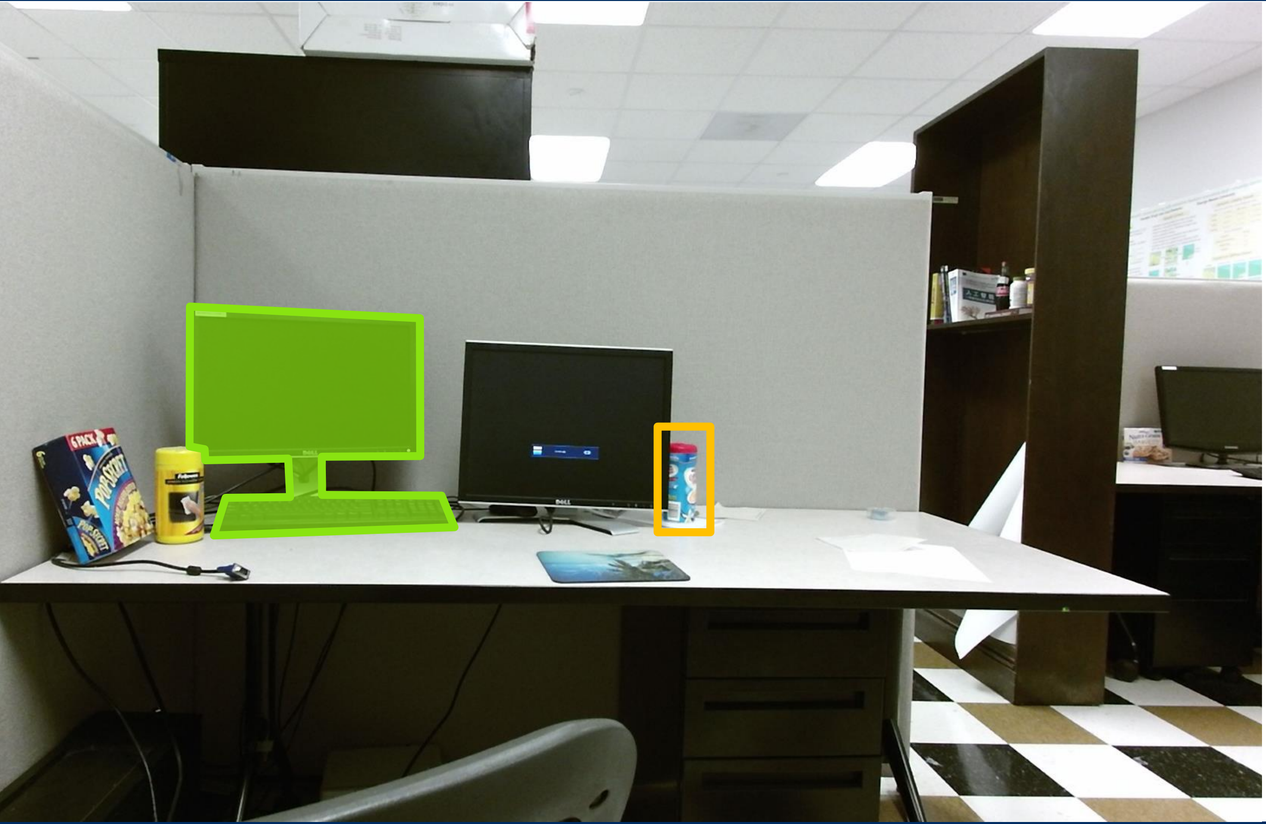
MapNet: An allocentric spatial memory for mapping environments

João F. Henriques, Andrea Vedaldi

Visual Geometry Group

UNIVERSITY OF OXFORD

Motivation



What we usually have:

- Object detections
- Segmentations
- 3D information (relative to camera)
- ...

⇒ Image-centric tasks

Motivation

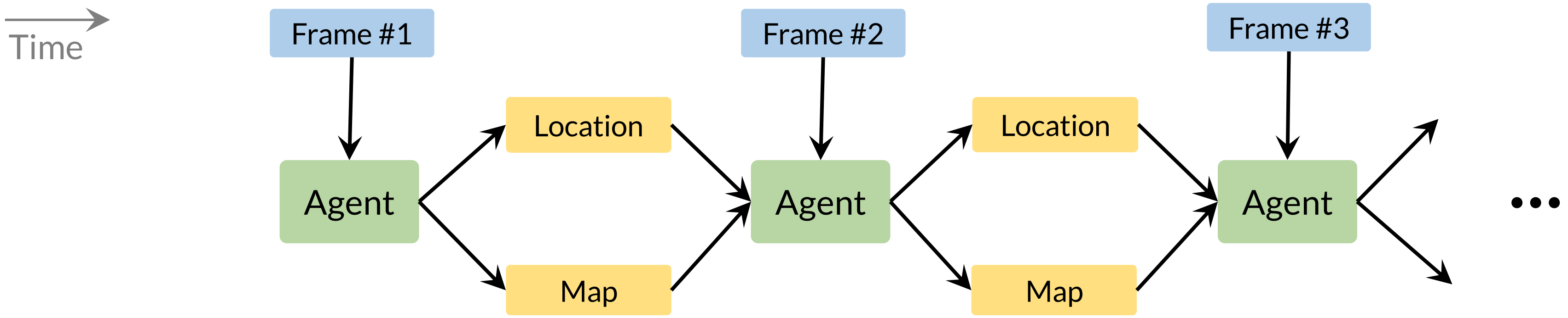
What we would like:

- Reason beyond image, into world
- Object permanence
- Eventually, long-term goals and planning

$\Rightarrow$ World-centric tasks

Simultaneous Localization And Mapping (SLAM)

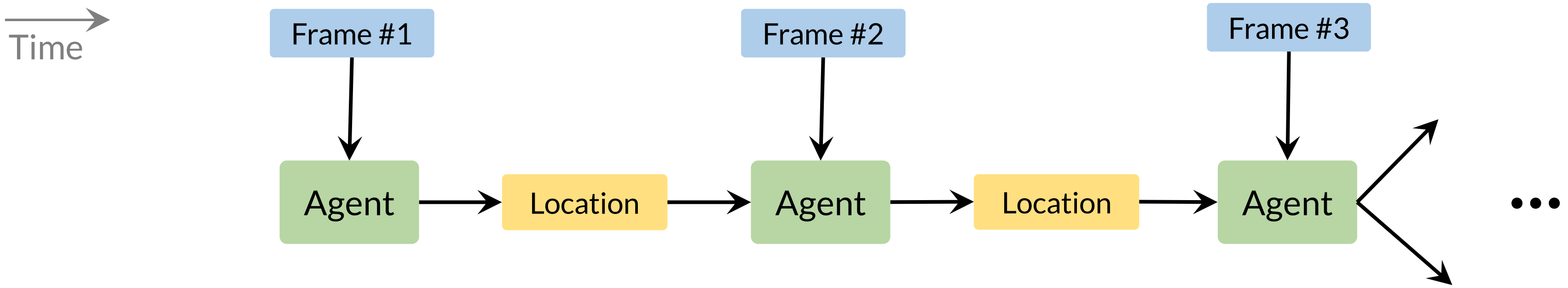
UNIVERSITY OF OXFORD



Classic SLAM
(No learning)

