iit PAVIS

ISTITUTO ITALIANO
DI TECNOLOGIA
PATTERN ANALYSIS
AND COMPUTER VISION

Semantic 3D Scene Understanding: from 3D Point Cloud to Graph representations

Alessio Del Bue

Fondazione Istituto Italiano di Tecnologia (IIT)

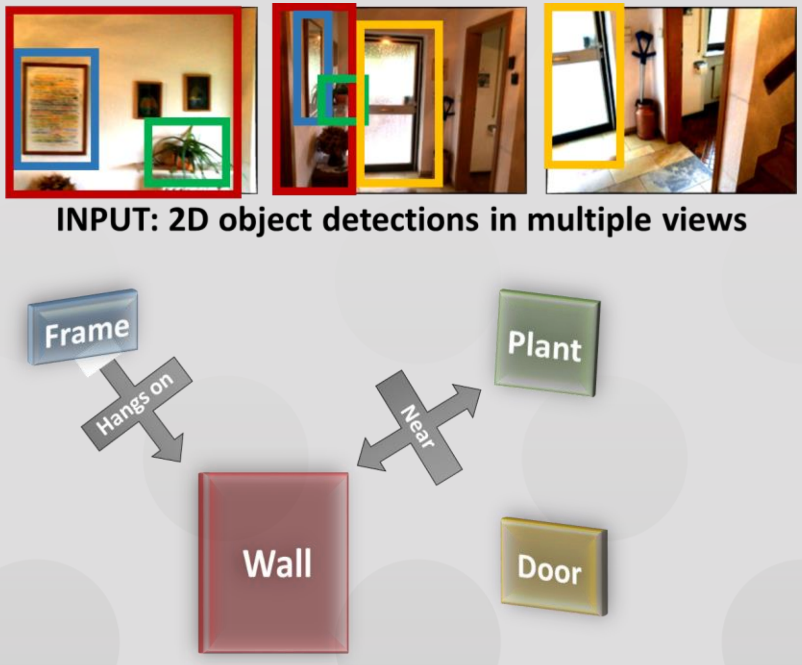
September 2 2023

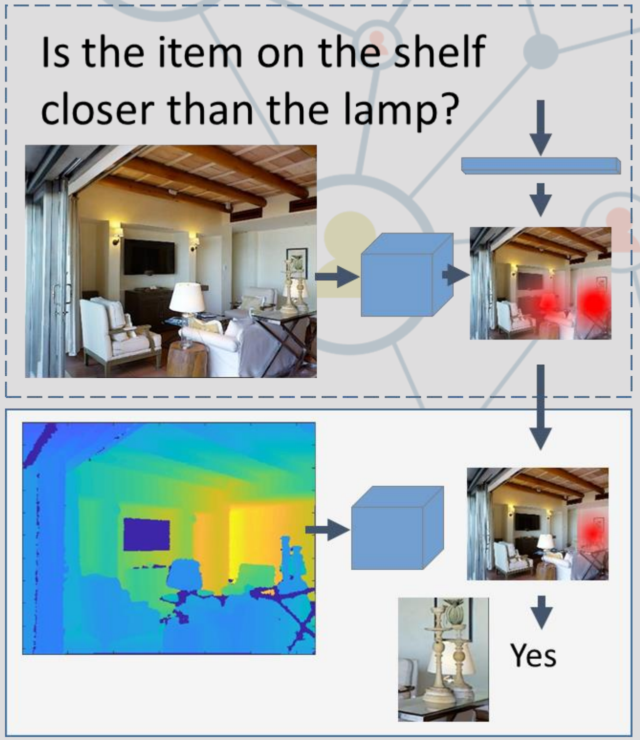
Padova, Italy - September 4-8, 2023

VISMAC 23

International Summer School on
Machine Vision

3D scene understanding





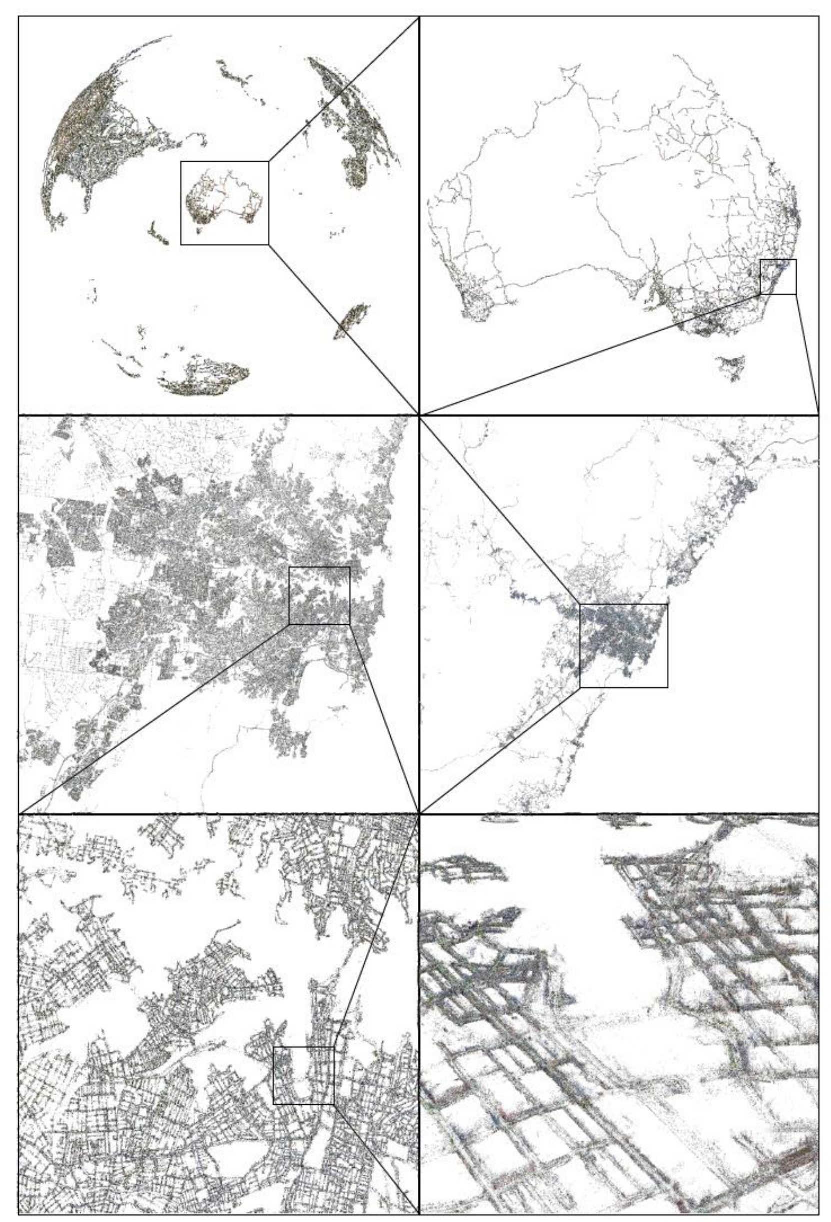
DIGITISATION & 3D MODELLING

REPRESENTATION (RIGID AND DYNAMIC)

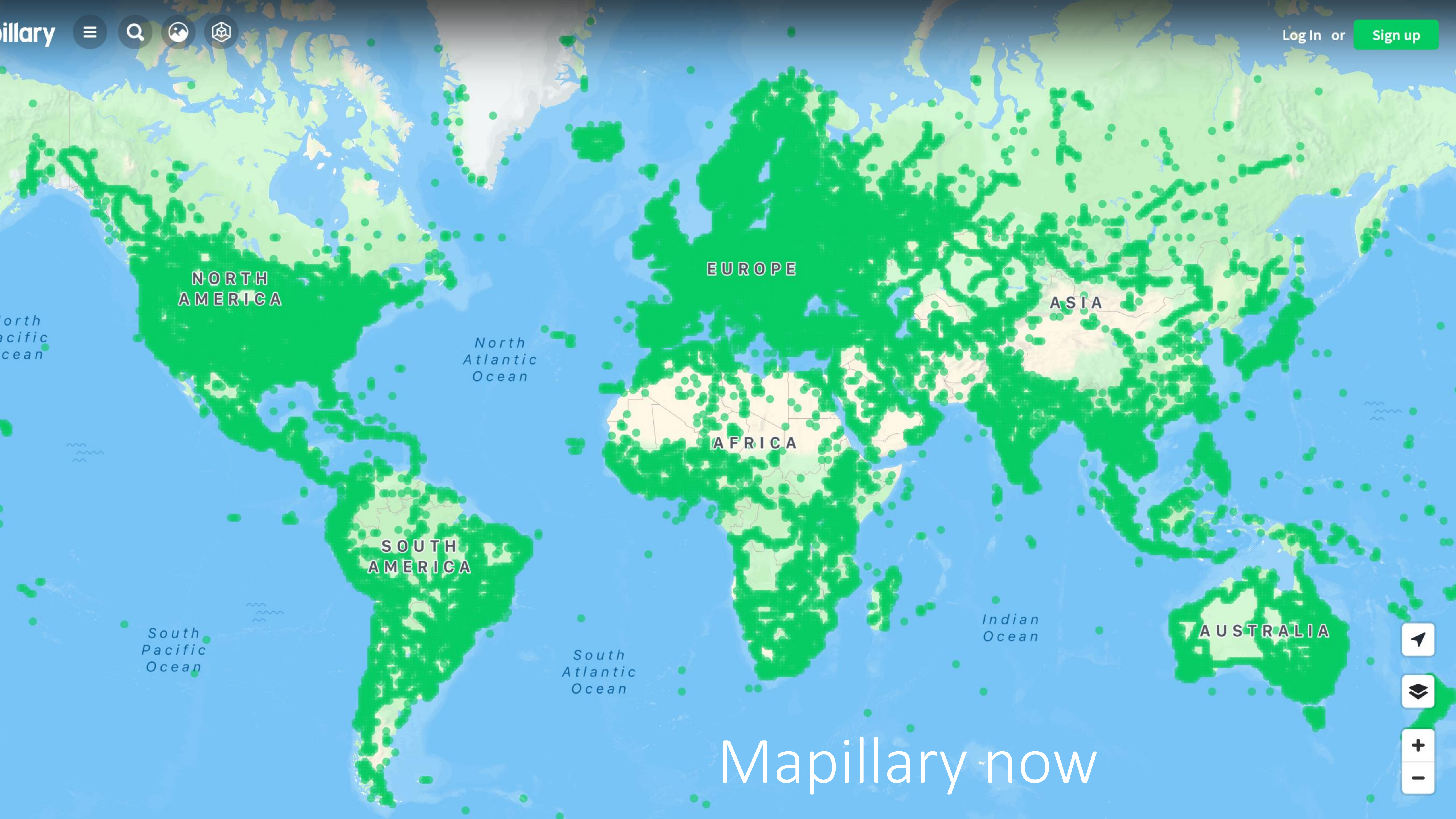
INTERPRETATION & UNDERSTANDING

Digitisation and 3D modelling so far…

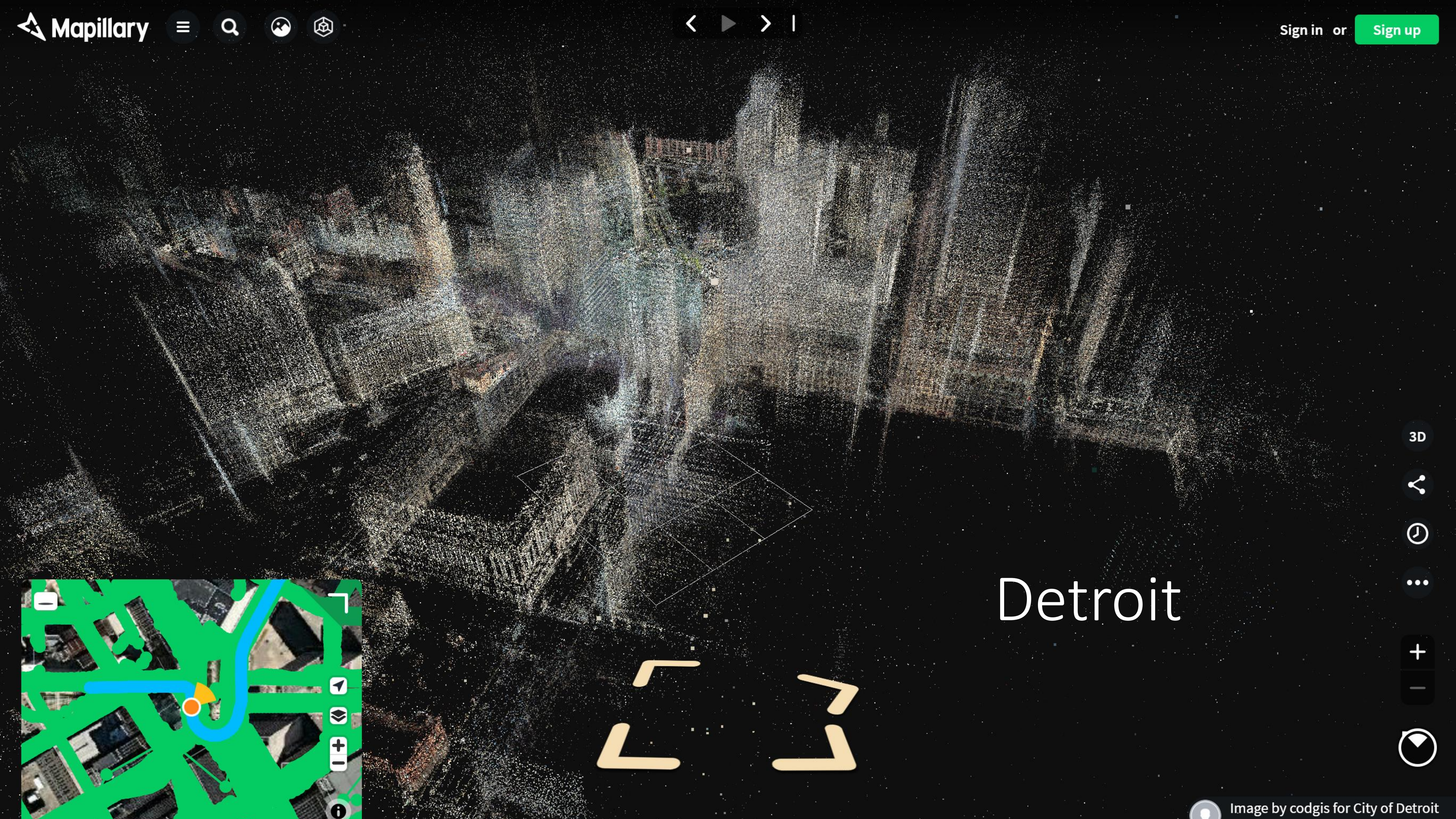
- An appearance-augmented point cloud comprising 404 billion tracked features, computed from Google street-level imagery.
- Every point in the cloud carries its local appearance descriptors from at least three different viewpoints.



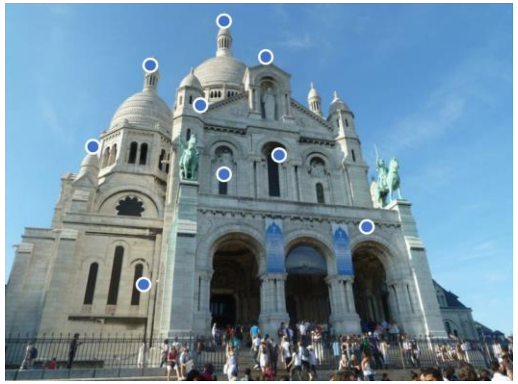
Klingner, Bryan, David Martin, and James Roseborough. "Street view motion-from-structure-from-motion." CVPR 2013.



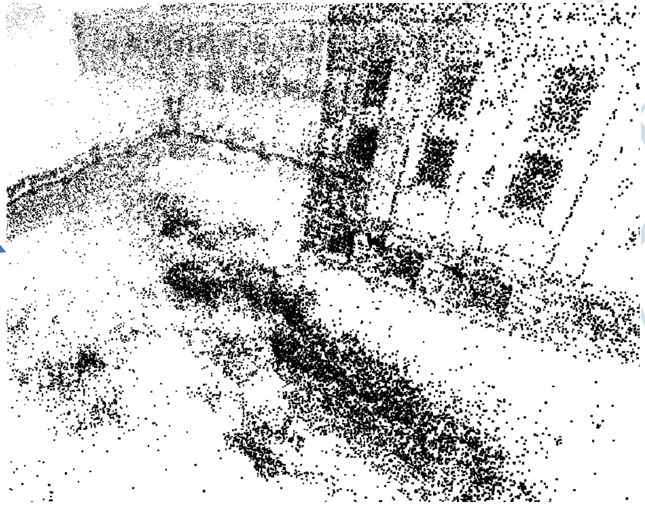
Mapillary now

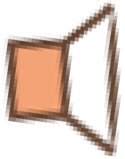

Mapillary
Sign in
or
Sign up
3D
Detroit
Image by codgis for City of Detroit

Multi-view 3D model (projective)

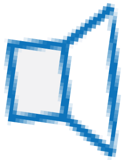


$$= K \left[\; R \mid t \;\right] \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$





3 x 3 matrix with intrinsic parameters (focal length, aspect ratio, principal point)



3 x 4 matrix of extrinsic parameters (camera rotation and translation)

4 x 1 homogeneous coordinates of the 3D point

Camera projection

From projective geometry to image plane coordinates (non-linear operator) $\Re^3 \longmapsto \Re^2$

$$\Pi\begin{pmatrix} u \\ v \\ \lambda \end{pmatrix} = \begin{pmatrix} \frac{u}{\lambda} \\ \frac{v}{\lambda} \end{pmatrix} \Rightarrow \begin{pmatrix} u' \\ v' \end{pmatrix} = \Pi\left(K\,[\,R \mid t\,]\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}\right)$$

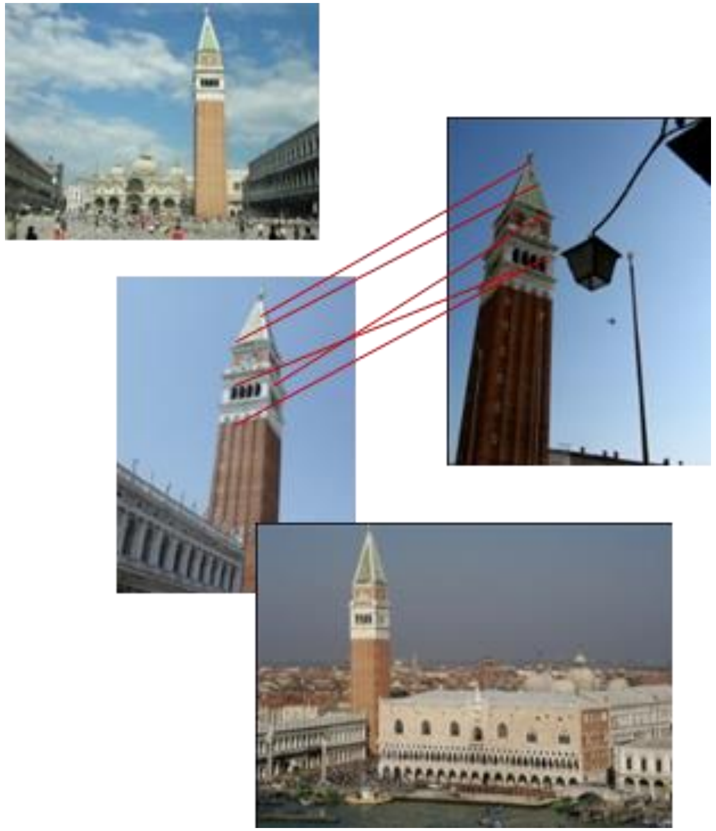
$$\begin{pmatrix} u \\ v \\ \lambda \end{pmatrix} = P\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix}$$

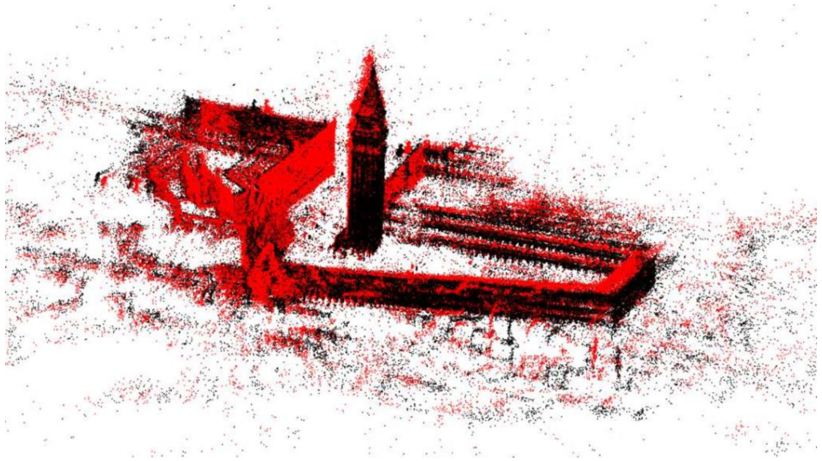
Other option is to estimate 3D structure and motion in the projective space and then upgrade later the reconstruction to metric.

Cost (loss) function

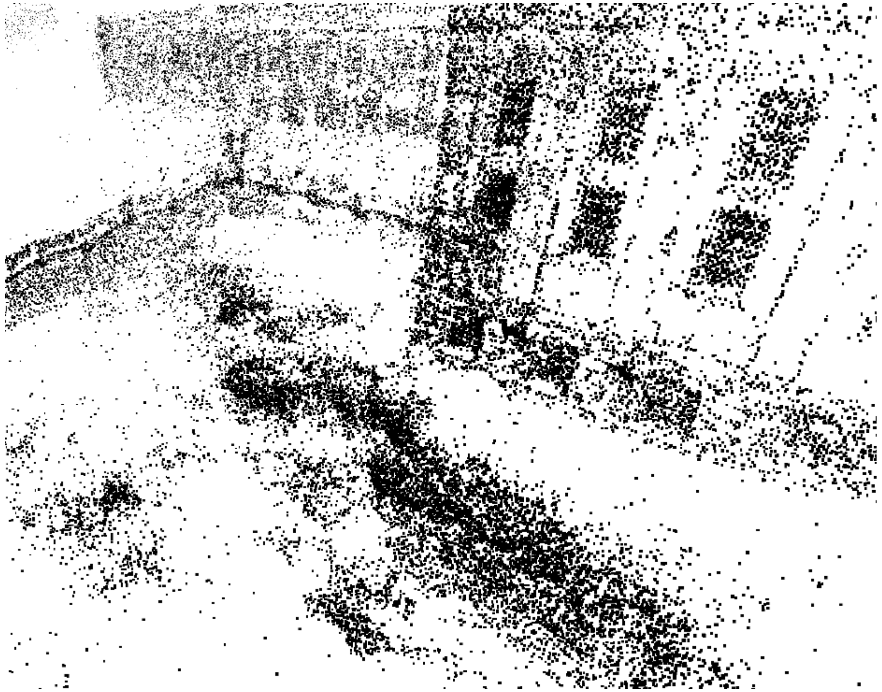
For every 3D point *i* and every camera *f* optimise for the model parameters:

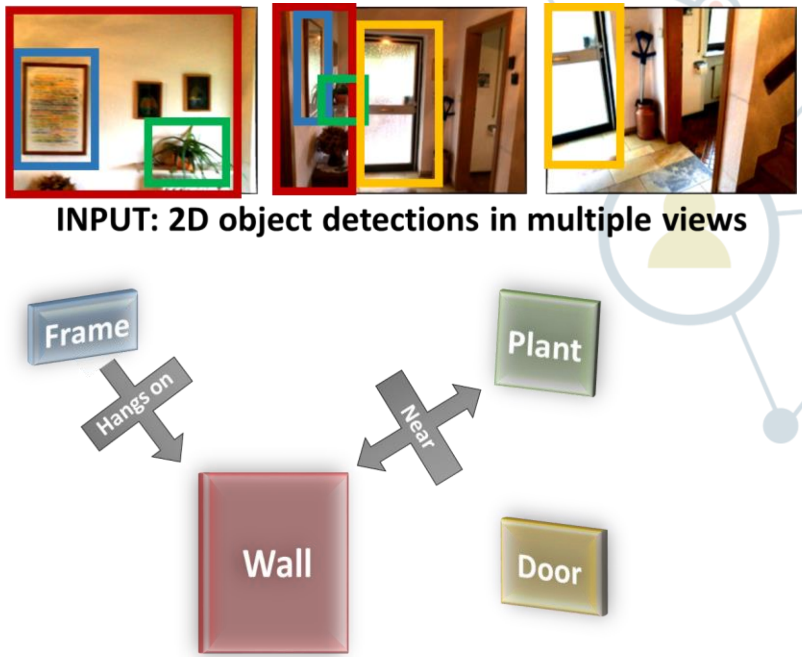
$$L_\theta = \sum_{i,f} \left\| \begin{pmatrix} u'_{i,f} \\ v'_{i,f} \end{pmatrix} - \Pi\left(K_f \, [\, R_f \mid t_f \,] \begin{pmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{pmatrix} \right) \right\|_2^2$$





From point cloud to semantic





INPUT: 2D object detections in multiple views

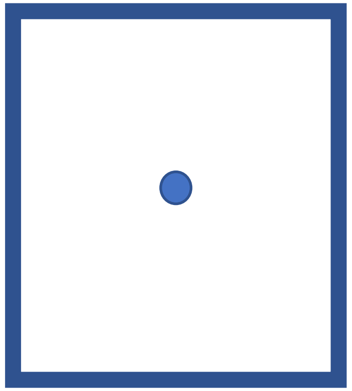
REPRESENTATION
(RIGID AND DYNAMIC)

