Robot Learning Laboratory

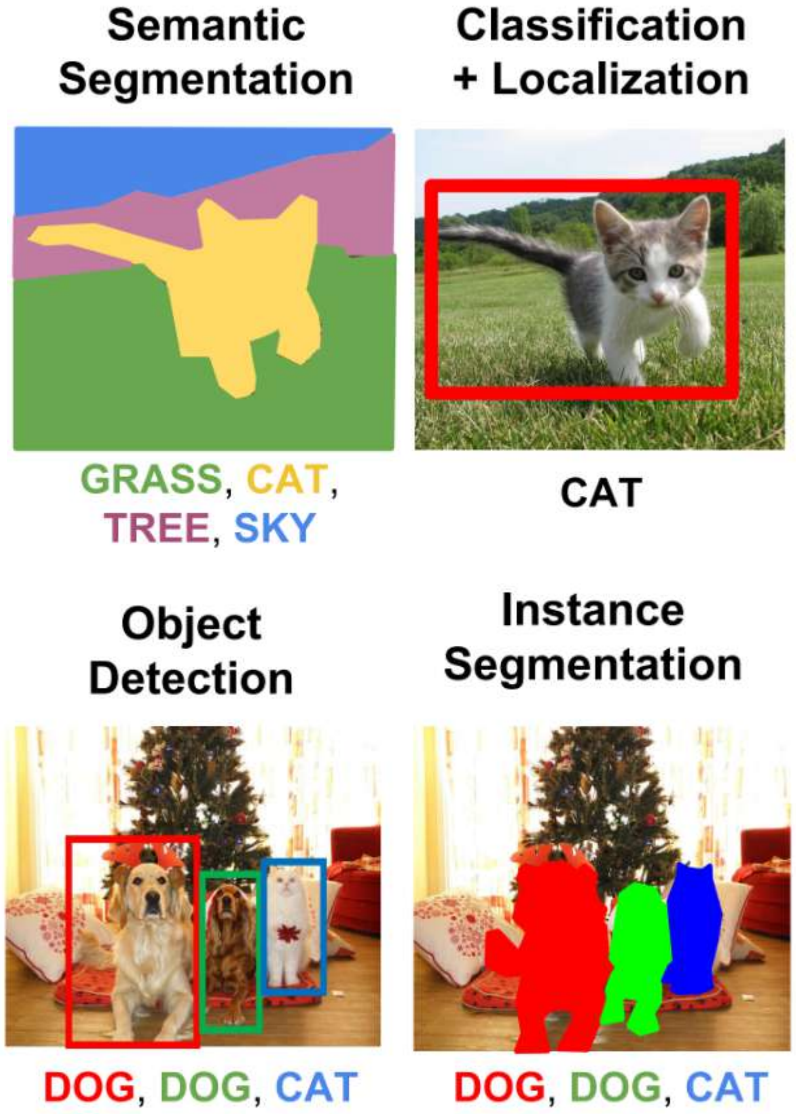
Semantic Visual Navigation for Embodied Agents: A Graph-Based Approach