- Hard to adapt to new environments (hand-tuning)
- No semantic information
- No use of priors to compensate for missing data

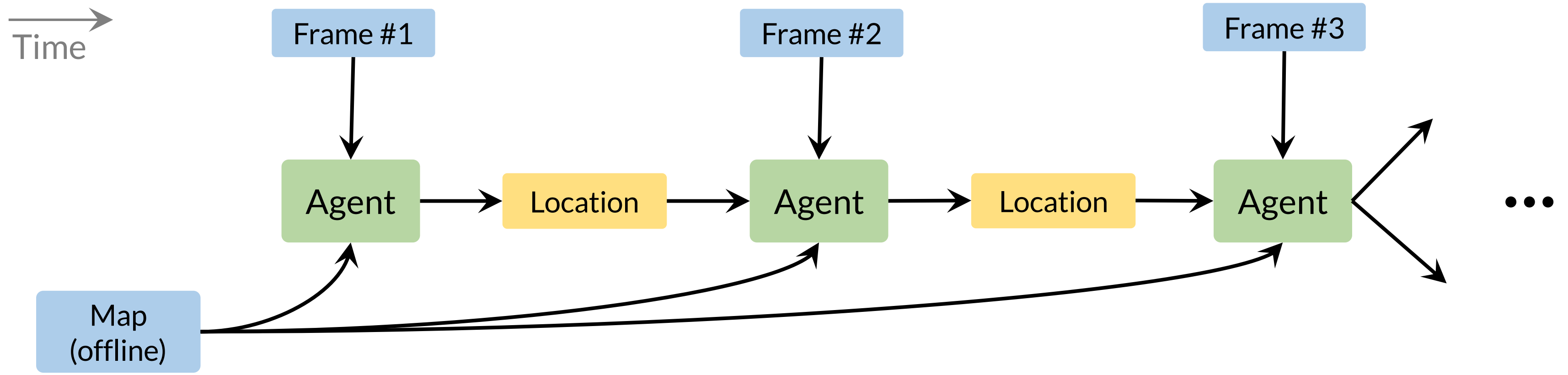
Related work – deep learning for SLAM



Egomotion predictors
- No map
- Cannot correct for inevitable drift

Costante'15, Clark'17, Zhu'17, Wang'17, ...

Related work – deep learning for SLAM

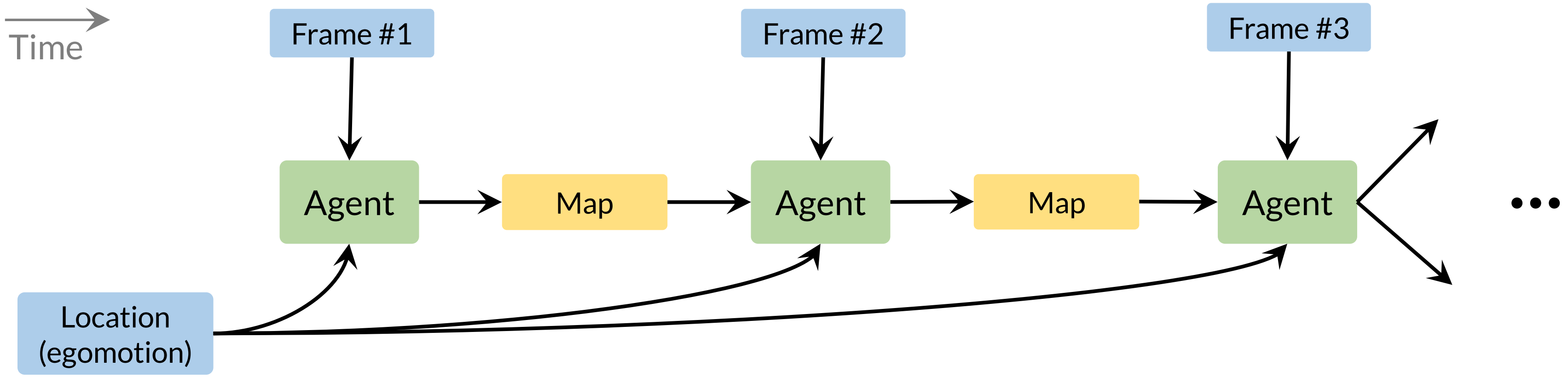


Offline-learned localization

- Map is stored in deep network's parameters
- New environments require re-training

Kendall'15, Mirowski'18, Brahmbhatt'18, ...

Related work – deep learning for SLAM

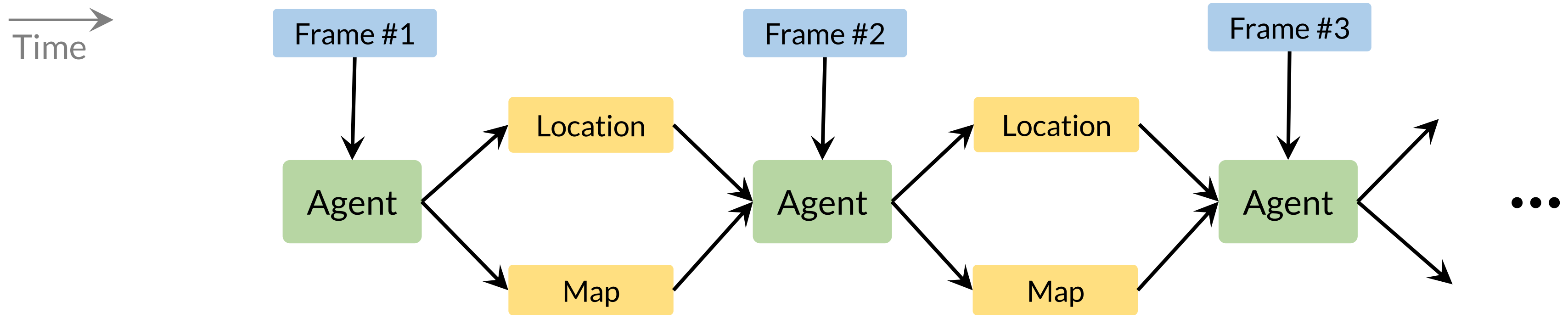


Online mapping, no localization

- Map is created on-the-fly as activations
- Perfect egomotion input is used for localization, not map
- Tested on synthetic environments (so far)

Kanitscheider'16, Gupta'17, Zhang'17, Parisotto'17, ...