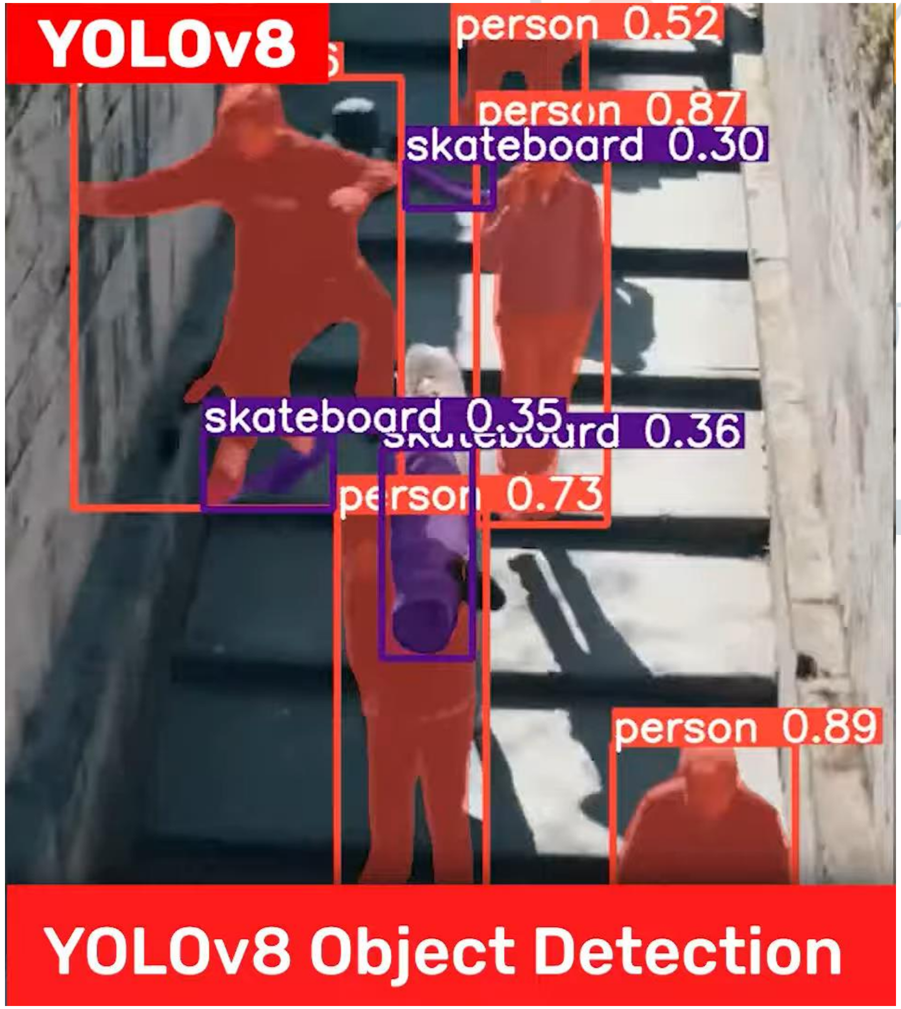
Observation: from points to detections

Source: Nvidia https://nvda.ws/2mCdYgx

Semantic Structure from Motion

- Bounding boxes (bbx) are geometrical instances but they convey semantics
- Degenerate bbx are just image points as in SfM

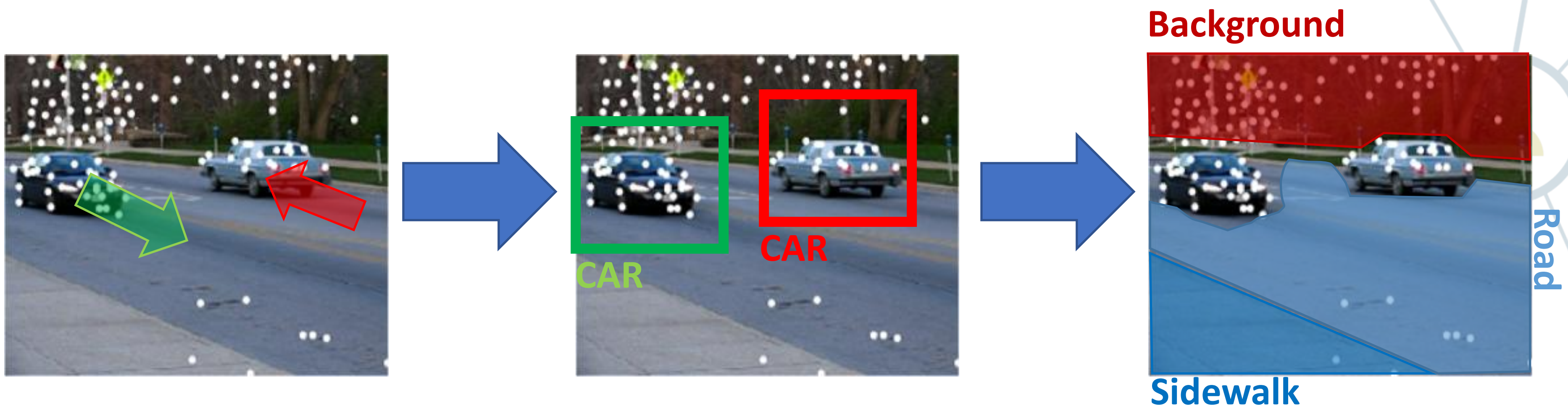




Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." *arXiv preprint* 2018.
YOLOv8: https://github.com/ultralytics/ultralytics

What can we gain from semantic?

How **high level semantic in 3D reconstruction** can help?



Generic scene

Points to objects relations

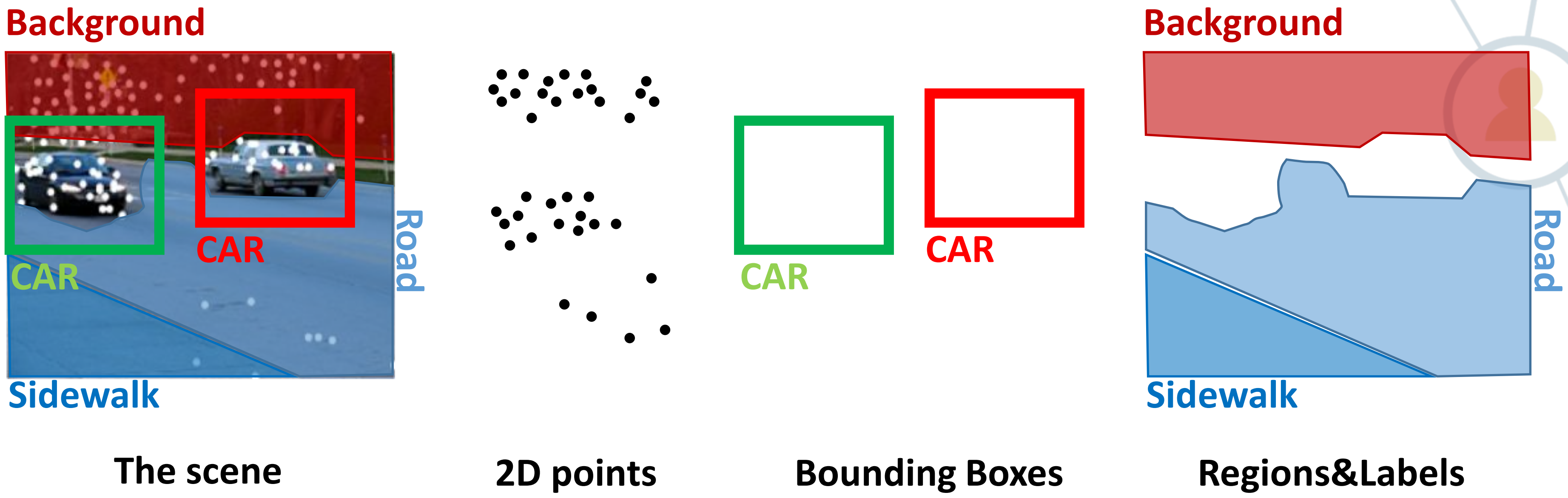
Points to regions relations

Including these relations **may help** to solve classical Motion Segmentation and Structure from Motion (SfM) problems.

Bao and Savarese, "Semantic structure from motion" *CVPR 2011*

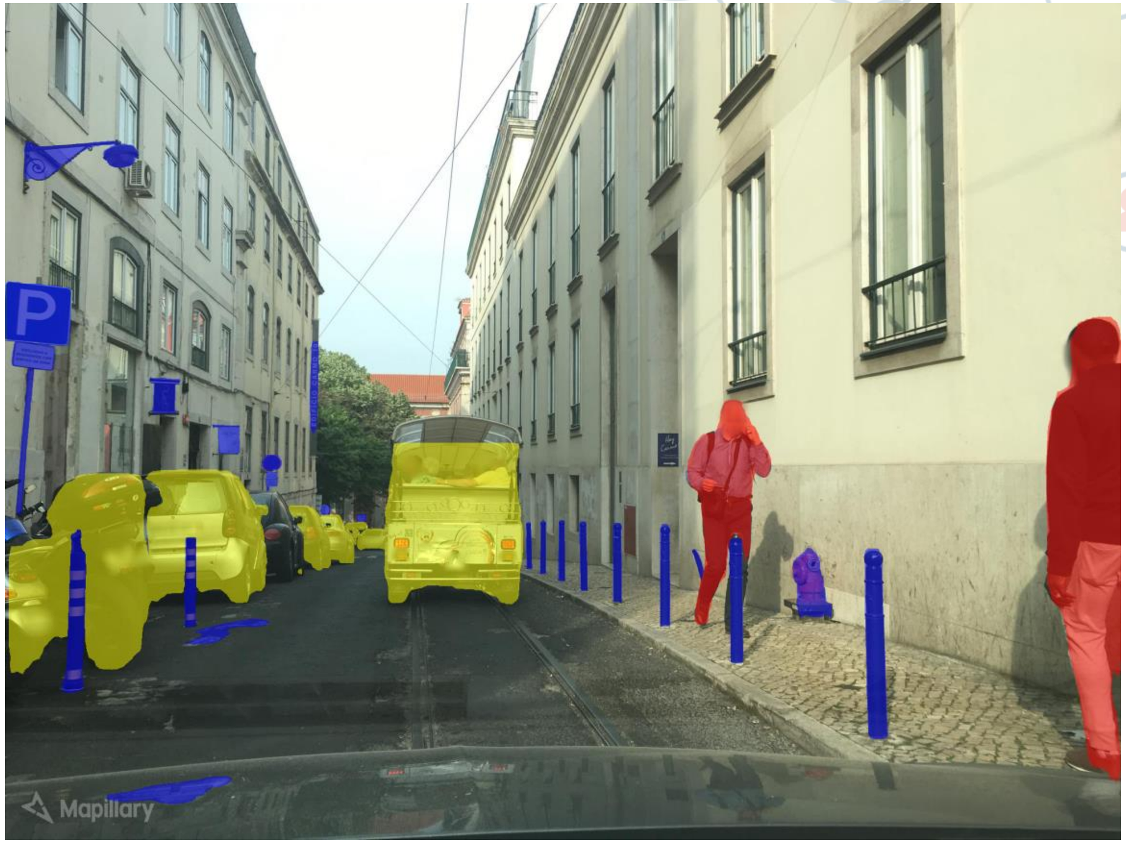
How to bridge the gap?

Different data that lives in different spaces (points, bbx, labels, etc.)

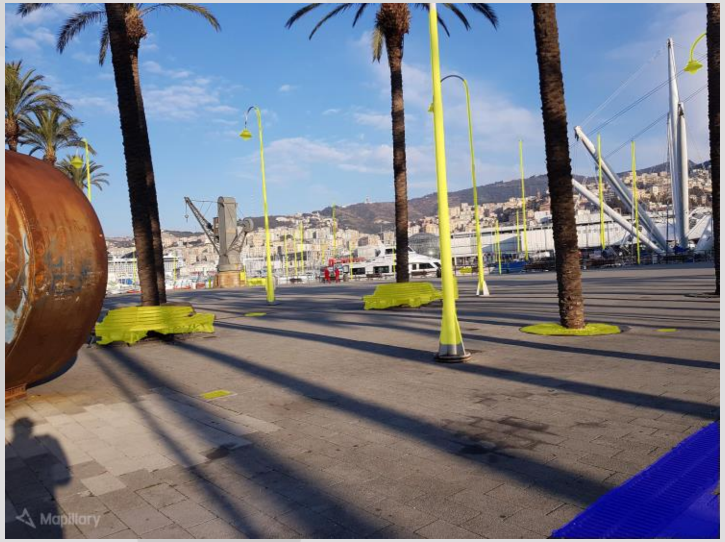


How to blend this information available from deep learning in a single computational approach? Is there a viable solution (i.e. efficient)?

Now semantic is ready to use



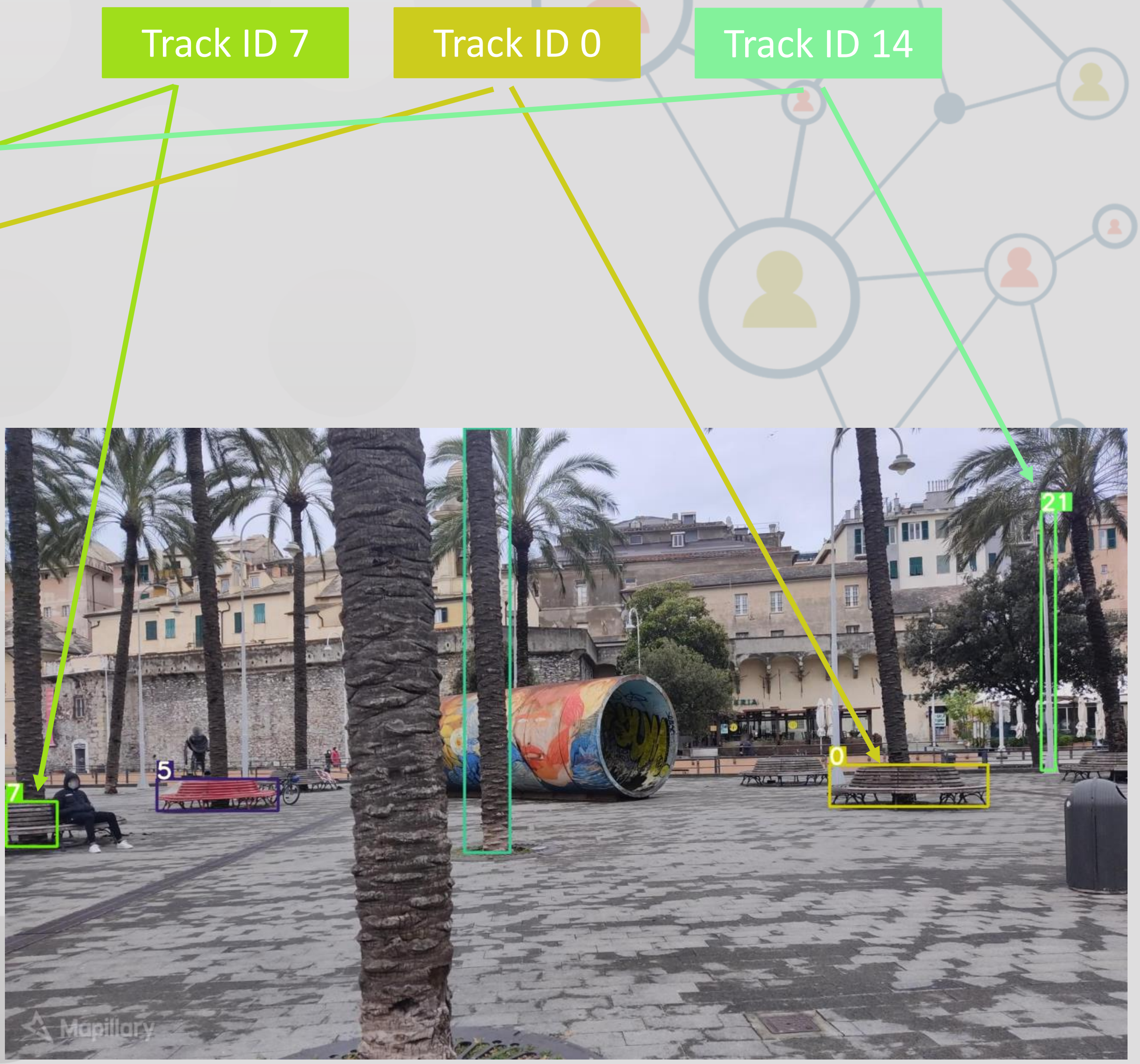
Genoa Porto Antico – Instance segmentation



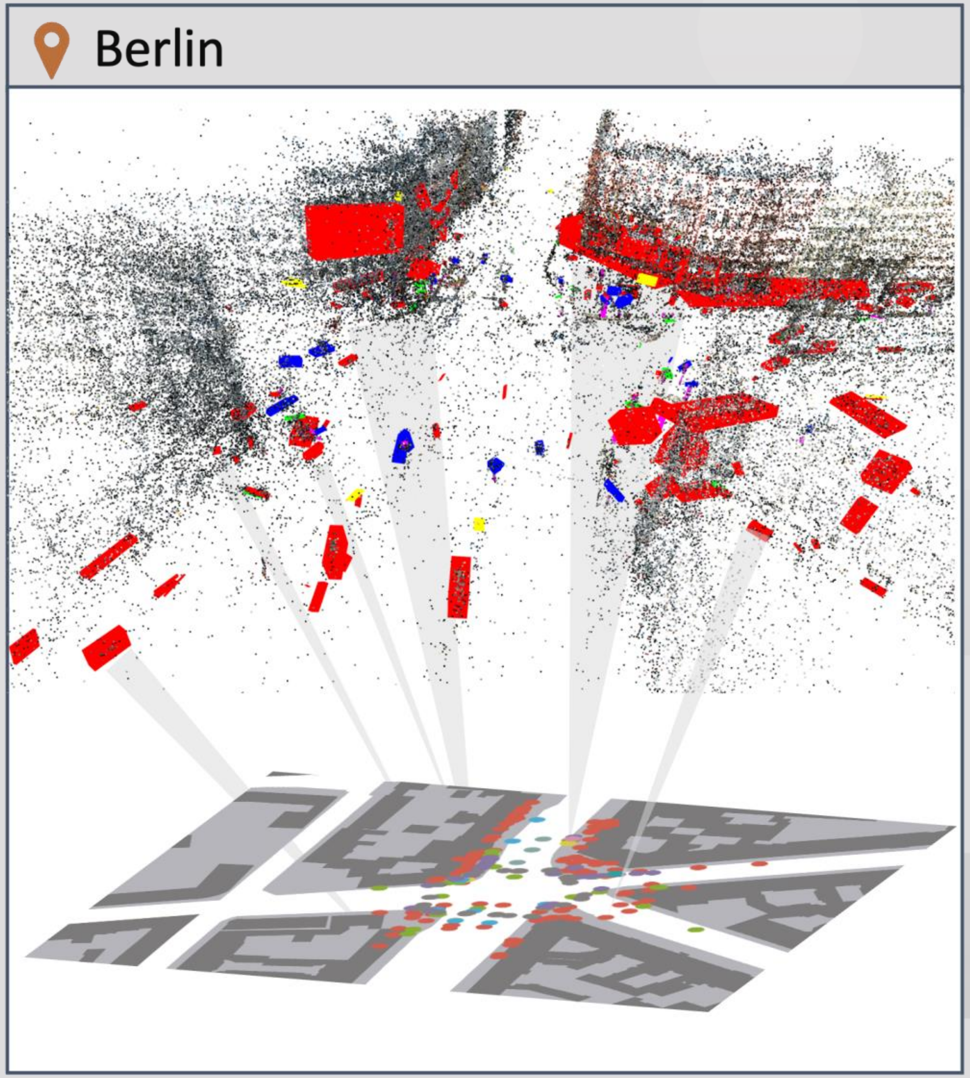
Red – people, animals
Blue – vehicles
Yellow – static objects:

- Benches
- Poles
- Manholes
- Banners
- Street lights
- ...

Genoa Porto Antico – Matched detections



Revisiting Semantic Structure from Motion



Crocco M, Rubino C, Del Bue A. Structure from motion with objects. CVPR 2016.

Rubino C, Crocco M, Del Bue A. 3d object localisation from multi-view image detections. TPAMI 2017

Gay P, Rubino C, Bansal V, Del Bue A. Probabilistic structure from motion with objects (psfmo). ICCV 2017.

Gay P, Stuart J, Del Bue A. Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning. ACCV 2018.

Giuliari et al. Spatial commonsense graph for object localisation in partial scenes. CVPR 2022.

Taiana et al. “PoserNet: Refining Relative Camera Poses Exploiting Object Detections”. ECCV 2022

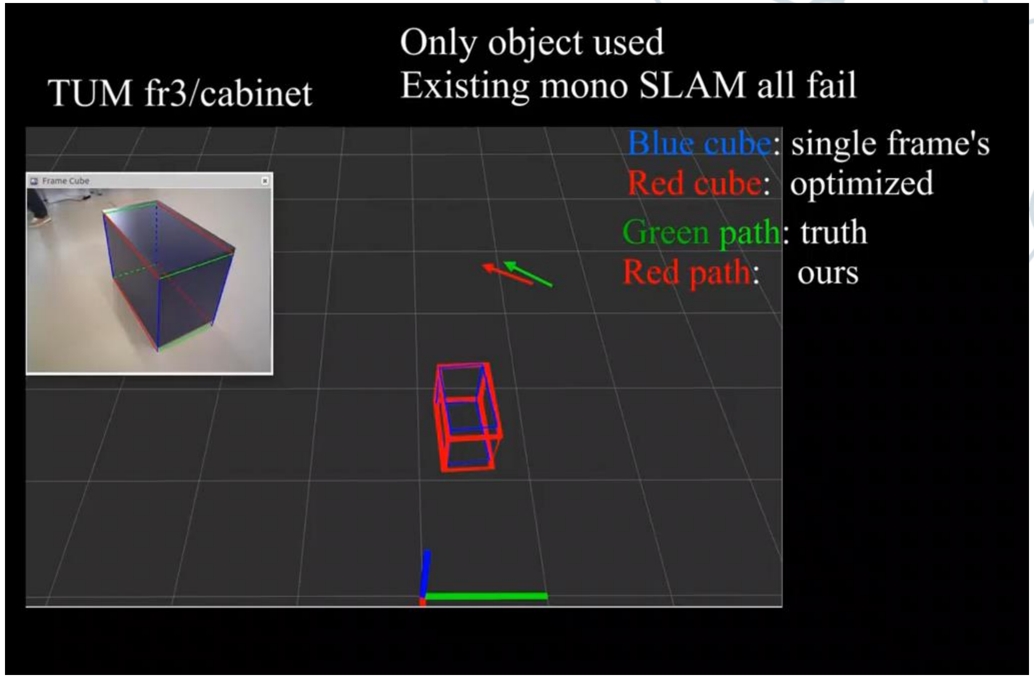
Toso et al. You are here! Finding position and orientation on a 2D map from a single image: The Flatlandia localization problem and dataset. arXiv 2023.

Key

Arrow Marking	Streetlight	Bench
Bicycle Symbol	Support Pole	Bike Rack
Sign	Traffic Light	Catch Basin
Traffic-Sign	Manhole	CCTV Camera
Trash can	Junction Box	Fire Hydrant

Including object detection in 3D vision

Gálvez-López, Dorian, et al. "Real-time monocular object slam." *RAS* 2016

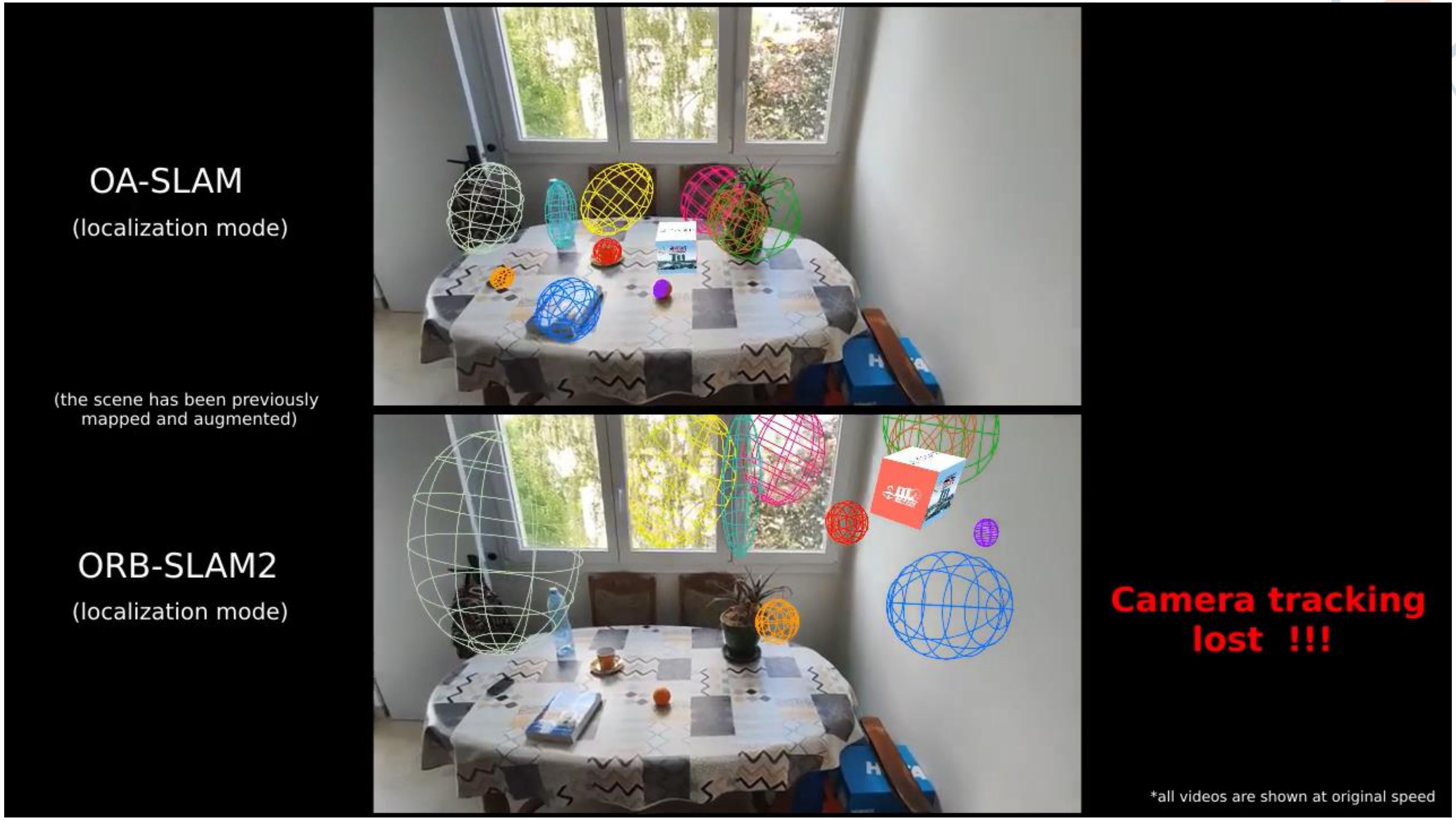


Yang, Shichao, and Sebastian Scherer. "Cubeslam: Monocular 3-d object slam.“ TRO 2019

Including object detection in 3D vision

Nicholson, Lachlan, Michael Milford, and Niko Sünderhauf. "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam." *RA-L* 2018

Including object detection in 3D vision



Matthieu Zins, Gilles Simon, Marie-Odile Berger. "OA-SLAM: Leveraging Objects for Camera Relocalization in Visual SLAM." *ISMAR 2022*

More examples on…

README.md

Awesome Object SLAM awesome

A curated list of Object SLAM papers and resources, inspired by awesome-implicit-representations.

