Learning high-level reasoning in vision, language and robotics

Christian Wolf Corentin Kervadec

May 28th, 2021

INSA LIRIS CNRS

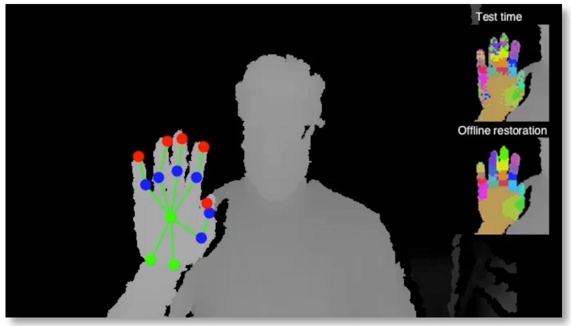
Our group

Christian Wolf
Chair in Research and Teaching in Artificial Intelligence at INSA-Lyon,
LIRIS UMR CNRS 5205
liris.cnrs.fr/christian.wolf

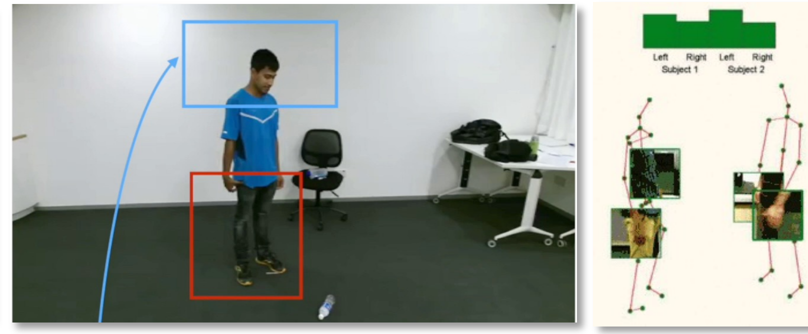
The group in Feb. 2020: Corentin Kervadec, Steeven Janny, Edward Beeching, Fabien Baradel, Théo Jaunet, Quentin Possamaï.

Learning vision & robotics

Gesture recognition



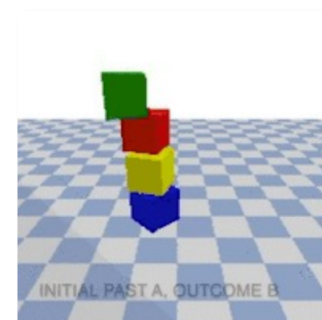
Pose estimation



Activity Recognition

Robot Perception and Navigation

H-C Interaction



Physics

What happened here?

Algebraically manipulating previously acquired knowledge in order to answer a new question

[Bottou, ML 2014]

Vision and Language Reasoning

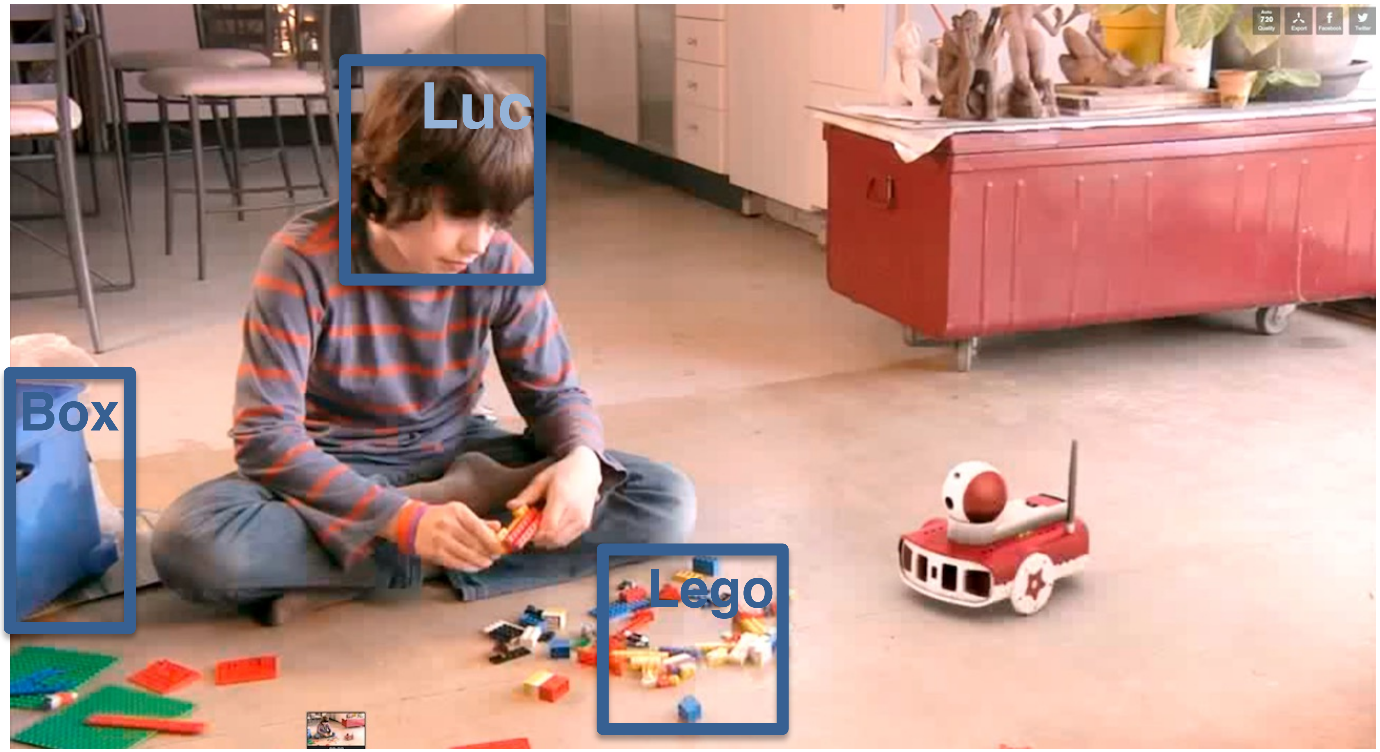
"How much money do I have in my hand?"

"What is in this jar?"

"Did I leave the door open?"

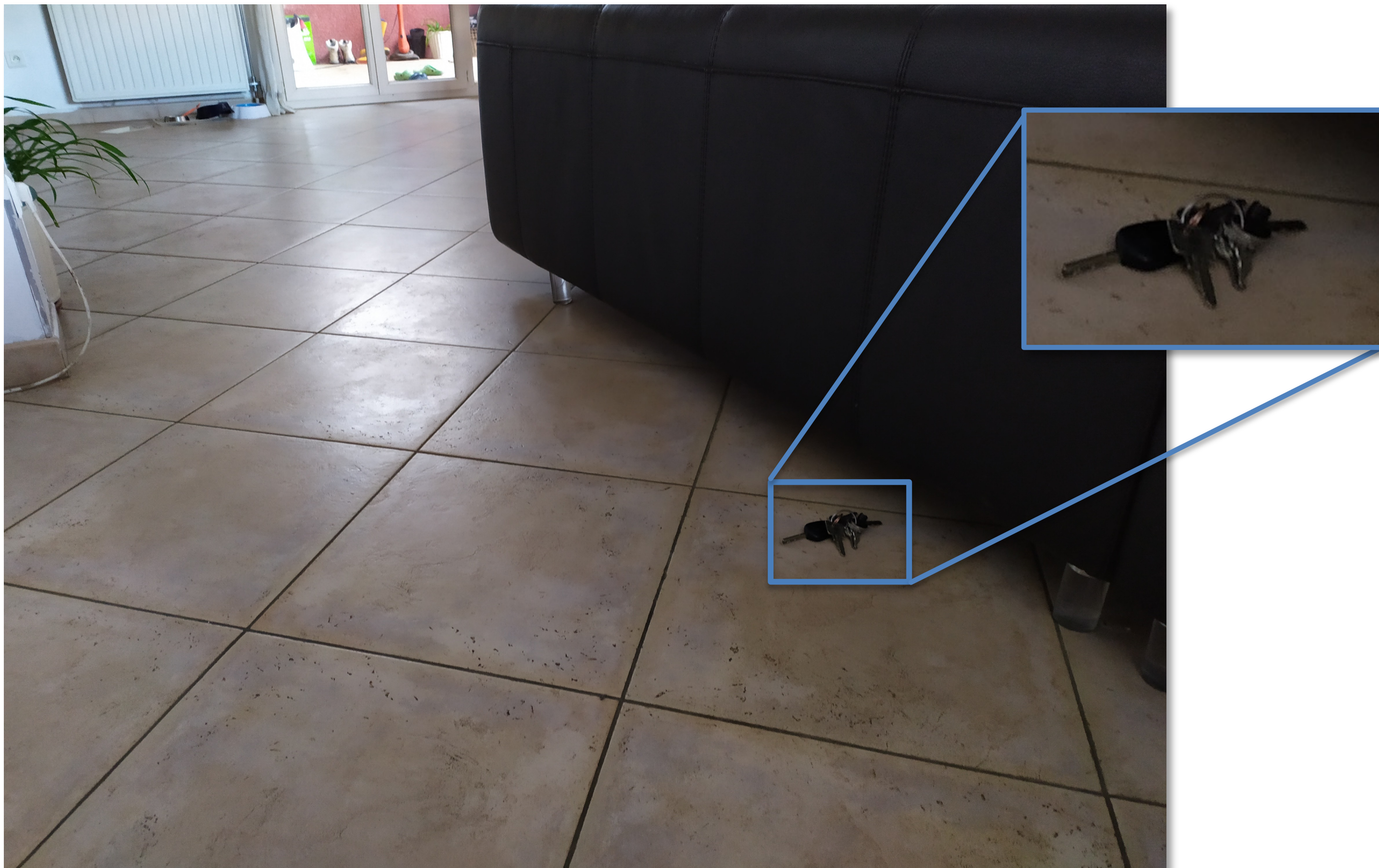
"Did I leave the lights on?"