Proposed method



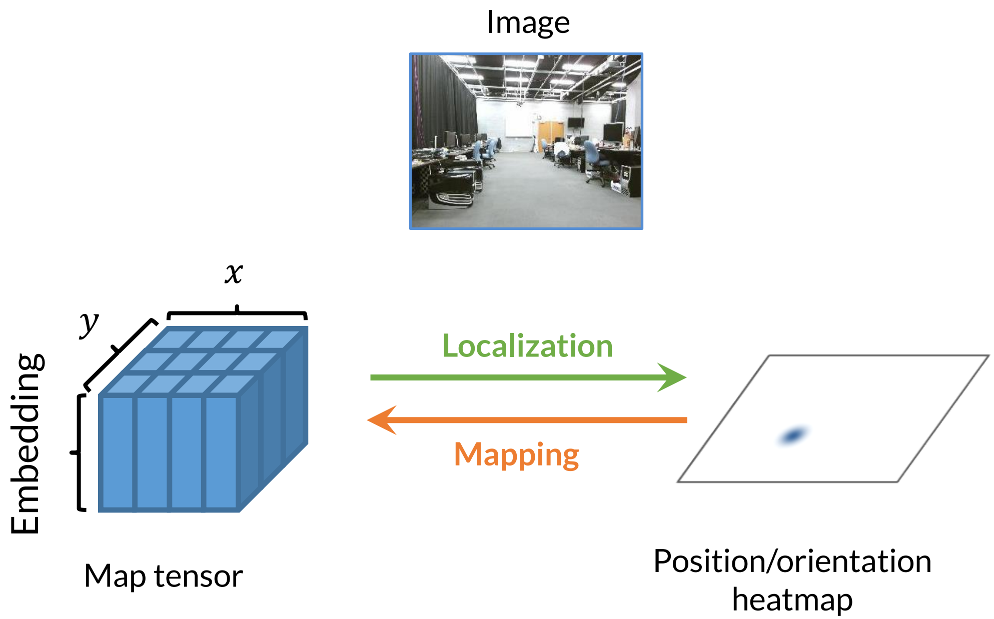
Our method (MapNet)

- Performs **both Mapping and Localization** with a deep net
- No egomotion information
- Fully online (mapping as we go)

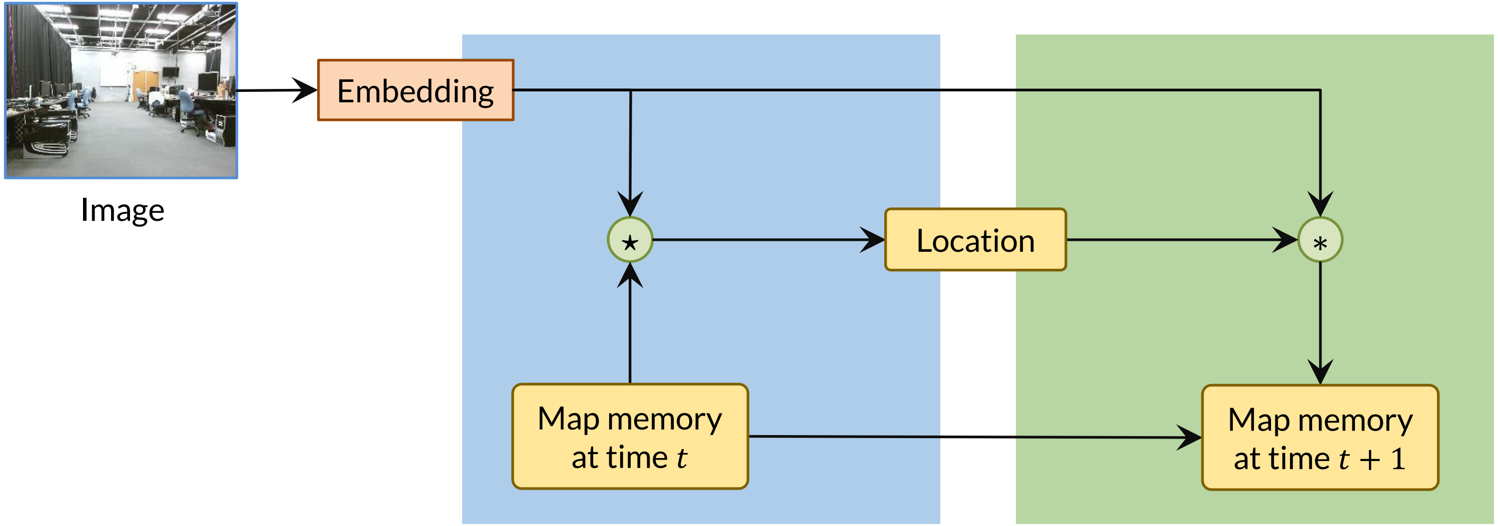
Allocentric map memory

Map model:

- Represent **ground plane** as 2D grid.
- Store one **embedding** per location.
- Allows associating semantics with *world coordinates.*