Disclaimer

This list *does not aim to be exhaustive*. In this list, we consider papers that **jointly** optimize robot (camera) and object states, where object states typically include object poses and object shape parameters.

For more general SLAM papers, please refer to awesome-visual-SLAM and Awesome-SLAM.

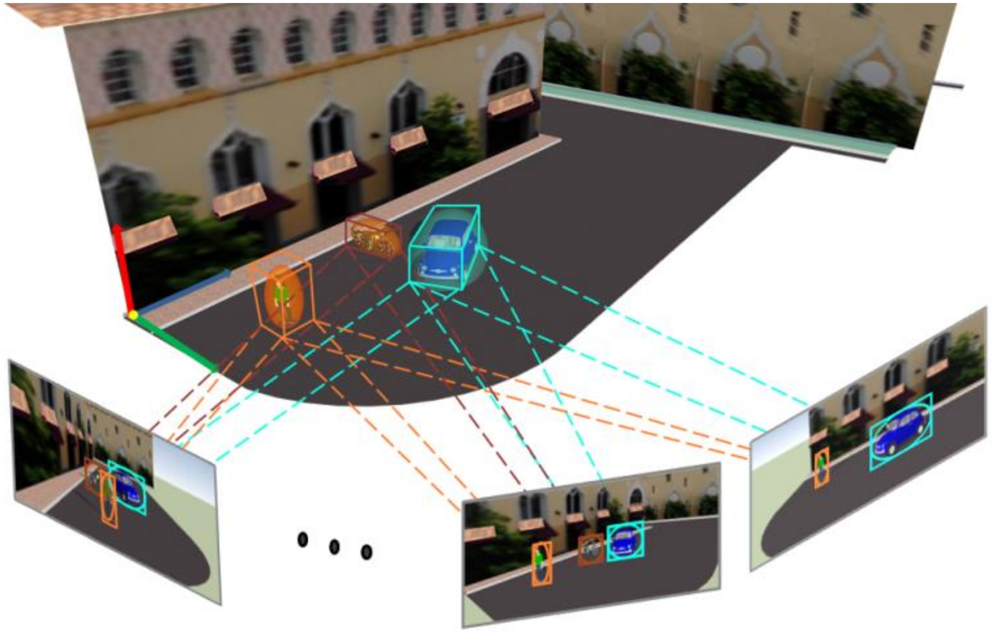
This repo is maintained by Ziqi Lu and Akash Sharma. You are very welcome to contribute to this repo. If you spot anything wrong or missing, please feel free to submit a pull request or contact the maintainers.

Table of Contents

https://github.com/520xyxyzq/awesome-object-SLAM

Localisation from Detections (LfD)

<u>Main goal:</u> to recover the **3D occupancy** of a set of objects given the **2D bounding boxes** from detections at each image frame.

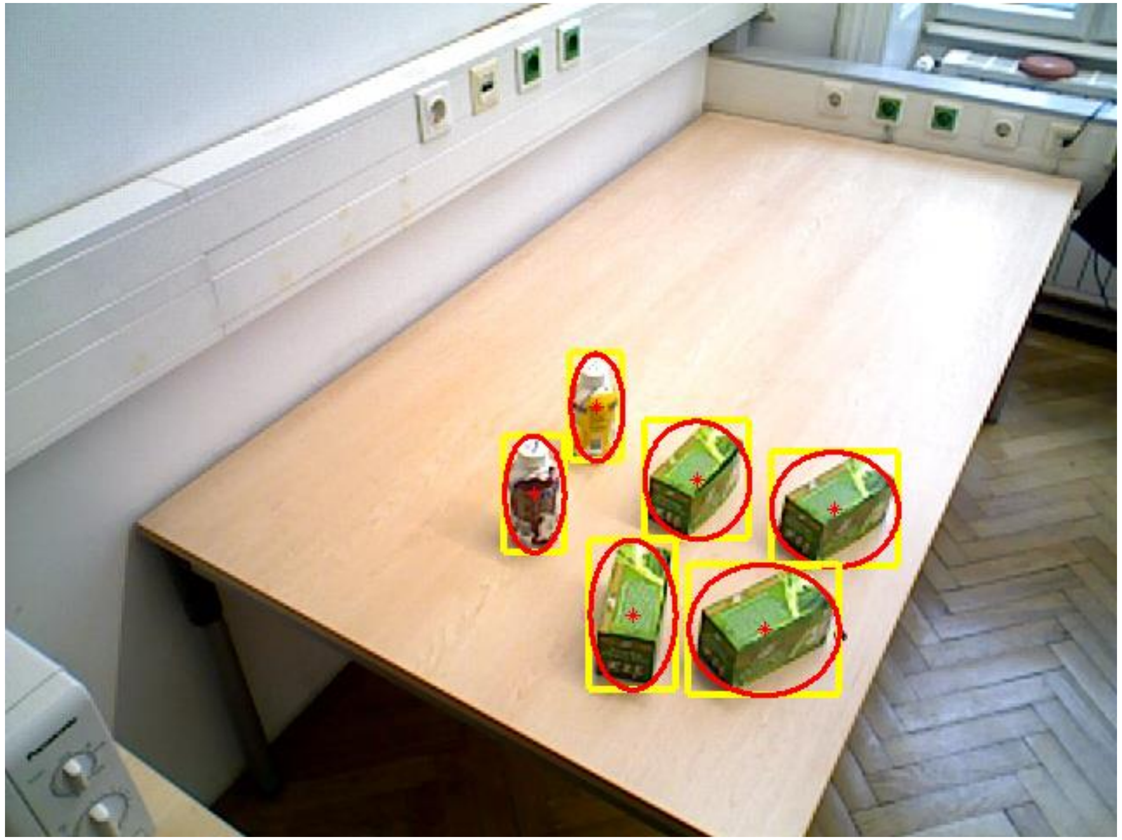


Features:

- Method working with **texture-less** objects: no 2D **interest points** required
- Just 2D bounding boxes position, size and aspect ratio are required

Rubino, Cosimo, Marco Crocco, and Alessio Del Bue. "3d object localisation from multi-view image detections." *TPAMI 2017*.

3D ellipsoids from bounding boxes



3D bounding box from set of **2D bounding boxes** (one per frame)

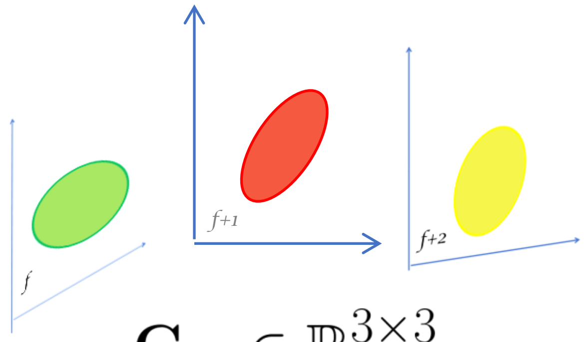
Not simple: piece-wise defined curves

3D ellipsoid from set of **2D ellipses** fitted to 2D bounding box

Algebraic solution available in closed form

Multiple conic to quadric reconstruction (1)

Conics from 2D bounding boxes at frames 1 ... *F* related to object *i*



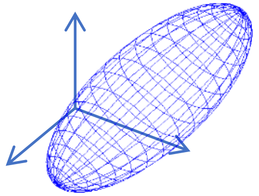
$$\mathbf{C}_{if} \in \mathbb{R}^{3\times 3}$$

Projection matrices for frames 1 ... *F*: can be estimated from the image background by standard multi-view relations

$$\mathbf{P}_f \in \mathbb{R}^{3\times 4}$$

$$\mathbf{Q}_i \in \mathbb{R}^{4\times 4}$$

Estimated Quadric (ellipsoid) representing 3D space occupancy of object *i*



[1] De Ma, S., Chen, X.: Reconstruction of quadric surface from occluding contour. In: Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on. Volume 1., IEEE (1994) 27–31
[2] G. Cross and A. Zisserman. Quadric reconstruction from dual-space geometry. ICCV 1998.

Multiple conic to quadric reconstruction (2)

Space of points in 2D and 3D

Primal space $\mathbf{C}_{if}$ $\mathbf{Q}_i$

No simple relation between quadric and conics in primal space

Adjoint operator

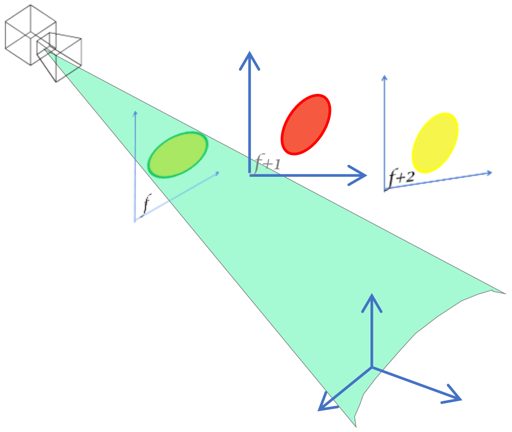
Dual space $\mathbf{C}^*_{if}$ $\mathbf{Q}^*_i$

$$\boxed{\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f \mathbf{Q}^*_i \mathbf{P}_f^\top}$$

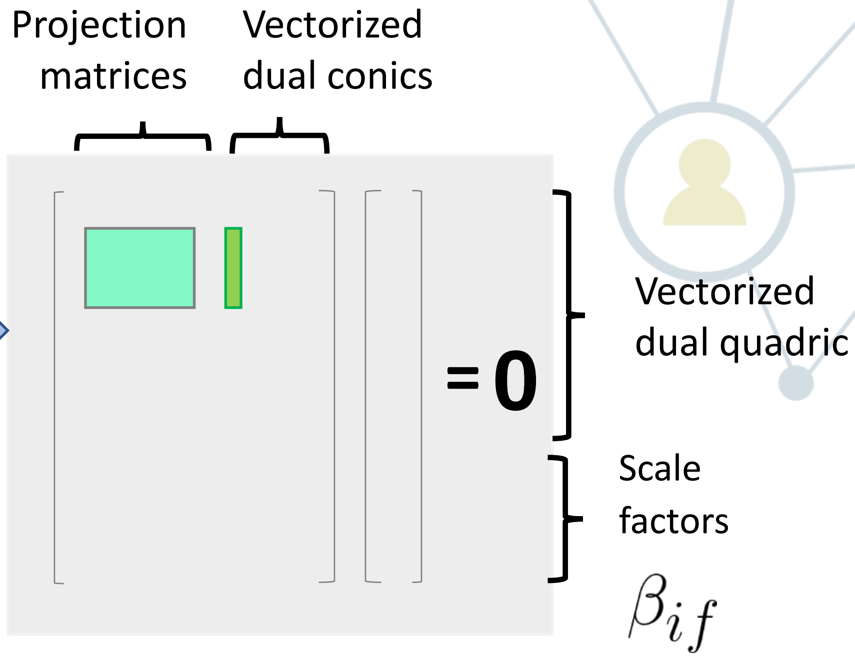
Space of lines and planes

Linear relation between quadric and conics in dual space

Multiple conic to quadric reconstruction (3)



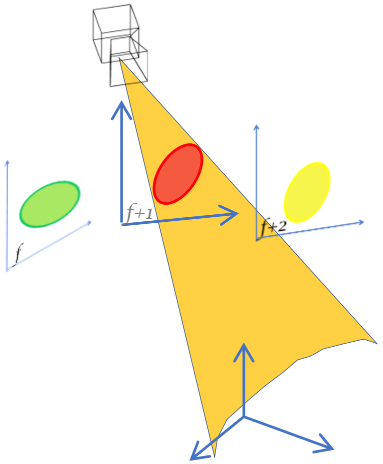
Vectorize dual conics and quadrics and stack over *f*



$$\boxed{\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f\mathbf{Q}^*_i\mathbf{P}_f^\top}$$

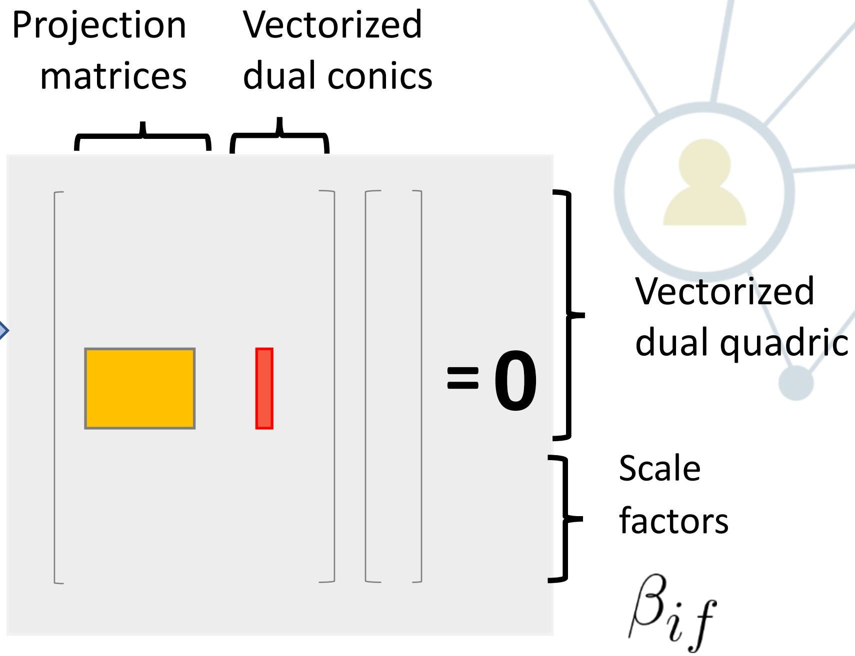
$$\mathbf{M}_i\mathbf{w}_i = \mathbf{0}_{6F}$$

Multiple conic to quadric reconstruction (3)



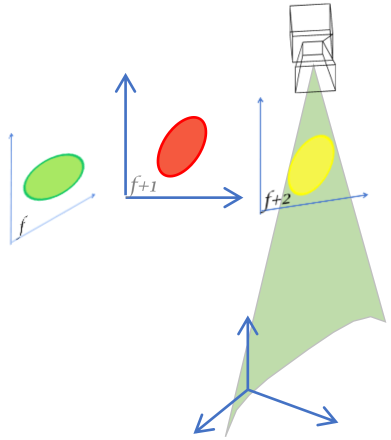
$$\boxed{\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f\mathbf{Q}^*_i\mathbf{P}_f^\top}$$

Vectorize dual conics and quadrics and stack over f



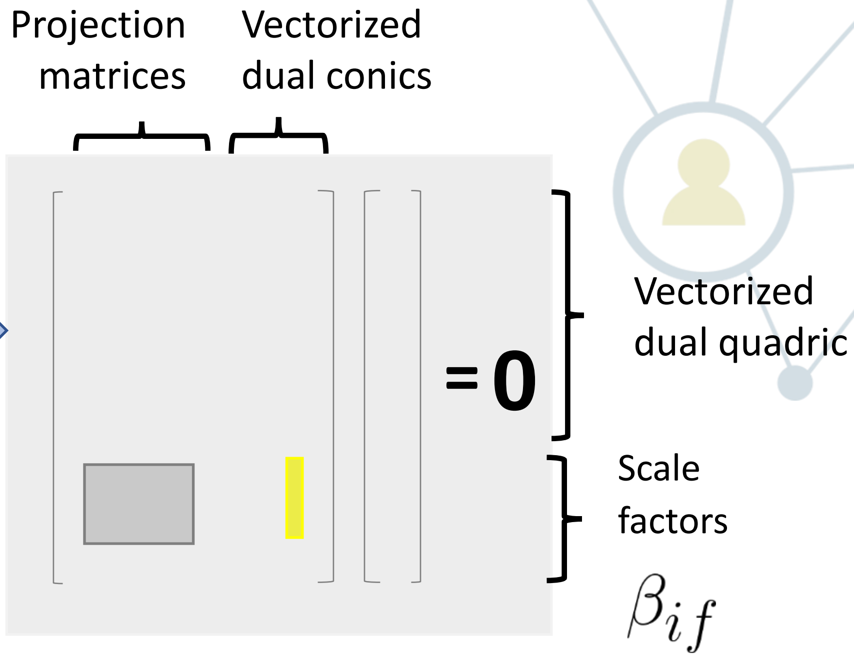
$$\mathbf{M}_i\mathbf{w}_i = \mathbf{0}_{6F}$$

Multiple conic to quadric reconstruction (3)



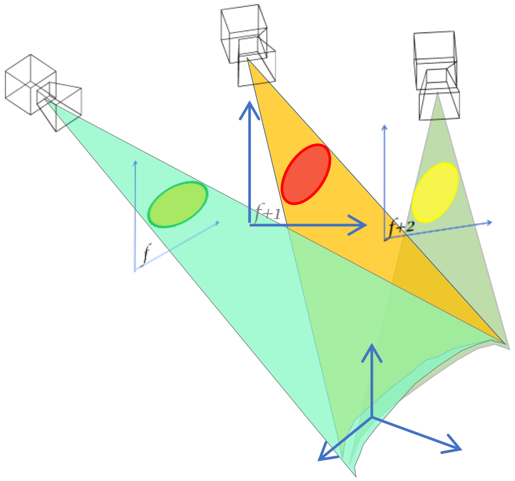
$$\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f\mathbf{Q}^*_i\mathbf{P}_f^\top$$

Vectorize dual conics and quadrics and stack over *f*



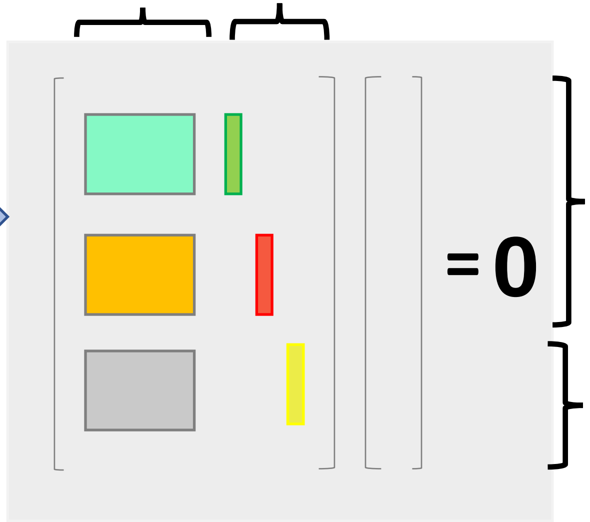
$$\mathbf{M}_i\mathbf{w}_i = \mathbf{0}_{6F}$$

Multiple conic to quadric reconstruction (3)



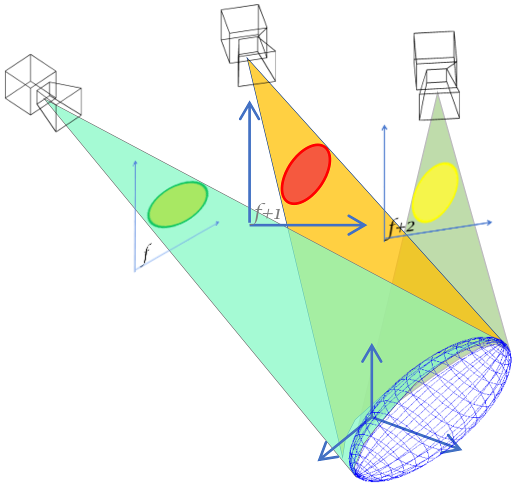
$$\boxed{\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f\mathbf{Q}^*_i\mathbf{P}_f^\top}$$

Vectorize dual conics and quadrics and stack over *f*



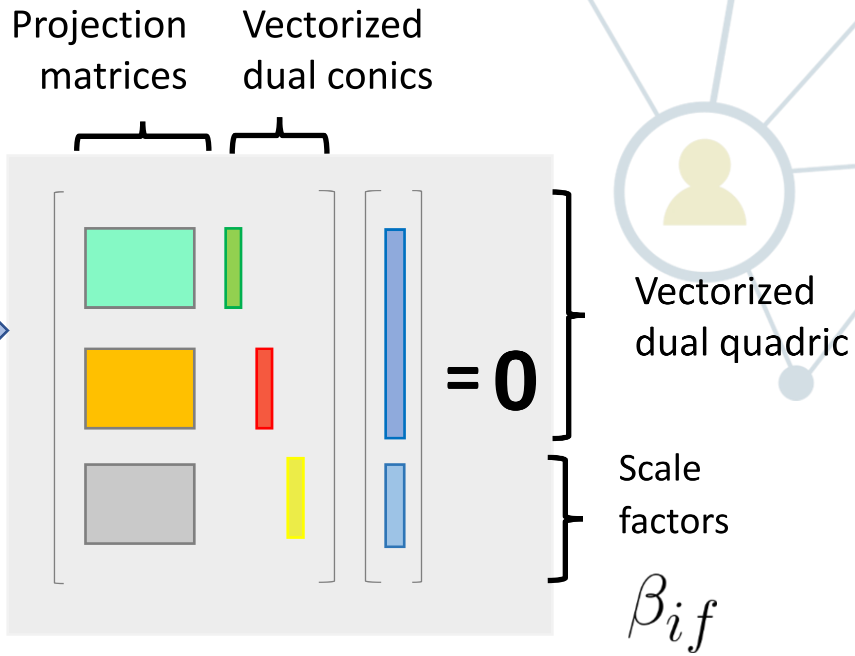
$$\mathbf{M}_i\mathbf{w}_i = \mathbf{0}_{6F}$$

Multiple conic to quadric reconstruction (3)



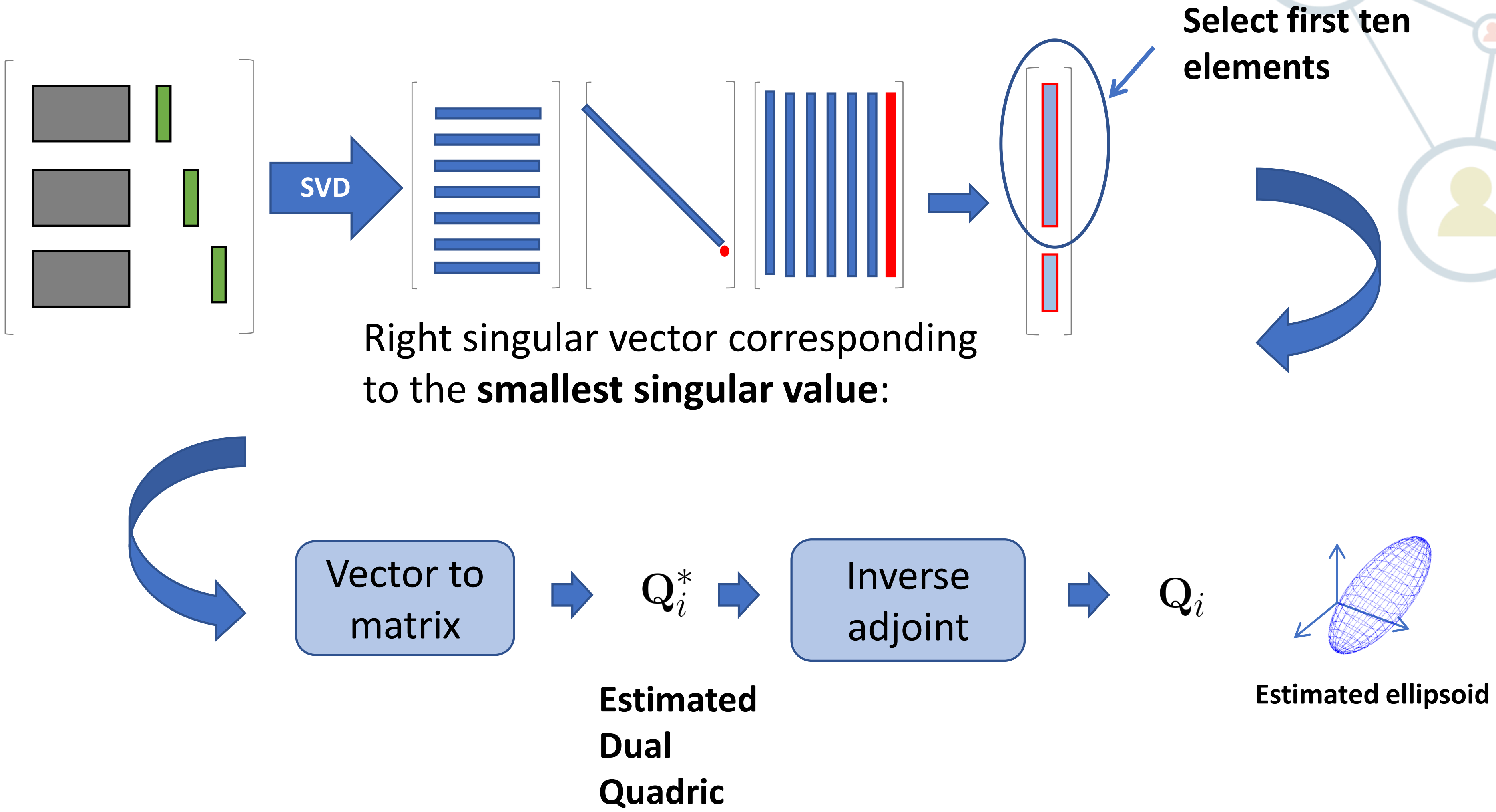
$$\boxed{\beta_{if}\mathbf{C}^*_{if} = \mathbf{P}_f\mathbf{Q}^*_i\mathbf{P}_f^\top}$$

Vectorize dual conics and quadrics and stack over *f*



$$\mathbf{M}_i\mathbf{w}_i = \mathbf{0}_{6F}$$

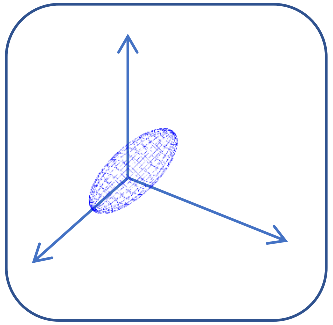
Reconstruction via SVD

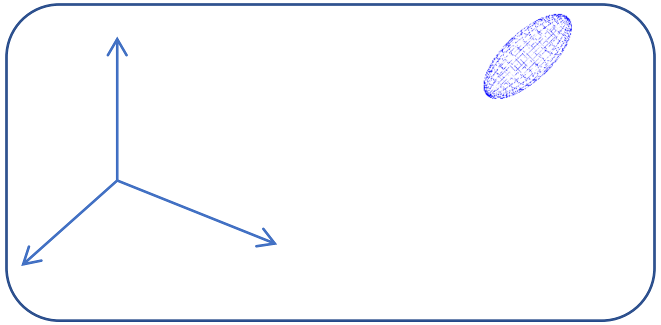


SVD approach: drawbacks

Problem solved in **dual space** may give **ill-conditioned** matrices: reverting to primal space by **matrix inversion** may result in (close to) **degenerate** ellipsoid or even other quadrics (i.e. hyperboloids).