Nuri Kim

Seoul National University

February 2, 2023

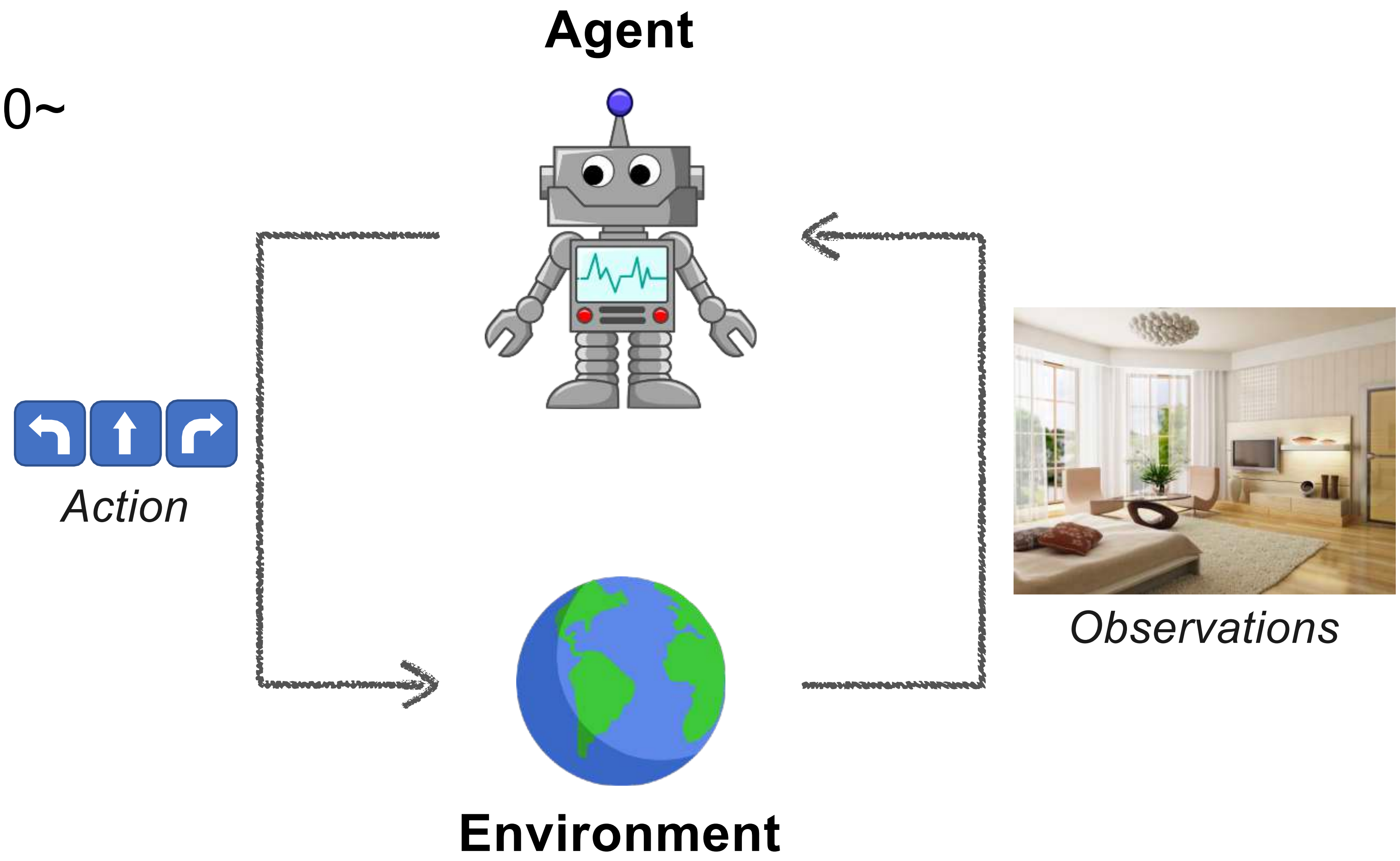
Visual Intelligence: Passive Learning



Visual Intelligence: Interactive Learning

RLLAB
Robot Learning Laboratory

From 1970~



Visual Intelligence: Interactive Learning

RLLAB
Robot Learning Laboratory

source: Habitat

Visual Intelligence: Interactive Learning

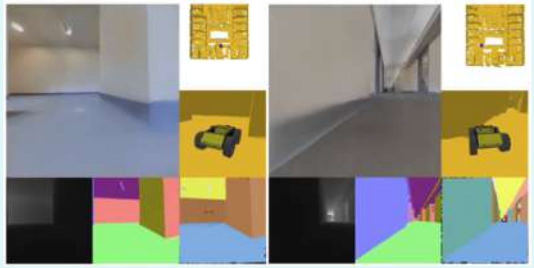
source: D. Klein, P.Abbeel

Visual Intelligence: Interactive Learning

RLLAB
Robot Learning Laboratory

Simulators

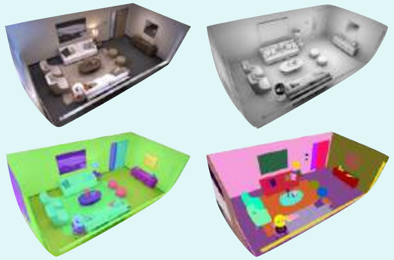
AI2-THOR (Kolve et al. 2017)



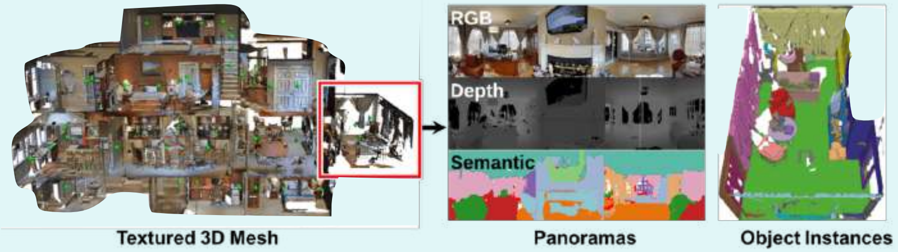
Gibson (Zamir et al. 2018)

Habitat (Savva et al. 2018)

Datasets



Replica (Straub et al. 2019)

