Machine Learning Dublin Meetup,
25 September 2017

Object Detection on Street View Images: from Panoramas to Geotags

Vladimir A. Krylov

in collaboration with Eamonn Kenny (TCD), Rozenn Dahyot (TCD)

Fondúreacht Eolaíochta Éireann
Dá bhfuil romhainn
Science Foundation Ireland
For what's next

Ireland's European Structural and Investment Funds Programmes 2014-2020
Co-funded by the Irish Government and the European Union

European Union
European Regional Development Fund

The ADAPT Centre is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

Object detection. Intro.

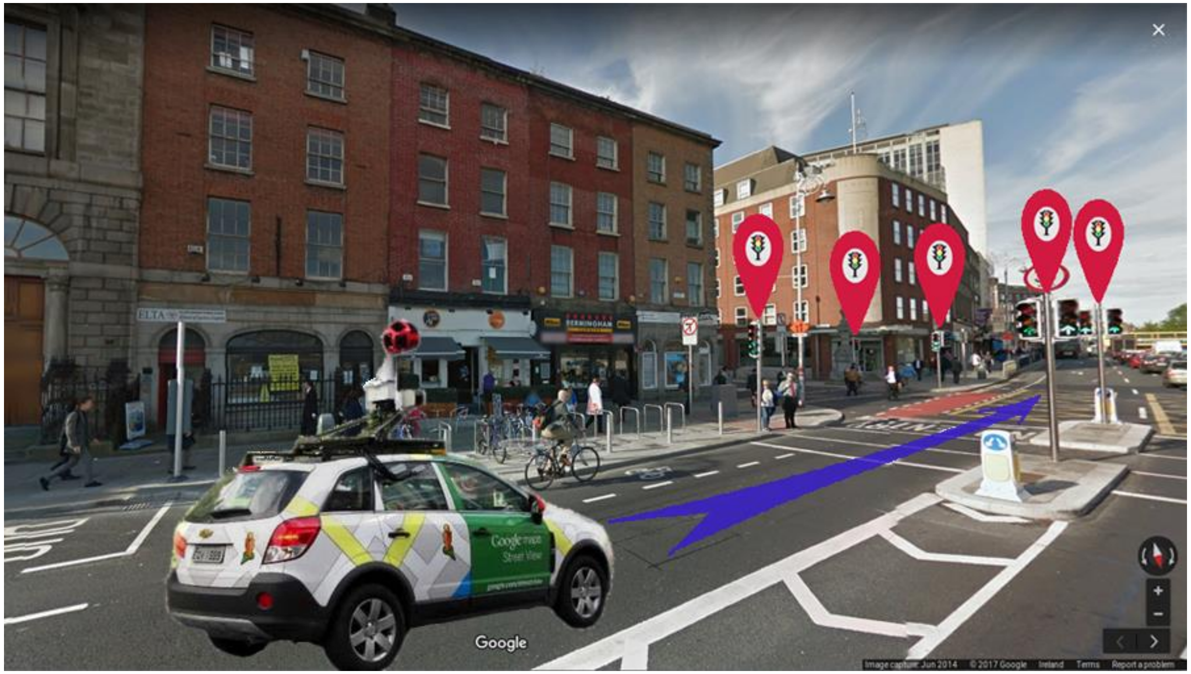
- *Motivation*. Billions of images (by ***Google***, *Bing*, *Mapillary*) covering mlns of kms of road.

~1 mln km coverage

>500 km

Object detection. Intro.

- *Motivation*. Billions of images (by *Google*, *Bing*, *Mapillary*) covering mlns of kms of road.
- *Target*. Automatic mapping of stationary recurring objects **from Street View**.



Object detection. Intro.

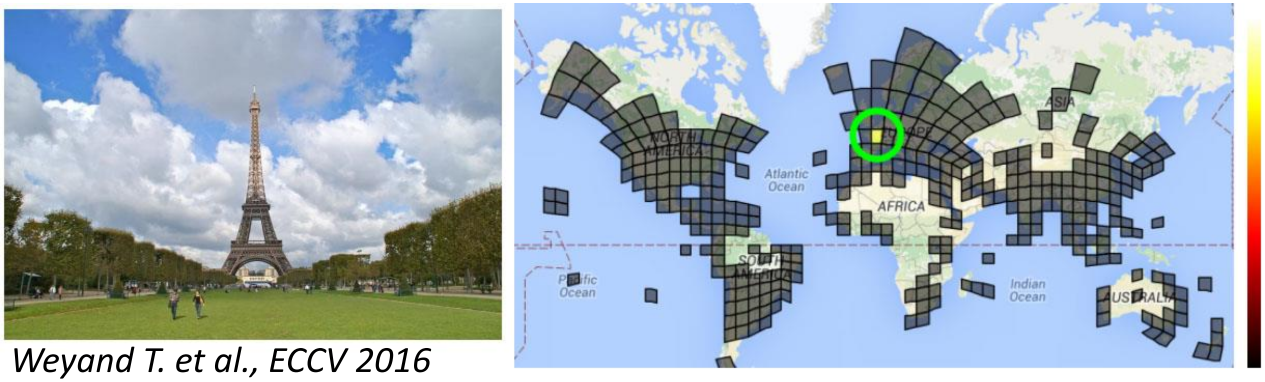
- *Motivation*. Billions of images (by *Google*, *Bing*, *Mapillary*) covering mlns of kms of road.
- *Target*. Automatic mapping of stationary recurring objects from Street View.
- *State-of-the-art*: **Object recognition.**