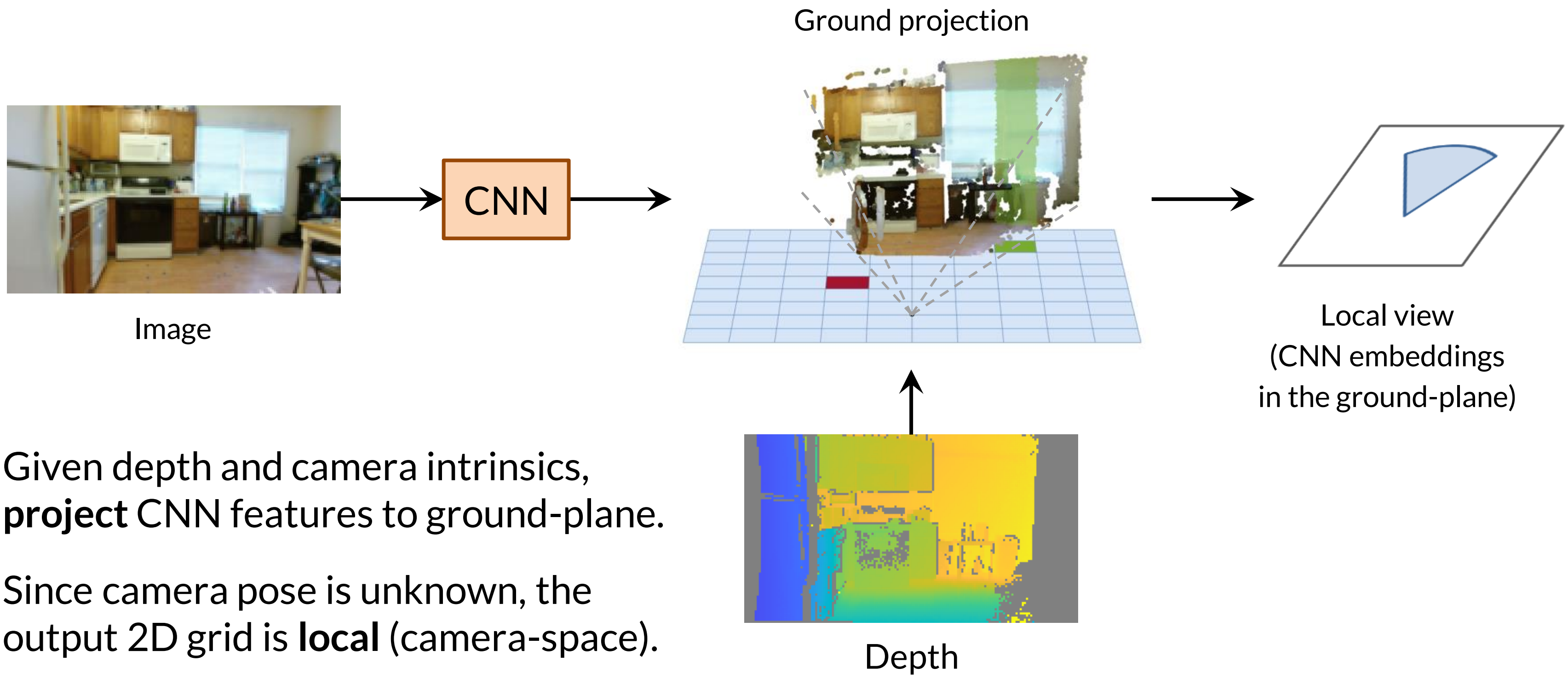


Localization and mapping as dual operators



Core insight: Localization ⇔ convolution Mapping ⇔ deconvolution

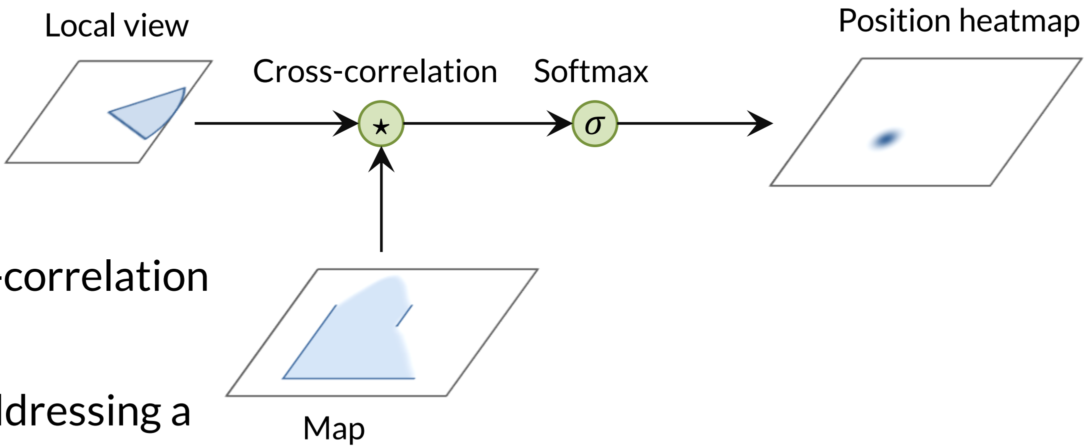
Ground projected CNN features



- Given depth and camera intrinsics, **project** CNN features to ground-plane.
- Since camera pose is unknown, the output 2D grid is **local** (camera-space).

Localization

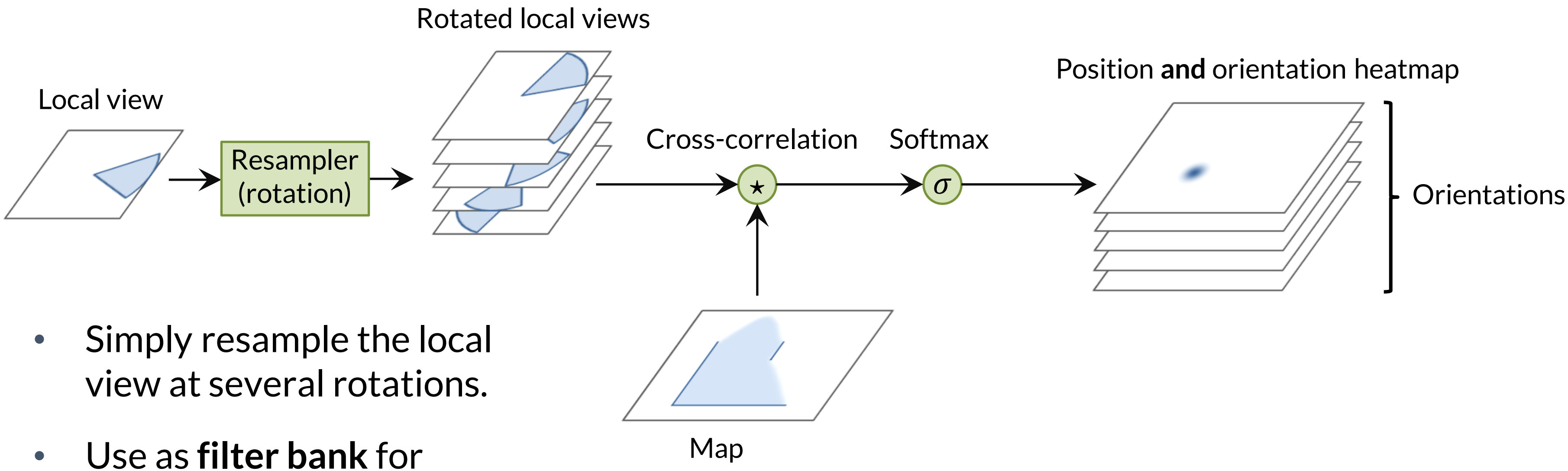
Localize by **dense matching** of the local view's embeddings to the map.



- Requires only **one** cross-correlation (convolution).
- Can be interpreted as addressing a **spatial associative memory**.

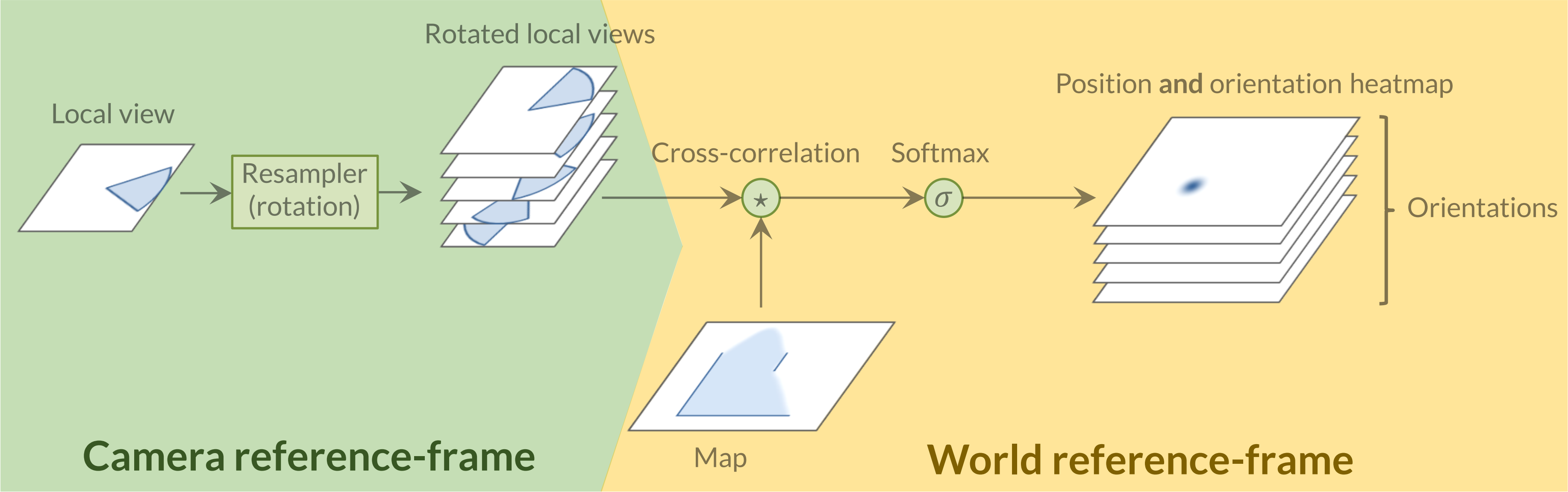
Localization

Also consider **camera orientation**:



- Simply resample the local view at several rotations.
- Use as **filter bank** for cross-correlation.