Companion robots



Awabot

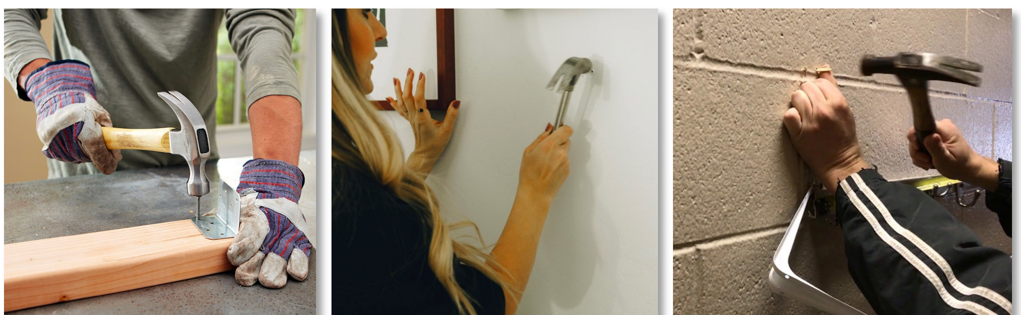
Discovering object affordances



Shortcuts in learning

Train for classification including
a class "to nail"

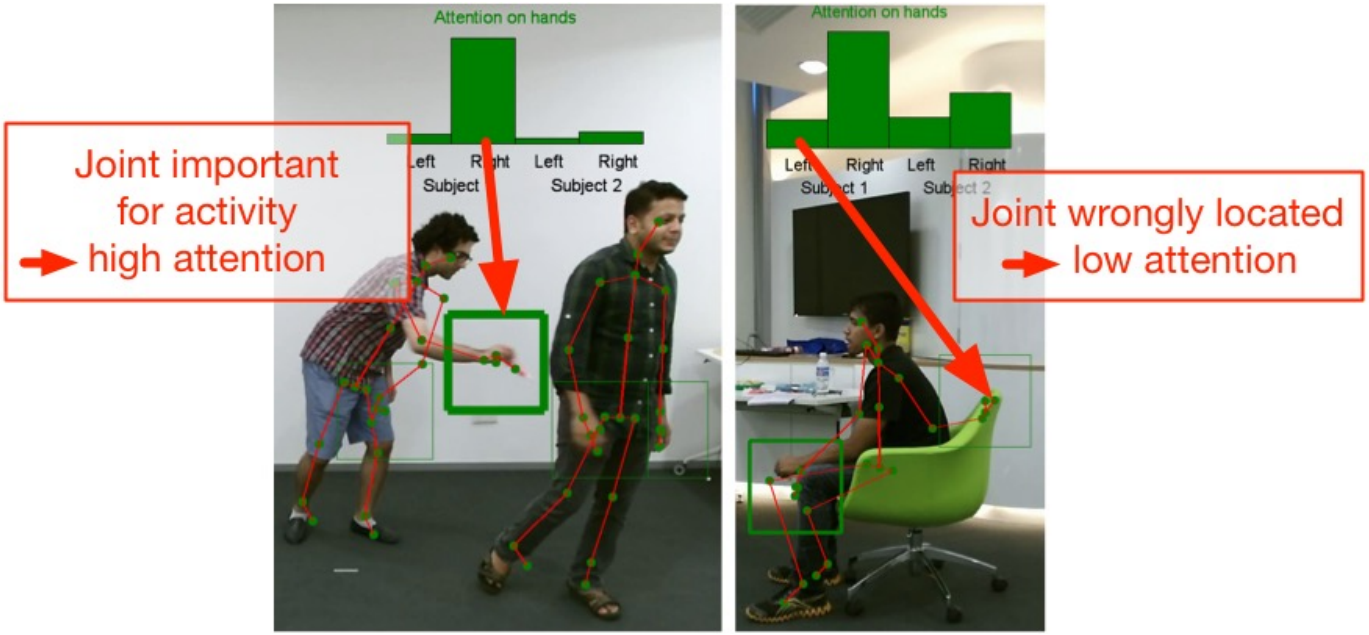
Test on data including new classes
"remove a nail", "store a hammer »



WYGISNWYE:

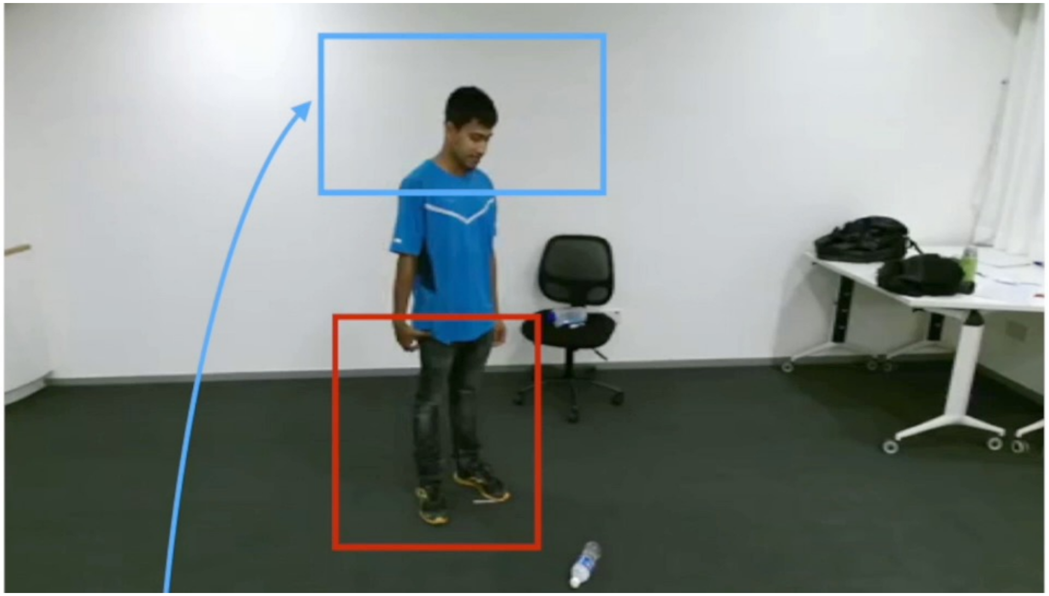
What you get is not what you expected!

WYGISNWYE



[Baradel, Wolf, Mille, BMVC 2018]

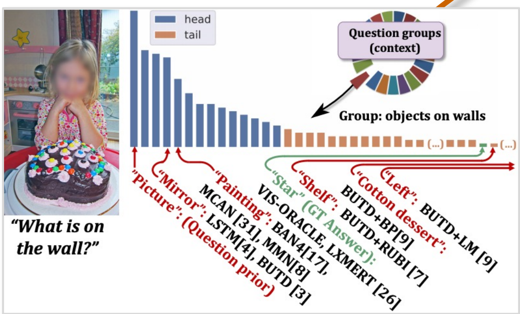
WYGISNWYE



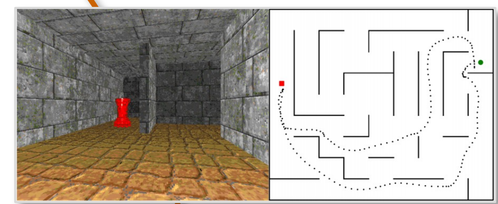
[Baradel, Wolf, Mille, Taylor, CVPR 2018]

Learning to reason

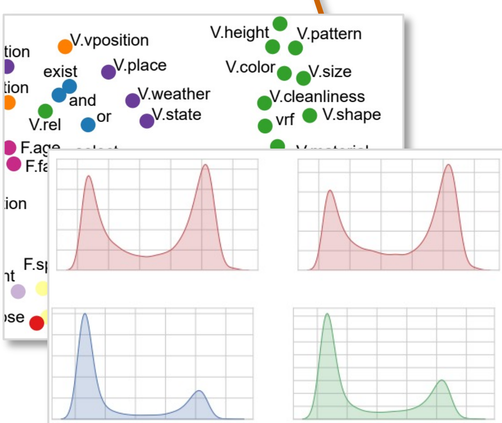
How can we evaluate biases in learning? (CVPR 2021a)



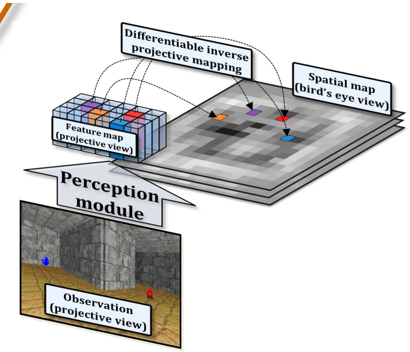
Which tasks favor emergence of reasoning?
(ICLR 2020, ECML-PKDD 2020)



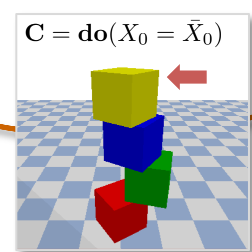
How can we visualize and transfer reasoning? (CVPR 2021b)



How can we create inductive bias for reasoning?
(ECCV 2018, ECCV 2020, ECAI 2020, ECML-PKDD 2020)



What are the causal links in the data? (ICLR 2020)