Mapillary Vistas Dataset

Object detection. Intro.

- *Motivation*. Billions of images (by *Google*, *Bing*, *Mapillary*) covering mlns of kms of road.
- *Target*. Automatic mapping of stationary recurring objects from Street View.
- *State-of-the-art*: Object recognition. **Image geolocation**.

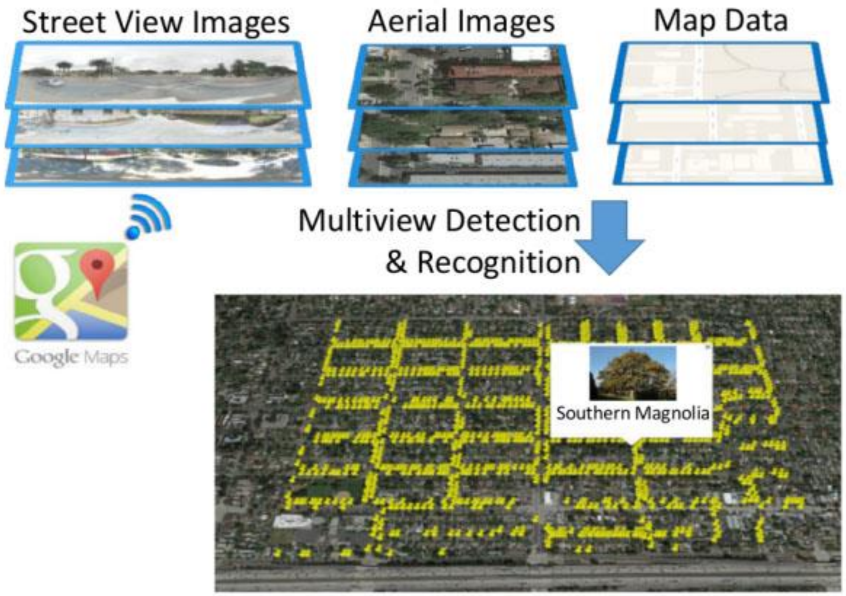
Lin T. et al., CVPR 2015



Weyand T. et al., ECCV 2016

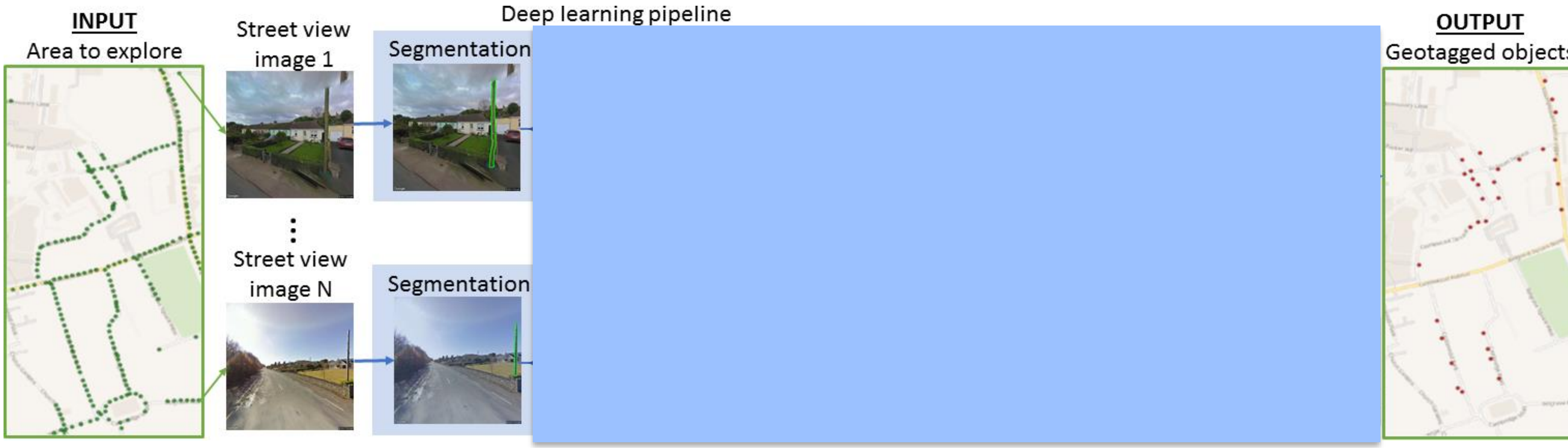
Object detection. Intro.