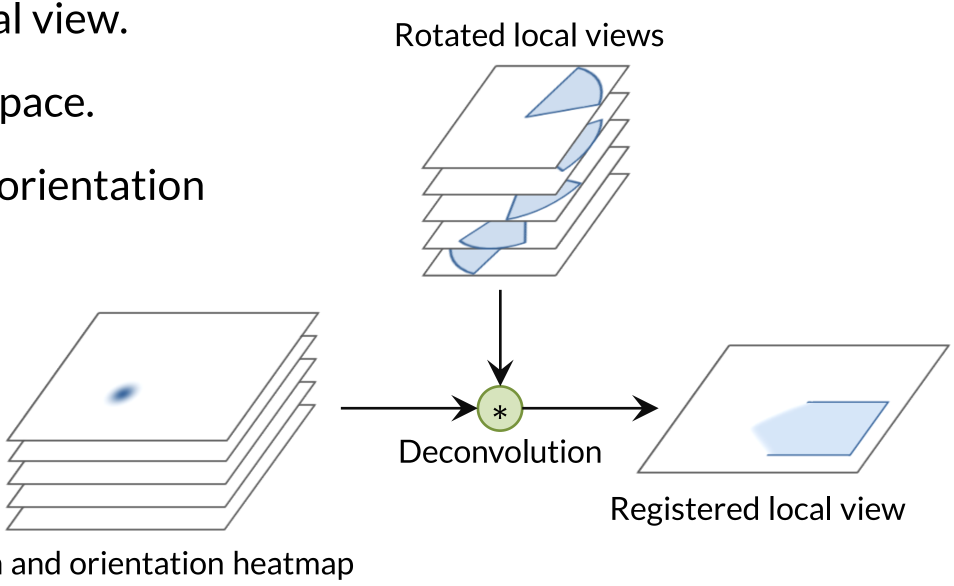
Localization



Mapping

The **mapping** step updates the map with the local view.

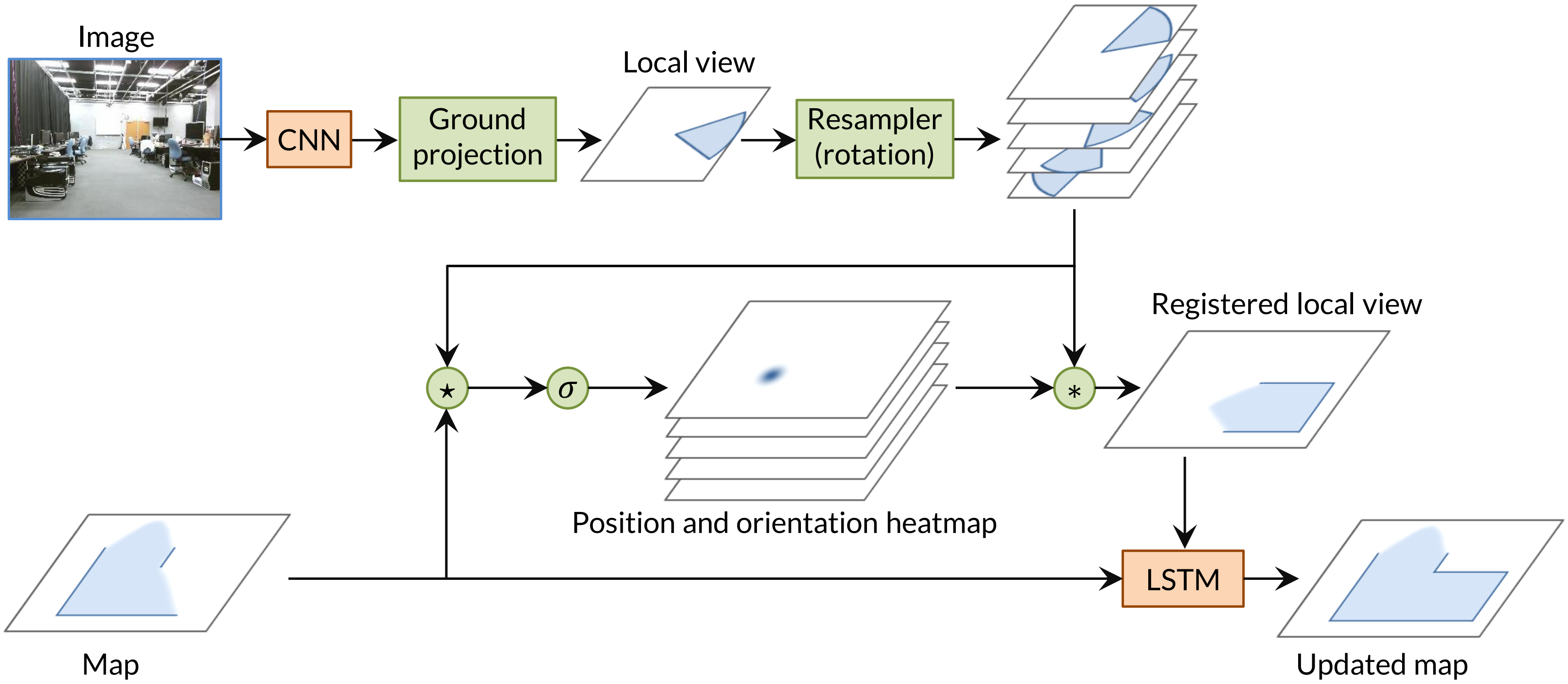
- The local view must be **registered** to world-space.
- Requires one **deconvolution** of the position/orientation heatmap, using the local views (filter bank).