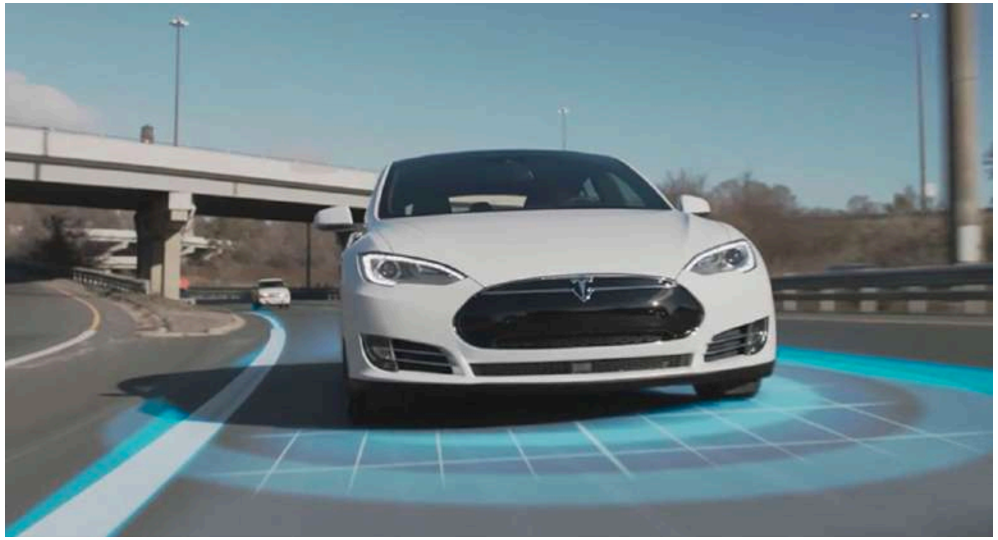


Matterport3D (Chang et al. 2017)

HM3D (Ramakrishnan et al. 2021)

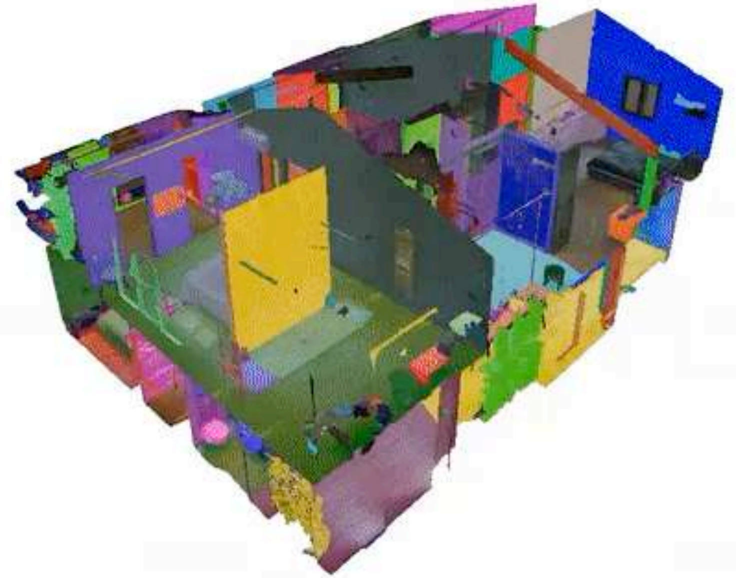
source: Habitat-Sim

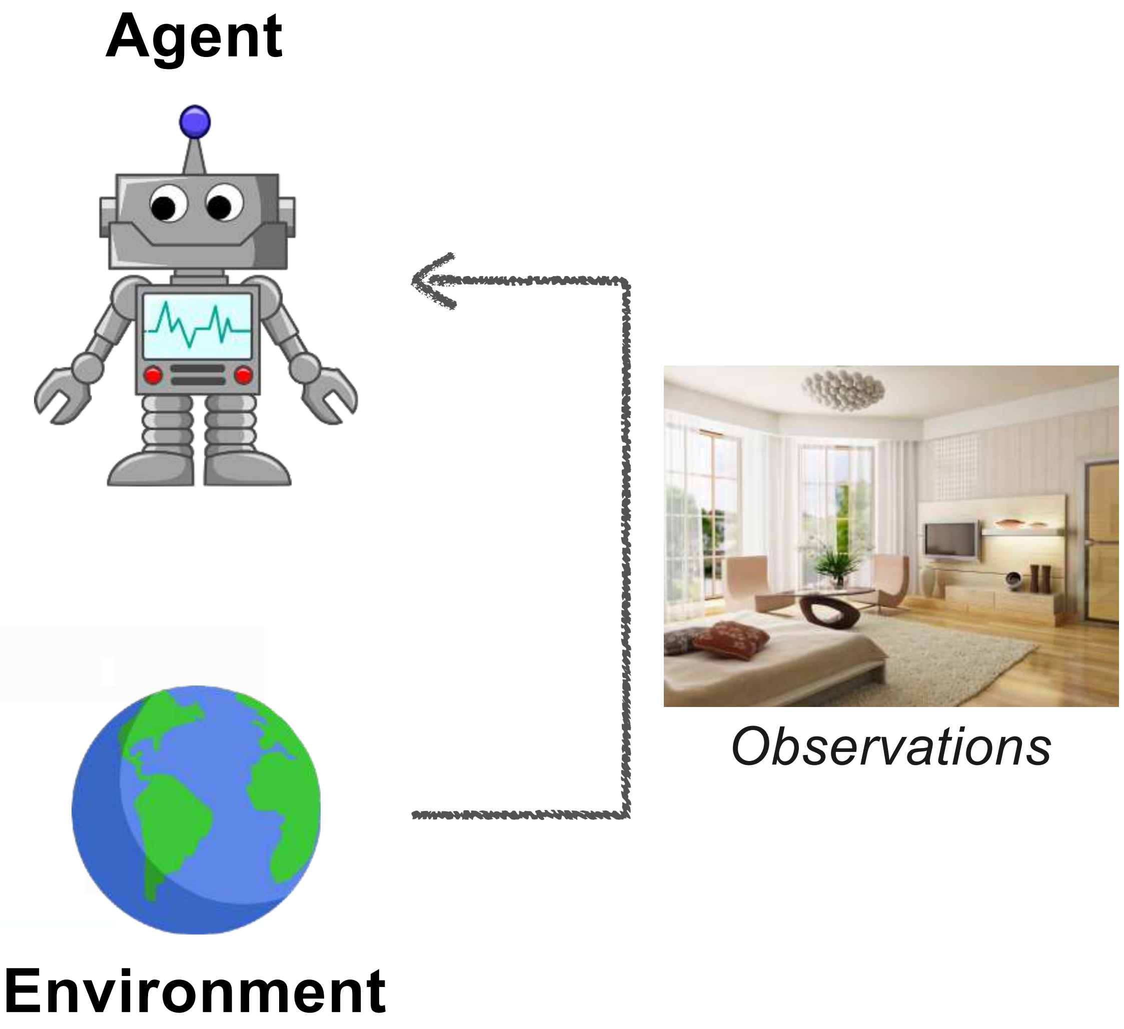
Visual Intelligence: Interactive Learning