- *Motivation*. Billions of images (by *Google*, *Bing*, *Mapillary*) covering mlns of kms of road.
- *Target*. Automatic mapping of stationary recurring objects from Street View.
- *State-of-the-art*: Object recognition. Image geolocation. **Object geolocation**.



Fine-Grained Geographic Tree Catalog
Wegner, J. et al., CVPR 2016

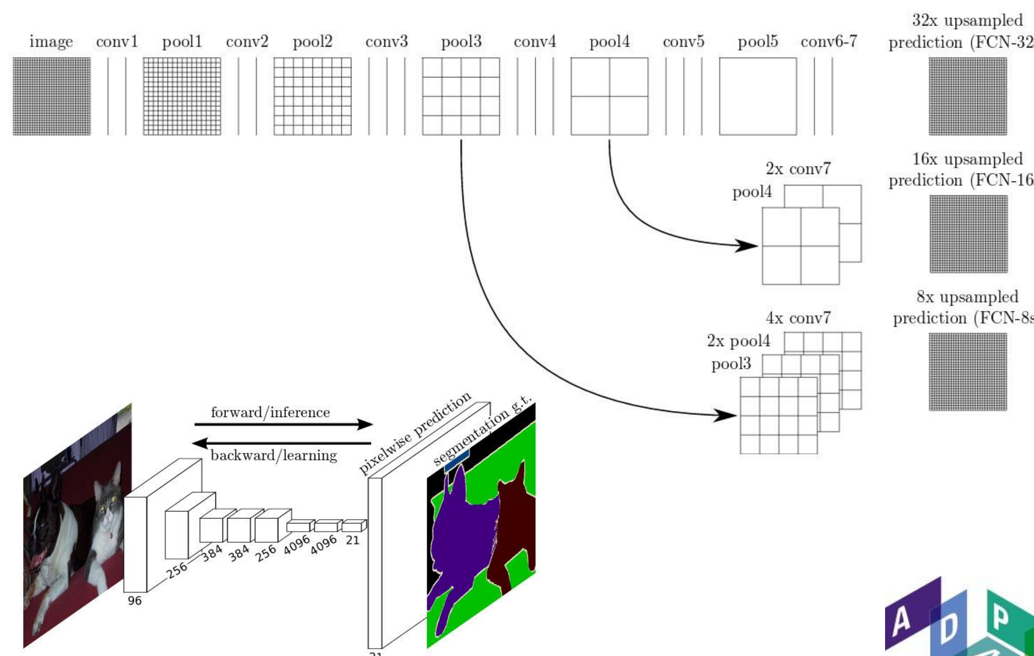
Processing pipeline: semantic segmentation



- ***Object detection:***
  Semantic segmentation
  with **Fully Convolutional NNs**:

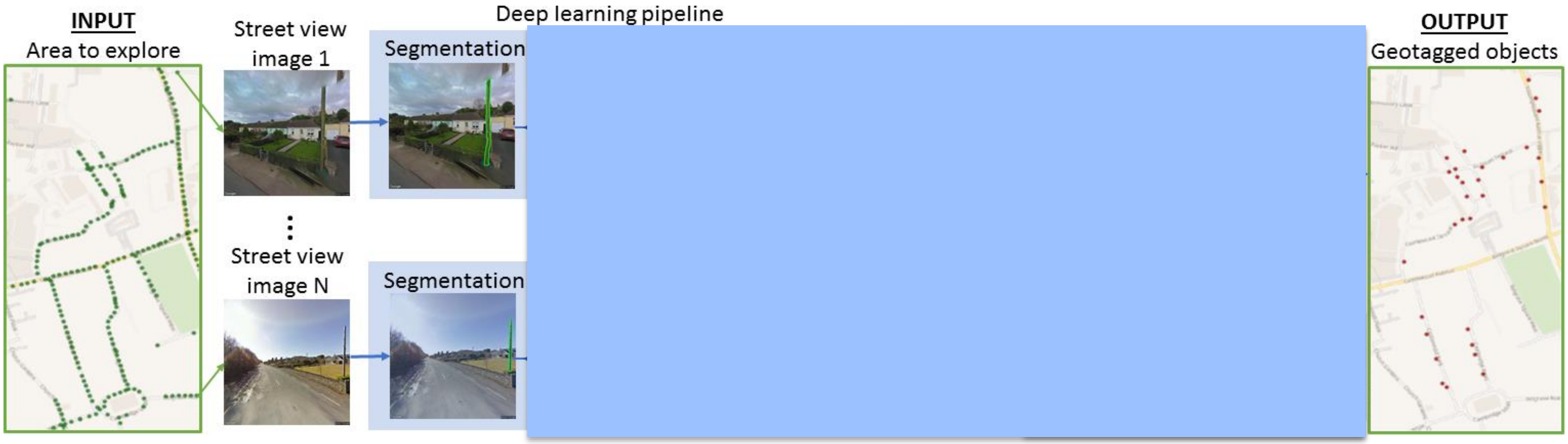
  - Introduce *extra FP penalty*

  - *Retrain* on one or multiple classes of objects: on *Mapillary Vistas, Cityscapes*



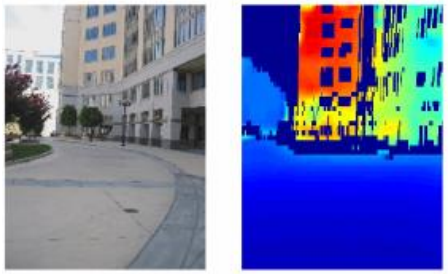
Shelhamer E. et al., IEEE T-PAMI 2017

Processing pipeline: monocular depth estimation



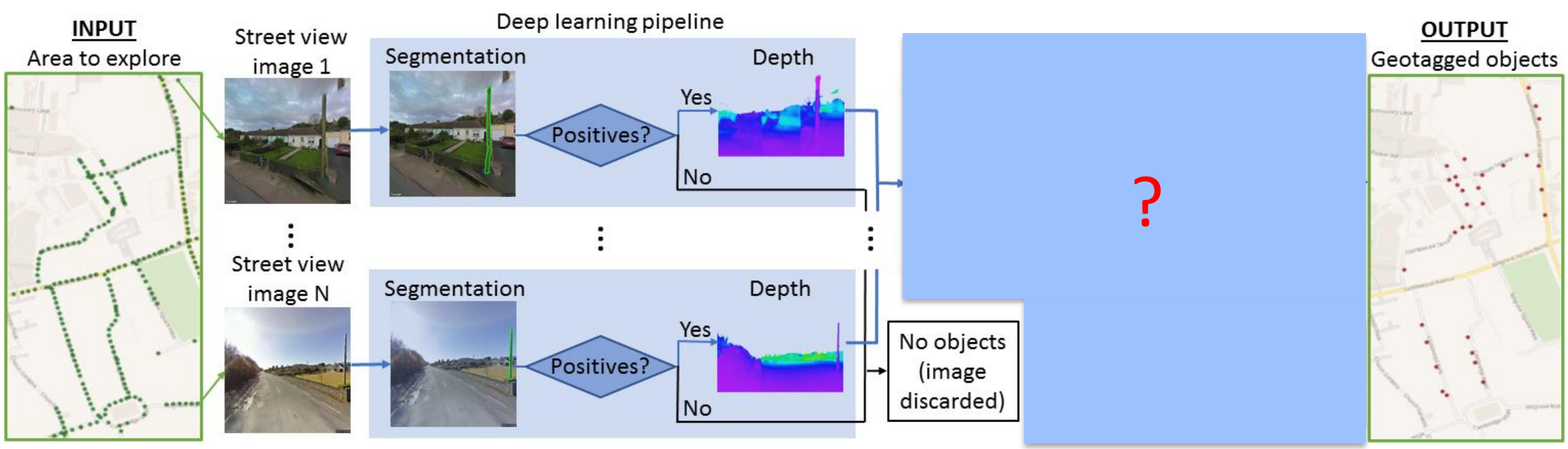
➢ *Spatial scene analysis:*

- Stereo-vision, Structure-from-Motion
  - *Requires more data, assumptions.*
- Monocular depth estimation
  - *Provides approximate accuracies;*
  - *Requires segmented objects.*

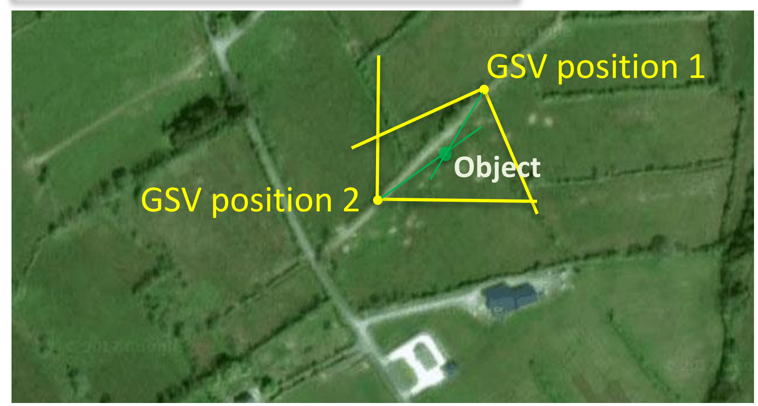


Laina I. et al., 3d Vision 2016

Processing pipeline: geotagging



- Strategies to estimate the position of objects from images:
  - Depth-based
  - Triangulation-based



