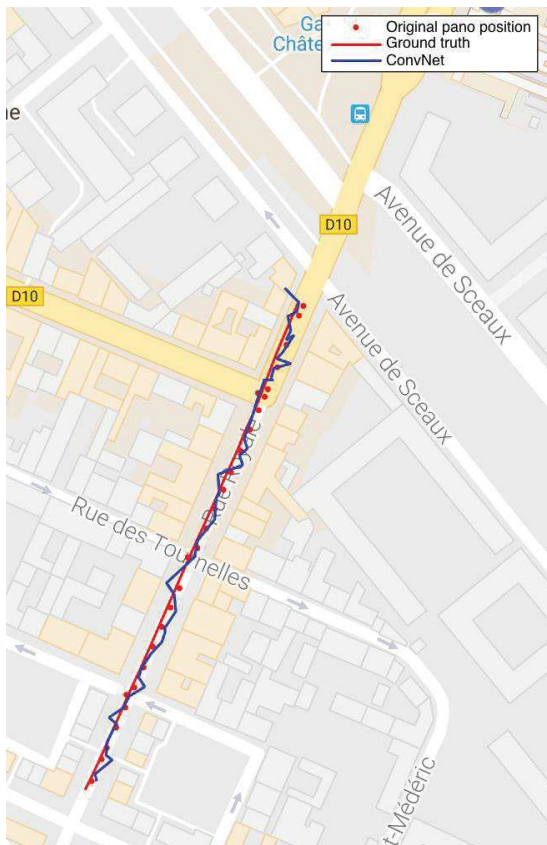
- GIS geo-tagged images
- Visual localization from GIS images using BoVW+RANSAC
- Visual localization with Deep-Learning
- **Visual Localization from GIS images using Deep-Learning**

Deep-Learning pose regression from GIS images

- Learn an only 3-DoF pose (x,y,θ)
- Start *transfer learning* from InceptionV3 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 my former PhD student Li YU

First results of Deep-Learning visual localization trained on GIS 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

**Localization errors (~ 7m and 23°)
larger than with BoVW+geometry**

BUT

**Error comparable to GPS, and much
faster to compute than using
BoVW+geometry**

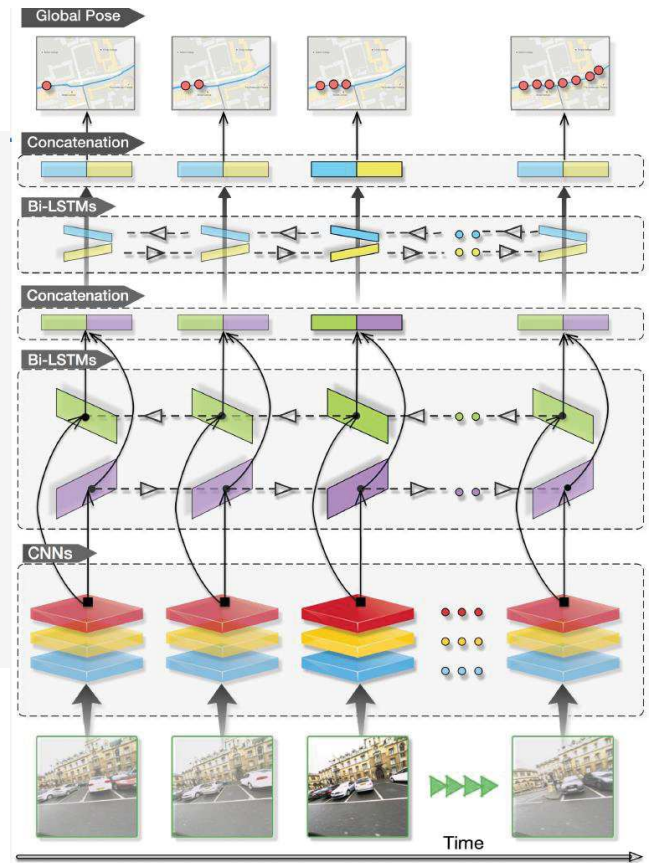
Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 39

Improvement perspectives for DL continuous pose regression

- Pre-train on much more data (from other places)?
- Use temporal continuity
(« video localization »)

Video localization with PoseNet+RNN

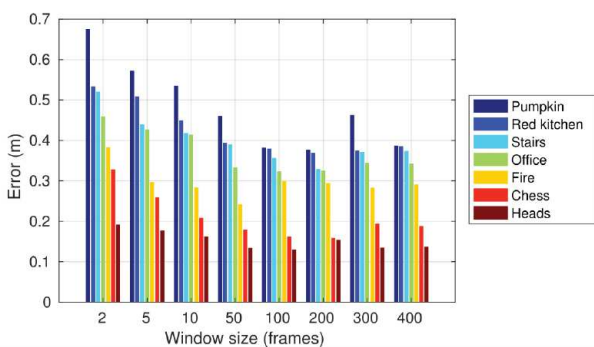
- PoseNet + Temporal Recurrent Neural Network
 - Learns dynamics of platform - temporal features
 - Bidirectional - analogous to “smoothing”
- Mixture of Gaussian output



Visual ego-Localization with Deep-Learning using GIS images, Pr. F. MOUTARDE, Center for Robotics, MINES ParisTech, Oct.2019 41

PoseNet+RNN results for video localization

- Outperforms smoothing baseline
- Diminishing returns using very long sequences



Clark et al., *VidLoc: A Deep Spatio-Temporal Model for 6-DoF Video-Clip Relocalization*. IEEE CVPR 2017.

Conclusions

- **Geo-tagged images from Geographical Information Systems (GIS) such as GoogleMaps+StreetView and BaiduMaps can be successfully leveraged for city-wide metric visual ego-localization of vehicles**
- **Machine-Learning approaches (in particular Deep-Learning pose regression) is a very interesting alternative to standard visual localization methods: currently still ~ 2 times less precise, but much less computer-intensive for online part**
- **The latter is therefore one of current « hot » research topics, and precision improvements are on the way**