Semantic Understanding

RILAB
Robot Learning Laboratory





Relationship of Data

RLLAB
Robot Learning Laboratory

Target

Building navigation agents capable of *semantic understanding* by learning *relationship* of data using *graphs*

Roadmap

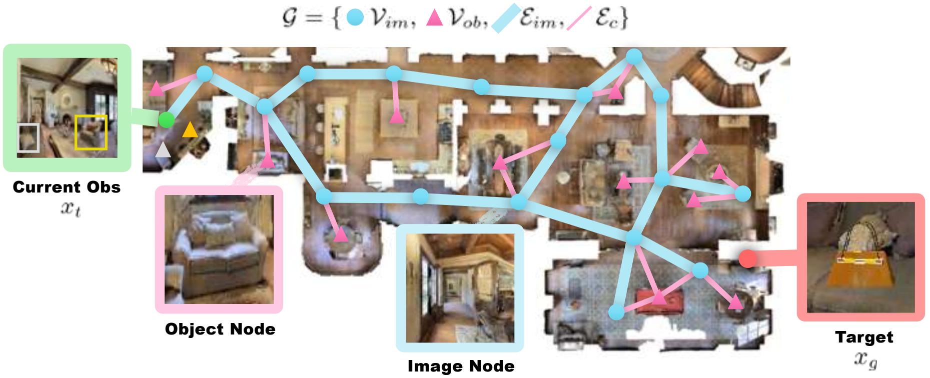
RLLAB
Robot Learning Laboratory

Passive Learning

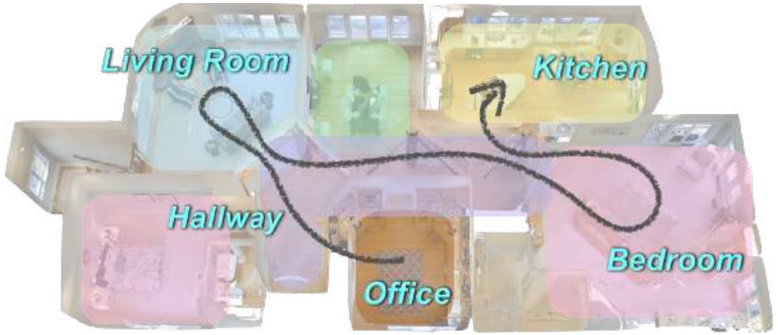
Instance-Aware Detection

Interactive Learning

Topological Semantic Graph Memory



Relational Semantic Visual Graph



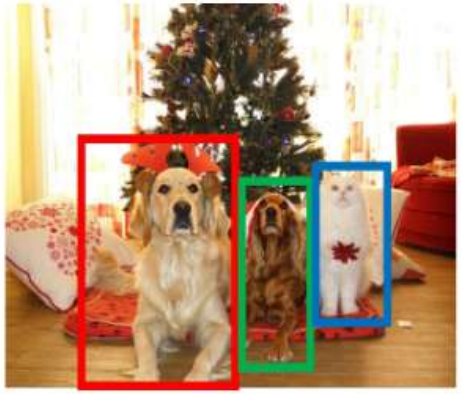
Roadmap

RILAB
Robot Learning Laboratory

Passive Learning

Interactive Learning

Object Detection



DOG, DOG, CAT

CVIU 2020

Image Goal

ICCV 2021

CoRL 2022 (oral)