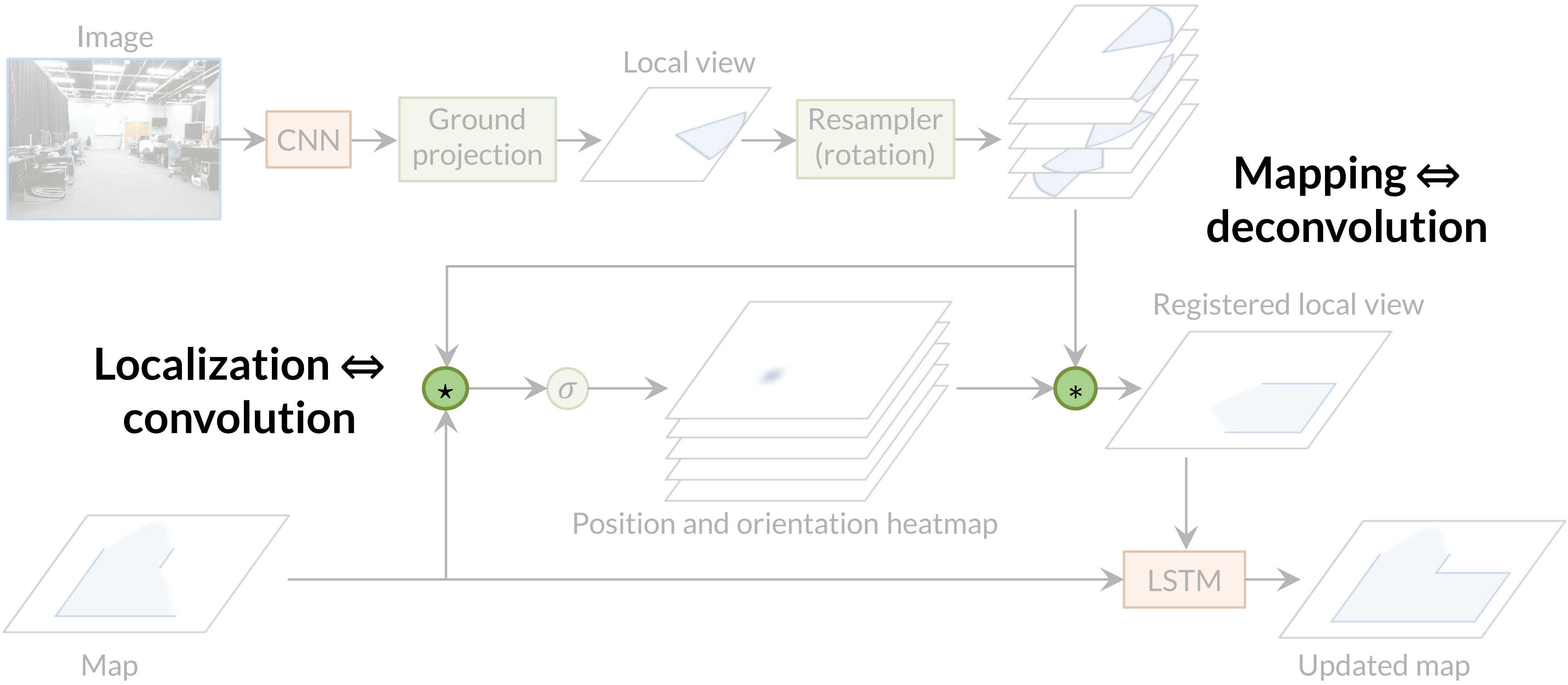
- After registration, the local view can be easily integrated into the map

  (e.g. by linear interpolation, or a convolutional LSTM)

Full pipeline



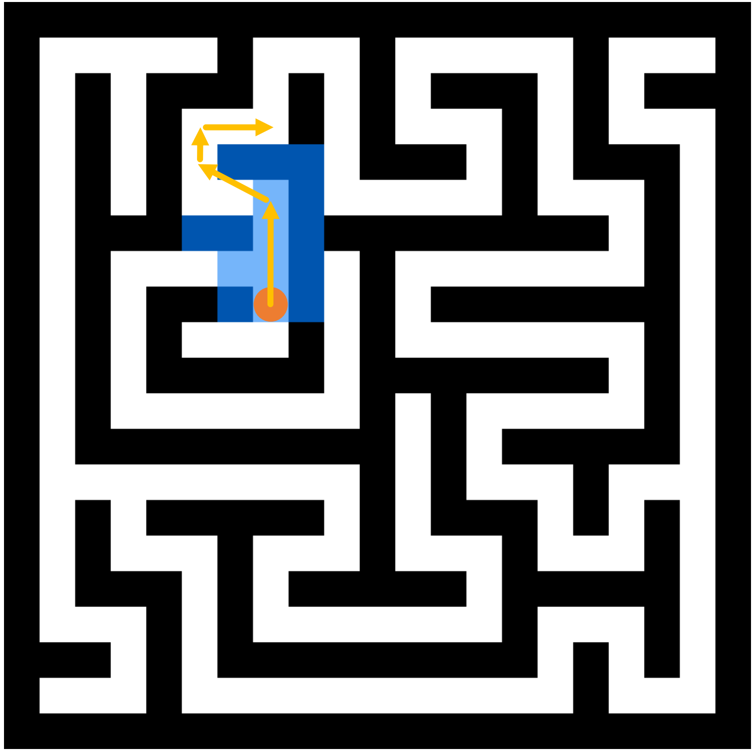
Full pipeline



Mapping ⇔ deconvolution

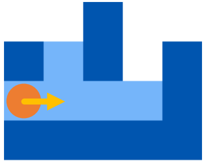
Localization ⇔ convolution

Experiments – 2D data



Toy problem setup

- 100,000 mazes
- Agent moves at random
- Limited, local visibility



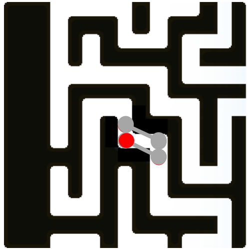
Local view

Training

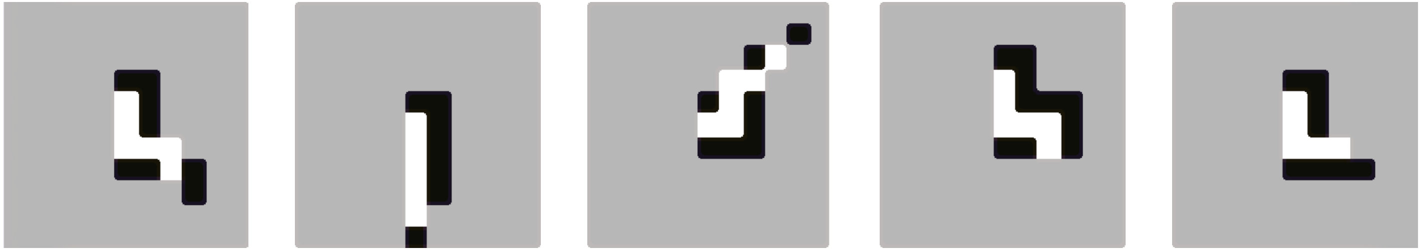
- Input sequences of 5 frames
- Position/orientation supervision
- Min. logistic loss of predicted position (heatmap)