- ✓ Single view: sensitivity
- ✓ Single view: false positives
- ✓ Low accuracy: up to 7m error

- ✓ High accuracy
- ✓ Multiple views
- ✓ Matching required

Processing pipeline: geotagging

- We define a **Markov Random Field** (MRF) model over the space of all view-rays intersections:
  - label *z=0* if not occupied by object
  - label *z=1* if occupied

- MRF configuration is characterized by its corresponding energy *U*. *Optimal = minimum of U.*

  Energy terms:
  - Unary term. *Consistency with depth.*

$$u_1(z) = z \sum_{j=1,2} \|\Delta_j - d_j\|$$

  - Pairwise term. *No occlusions. No spread.*

$$u_2(z) = z \sum_k z_k \, \|x - x_k\|$$

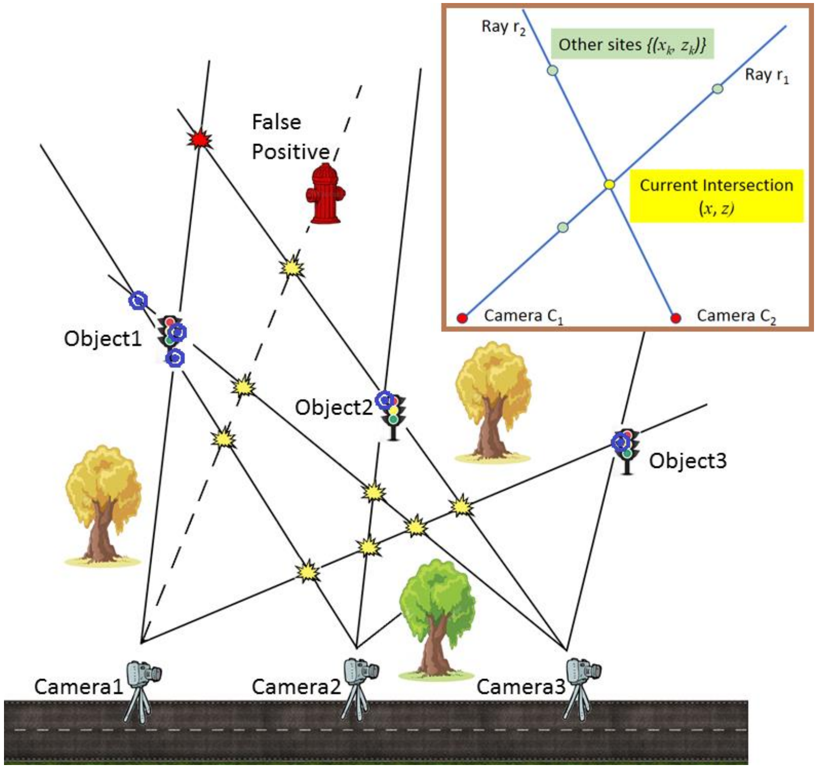
  - Ray term. *Penalize not matched rays.*

$$u_3(z) = (1-z) \prod_k (1 - z_k)$$

Total energy:

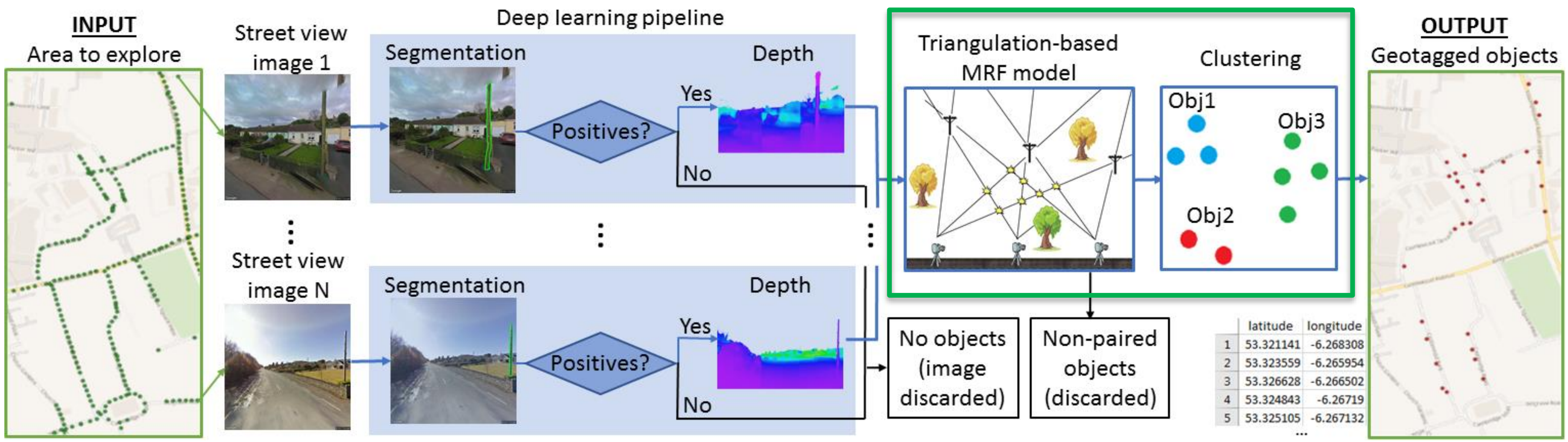
$$\mathcal{U}(\mathbf{z}) = \sum_{i=1}^{N_Z} \left[\alpha \, u_1(z_i) + \beta \, u_2(z_i) + (1-\alpha-\beta) \, u_3(z_i) \right]$$