Visual navigation and spatial reasoning

Office space

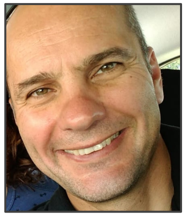
Homes

Hospitals

Edward Beeching

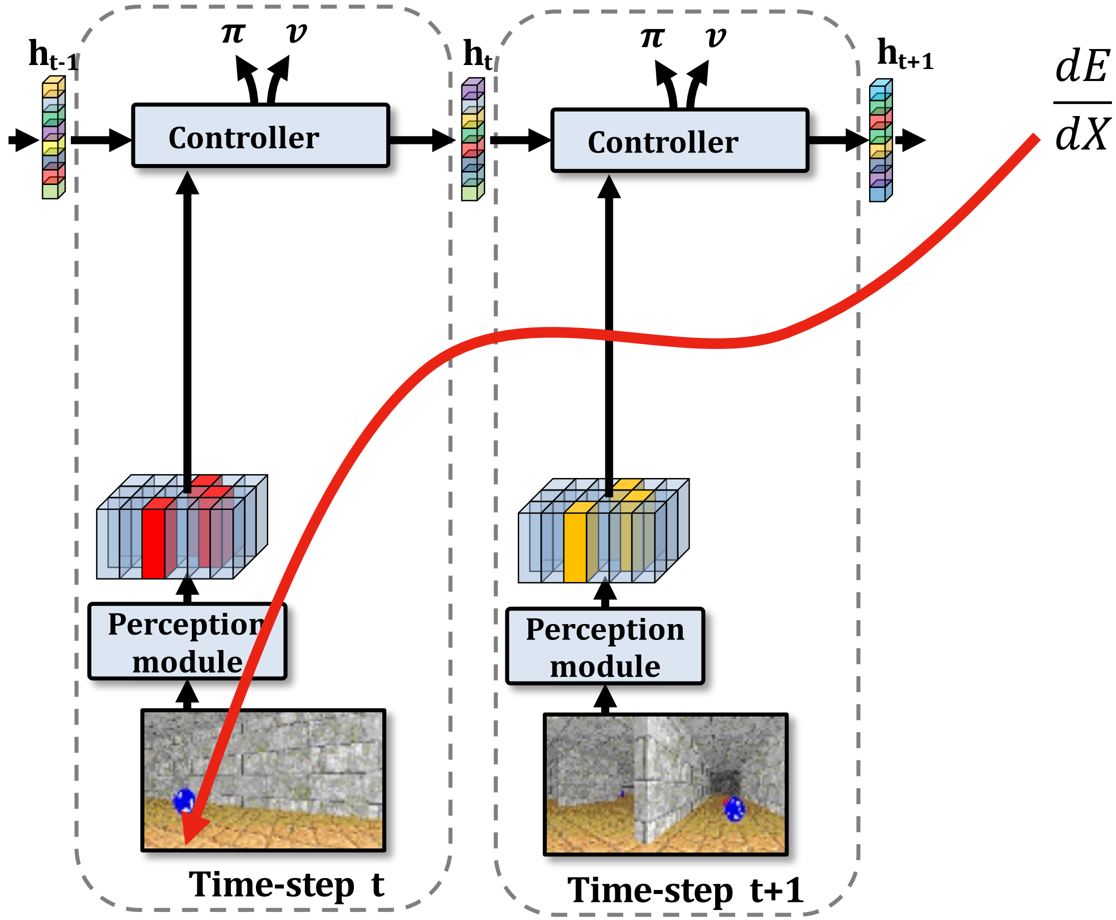
Jilles Dibangoye

Olivier Simonin

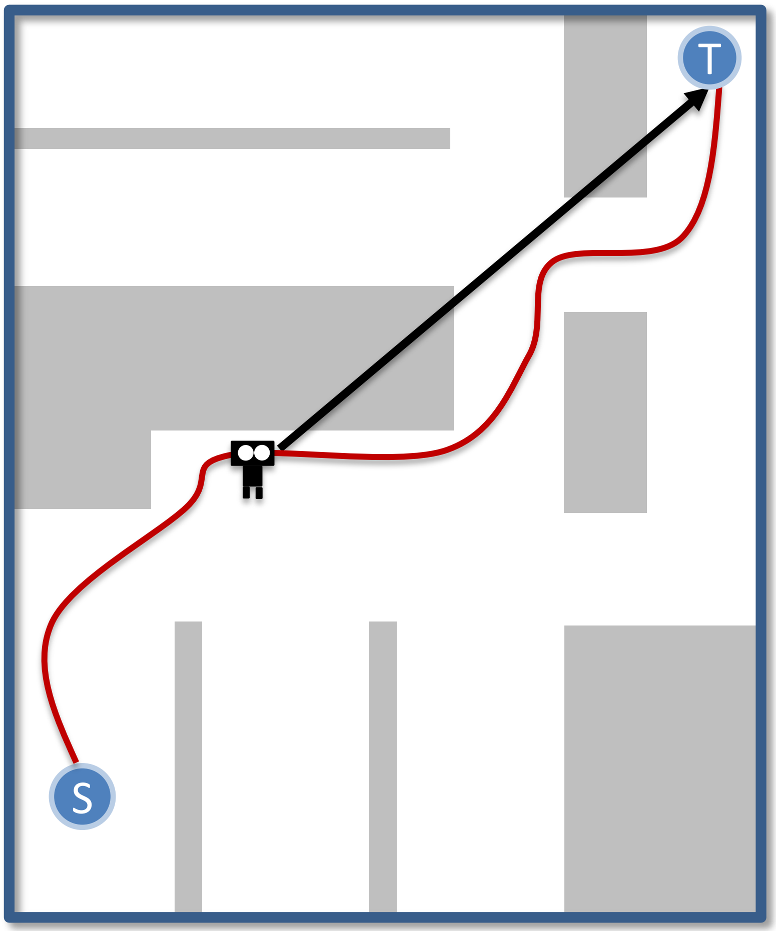


Christian Wolf

A Deep-RL baseline



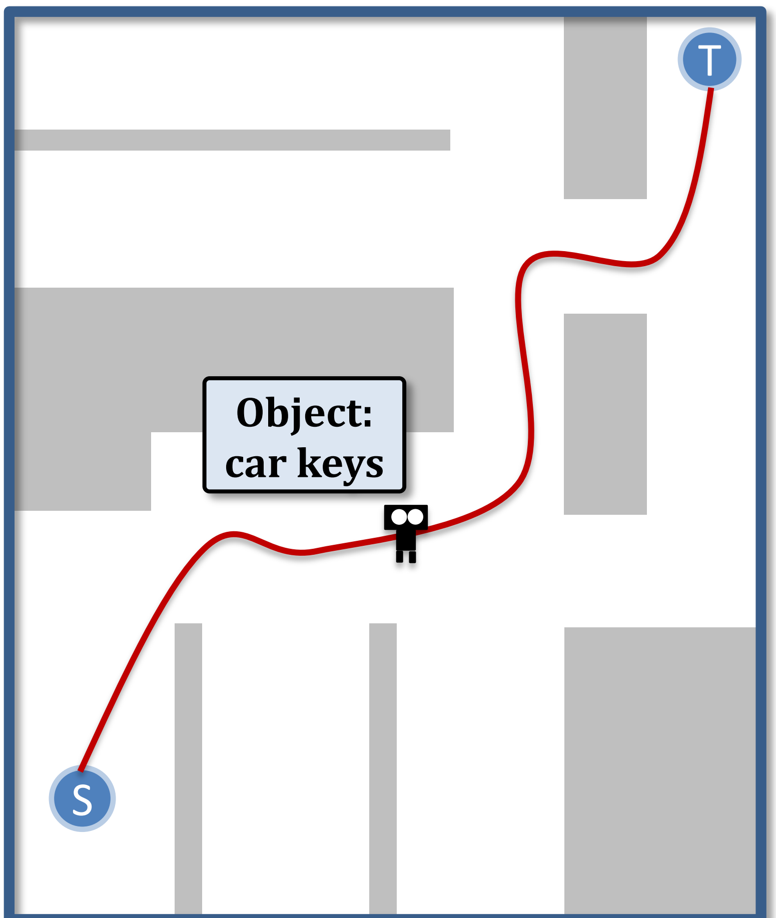
Task: PointGoal (+GPS)



Task:
Find a target given coordinates, receive a direction vector et each step

Required reasoning:
- Recognize free navigation space
- Follow the direction vector
- Learn how to overcome obstacles (difference between Euclidean and geodesic path)

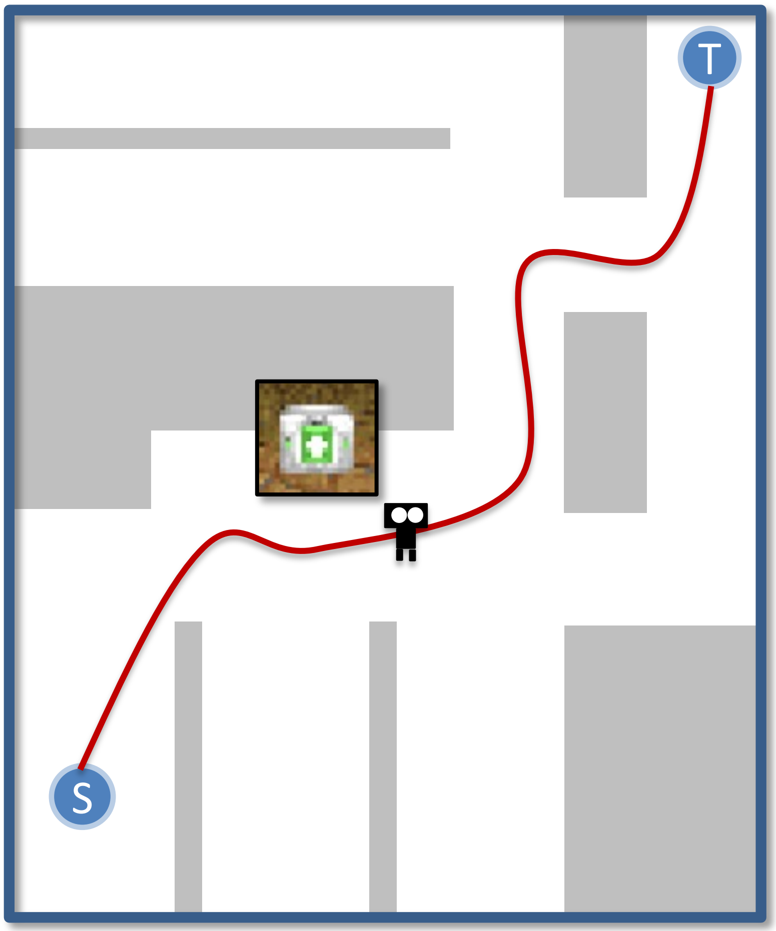
Task: ObjectNav



Task:
Find a target object given its object class.