Experiments – 2D data

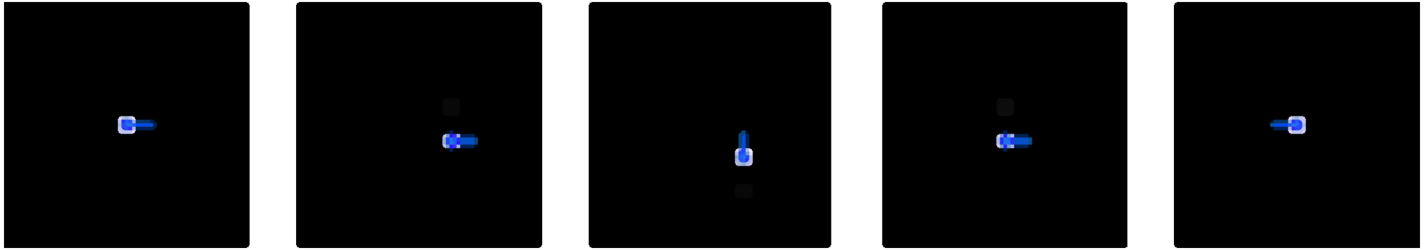
Global view



Local view (always facing right)

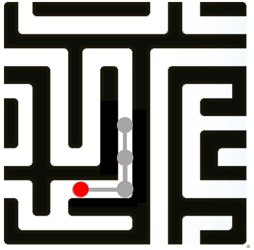


Predicted heatmap (blue – ground truth)

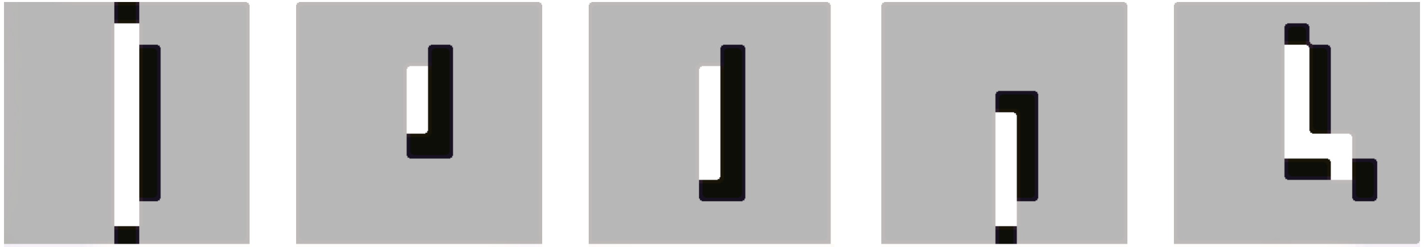


Experiments – 2D data

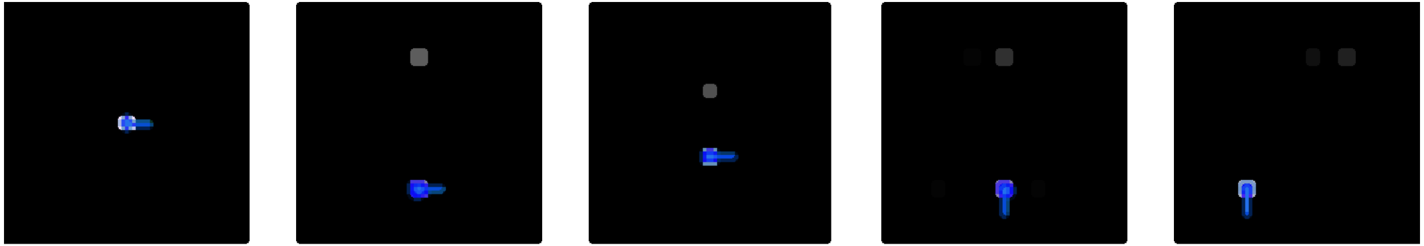
Global view



Local view (always facing right)



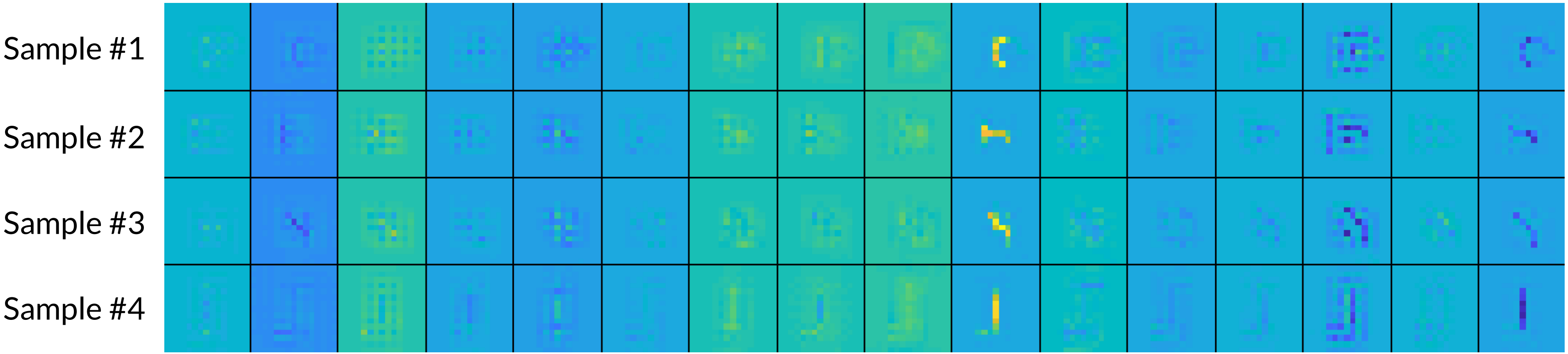
Predicted heatmap (blue – ground truth)



Experiments – 2D data

UNIVERSITY OF OXFORD

Map tensor (one channel per column)



⇒ Several local views are integrated into a **larger** map.

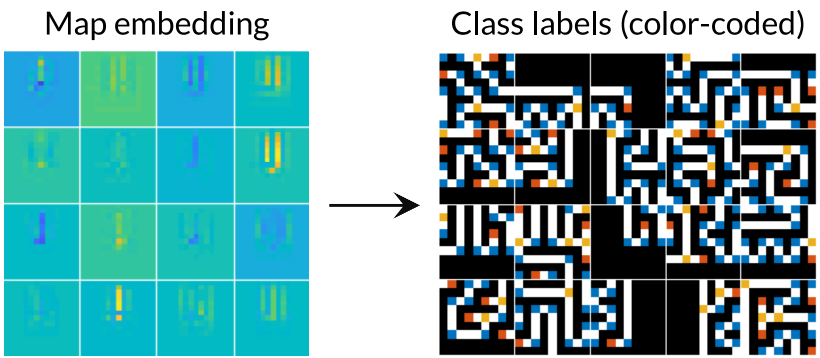
Experiments – 2D data

Is this map semantic? → Yes!

- Assigned class labels to maze cells (corridors, turns, dead-ends...).
- Class label is correctly predicted from a cell's embedding most of the time.