$$\alpha, \beta \geqslant 0, \ \alpha + \beta \leqslant 1.$$



Δ – depth estimates
d – triangulated distances
x – Euclidean intersections

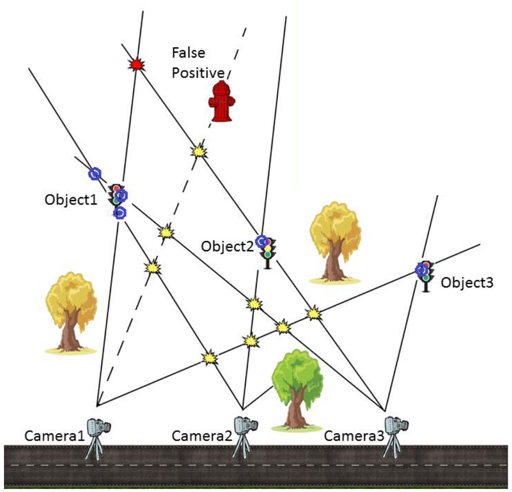
Processing pipeline: geotagging



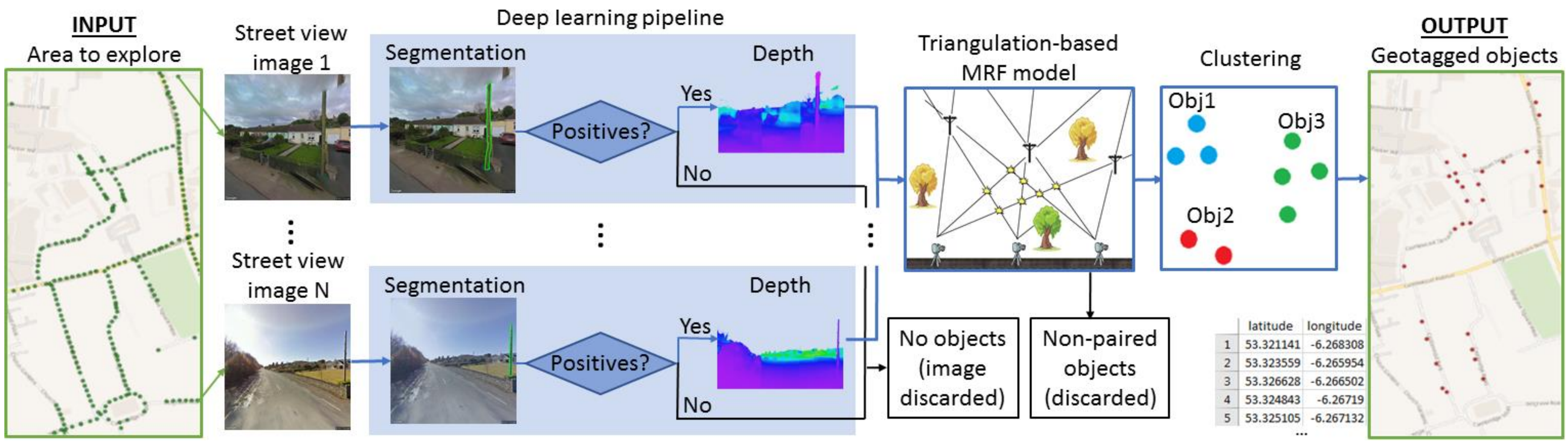
	latitude	longitude
1	53.321141	-6.268308
2	53.323559	-6.265954
3	53.326628	-6.266502
4	53.324843	-6.26719
5	53.325105	-6.267132

...