Required reasoning:
- Recognize free navigation space
- *Explore the environment*
- *Recognize the object when seen and move towards it.*

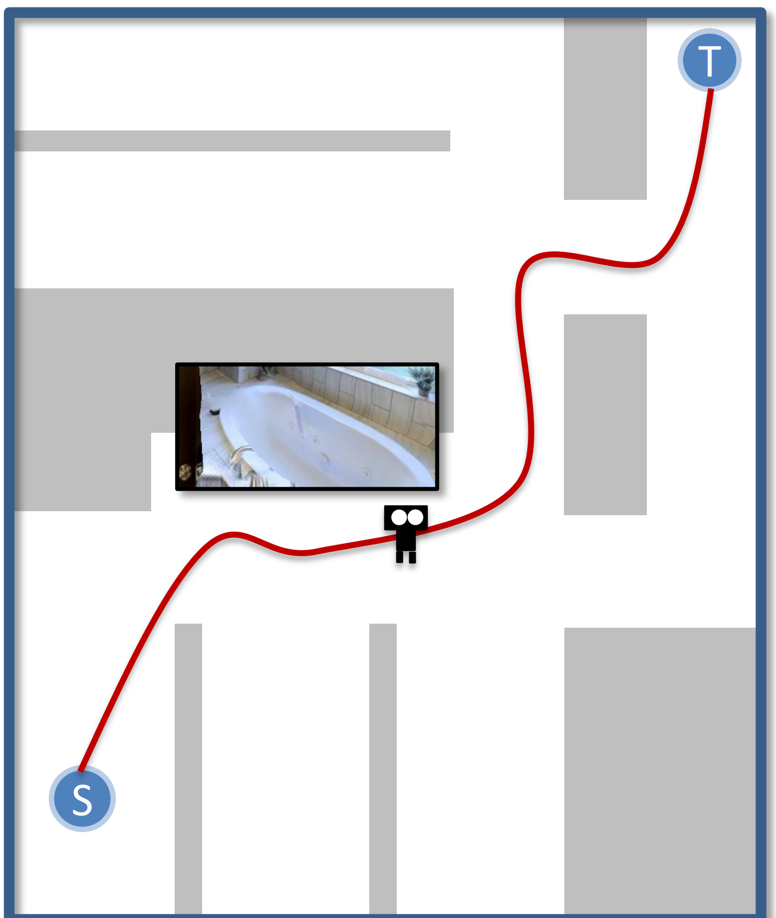
Task: ImageGoal (object)



Task:
Find a target object given its visual appearance (image).

Required reasoning:
- Recognize free navigation space
- Explore the environment
- Recognize the object when seen *by comparing it to a target image* and move towards it.

Task: ImageGoal (location)



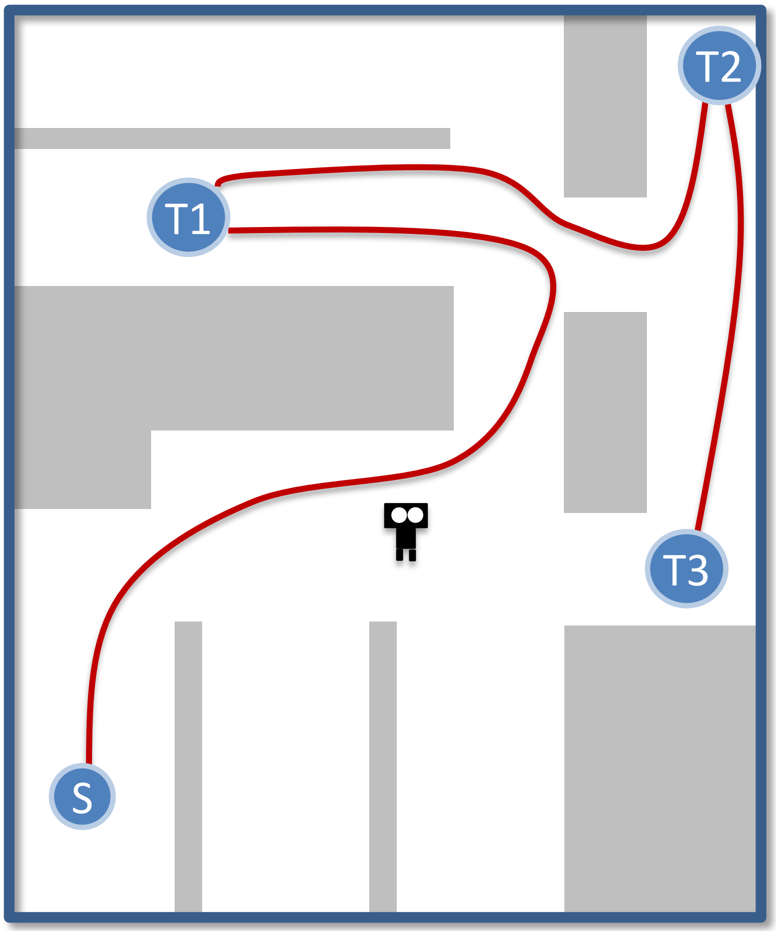
Task:

Find a target location given its visual appearance (image).

Required reasoning:

- Recognize free navigation space
- Explore the environment
- *Exploit spatial regularities, eg. room layouts*
- Recognize the location when seen and move towards it.

Task: K-item scenario



Task:
Navigate to a list of objects sequentially in the right order.

Required reasoning:
- Recognize free navigation space
- Explore the environment
- *Map an object if I need to find it later (!)*
- Recognize the object when seen and move towards it.

[Beeching, Dibangoye, Simonin, Wolf, ECML-PKDD 2020]

[Beeching, Dibangoye, Simonin, Wolf, ICPR 2020]

What do we want the model to learn?

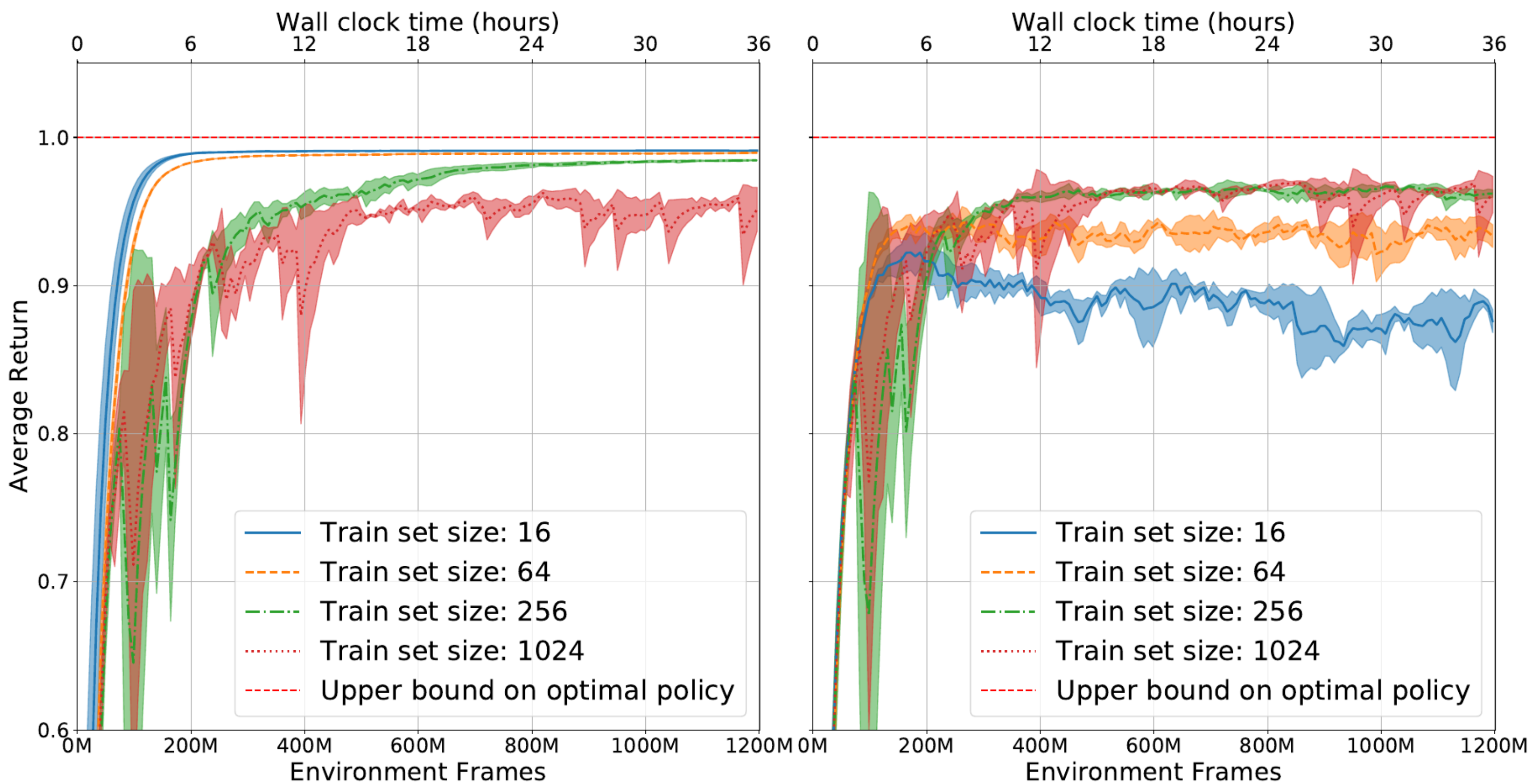
Generalization to new environments requires:

	Agent memory	Network parameters
Information on the (seen) environment, e.g. position of a couch	✓	✗
Positions of objects placed in the environment	✓	✗
Regularities of the environment (eg. bath tubs are in bathrooms; toilets are accessible from an aisle, not the living room)	✗	✓
Object affordances	✓	✓

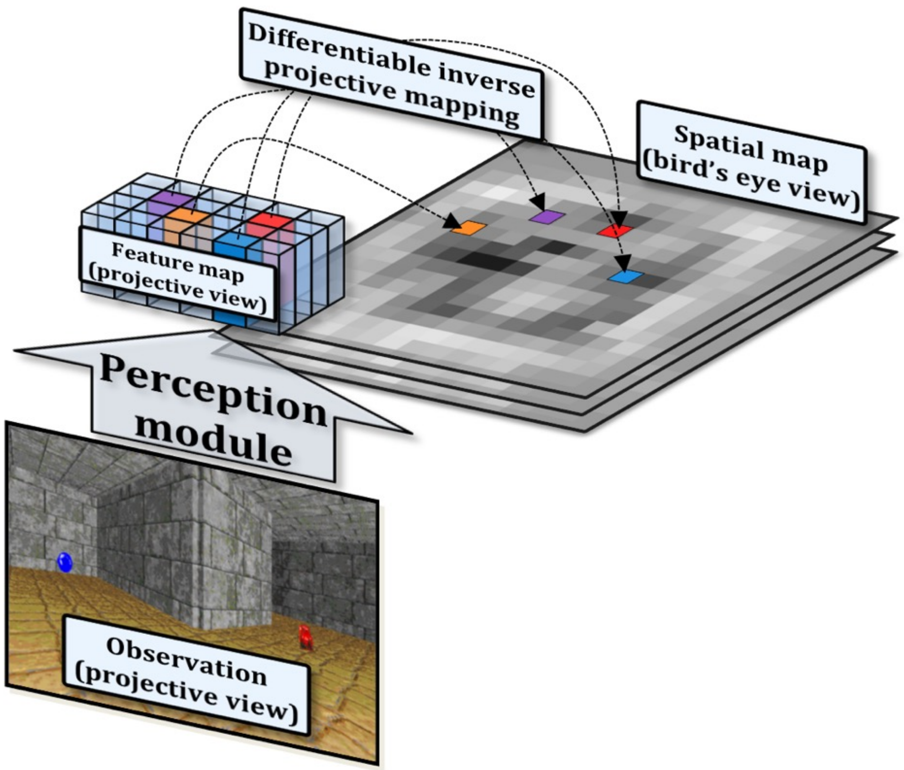
The task formulation decides how learned information is stored!

Tasks, regularities and generalization

What does my network learn:
A reasoning strategy, or the environment?

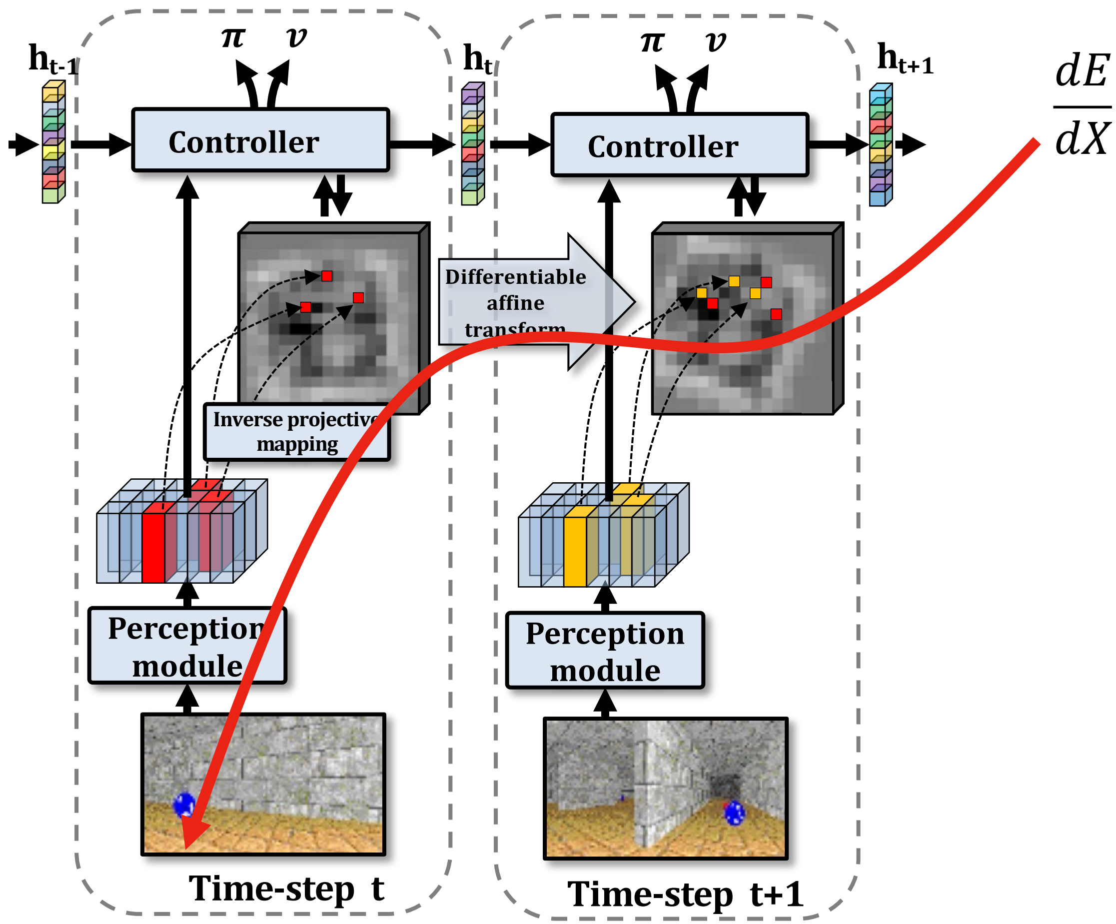


Inductive bias for projective mapping



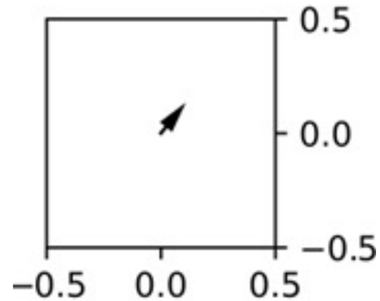
[Beeching, Dibangoye, Simonin, Wolf, ECML-PKDD 2020]