Object Goal

Chair

TV

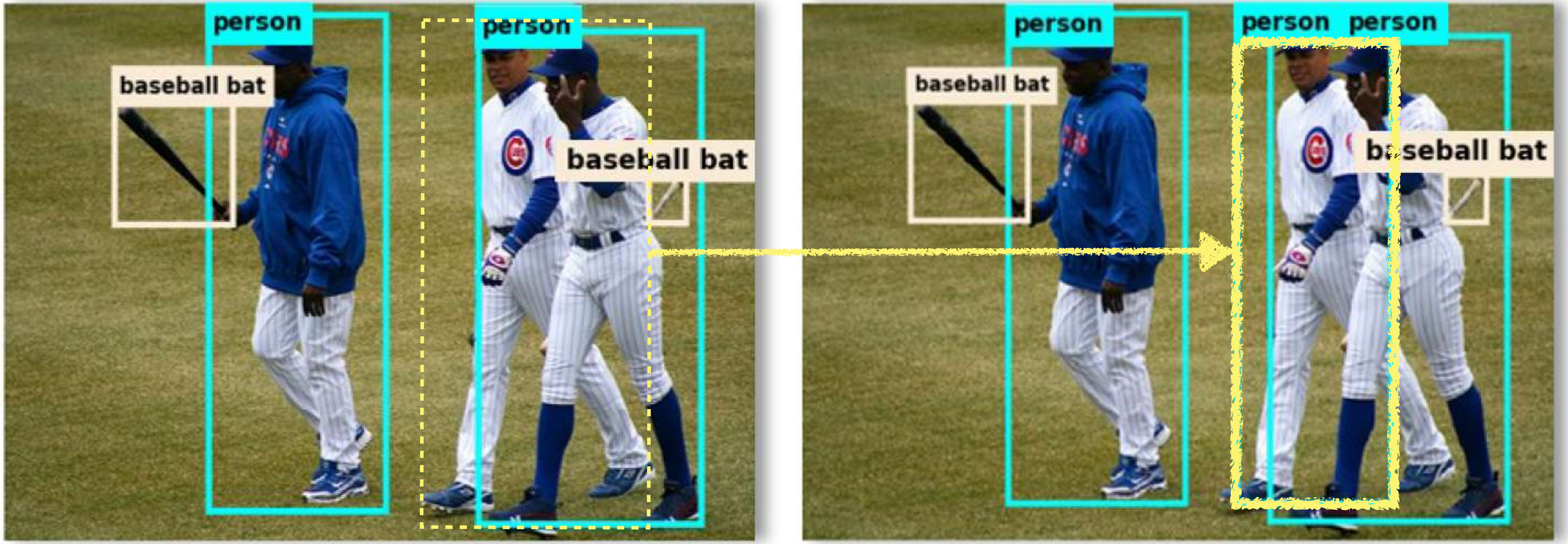
Sofa

CVPR 2023 (submitted)

Nuri Kim, Donghoon Lee, and Songhwai Oh., "**Learning Instance-Aware Object Detection Using Determinantal Point Processes**," Computer Vision and Image Understanding (CVIU-20)

Detection on Crowd Scene

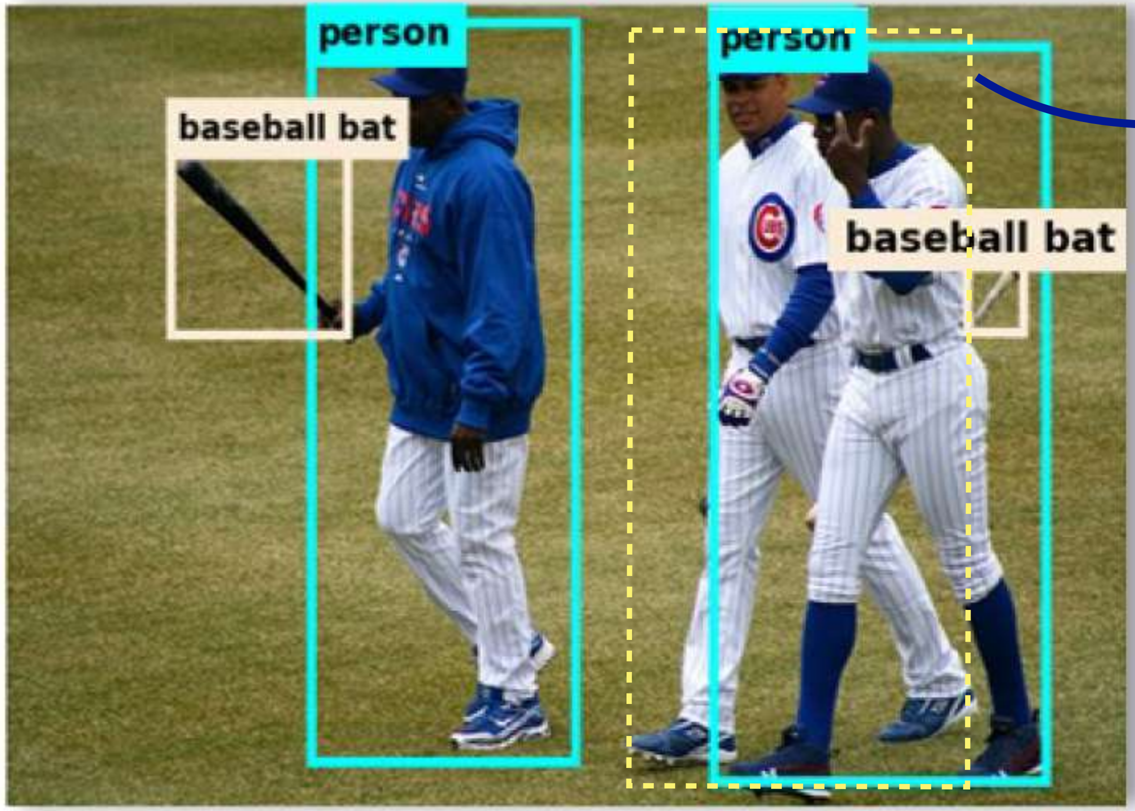
Goal: Find **individual instances** when they are ***overlapped***