Ill-conditioning is increased for objects **far** from the 3D coordinate origin, where **translation** terms are **dominant** in the quadric.



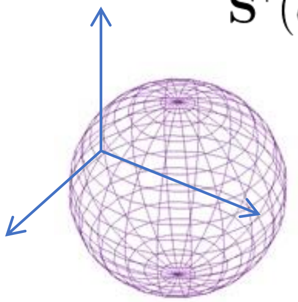


Rcond= 0.216 $Q_i^* = \begin{bmatrix} -0.43 & -0.18 & -0.26 & 0.0 \\ -0.18 & -0.43 & -0.26 & 0.0 \\ -0.26 & -0.26 & -0.62 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.25 \end{bmatrix}$

Rcond= 0.000037 $Q_i^* = \begin{bmatrix} 24.5 & 24.8 & 24.7 & 2.5 \\ 24.8 & 24.5 & 24.7 & 2.5 \\ 24.7 & 24.7 & 24.3 & 2.5 \\ 2.5 & 2.5 & 2.5 & 0.25 \end{bmatrix}$

Regularization

<u>Strategy:</u> enforce **prior** on (known) objects **aspect ratio**: penalize the distance of ellipsoids from a **sphere of given radius and center.**



$$\mathbf{S}^*(a, t_1, t_2, t_3) = \begin{bmatrix} 1 & 0 & 0 & t_1 \\ 0 & 1 & 0 & t_2 \\ 0 & 0 & 1 & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a & 0 & 0 & 0 \\ 0 & a & 0 & 0 \\ 0 & 0 & a & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ t_1 & t_2 & t_3 & 1 \end{bmatrix}$$

$$a > 0$$

Arbitrary center

Axes equal

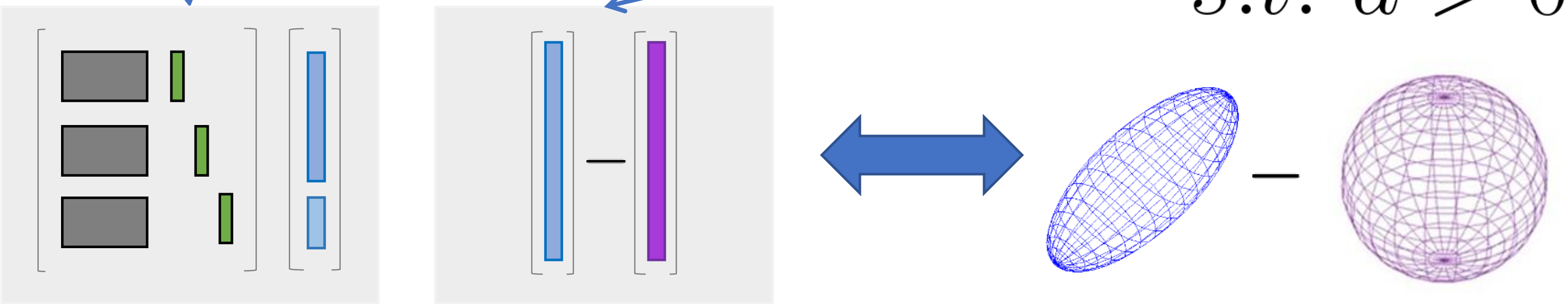
Feasible radius

<u>Sphere in dual space</u>

Regularization: cost function

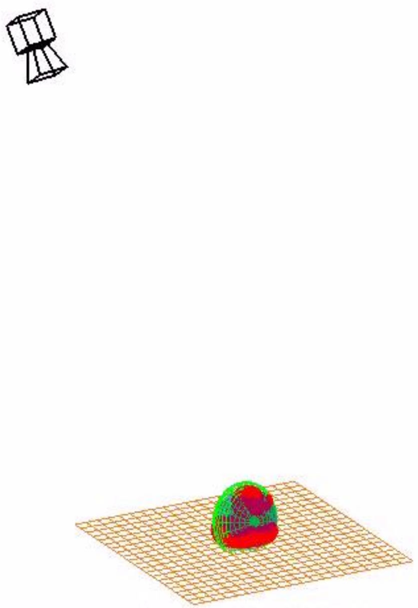
$$\|\mathbf{M}\mathbf{w}_i\|_2^2 + \lambda\|\mathbf{v}_i^* - vech\left(\mathbf{S}^*(a, t_1, t_2, t_3)\right)\|_2^2$$

$$s.t.\ \ a > 0$$

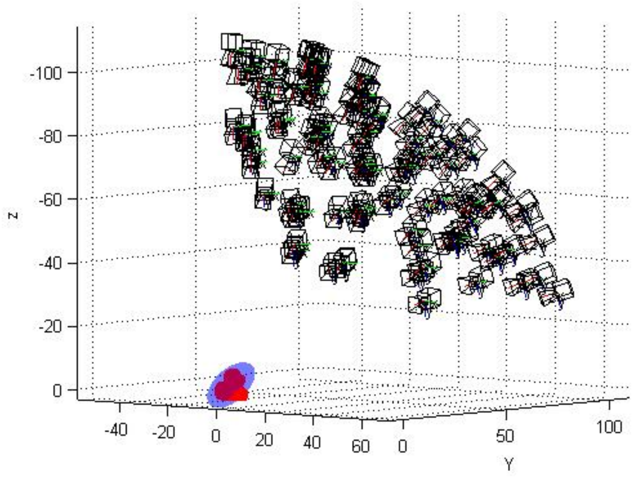


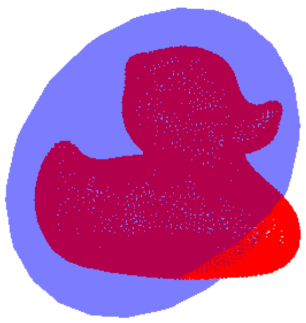
Problem solvable with nonlinear Least Squares with boundary constraints

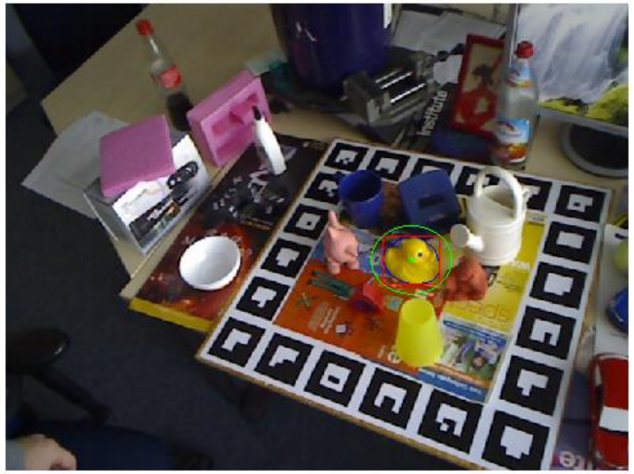
ACCV sequence (Hinterstoißer dataset)



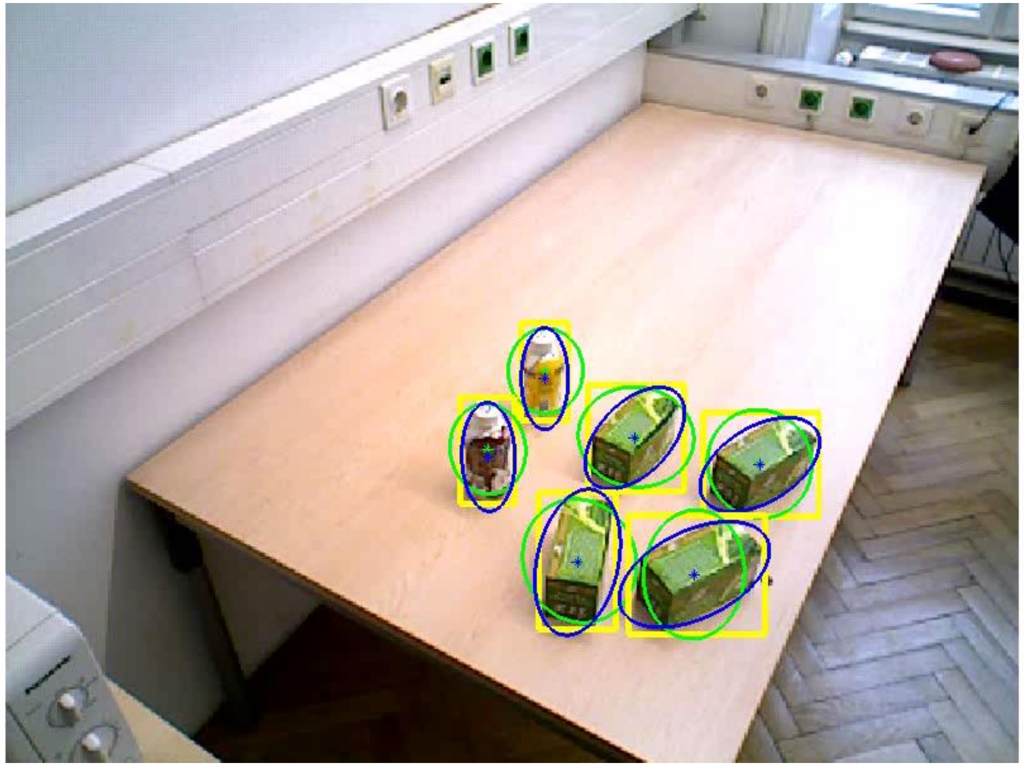
ACCV sequence (Hinterstoißer dataset)

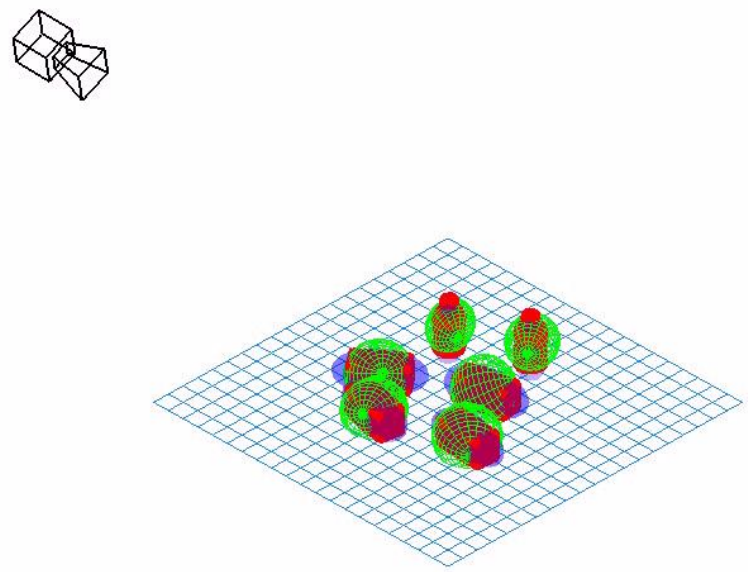




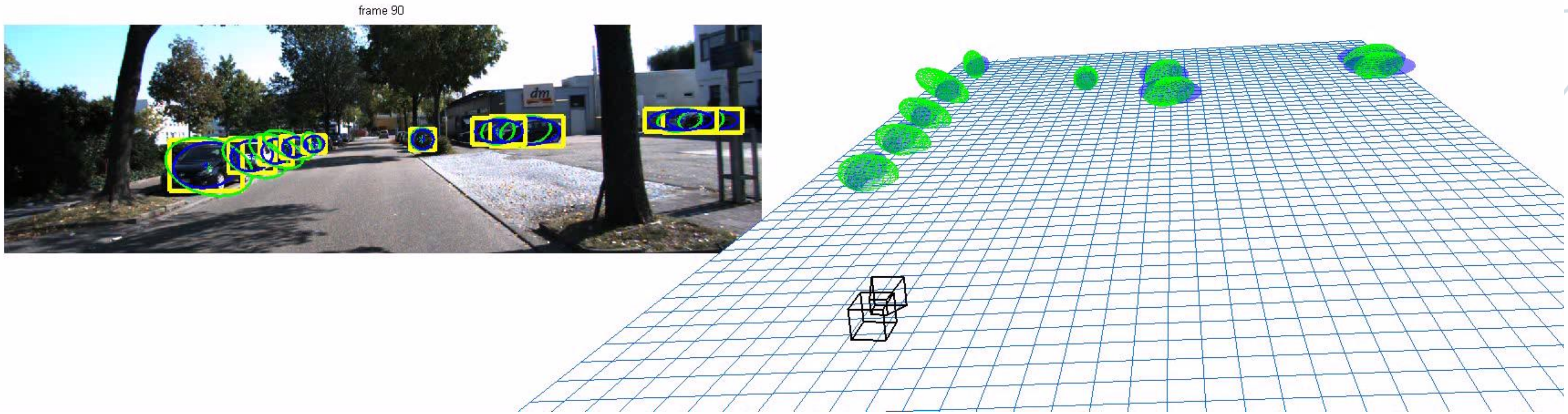


TUW dataset sequence





KITTI sequence



Problem 1: Ellipse from BB mismatch

- Bounding boxes (green) from generic object detectors are not precisely matched with their ground truth -> **ellipses from BB** (red) are not exactly matched as well with their **ground truth** counterpart (blue)

- Also with perfect BBs, ellipse axes are aligned to the image axes -> **high rotation and size error** in respect to GT ellipse if the object is rotated with respect to the image axes

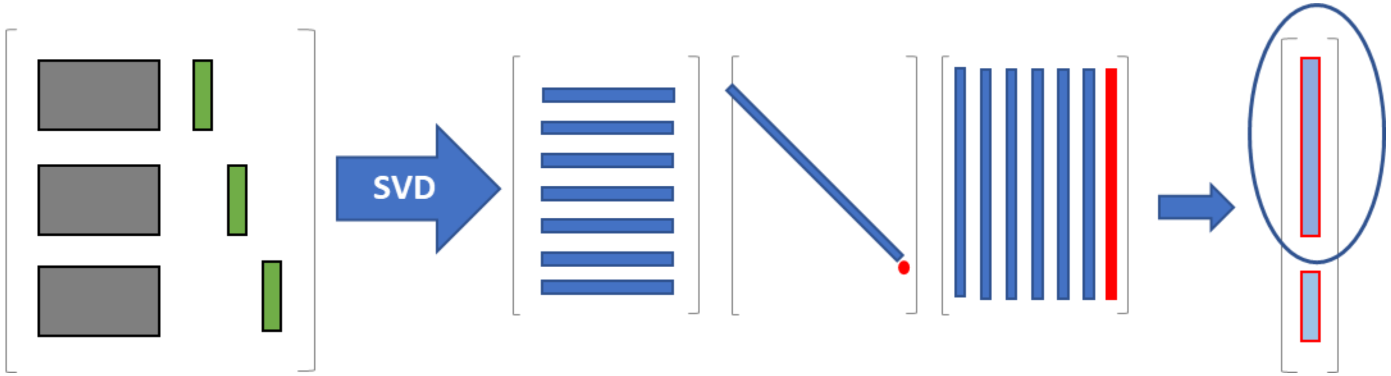
- Further works extend ellipses fitting to objects or use instance segmentation.

Example of mismatch between ground truth ellipse (blue) and ellipse from bounding box (red).

Dong W, Roy P, Peng C, Isler V. Ellipse R-CNN: Learning to infer elliptical object from clustering and occlusion. IEEE TIP 2021.

Problem 2: Analytical solution

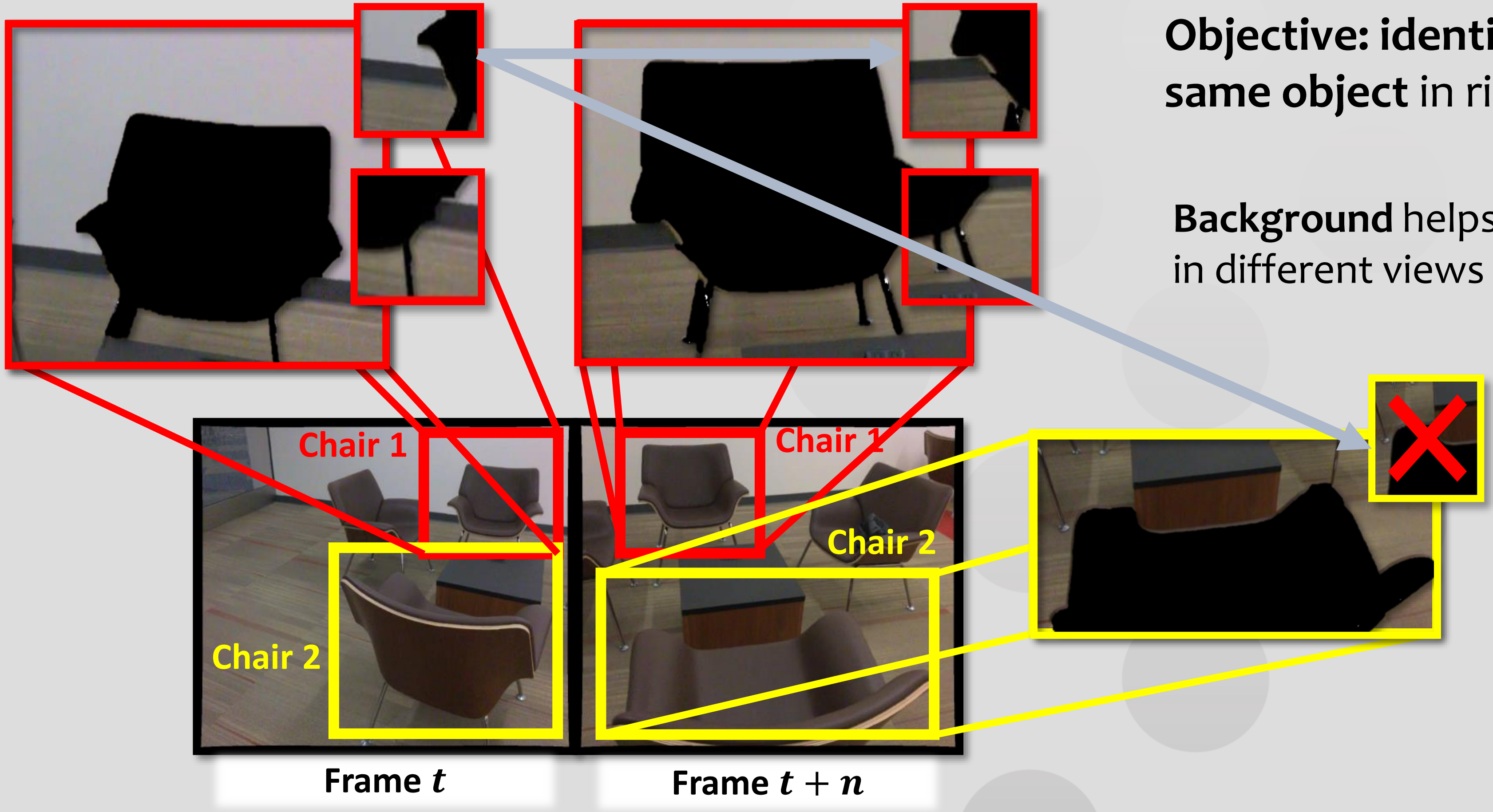
The solution computes the quadric from a collection of ellipses:



Ideally, as it happens in SfM we achieve better results and we avoid degenerate solution by minimising a reprojection error.

However: $O_{2D} = \frac{1}{J}\sum_{f=1}^{F}\sum_{i=1}^{N} I(i,f)\frac{\mathcal{C}_{if} \cap \tilde{\mathcal{C}}_{if}}{\mathcal{C}_{if} \cup \tilde{\mathcal{C}}_{if}}$ ☹

Problem 3: the same object everywhere!



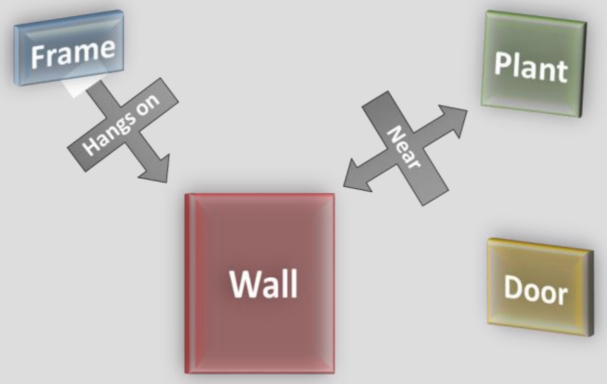
Objective: identifying multiple instances of the **same object** in rigid scenes

Background helps in identifying the same instance in different views

V. Bansal, S. James, and A. Del Bue, re-OBJ: Jointly learning the foreground and background for object instance re-identification. In *Proceedings of the International Conference on image analysis and processing,* 2019.