Spatial memory in Deep-RL

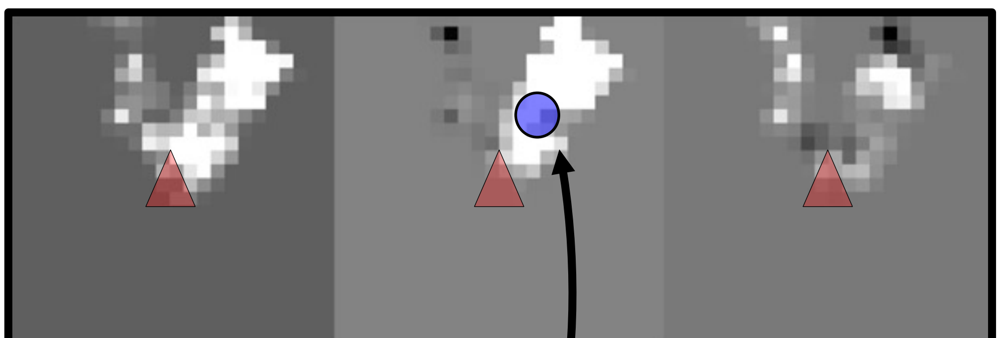


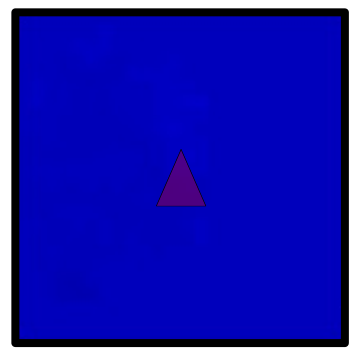
6 item scenario: time-step 005

Projective mapping of blue object



EgoMap: Query position



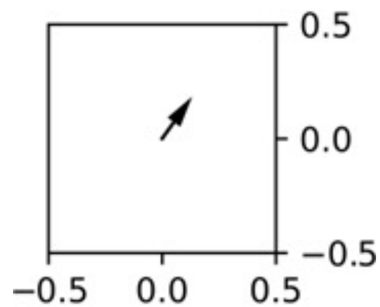


EgoMap: Attention

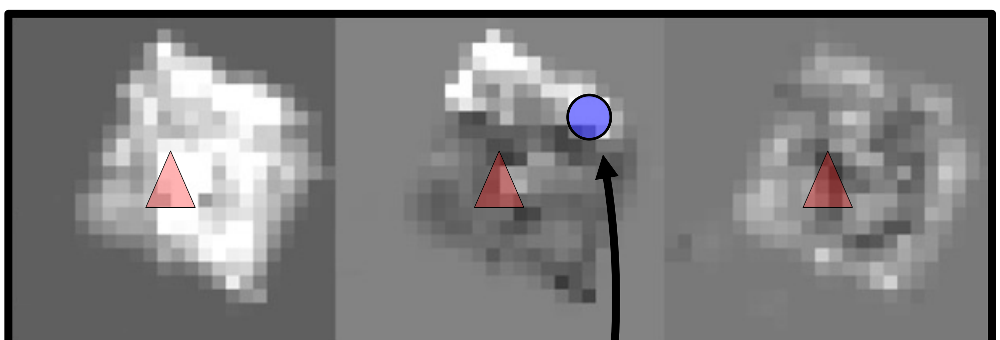
EgoMap: 3 Largest principal components

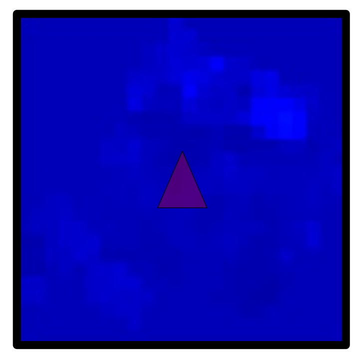
Object features retained in map

6 item scenario: time-step 105



EgoMap: Query position



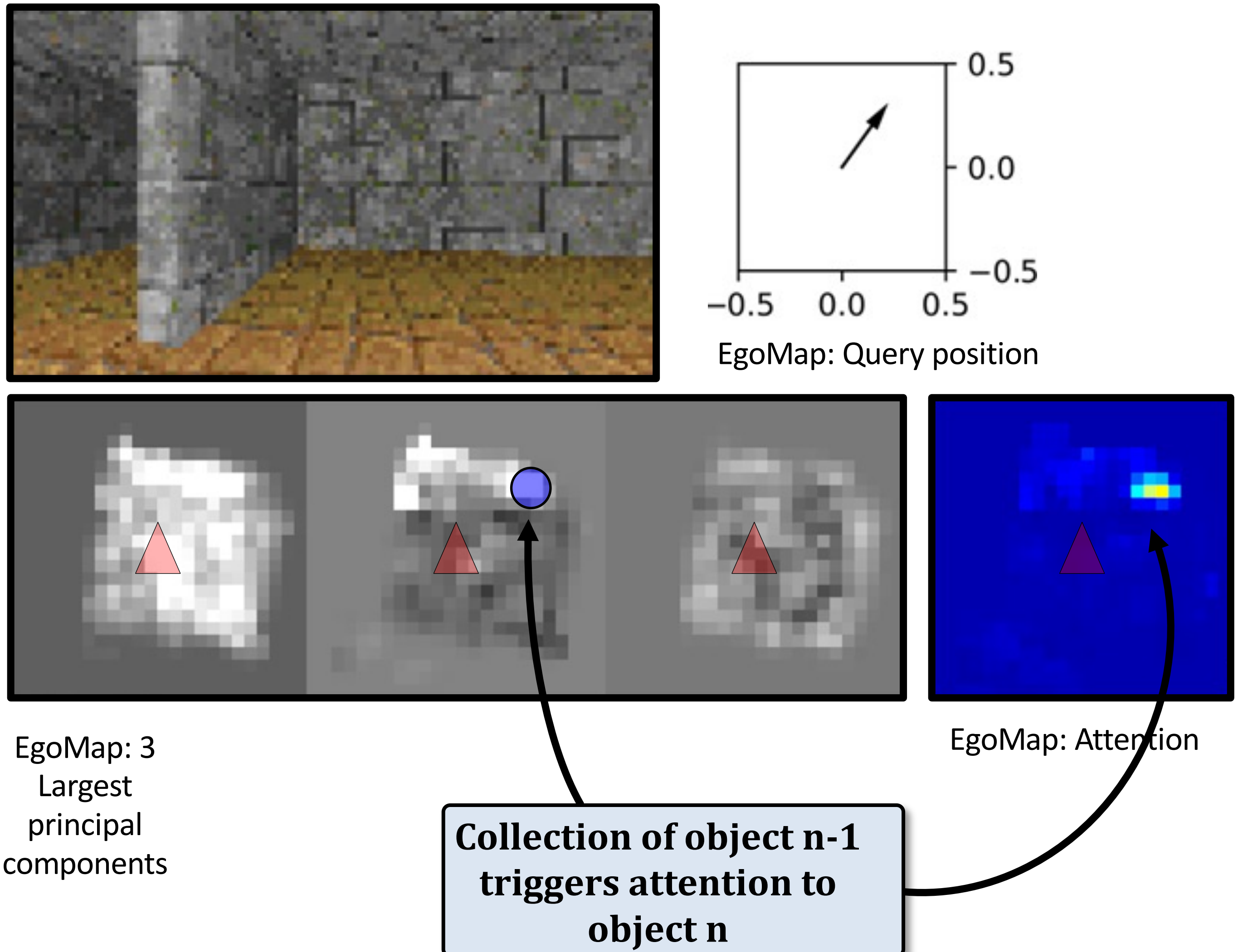


EgoMap: 3 Largest principal components

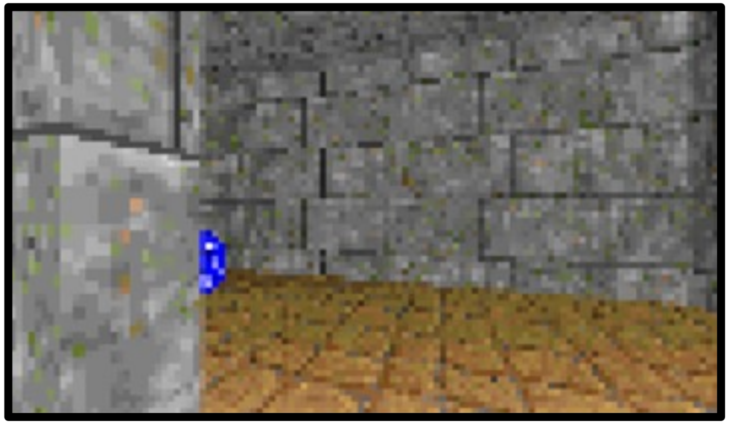
Object features retained in map

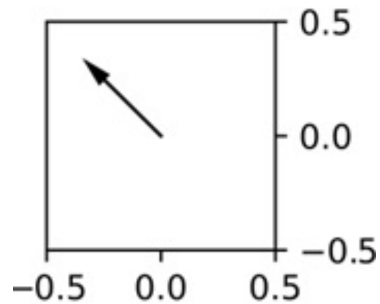
EgoMap: Attention

6 item scenario: time-step 108

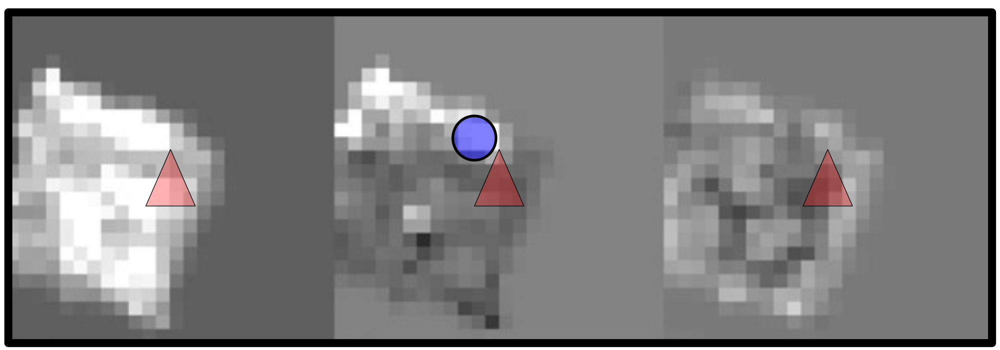


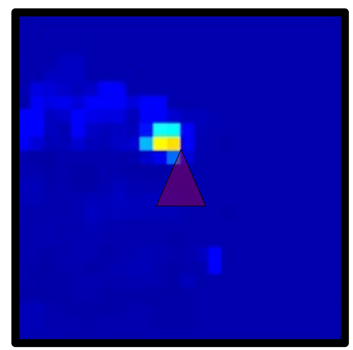
6 item scenario: time-step 134





EgoMap: Query position

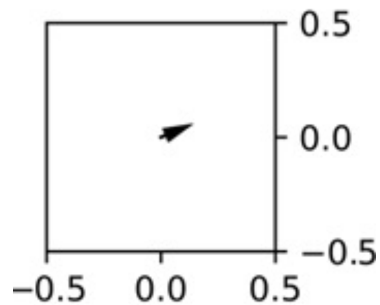




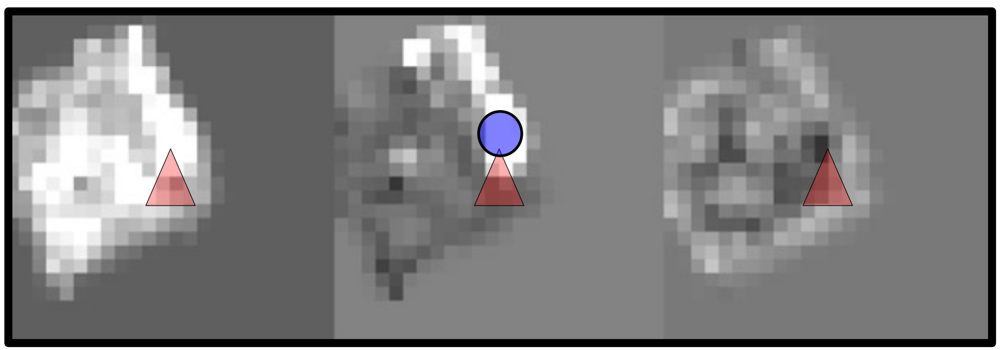
EgoMap: 3 Largest principal components

EgoMap: Attention

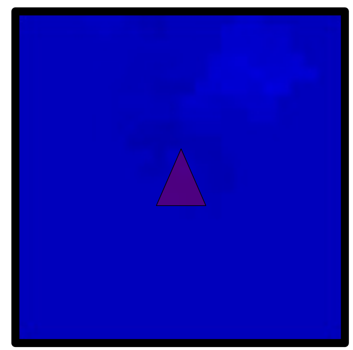
6 item scenario: time-step 140



EgoMap: Query position