Detection on Crowd Scene

RILAB
Robot Learning Laboratory

Non Maximum Suppression (NMS[1]) : Not robust for detecting overlapped objects



Detection results from an object detector with ***NMS***

Missing detections *due to overlapped bounding boxes.*

Algorithm 1 Non-Max Suppression

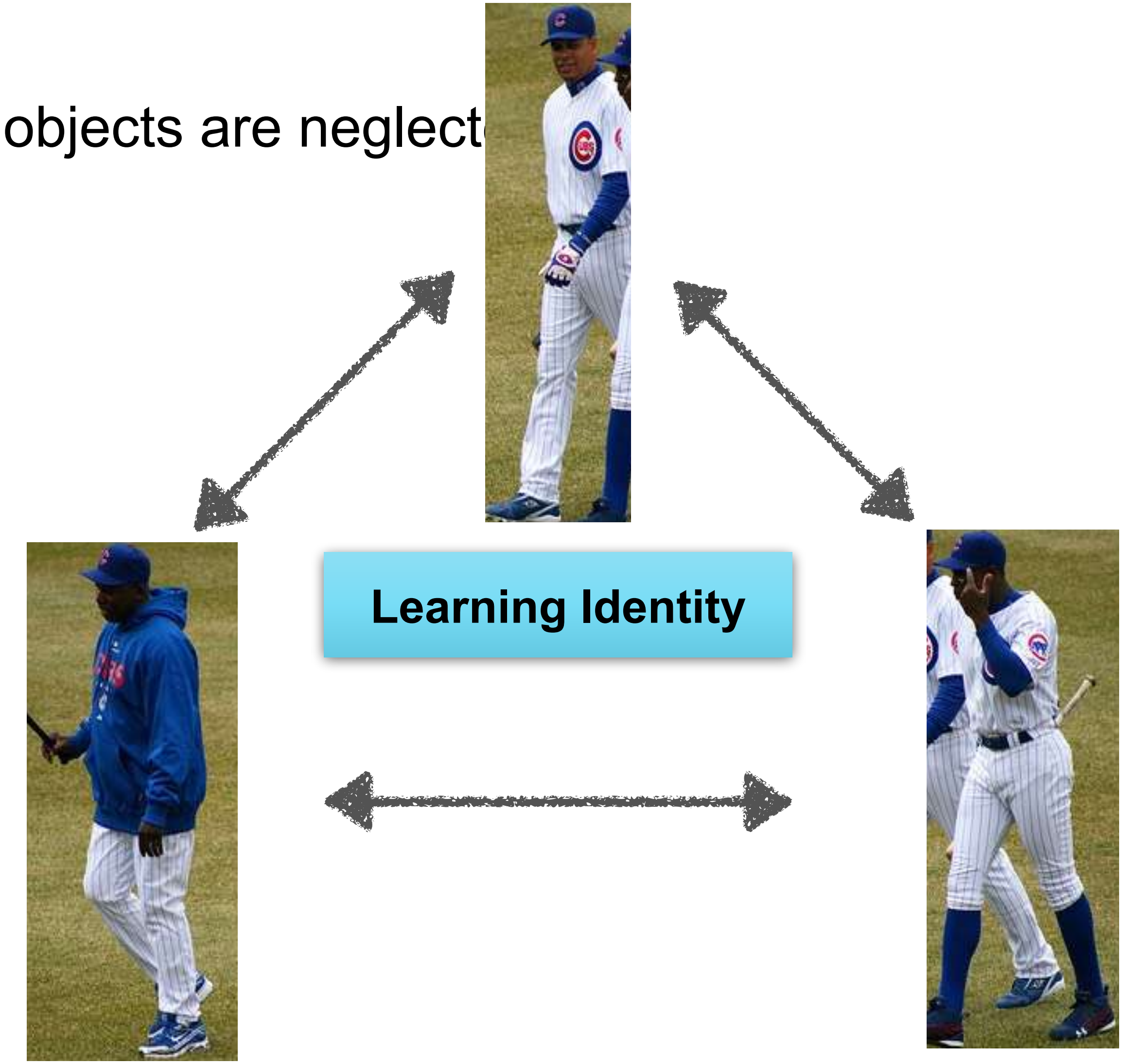
procedure NMS(B,c)
  $B_{nms} \leftarrow \emptyset$
  for $b_i \in B$ **do**
    $discard \leftarrow$ False
    for $b_j \in B$ **do**
      if same$(b_i, b_j) > \boldsymbol{\lambda}_{\mathbf{nms}}$ **then**
        if score$(c, b_j) >$ score(c, b_i) **then**
          $discard \leftarrow$ True
    if not $discard$ **then**
      $B_{nms} \leftarrow B_{nms} \cup b_i$
  return B_{nms}

[1] Neubeck, Alexander, and Luc Van Gool. "**Efficient non-maximum suppression.**" International Conference on Pattern Recognition (ICPR). 2006.