3D scene understanding

INPUT: 2D object detections in multiple views



REPRESENTATION
(RIGID AND DYNAMIC)

- How to compute 3D representations with higher level of representation
- From 2D to 3D scene graphs
- Application scenarios

Back to 2D…

Images can be described as graph encoding objects and their relationships in the scene

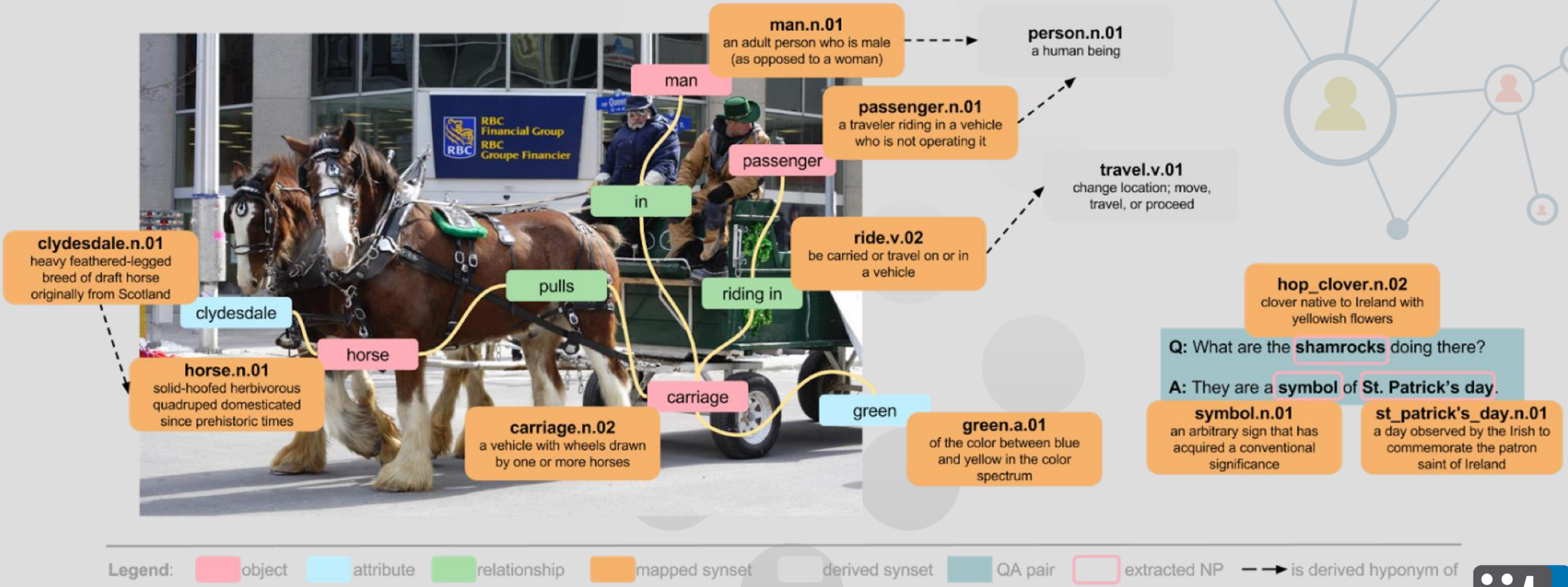
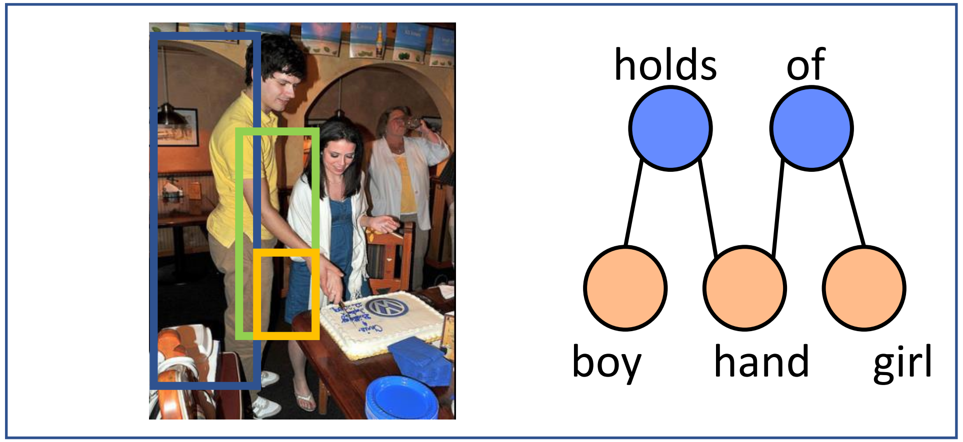


Image from the first workshop on graph based learning in computer vision, held in conjunction with ICCV 2019

Previous work

- Recent architectures generate scene graphs from single images
- Explicit modelling of object and relations with Graph Neural Network (GNN) [1,2].
    - Global optimisation with message passing on a bi-partite graph.
    - Dealing with chain of relations



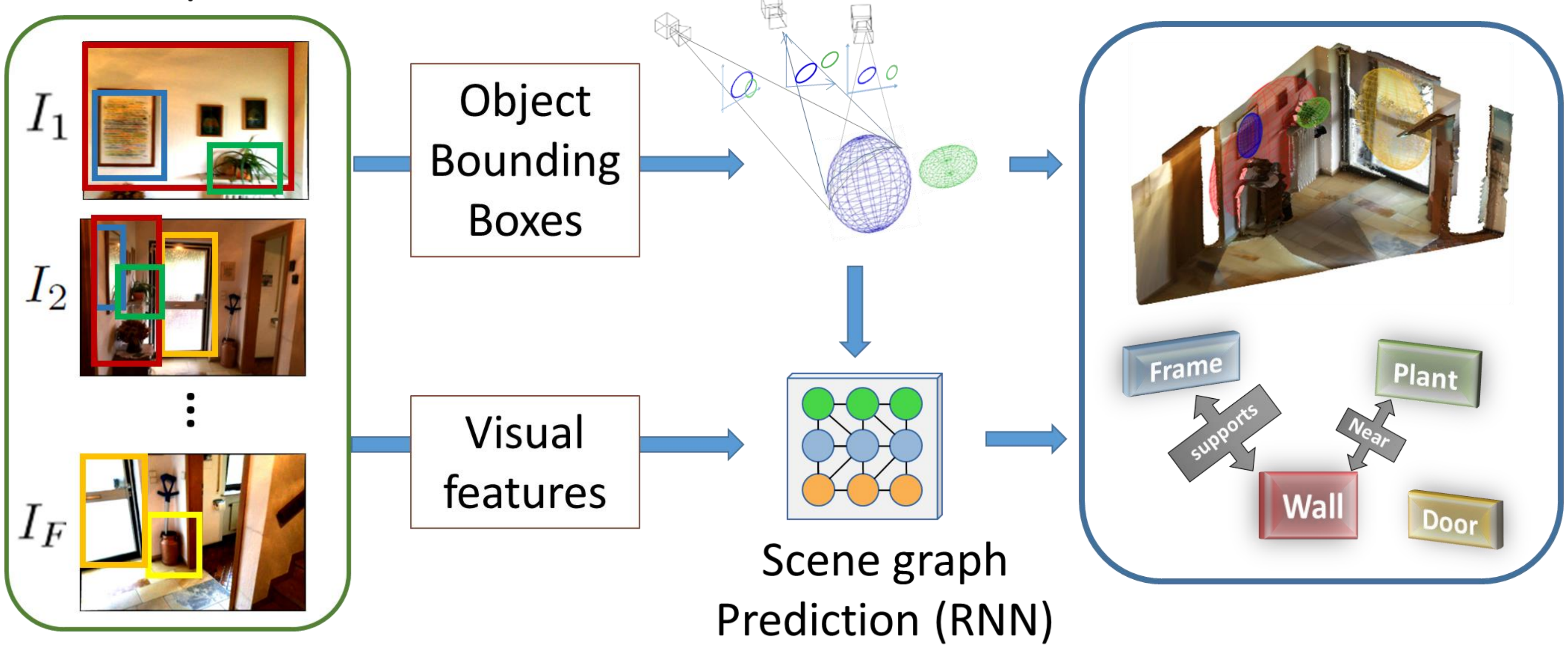
[1] Danfei et al, Scene Graph Generation by Iterative Message Passing, CVPR 2017
[2] Yikang et al, Scene Graph Generation from Objects, Phrases and Region Captions, ICCV 2017

From 2D to 3D scene graphs

Input: object detections in multiple views

3D object position from bounding boxes

Output: 3D Scene Graph

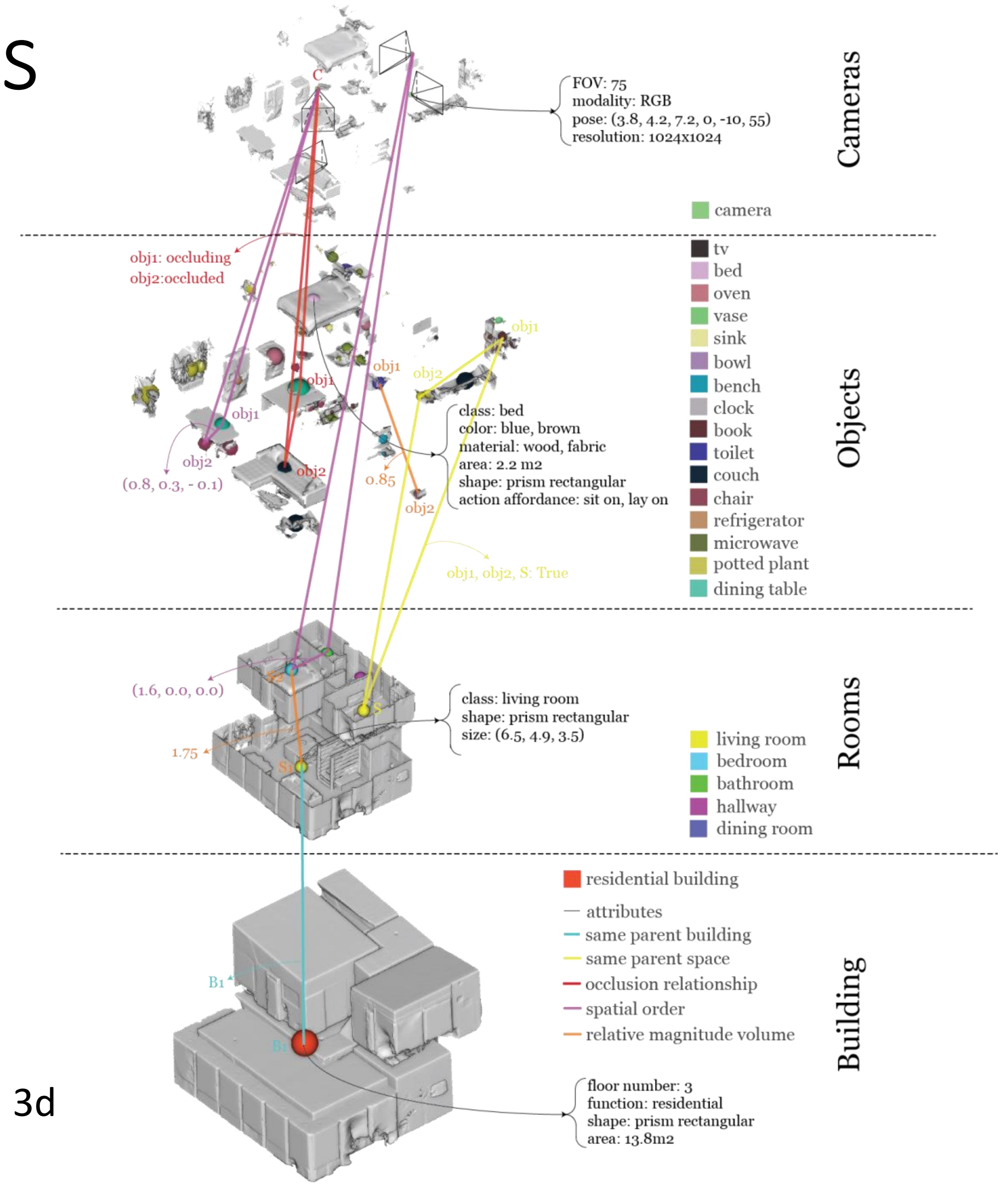


Gay, P., James, S., & Del Bue, A. “Visual Graphs from Motion (VGfM): Scene understanding with object geometry reasoning.” *ACCV 2018.*

From 2D to 3D scene graphs

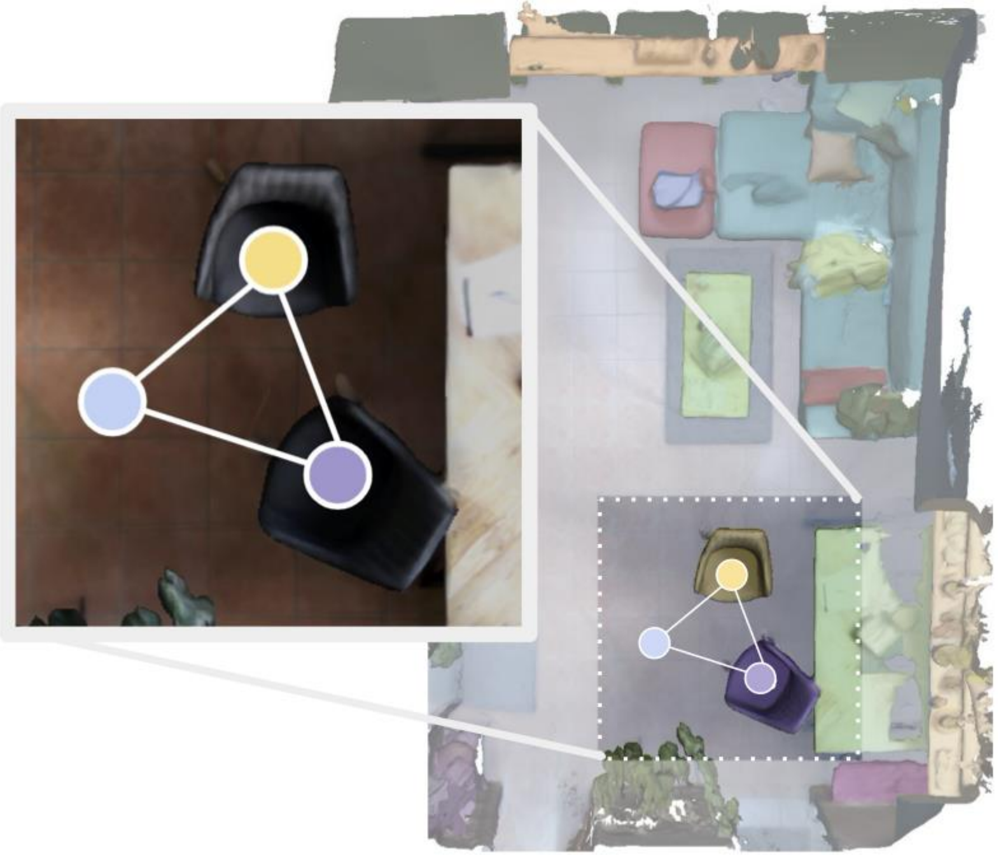
- Scene graphs include semantics on objects (e.g., class, material, and other attributes), rooms (e.g., scene category, volume, etc.) and cameras (e.g., location, etc.)

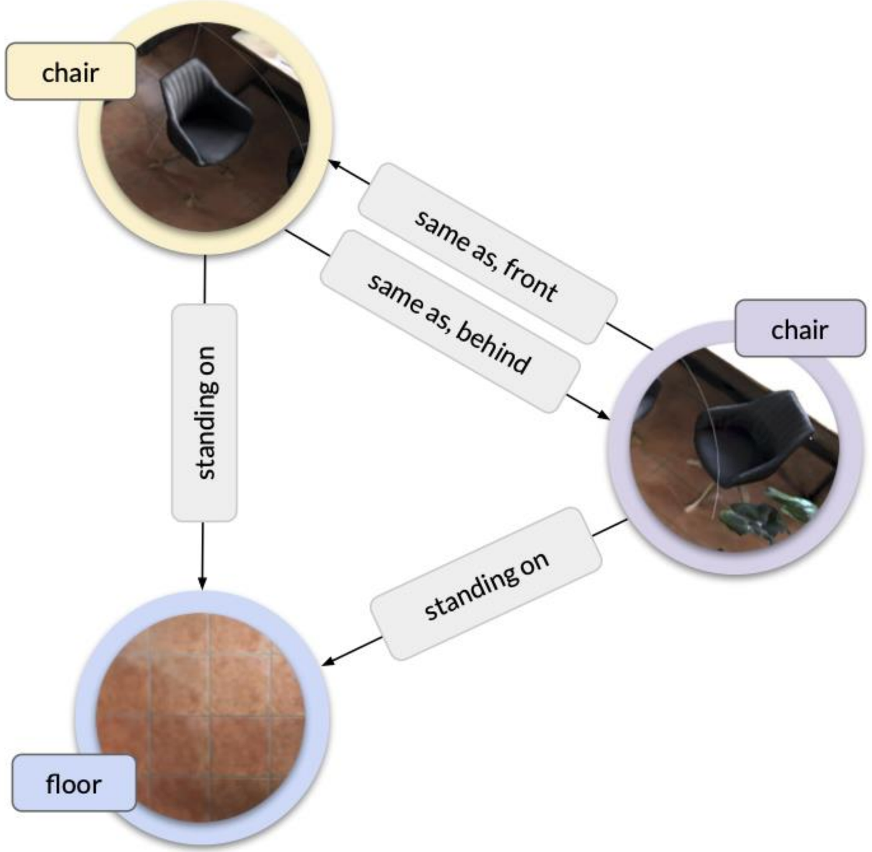
- The representation can have different levels of hierarchy, from a building to a city…



Armeni, Iro, et al. "3d scene graph: A structure for unified semantics, 3d space, and camera." *ICCV* 2019.

Learning graphs from RGBD indoor dataset



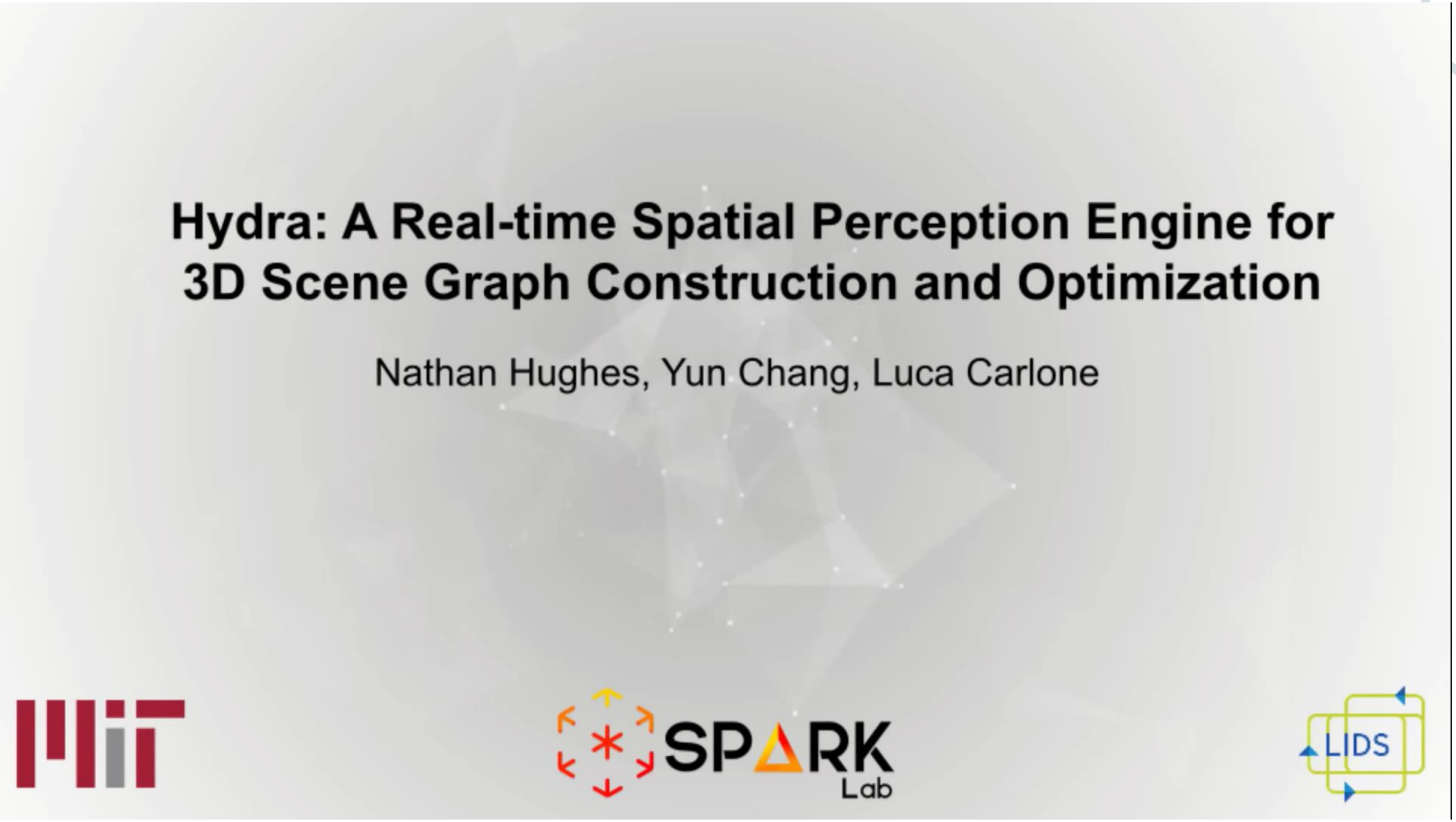


Wald, Johanna, et al. "Learning 3d semantic scene graphs from 3d indoor reconstructions." *CVPR* 2020.

Learning graphs from RGBD indoor dataset

Wald, Johanna, et al. "Learning 3d semantic scene graphs from 3d indoor reconstructions." *CVPR* 2020.

Learning graphs in real-time

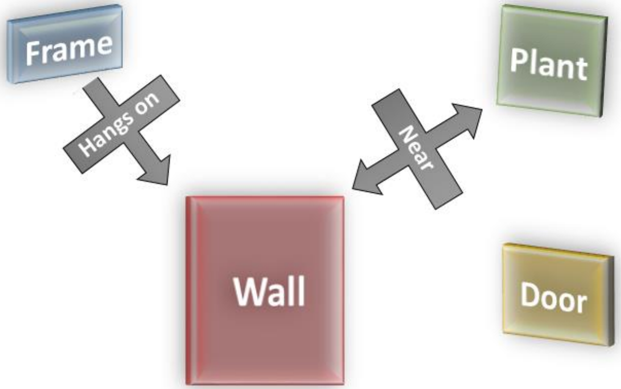


Hughes N, Chang Y, Carlone L. Hydra: A real-time spatial perception system for 3D scene graph construction and optimization. RSS 2022.

3D scene graph from multi-view

Goal: Build **3D Scene graph** to obtain a **rich** and **grounded semantic** information of the scene using only images and object detections.

INPUT: 2D object detections in multiple views



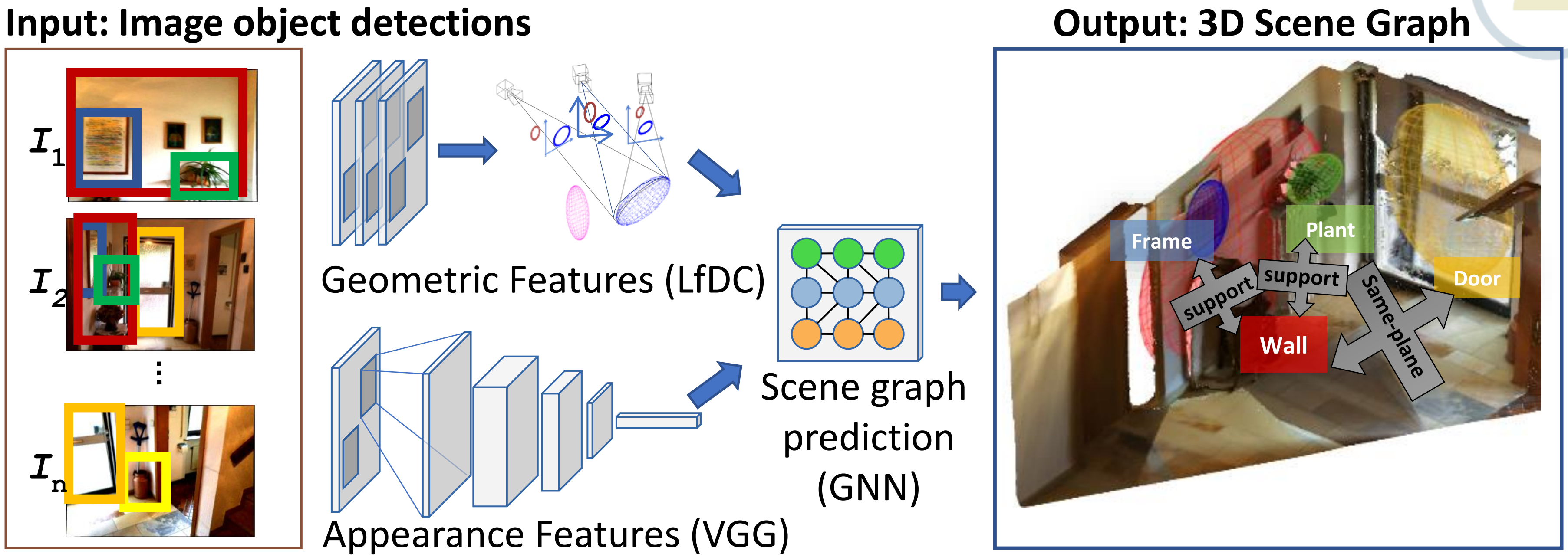
OUTPUT: 3D Scene Graph

OUTPUT: 3D object localisation

Gay, P., James, S., & Del Bue, A. “Visual Graphs from Motion (VGfM): Scene understanding with object geometry reasoning.” *ACCV 2018.*

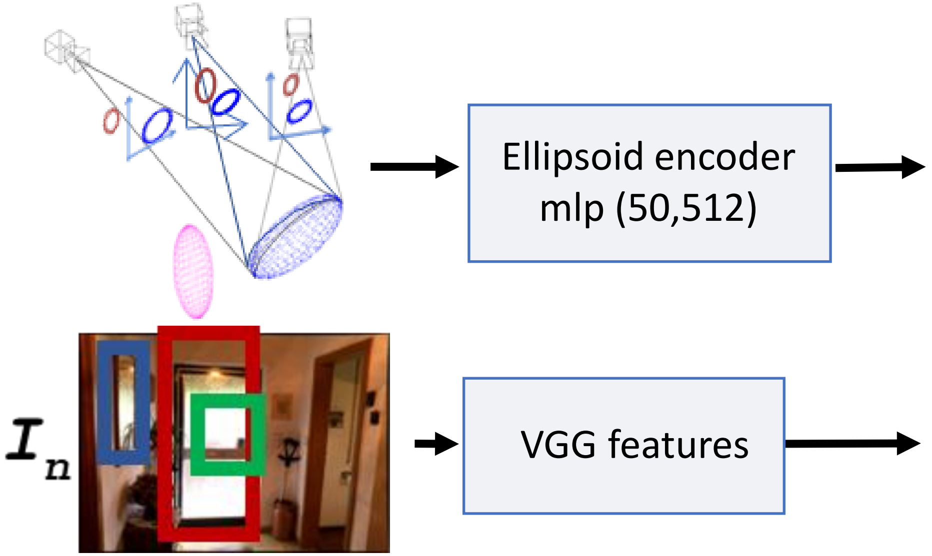
How to build a 3D scene graph?

- Extraction of the observations: Visual and geometric features
  - Ellipsoids computed from multiple views.
  - Visual features extracted from each image.
- Scene graph prediction with a tri-partite Graph Neural Network (GNN)



Use of the tri-graph for scene graph generation

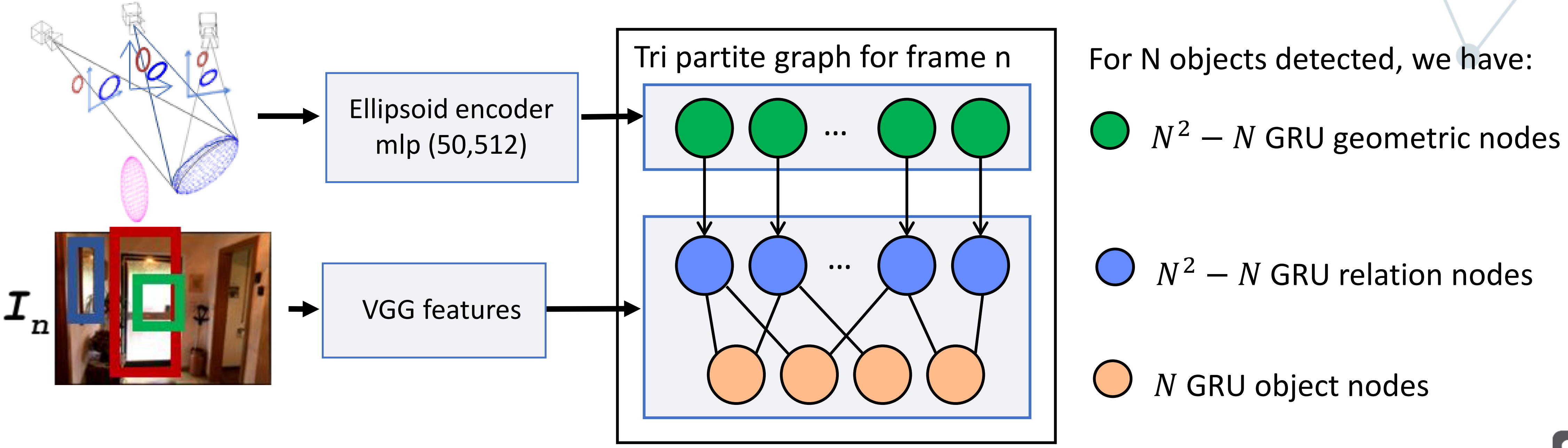
- Visual appearance representation : VGG features are extracted from each bounding box.
- Geometric representation: MLP encodes each pair of ellipsoids.
- These features are then used as initialisation of the node states.