EgoMap: 3 Largest principal components



EgoMap: Attention

When the object is not occluded, the agent does not attend to it

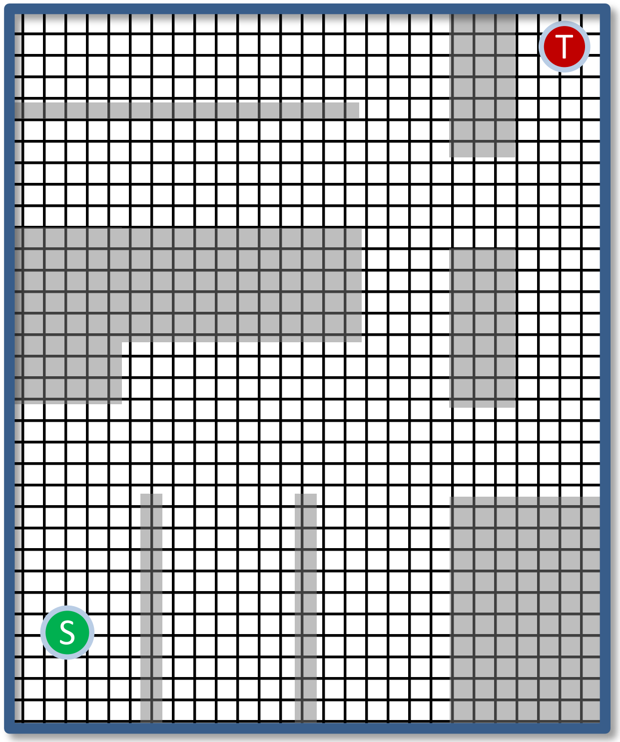
Quantitative results

	Scenario							
	4 item		6 item		Find and Return		Labyrinth	
Agent	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Random	-0.179	-0.206	-0.21	-0.21	-0.21	-0.21	-0.115	-0.086
Baseline	2.341 ± 0.026	2.266 ± 0.035	2.855 ± 0.164	2.545 ± 0.226	0.661 ± 0.003	0.633 ± 0.027	0.73 ± 0.02	0.694 ± 0.009
Neural Map	2.339 ± 0.038	2.223 ± 0.040	2.750 ± 0.062	2.465 ± 0.034	0.825 ± 0.070	0.723 ± 0.026	**0.769 ± 0.042**	0.706 ± 0.018
EgoMap	**2.398 ± 0.014**	**2.291 ± 0.021**	**3.214 ± 0.007**	**2.801 ± 0.048**	**0.893 ± 0.007**	**0.848 ± 0.017**	0.753 ± 0.002	**0.732 ± 0.016**
Optimum	2.5	2.5	3.5	3.5	1	1	1	1

Mapping objects required

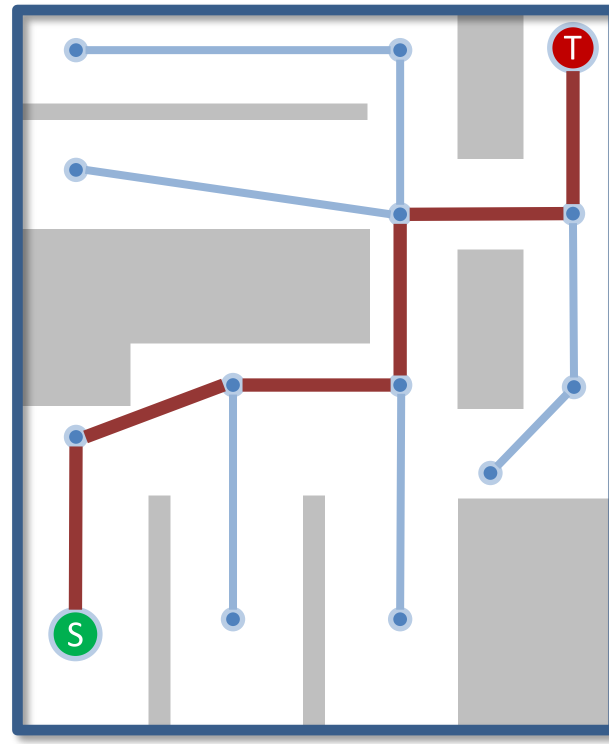
Not required

Spatial maps in robotics



Metric map
(=2D or 3D Grid)

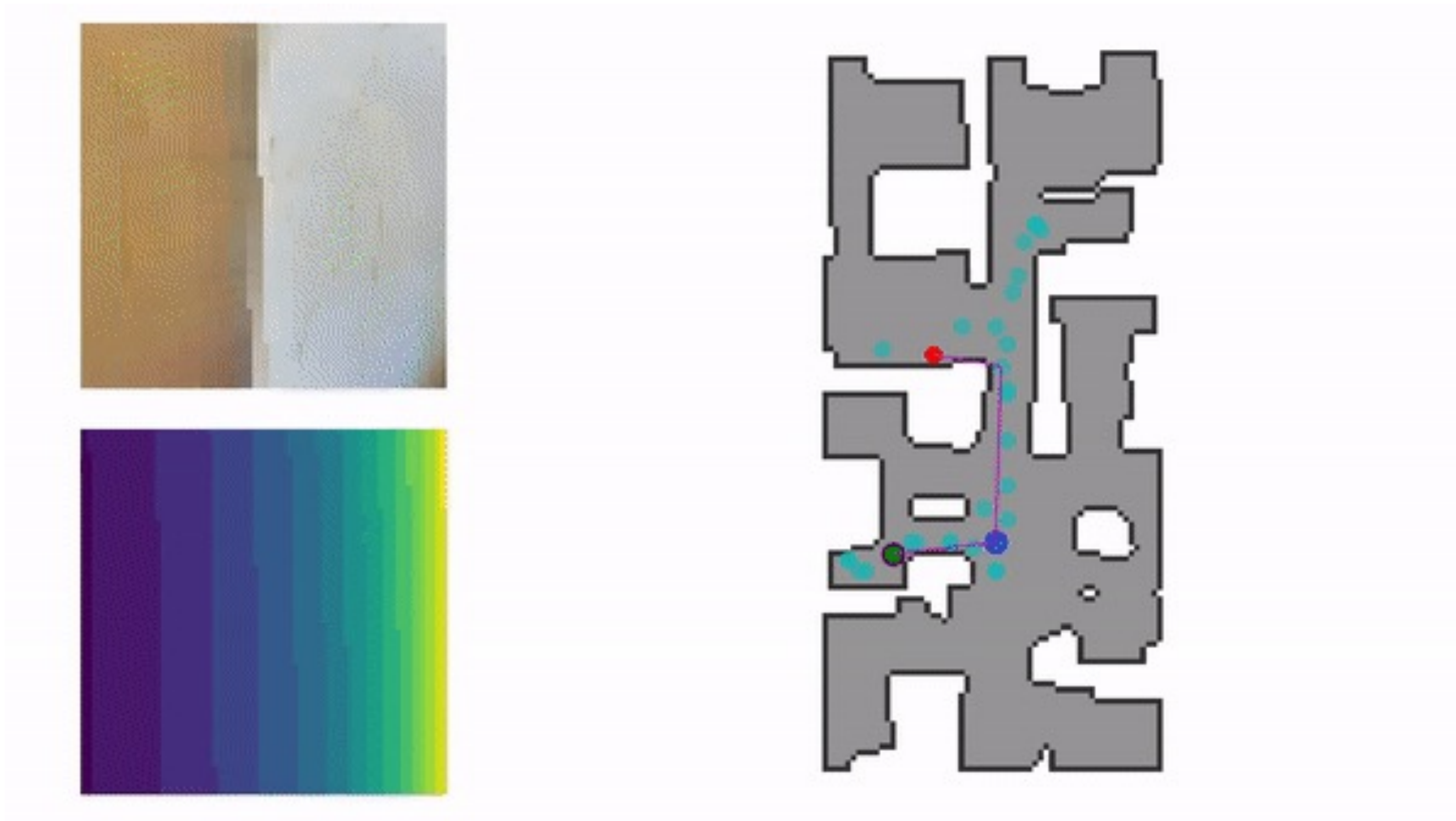
Beeching, Dibangoye, Simonin, Wolf,
EgoMap: Projective mapping and structured egocentric memory for Deep RL,
ECML-PKKD 2020



Topological map
(=Graph)