- The geotagging is performed as follows:
  - ✓ Calculate the space of all intersections;
  - ✓ Optimize the MRF model;
  - ✓ Discard non-paired instances;
  - ✓ Cluster the results. Take intra-cluster averages:
    - *Sparsity assumption*.



Processing pipeline: OVERVIEW



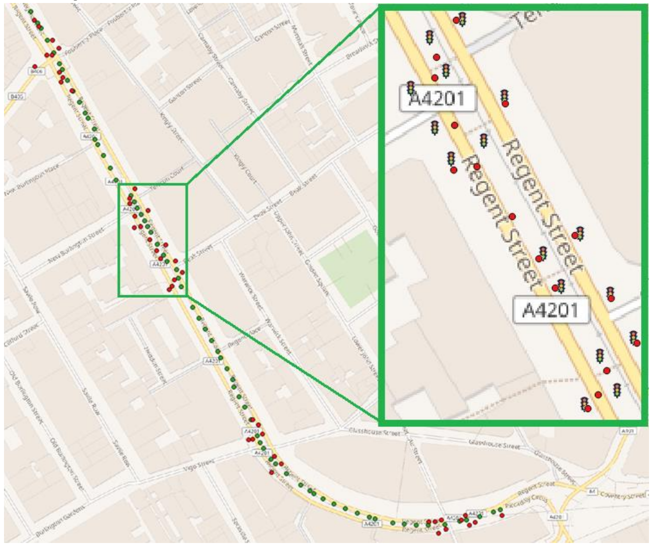
Object detection pipeline:

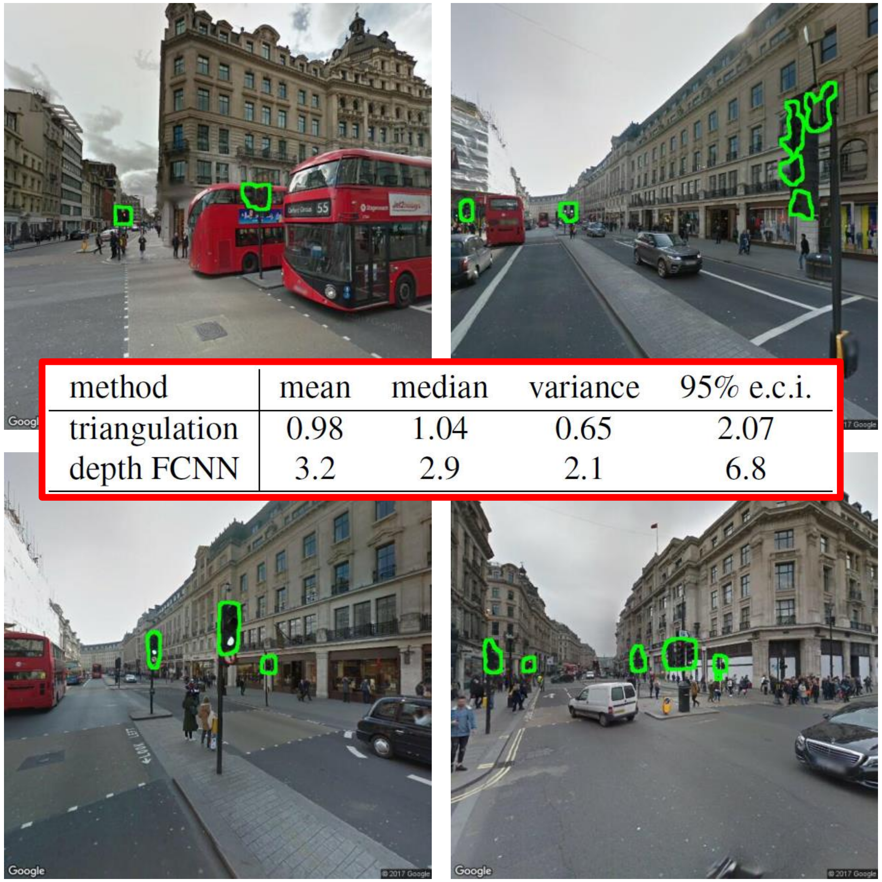
- DL: pixel-level **segmentation** to identify objects;
- DL: monocular **depth** (camera-to-object distance) estimation:
  - *max distance from camera: 25m;*
- **GPS-tagging** based on triangulation and Markov Random field model:
  - *mild object sparsity assumption - 1m apart;*
- Clustering.