Gay, P., James, S., & Del Bue, A. "Visual Graphs from Motion (VGfM): Scene understanding with object geometry reasoning." *ACCV 2018.*

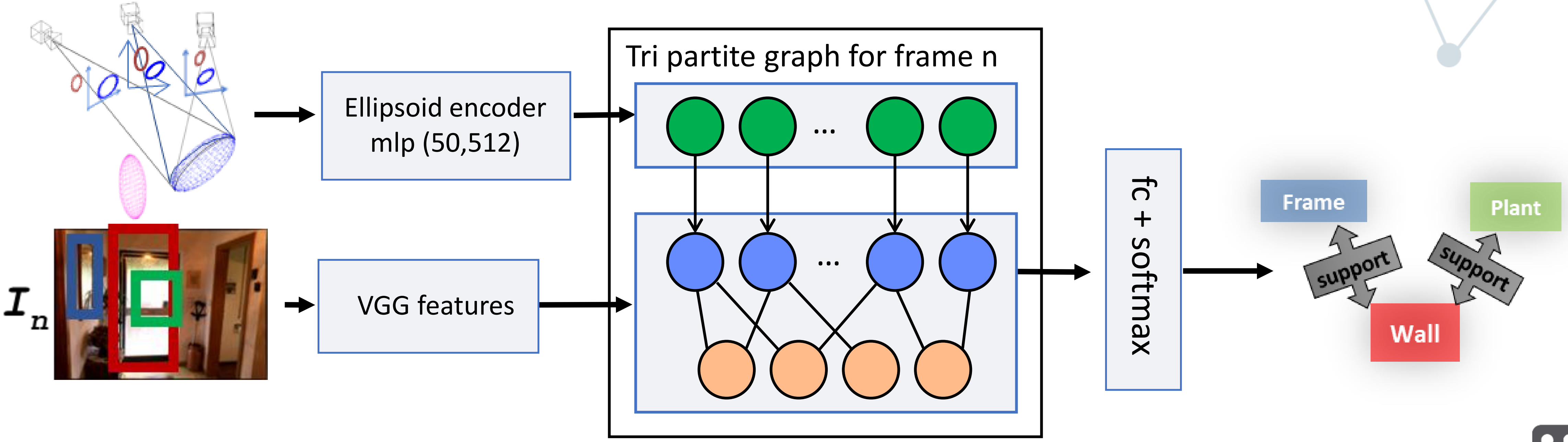
Use of the tri-graph for scene graph generation

- Tri-graph containing every possible relation between the N objects.
- Relation (blue) and object nodes (orange) are RNNs with GRU units:
  - Update of the GRU hidden state thanks to incoming messages;
  - Geometric nodes (green) are observations, i.e. they only send messages.



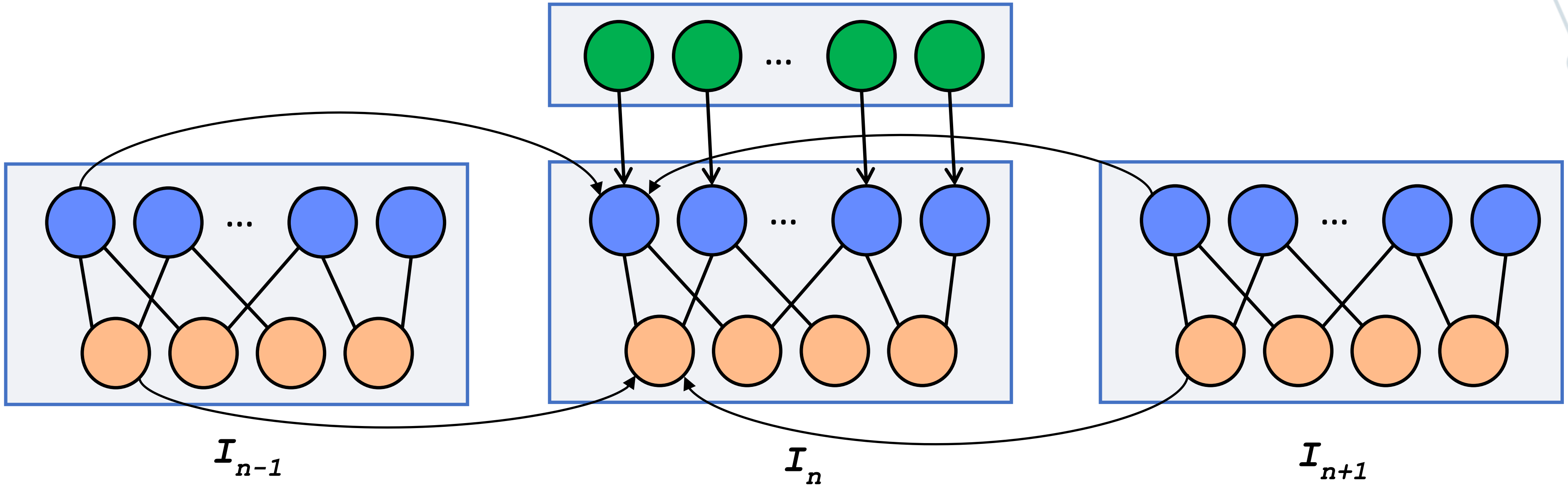
Use of the tri-graph for scene graph generation

- Lastly, a softmax layer gives the graph prediction from the hidden states:
  Probability distribution for each relation and object nodes over its set of labels.
- We now have a scene graph given a single image and geometry:
  But given a video, how to exploit the visual appearance from the other images?



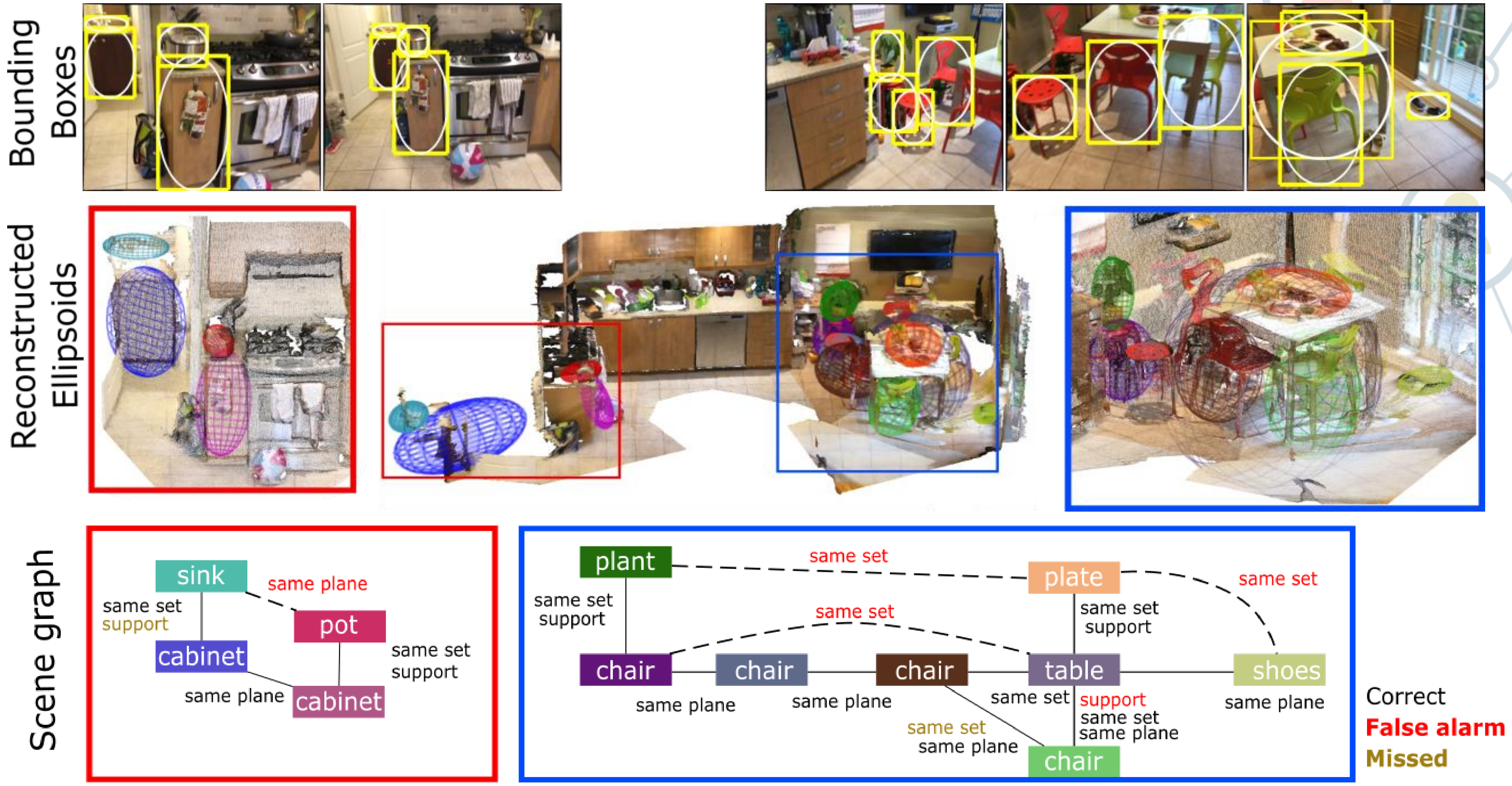
Fusion of the visual features with multiple frames

- Message passing among multiple frames
  - Fusion of the appearances of the object through the video.
- One tri-graph is instantiated in parallel for each image to process.



Gay, P., James, S., & Del Bue, A. "Visual Graphs from Motion (VGfM): Scene understanding with object geometry reasoning." *ACCV 2018.*

Qualitative results on ScanNet



Some final considerations

Considerations:

- Structure from Motion and 3D reconstruction methods are mature, but what use can have a trillion of 3D points?
- Scene graphs are an accessible, human interpretable, way to visualise the information.

Ongoing problems:

- Annotating 3D data is already hard, scaling to annotate pairwise relations between objects is too time consuming -> self-supervision!
- We might build efficient rigid scene graphs, what happens when the scene is dynamic? (check: Rosinol, A., Gupta, A., Abate, M., Shi, J., & Carlone, L. (2020). 3D Dynamic Scene Graphs: Actionable Spatial Perception with Places, Objects, and Humans. RSS 2020.)

Applications

MEMEX

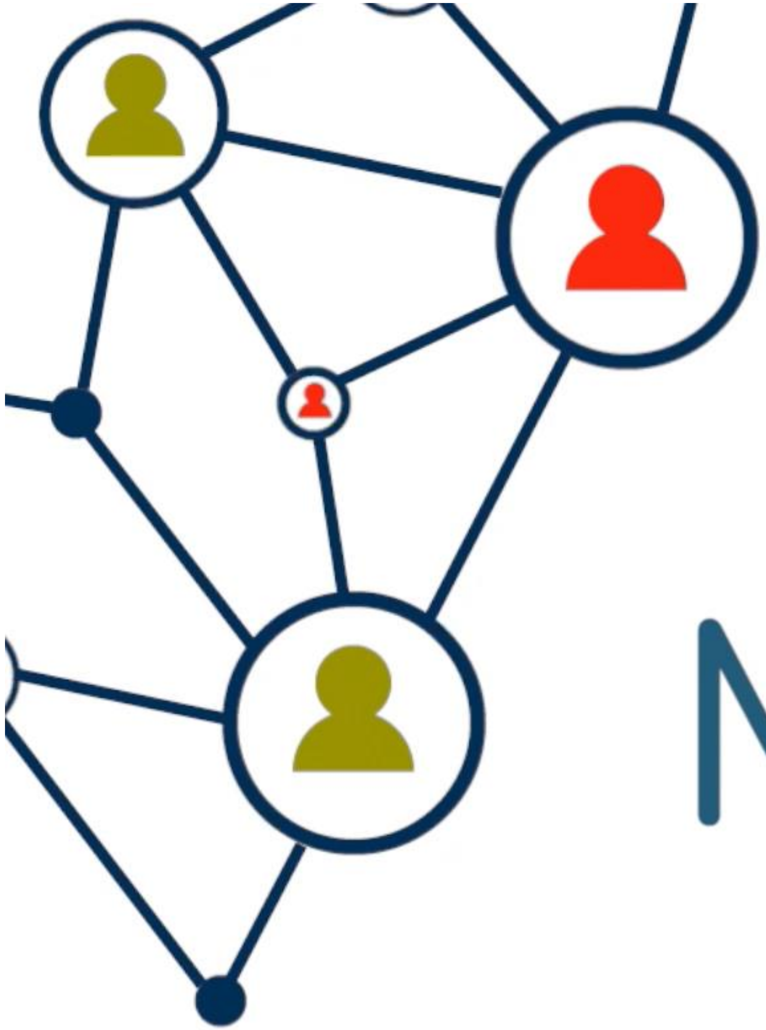
Linking digital information to 3D and 2D maps

Matteo Taiana, Matteo Toso, Stuart James, Alessio Del Bue
Istituto Italiano di Tecnologia (IIT)

2019-2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870743

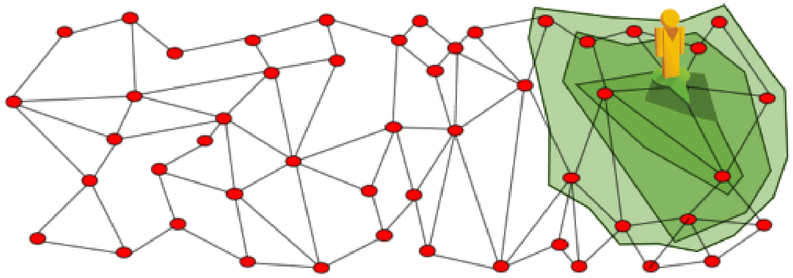

TO LET
Cultural Heritage in Augmented Reality
Phil Lynott Statue
Paper ID: 1116



MEMEX

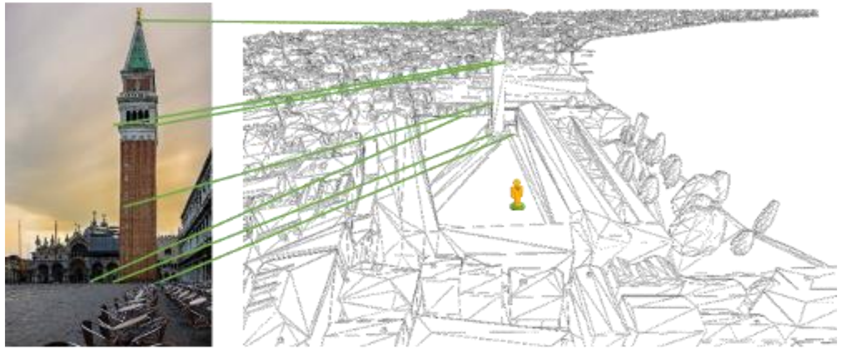
Artificial Intelligence enhances Technology

Researching and developing new technologies to facilitate the increase collaboration and inclusion of communities. MEMEX focuses on three core reusable technologies:



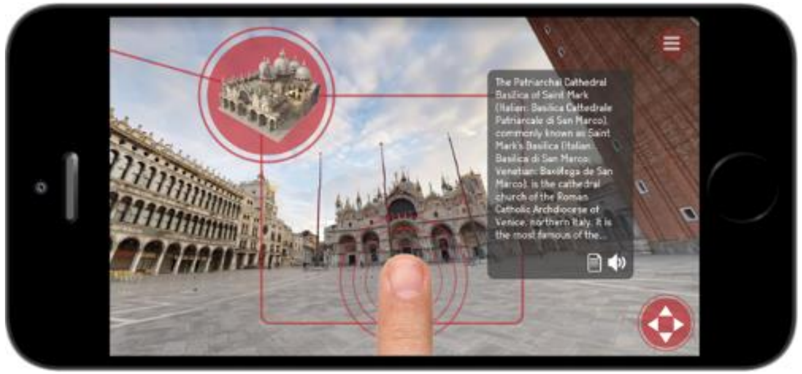
Knowledge Graph

Creating new infrastructure for geolocalised Cultural Heritage to reason on.



Localisation

Computer vision based automatic localisation of users and objects.



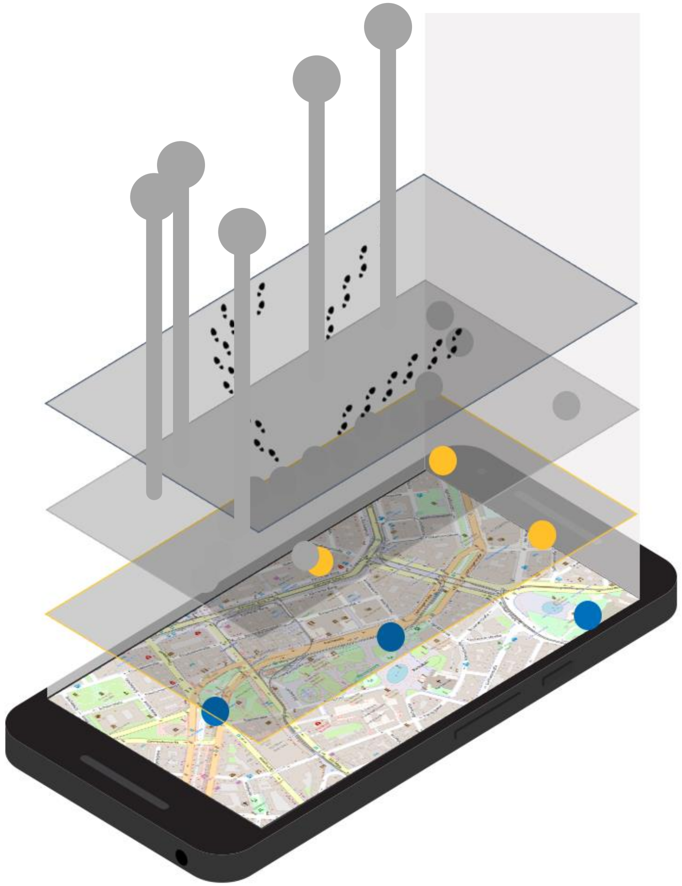
Storytelling through Augmented Reality

Assisted story creation and visualisation using advanced AR technologies.

MEMEX

Combining information from multiple sources

- Open maps
- Digital GLAM's content
- Street level images
- Users stories and objects