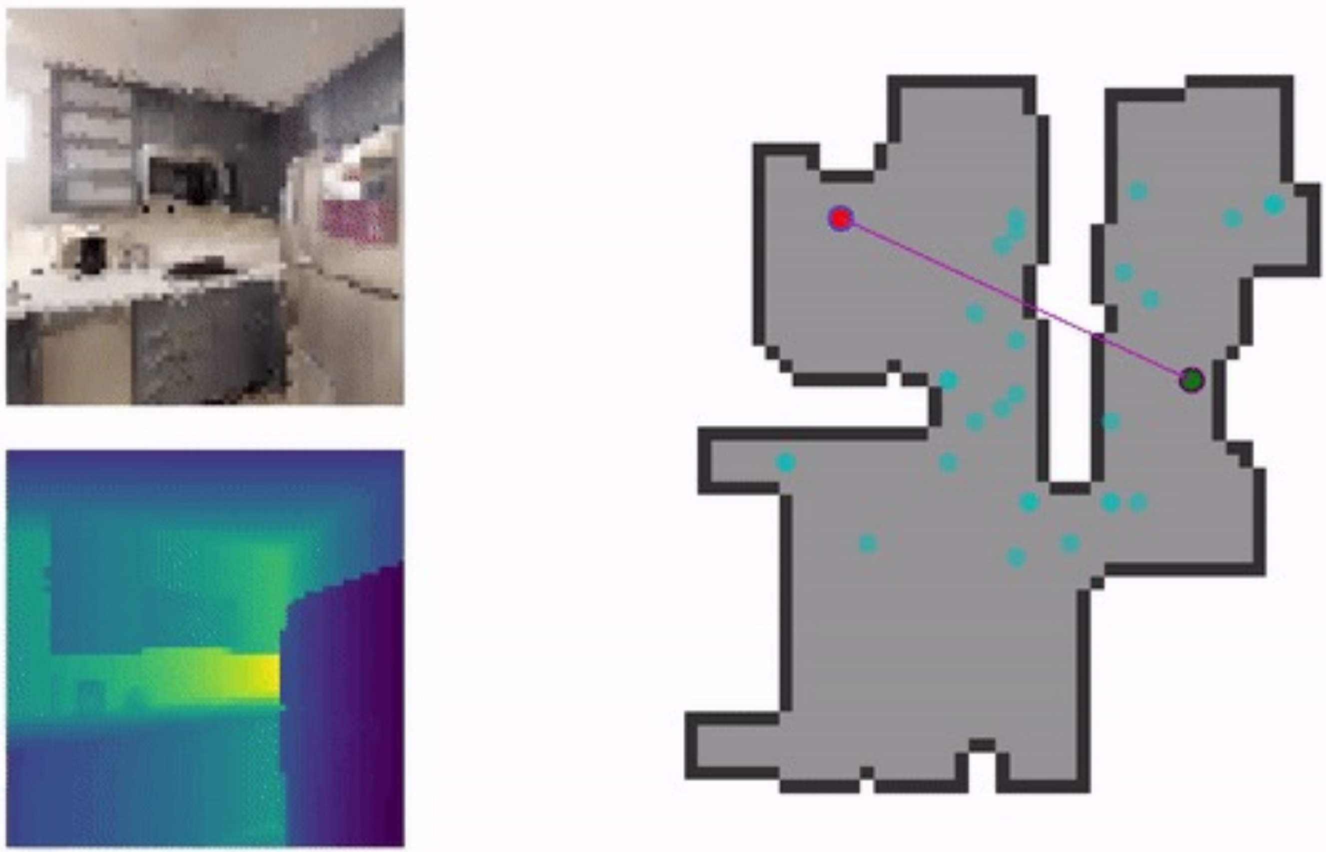
Beeching, Dibangoye, Simonin, Wolf,
Learning to plan with uncertain topological maps,
ECCV 2020

Hierarchical planning and control



[Beeching, Dibangoye, Simonin, Wolf, ECCV 2020]

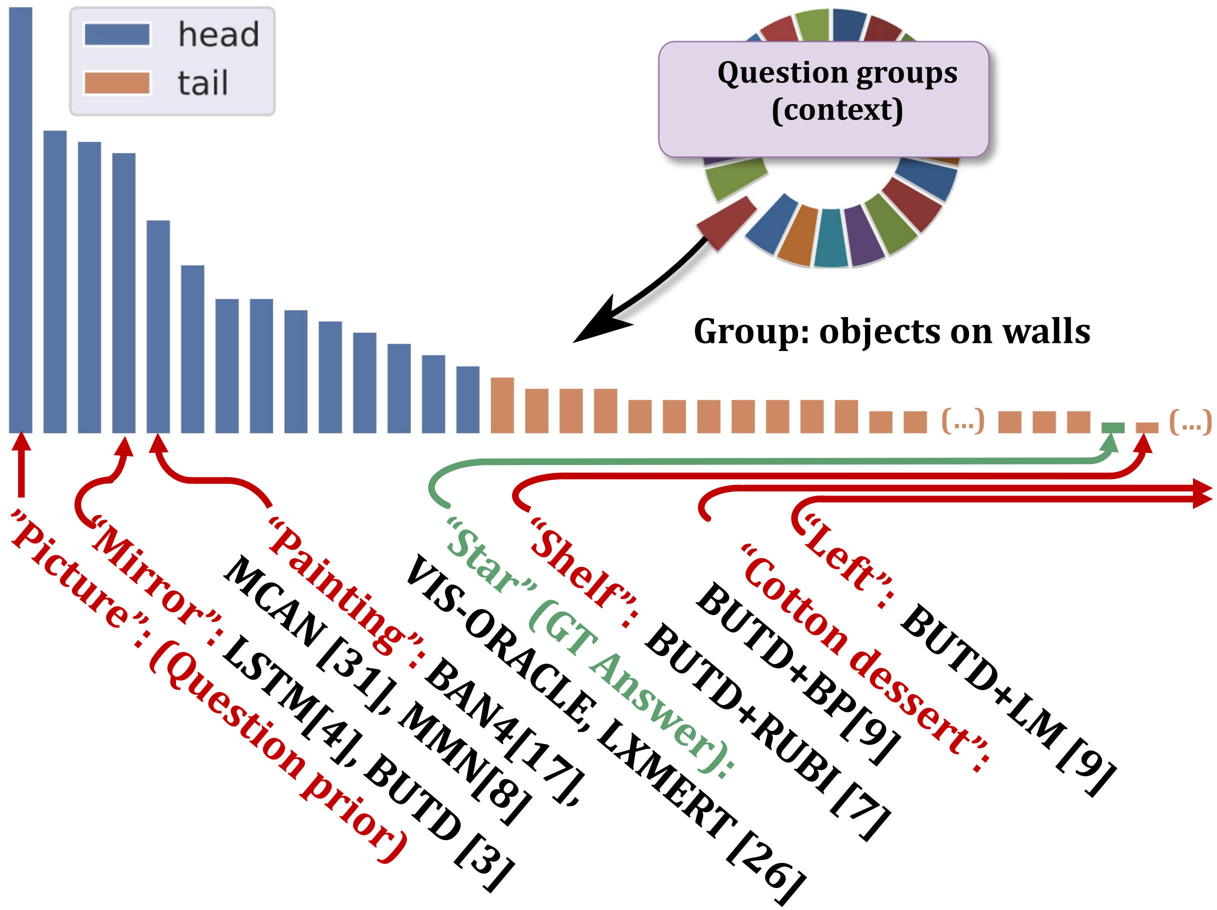
Failure Case



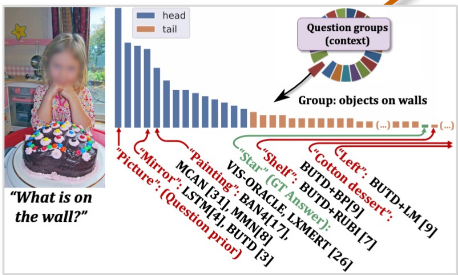
[Beeching, Dibangoye, Simonin, Wolf, ECCV 2020]

Visual Question Answering

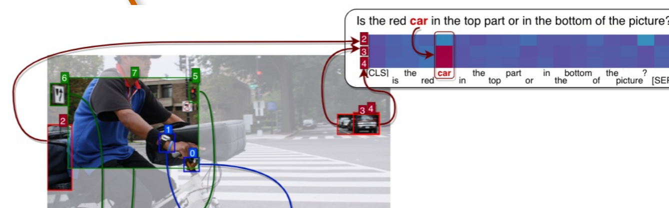
“What is on the wall?”



How can we evaluate biases in learning? (CVPR 2021a)

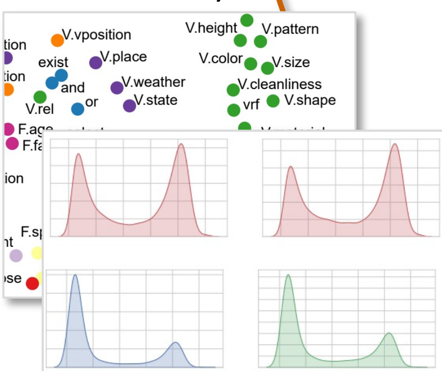


Can we ground object detection through language (under preparation)



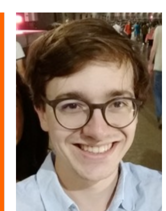
VQA

How can we visualize and transfer reasoning? (CVPR 2021b)



Can we weakly supervise word-object alignment? (ECAI 2020)

Can we supervise reasoning programs? (under preparation)



Corentin Kervadec

Grigory Antipov

Moez Baccouche

Christian Wolf

(Intermediate) conclusion

- Our objective is to train agents to reason from large-scale datasets, avoiding shortcuts:
  - Creation of tasks and auxiliary losses
  - We imbue neural networks with inductive biases
  - Visualization of reasoning patterns
  - Multi-modal inputs
  - Learned spatial representations