Corridor	Turn	Dead end	Fork	Crossroad	All
76.1%	73.3%	69.8%	68.8%	62.3%	71.3%

Balanced dataset prediction accuracy (chance: 50%)



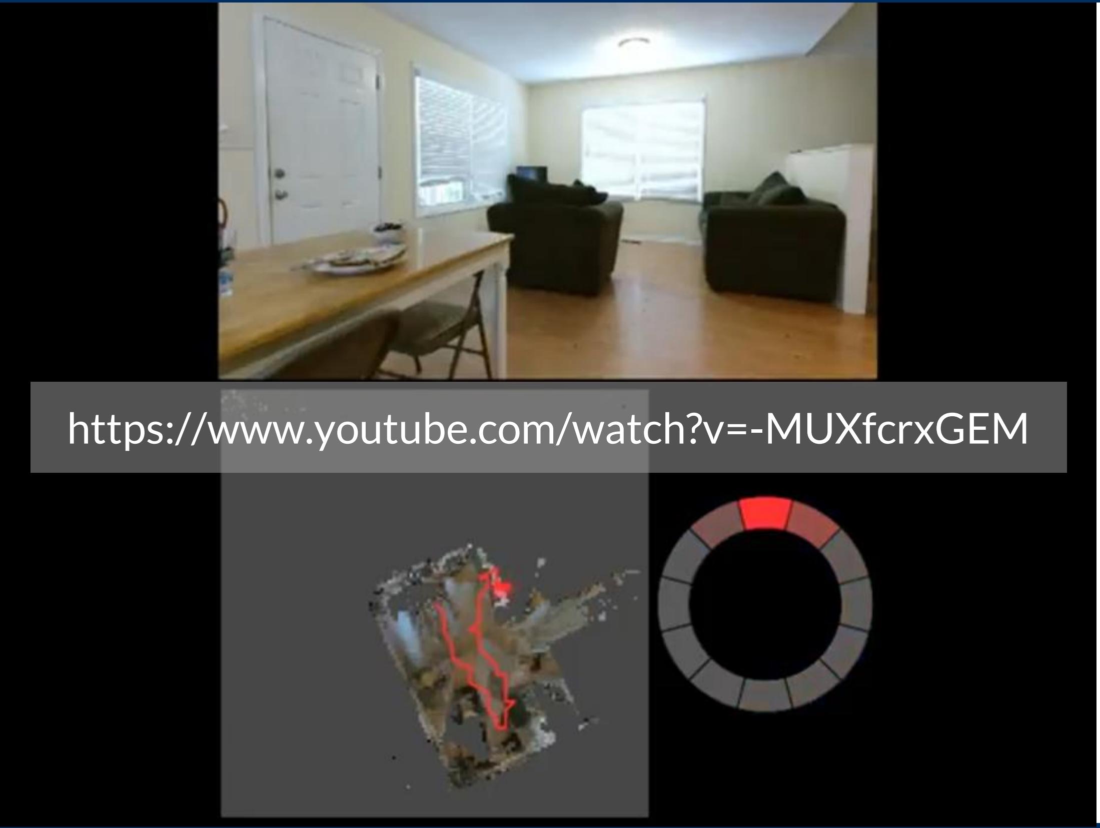
Experiments – 3D game data

https://www.youtube.com/watch?v=mInSO7YW1EU

ResearchDoom Dataset

- 4 recorded speed-runs through the whole game
- 6 hours of gameplay
- Challenging, large hand-crafted levels

Experiments – 3D real data



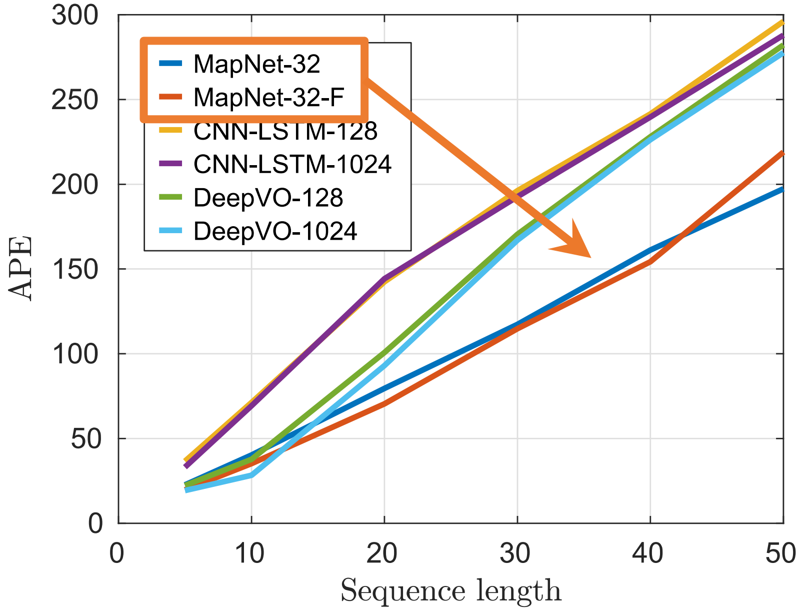
https://www.youtube.com/watch?v=-MUXfcrxGEM

Active Vision Dataset

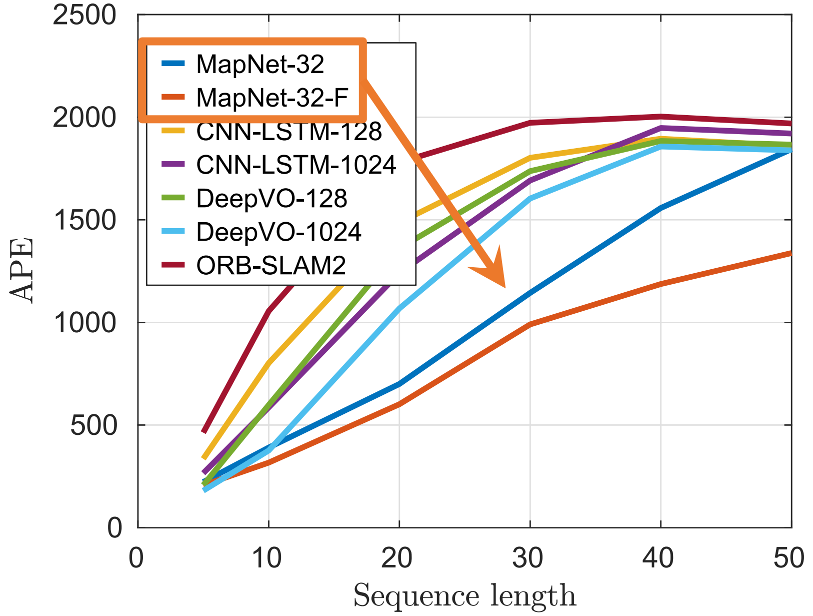
- Robot platform in 19 indoor scenes
- Images collected at all positions/orientations
- Can be composed into unlimited sequences

Experiments – 3D data quantitative results

ResearchDoom Dataset



Active Vision Dataset



Conclusions

- We perform SLAM **entirely online** using an end-to-end learned architecture.
- **Localization** and **Mapping** are a dual pair of **convolution/deconvolution**.
- **Semantic** embeddings of the World arise from the self-localization objective.
- **Next step:** navigation and long-term goals.

Project page with code:
`www.robots.ox.ac.uk/~joao/`