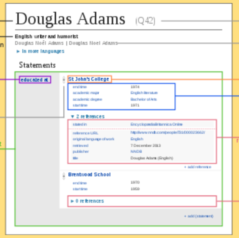


Map registration problem

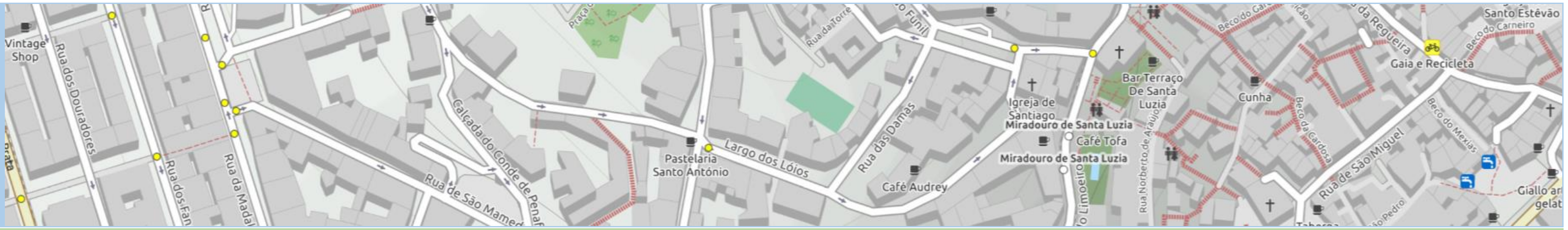
MEMEX

Localised data sources

WIKIDATA



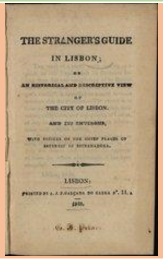
OpenStreetMap

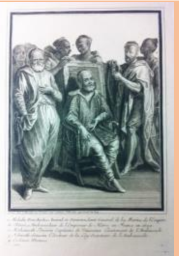


Mapillary

Databases:
Local archives and Museums,
+

europeana
think culture




MEMEX

Localised data sources

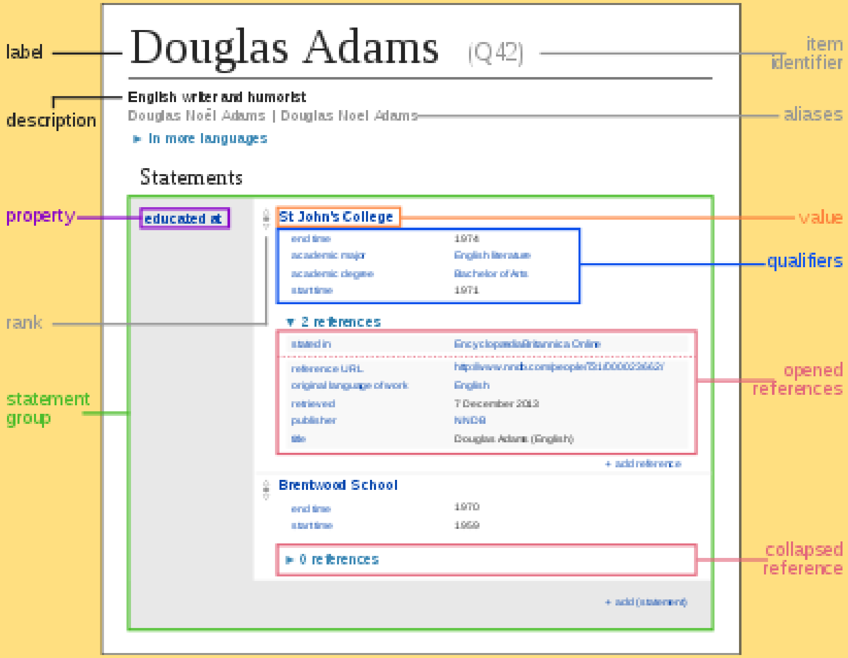
WIKIDATA

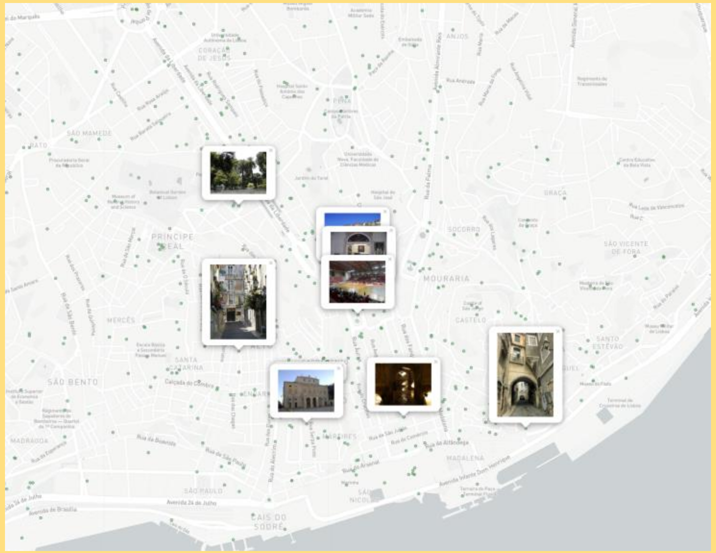
OpenStreetMap

Mapillary

Databases:
Local archives and Museums,
+

europeana
think culture





MEMEX

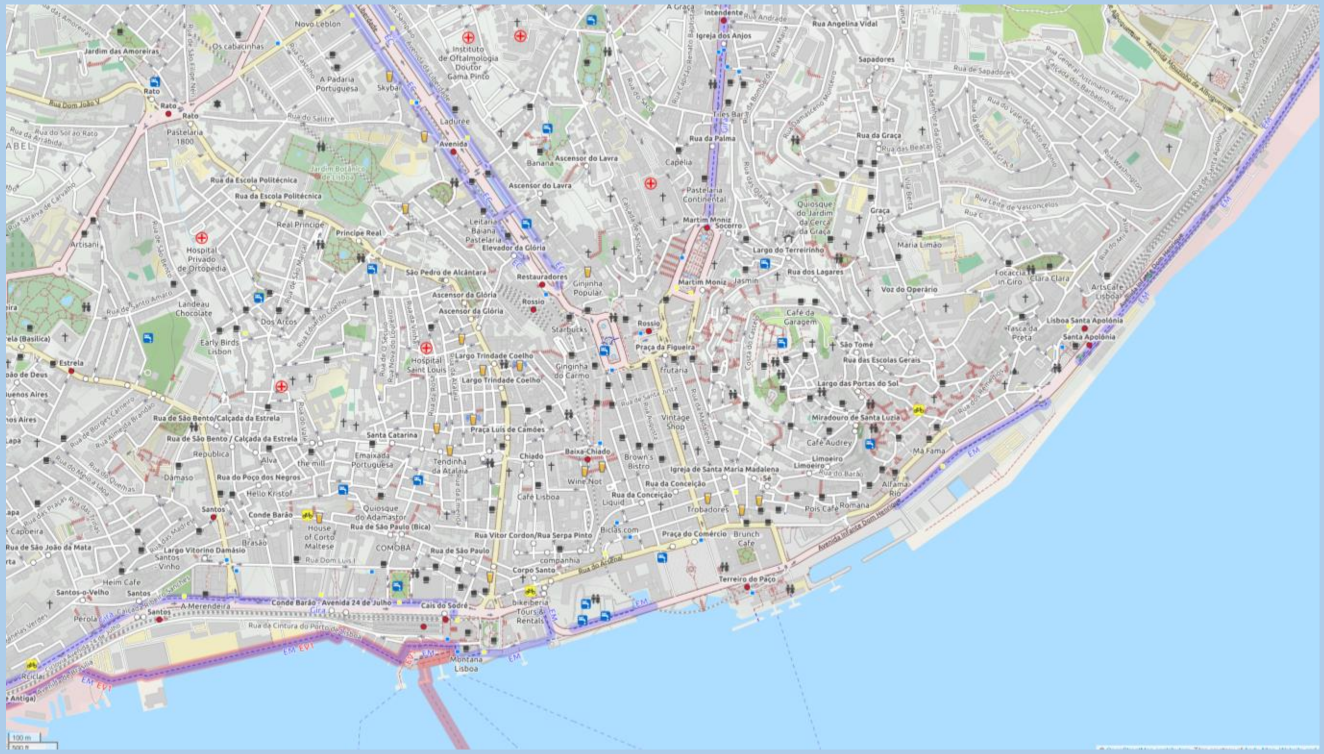
Localised data sources

WIKIDATA

OpenStreetMap

Mapillary

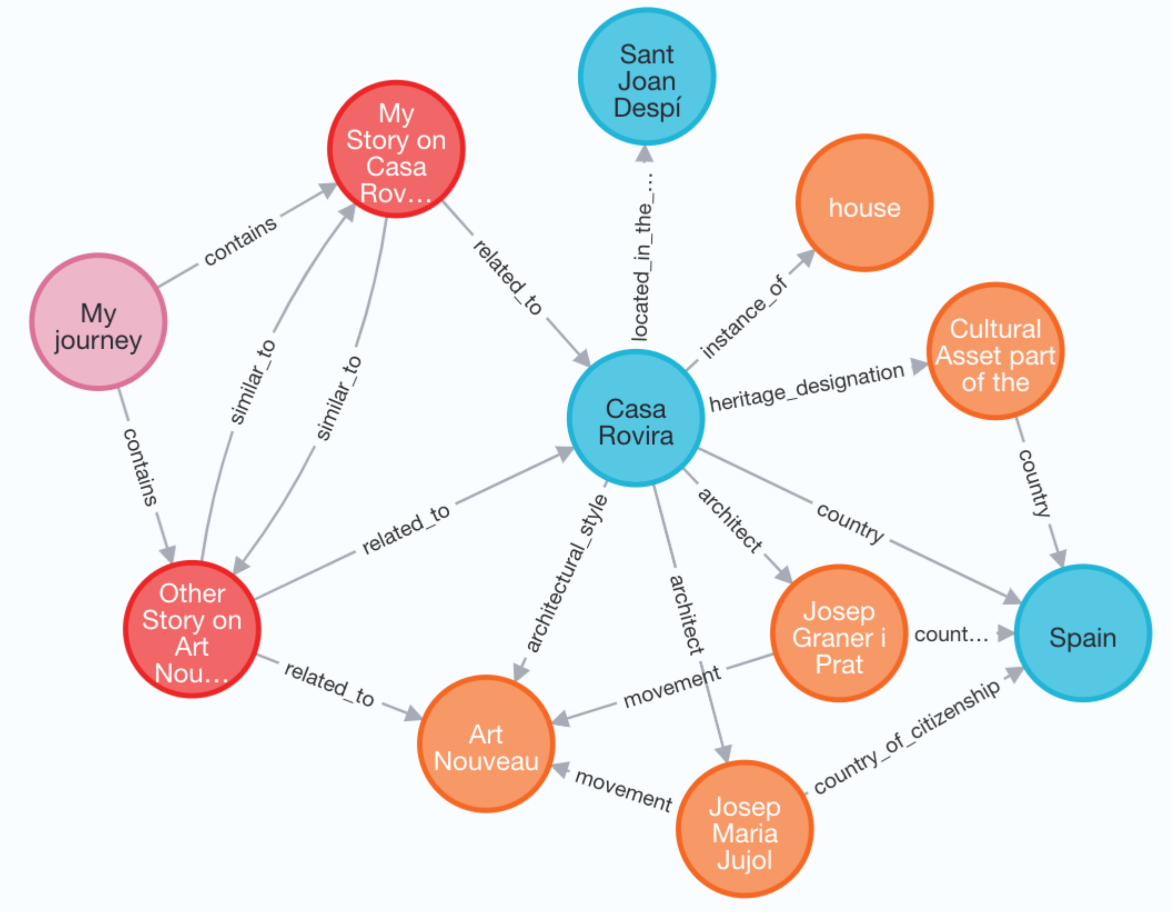
Databases:
Local archives and Museums,
+
europeana
think culture



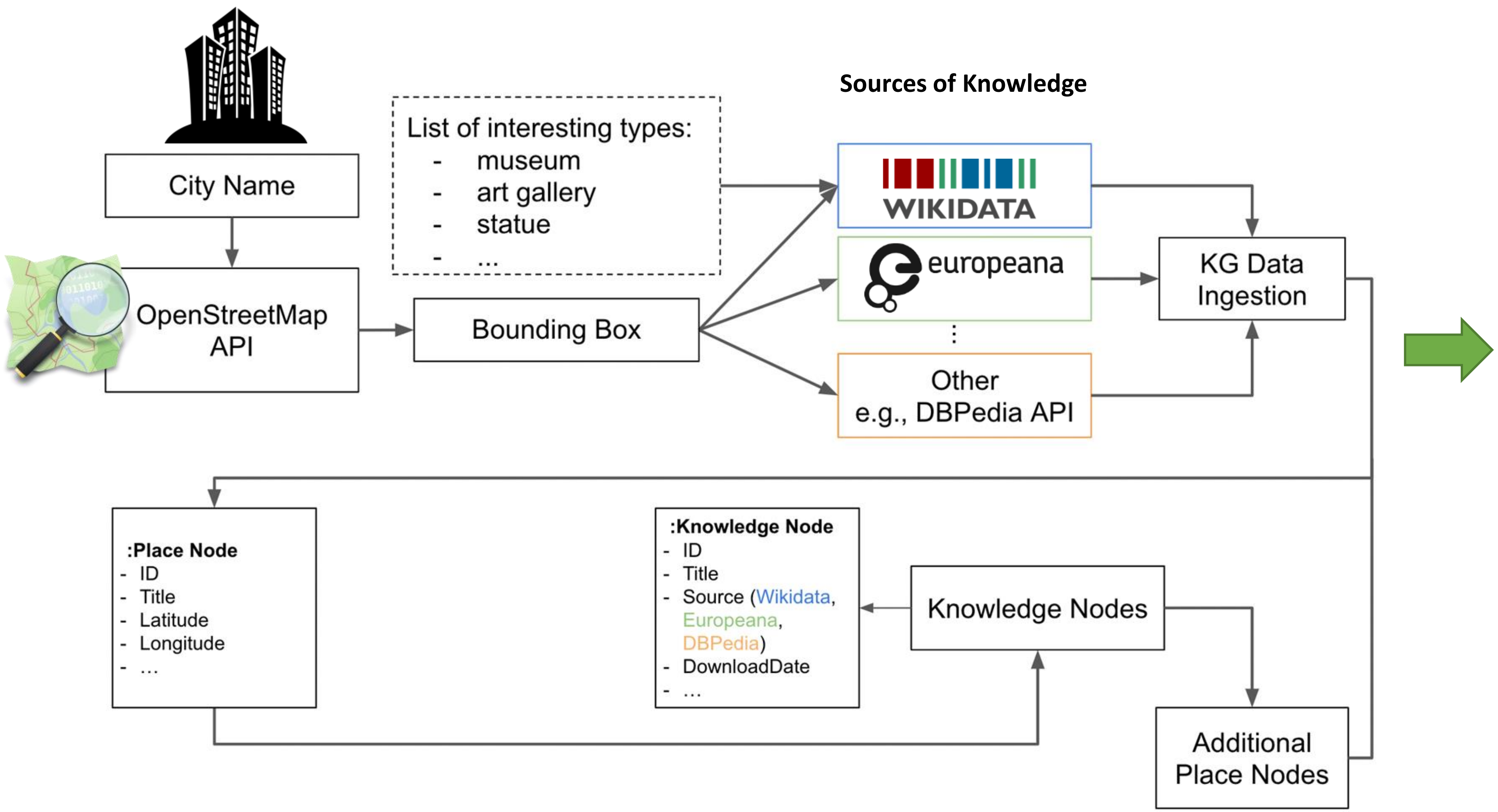
MEMEX

Knowledge Graph construction

The MEMEX-KG Structure:



Knowledge Graph construction



Lisbon

10,209 entities

18,860 relations

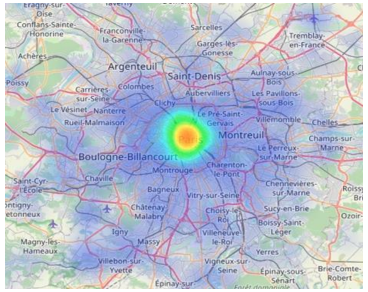
326 relation types

Barcelona

16,832 entities

43,934 relations

394 relation types



Paris

62,211 entities

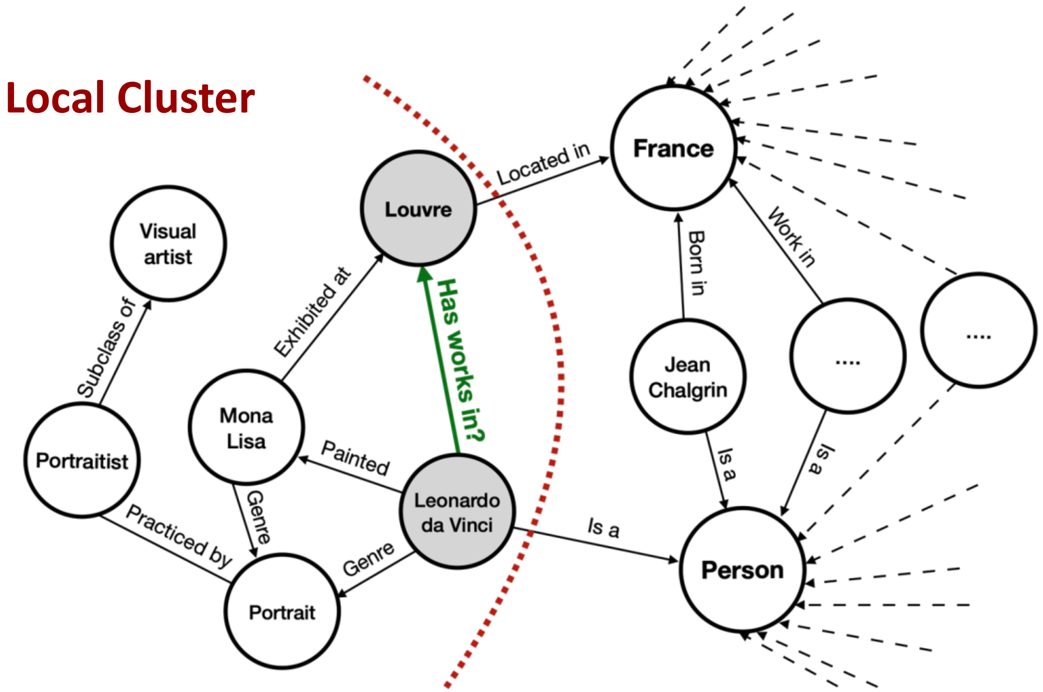
192,036 relations

562 relation types

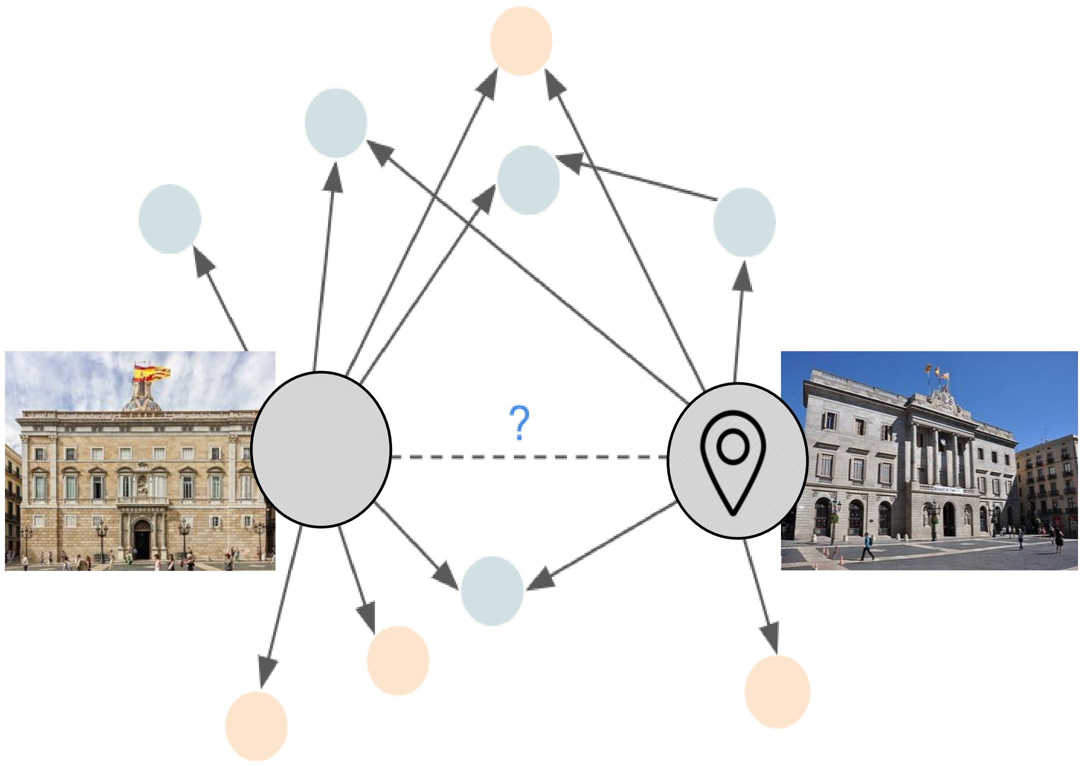
Knowledge and connecting knowledge

KGs suffer from **incompleteness** - especially when ingesting data from heterogeneous sources of information

Link Prediction:
inferring missing links between nodes in the KG[1]



Geolocalisation of CH using Graph Embeddings[2]

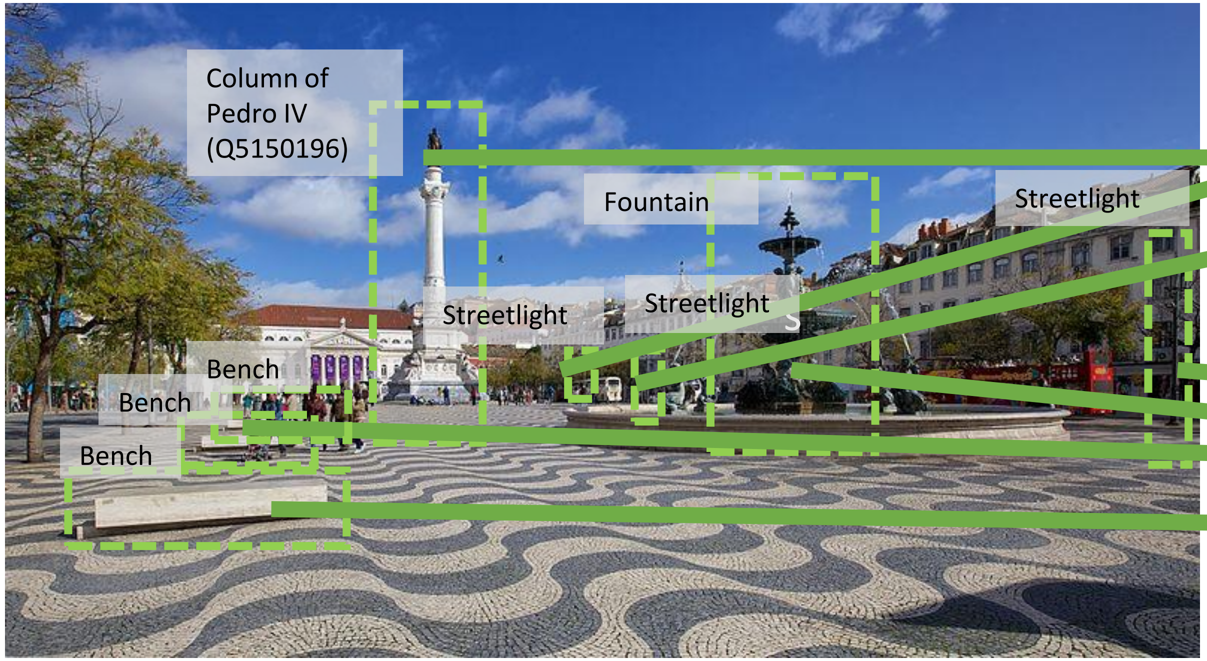


1. H. M. et al. "Locality-aware subgraphs for inductive link prediction in knowledge graphs" - Submitted to PRL2022.

2. H. M. et al. "Geolocation of Cultural Heritage using Multi-View Knowledge Graph Embedding PatReCH 2022

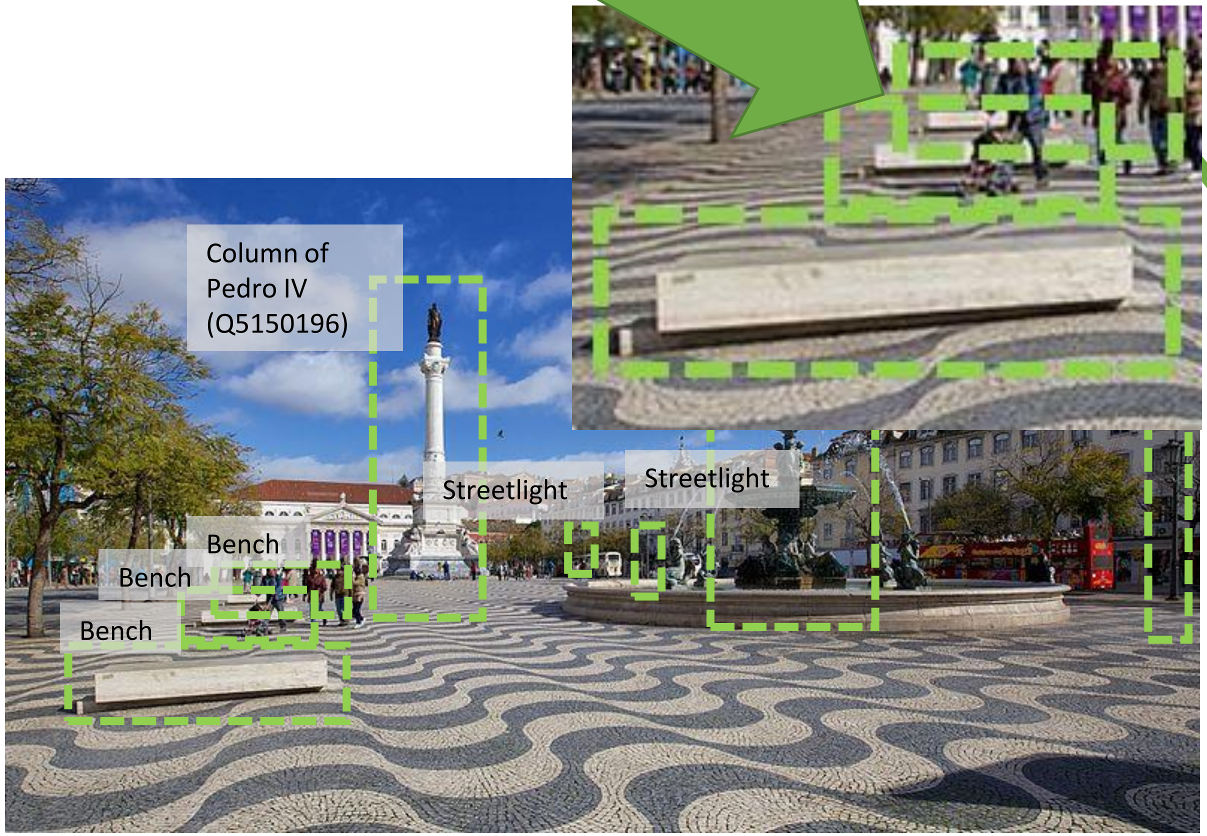
Why do we care about benches?

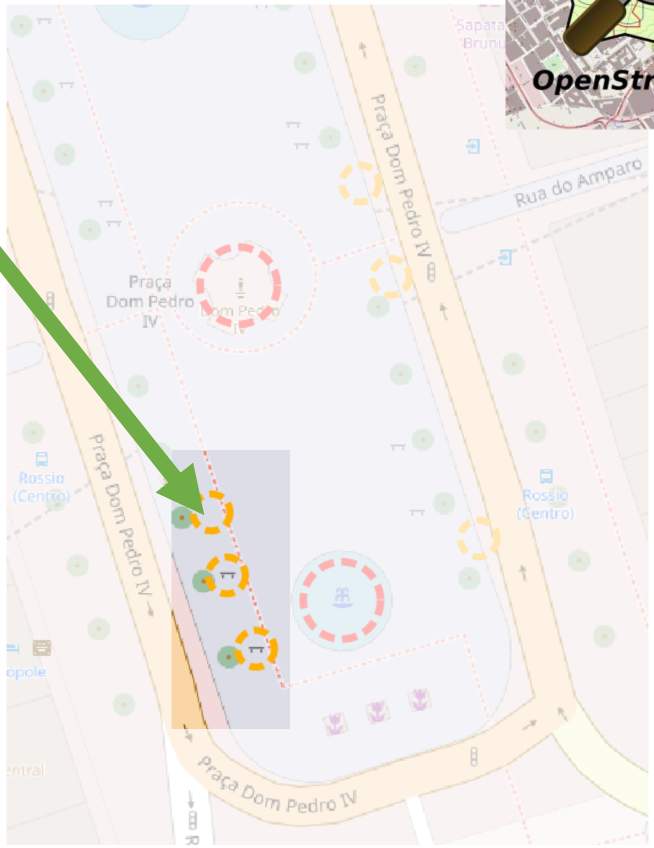
Target scenario



MEMEX

Why do we care about benches?

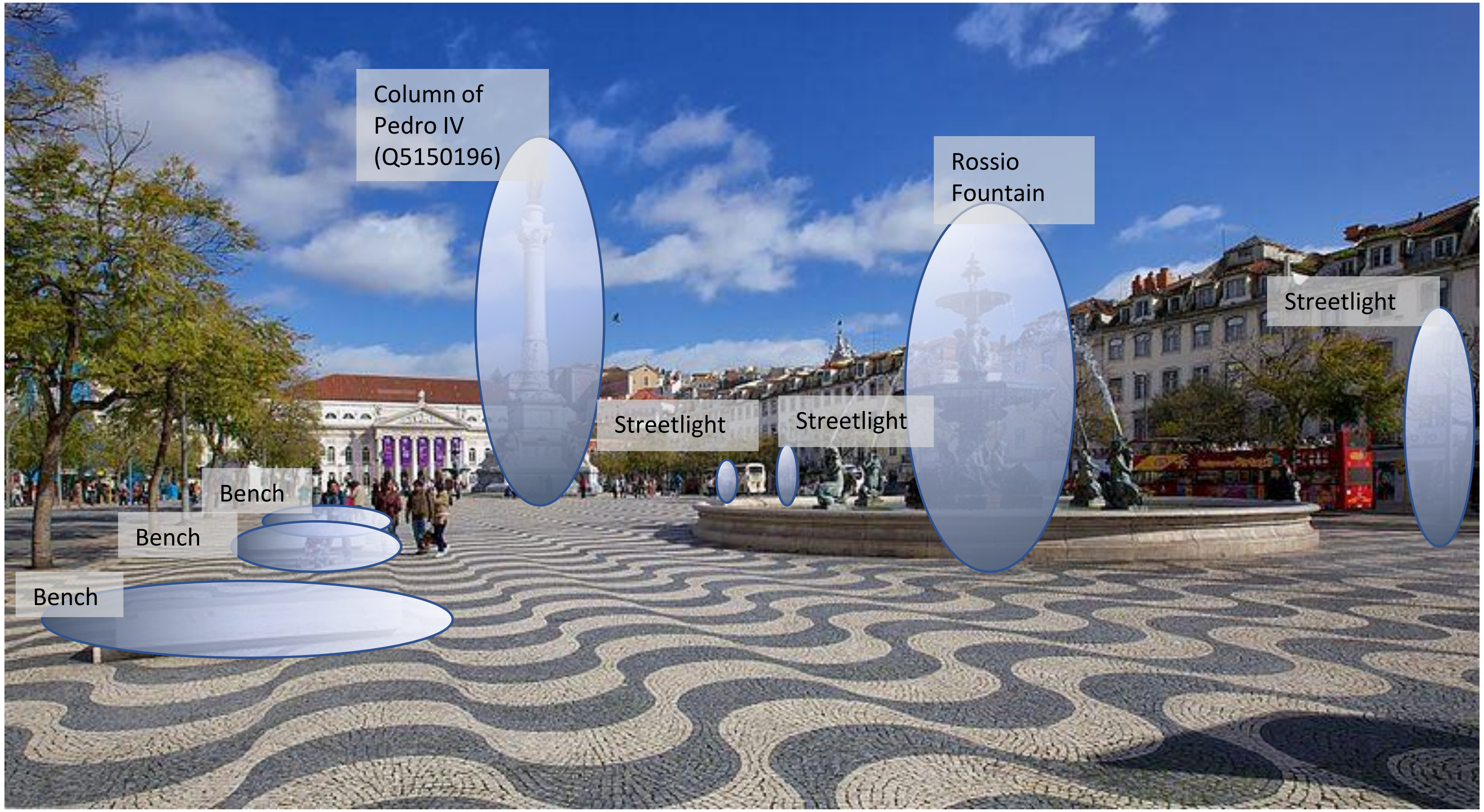




MEMEX

Where does this take us?