Results: traffic lights

- Geotagging of **traffic lights** in Regent str., London, UK:
  - 87 GSV panoramas, 47 out of 50 objects discovered (94% recall)

Map view:





method	mean	median	variance	95% e.c.i.
triangulation	0.98	1.04	0.65	2.07
depth FCNN	3.2	2.9	2.1	6.8

Quantitative performance:

Object	#Actual	#Detected	TP	FP	FN	Recall	Precision
	50	51	47	4	3	94.0%	92.2%

Results: DEMO

- Geotagging of **telegraph poles** over a 2km road, co. Kildare:
  - 170 GSV panoramas, 37 out of 38 objects discovered (97.4% recall)

- We gratefully acknowledge financial support and expertise of ***eir*** in producing these results

Conclusions

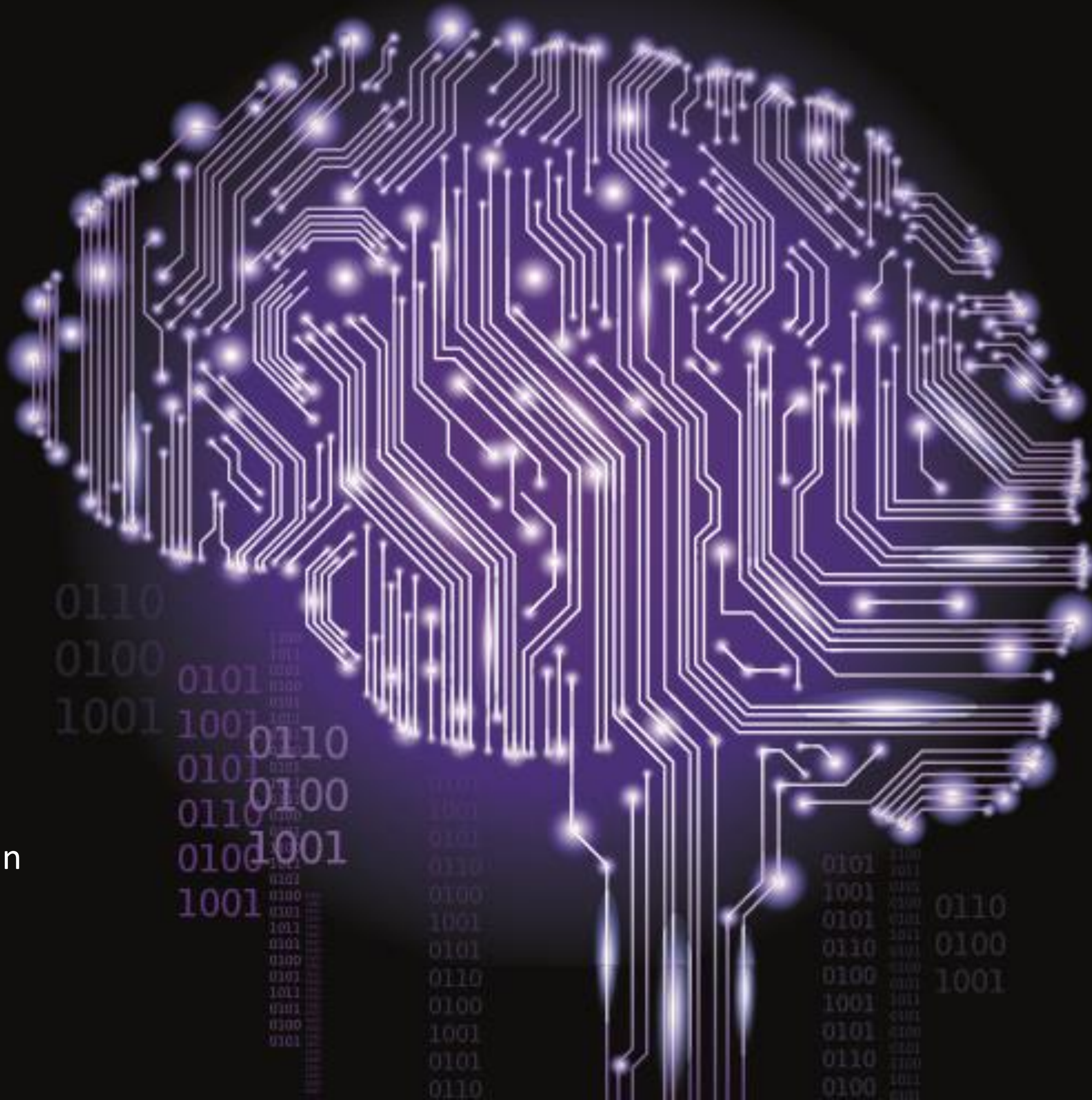
We have developed an image processing pipeline that:

- Is fully **automatic**;
- The geotagging accuracy **comparable** with commercial-range GPS-unit;
- Detects and geotags objects at approx. 1.1 GSV panorama per second rate (***~3.000 km*** in 24h on a desktop PC with 2 GPUs);
- Can accommodate custom detection and depth estimation modules.

ADAPT

Engaging Content
Engaging People

Thank you!



Contact Us
O'Reilly Building
Trinity College Dublin
Dublin 2
Ireland

adaptcentre.ie