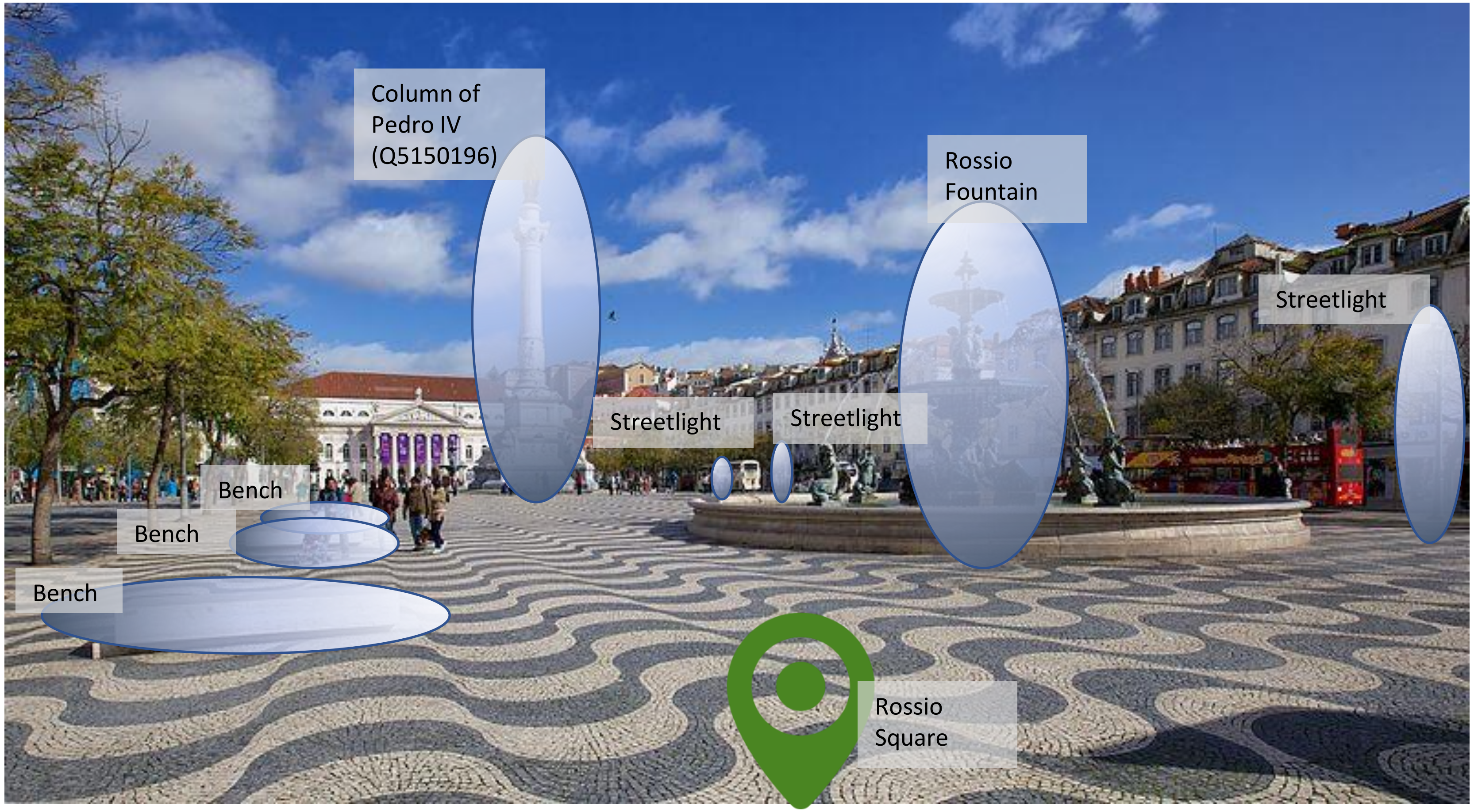
Self-localization using common landmarks!



MEMEX

Where does this take us?

Self-localization using common landmarks!

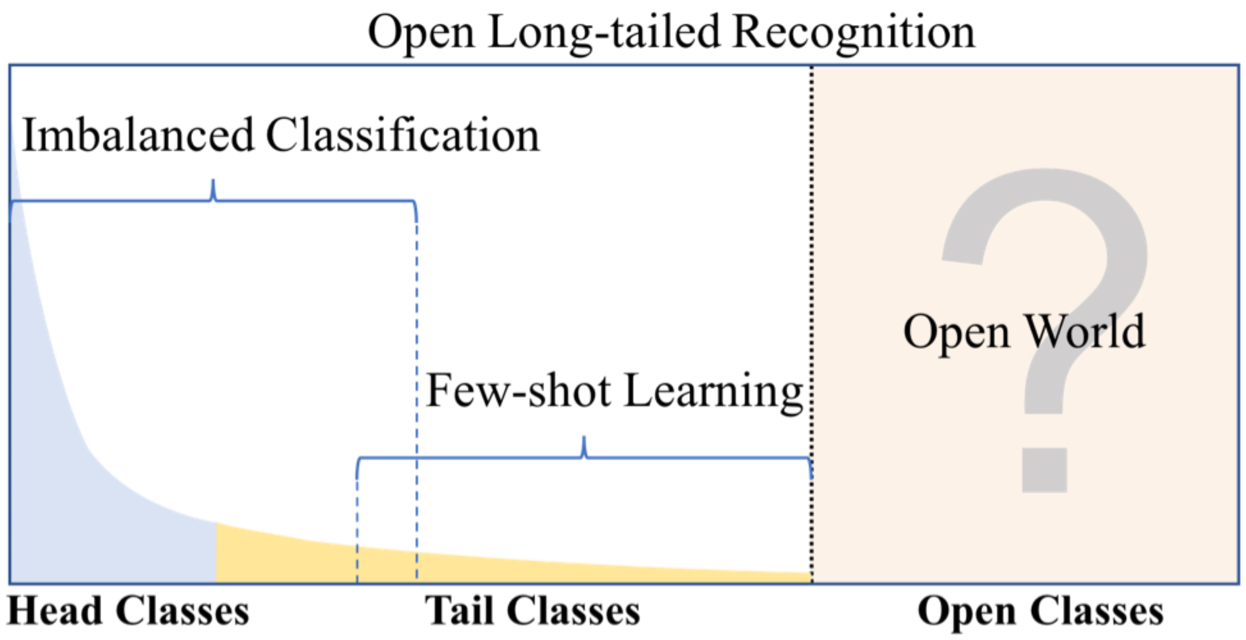


MEMEX

How do we get to Annotations?

MEMEX

How do we get to Annotations?



Liu et al. “Large-Scale Long-Tailed Recognition in an Open World” CVPR’19

MEMEX

Position-based Object Retrieval

MEMEX

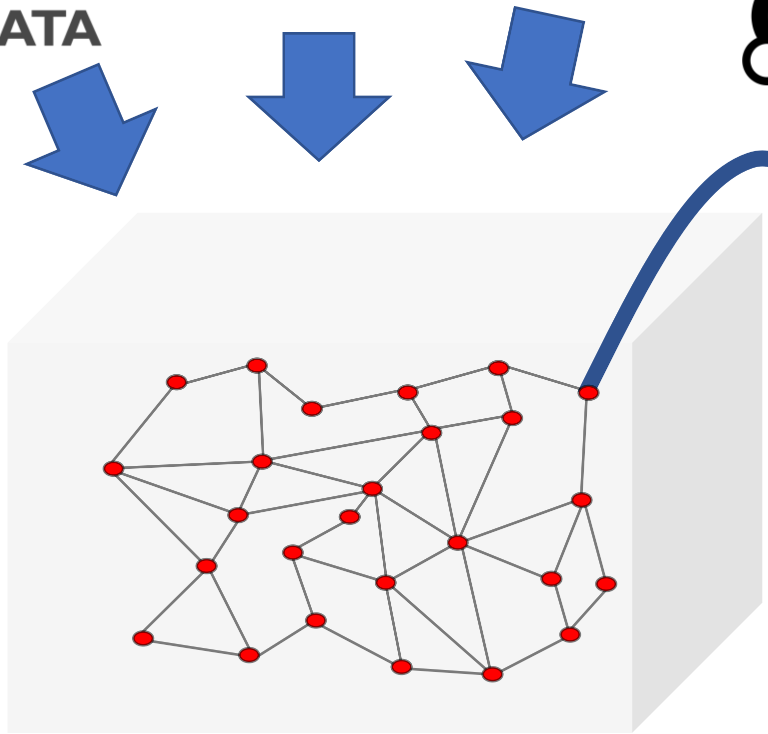
Multi-modal Object Retrieval

WIKIDATA

OpenStreetMap

Mapillary

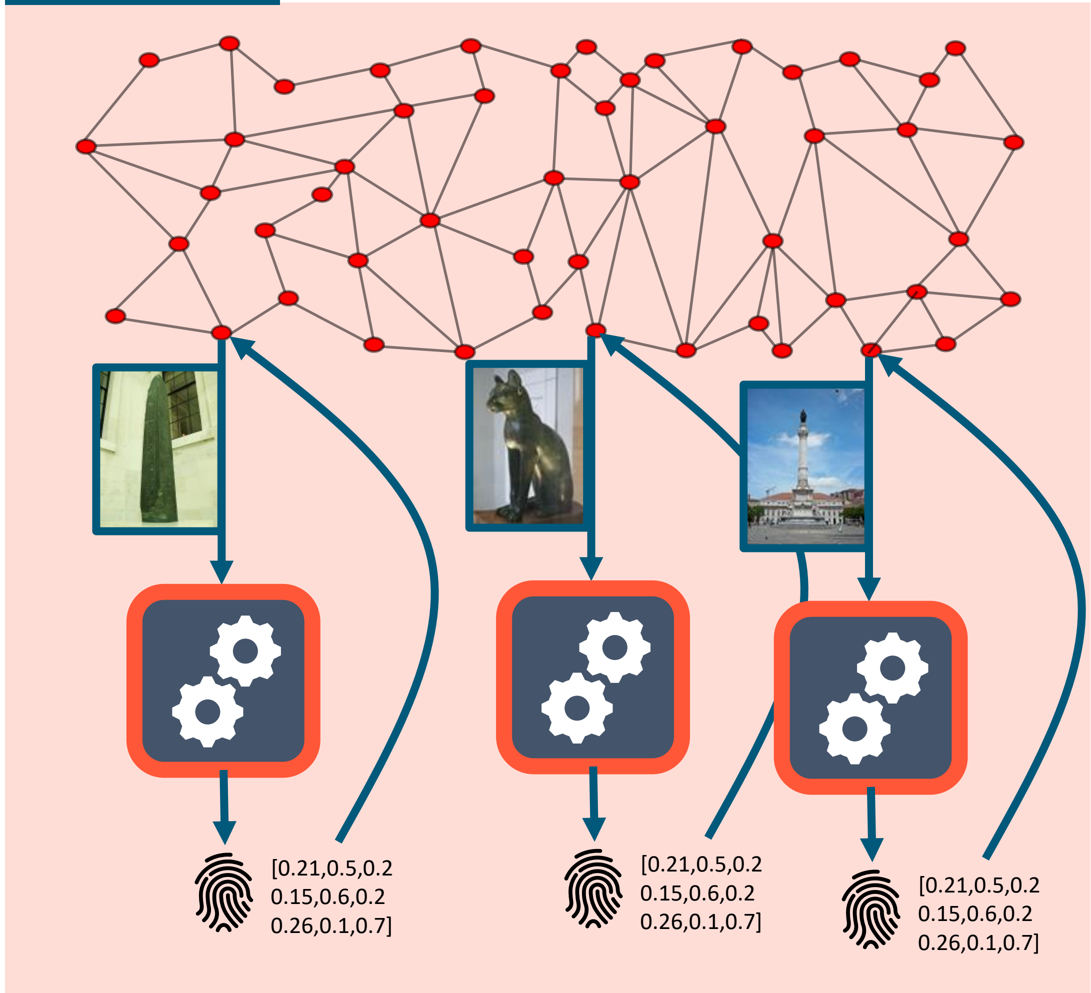
europeana
think culture



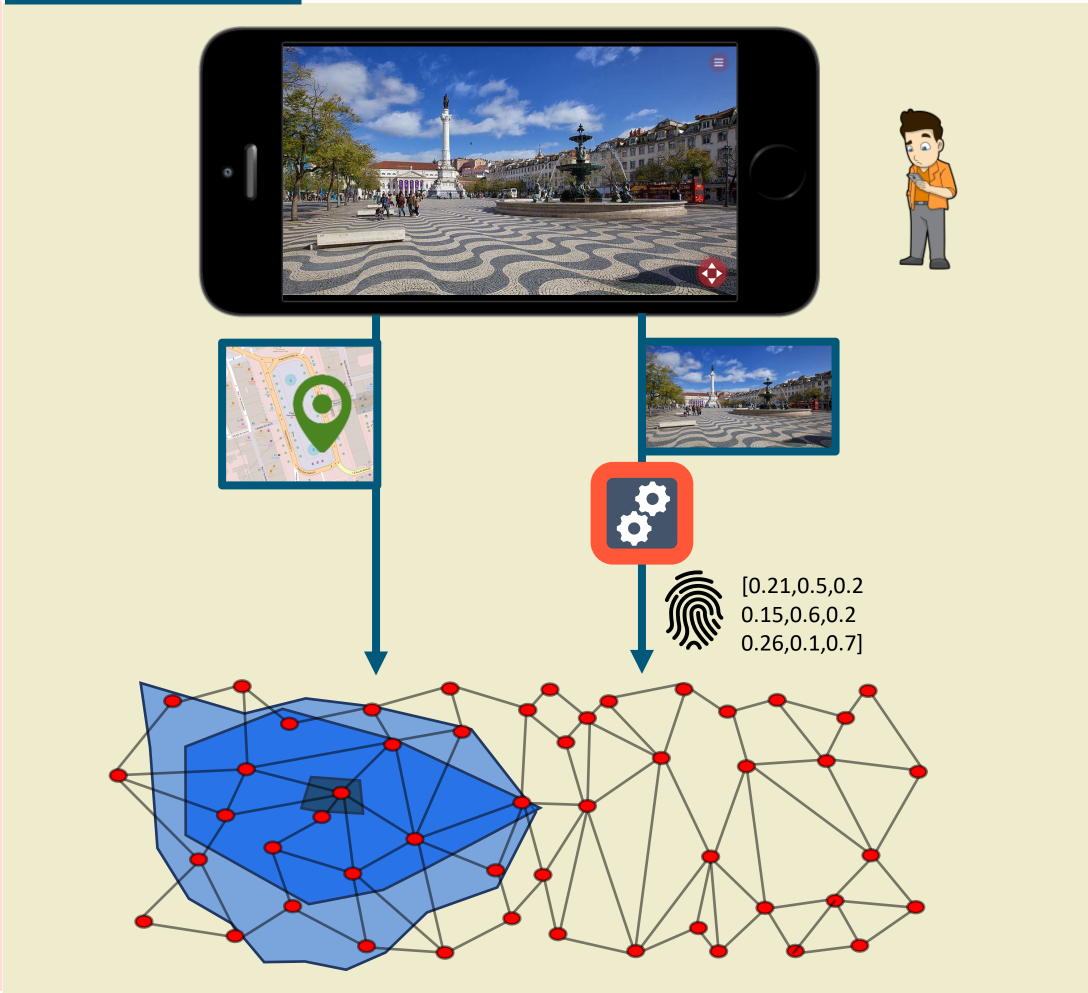
MEMEX

Multi-Modal Retrieval

Offline



At query time



Multi-Modal Retrieval

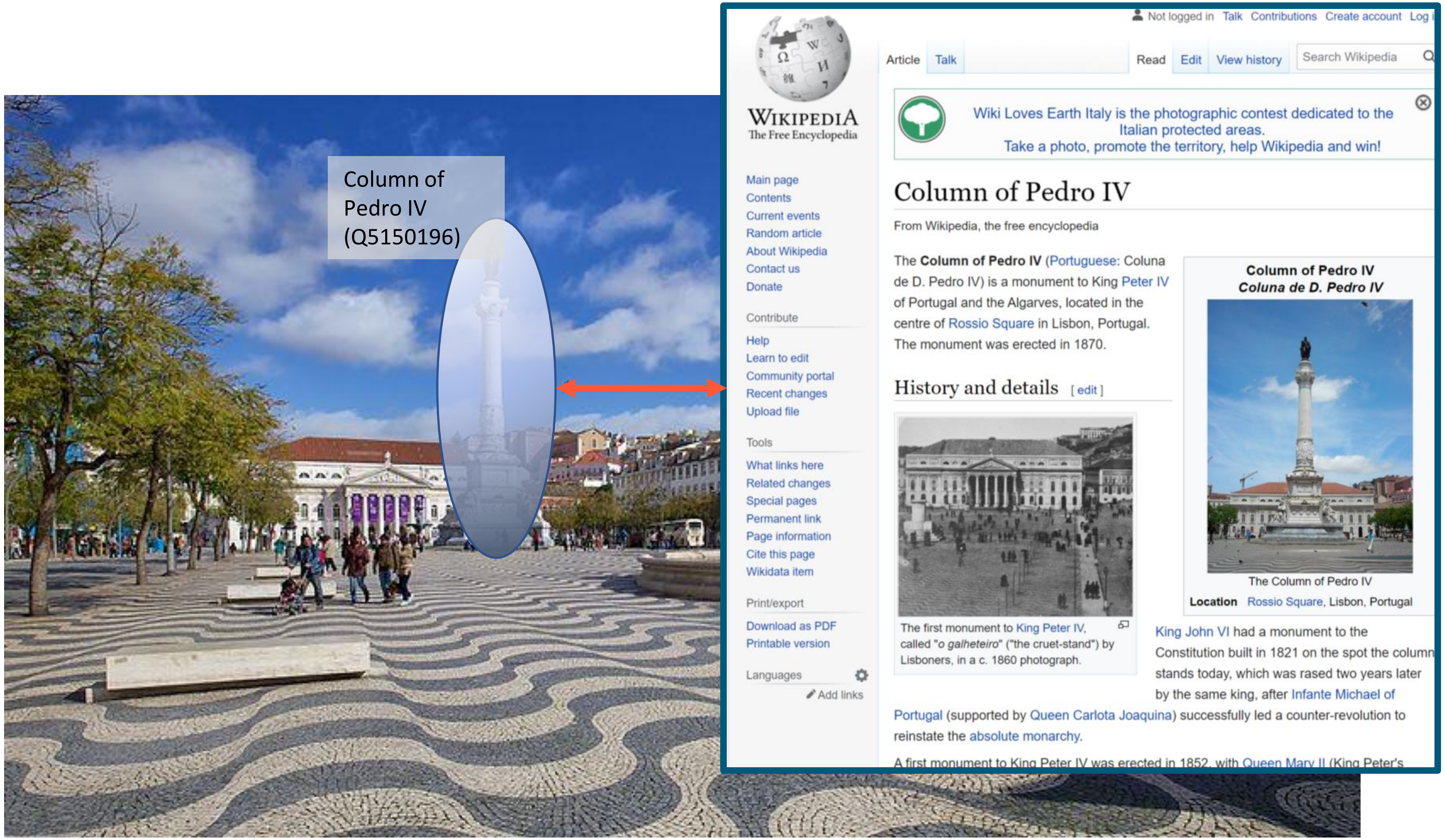
Location Only

Location and Image

MEMEX

Connecting real world to Knowledge Graph

Connected representations



You are here!

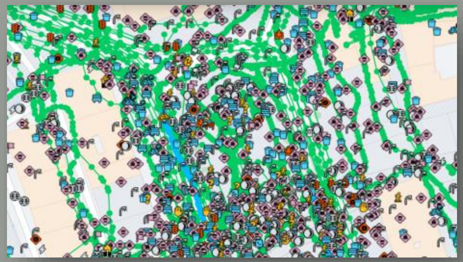
Finding position and orientation on a 2D map from a single image: The Flatlandia localization problem and dataset

iit PAVIS

3DoF visual localization from object detections

Why 2D maps?

Most static objects do not change appearance in time



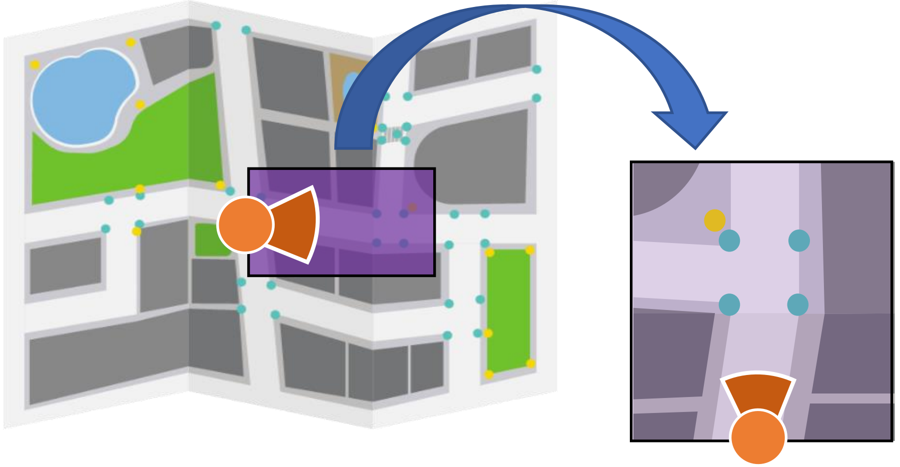
We have large-scale object maps

Maps are Concise to store in contrast to Street-level images

No privacy concerns over stored or transmitted data

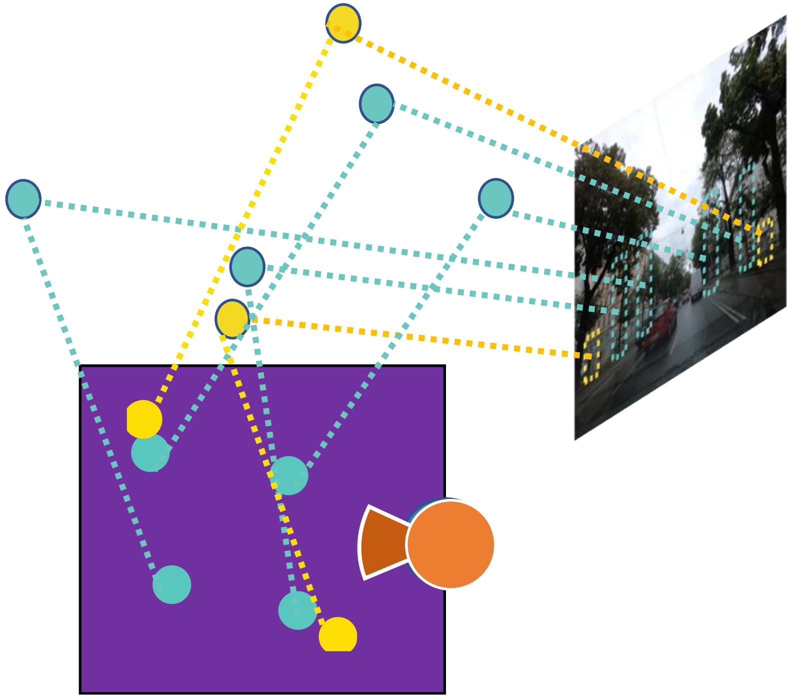
Local Map Generation (phone view)

GT-based



Reference map object locations, as seen by the camera

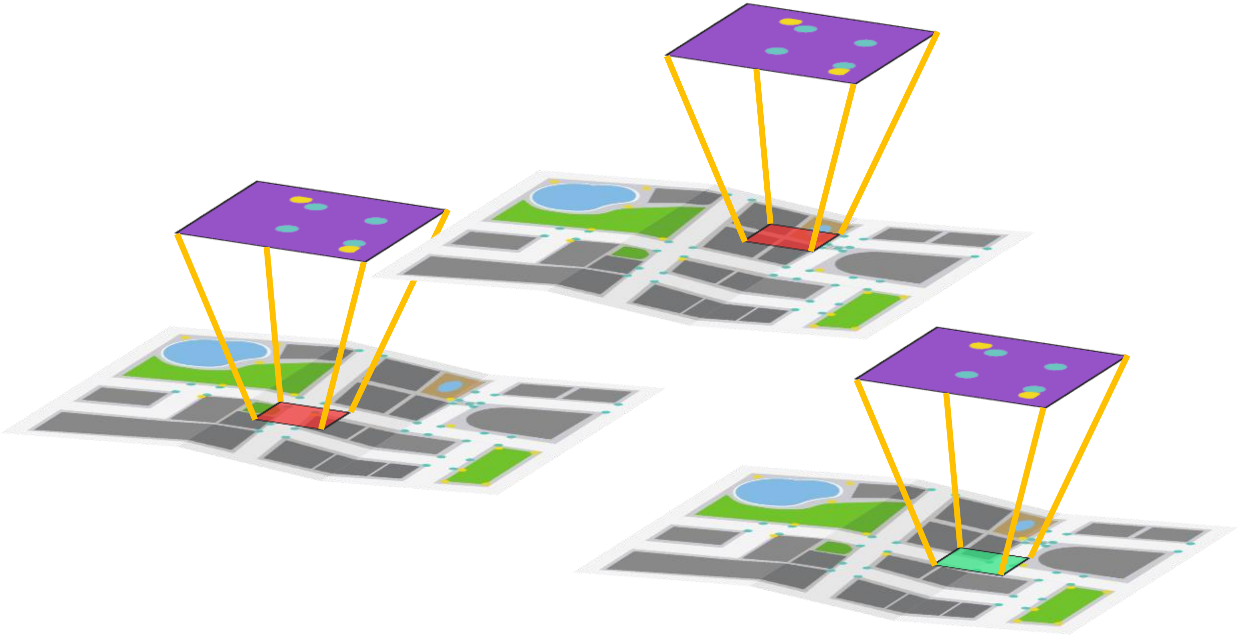
Depth-based



Detections projected using monocular depth estimation

3DoF visual localization from object detections

Coarse Map Localization



Where on the reference map can I find the objects arranged like the ones I am seeing?

Fine-grained 3DoF Localization

Given a candidate reference map region, where is the camera on the reference map?

Coarse Map Localization - Results

GT Local Maps			
Model	**Precision**	**Recall**	**Success**
Similarity	0.151	0.586	0.717
Triplet	0.161	0.559	0.693
Triplet + Similarity	0.206	0.617	0.804

Depth Local Maps			
Model	**Precision**	**Recall**	**Success**
Similarity	0.149	0.594	0.720
Triplet	0.160	0.570	0.740
Triplet + Similarity	0.184	0.630	0.808

More info here:

Paper: https://arxiv.org/abs/2304.06373

Github: https://github.com/IIT-PAVIS/Flatlandia

Thank you!