Learning Identity

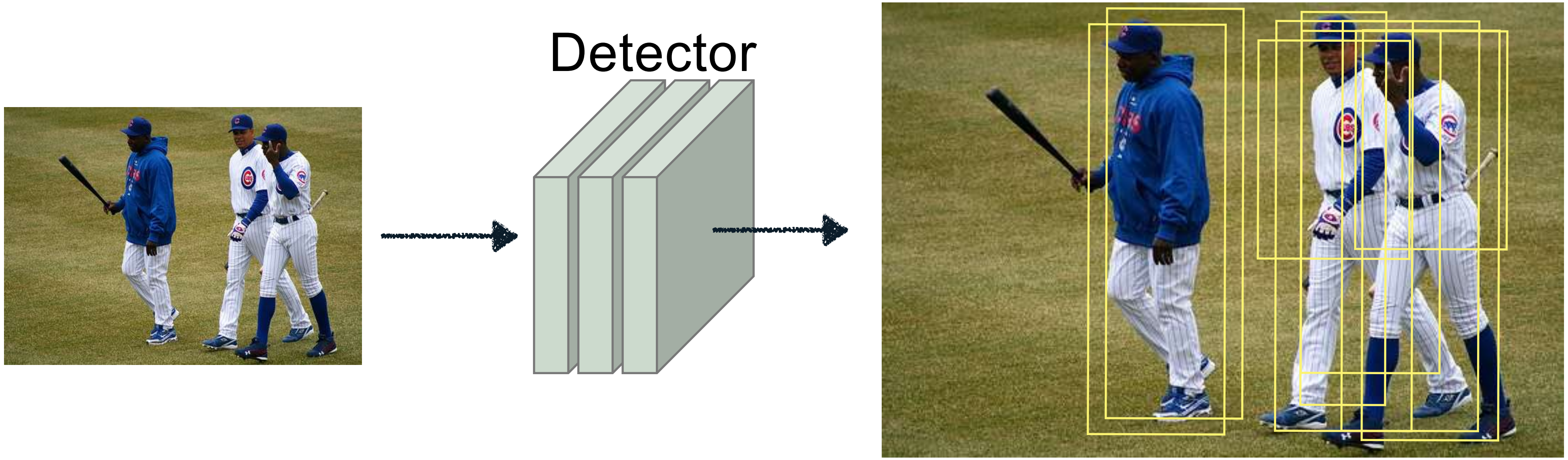
Robot Learning Laboratory

Non Maximum Suppression (NMS): Overlapped objects are neglect



Learning Identity

RLLAB
Robot Learning Laboratory

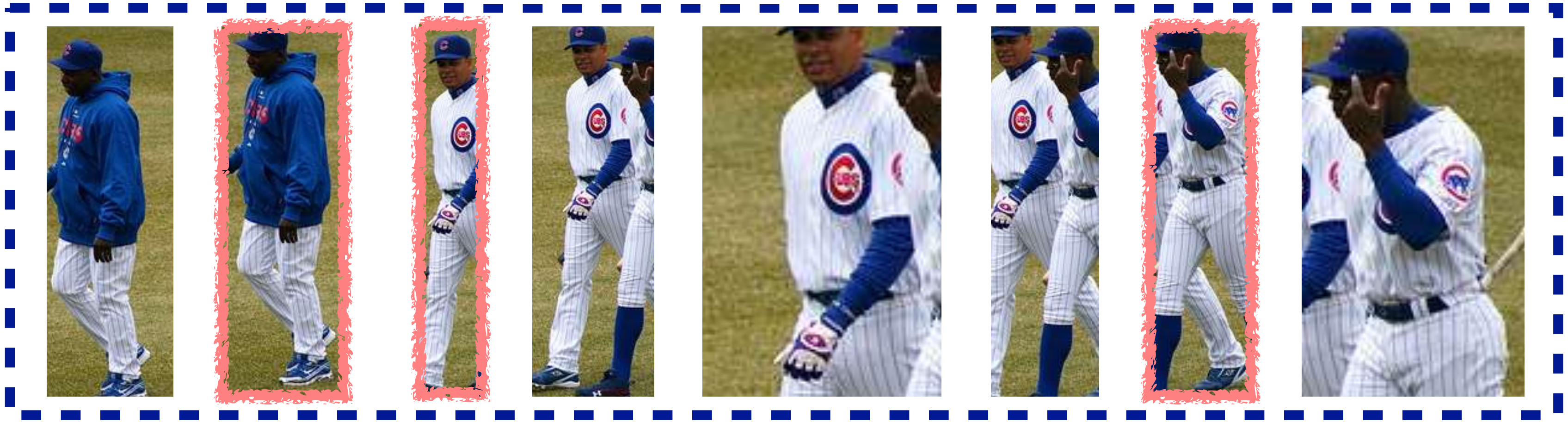


Object Candidates

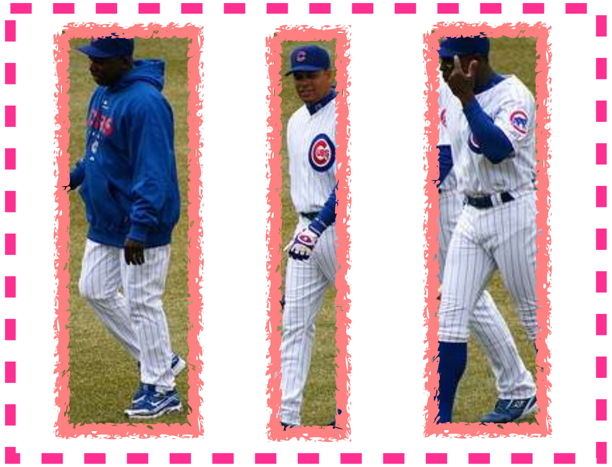
Learning Identity

RLLAB
Robot Learning Laboratory

Object Candidates ($\mathcal{Y}$)



Y_{rep}



representative set

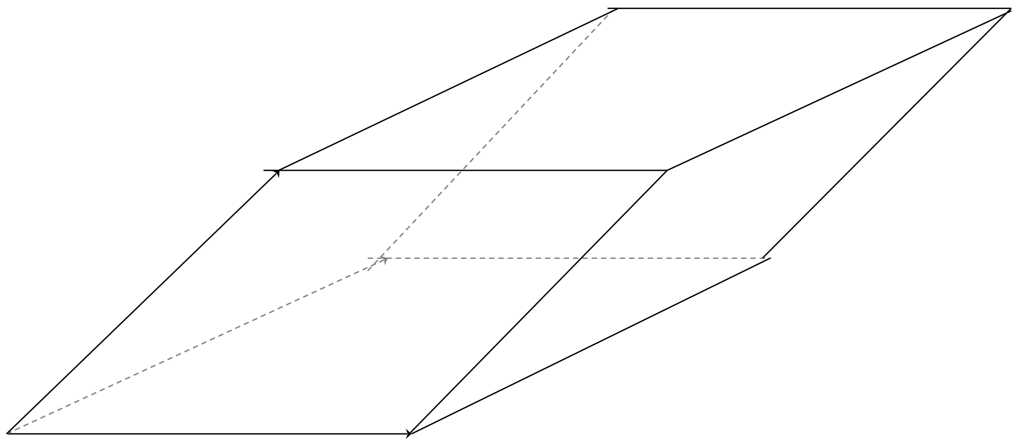
Learning **Instanceness** by increasing the **DPP** probability of **representative set**.

Learning Identity

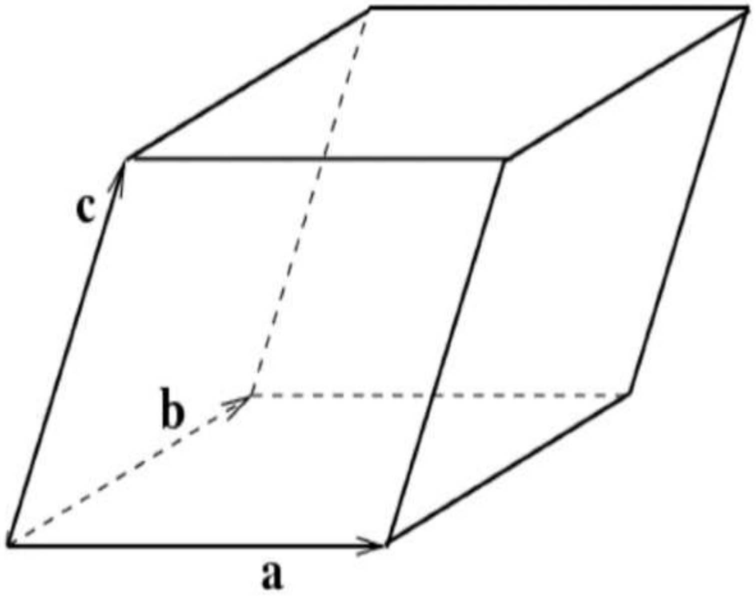
RILAB
Robot Learning Laboratory

Determinantal Point Processes

For selecting diverse and qualitative objects



<



The volume of vectors are bigger as the vectors are **diverse** and **qualitative**.

Learning Identity

RLLAB
Robot Learning Laboratory

- Determinantal point process (DPP) defines probability to every subset of a finite set $\mathcal{S} = \{1,\ldots,N\}$ of cardinality $|\mathcal{S}| = N$.
- The kernel $\mathbf{L}$ is defined using quality (Q) and similarity (S) matrices.

$$\mathbf{L} = Q \odot S = qq^T \odot \phi\phi^T,$$

where $q \in \mathbb{R}_+^N, \phi \in \mathbb{R}^{N\times D}$ and S is a cosine similarity matrix.

- Based on a positive semi-definite kernel $\mathbf{L} \in \mathbb{R}^{N\times N}$, the probability of selecting a set Y is,

$$\mathcal{P}(Y) = \frac{\det(\mathbf{L}_Y)}{\sum_{A\subseteq S}\det(\mathbf{L}_A)} = \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{L}+\mathbf{I})},$$

☑ **Exponential to polynomial**

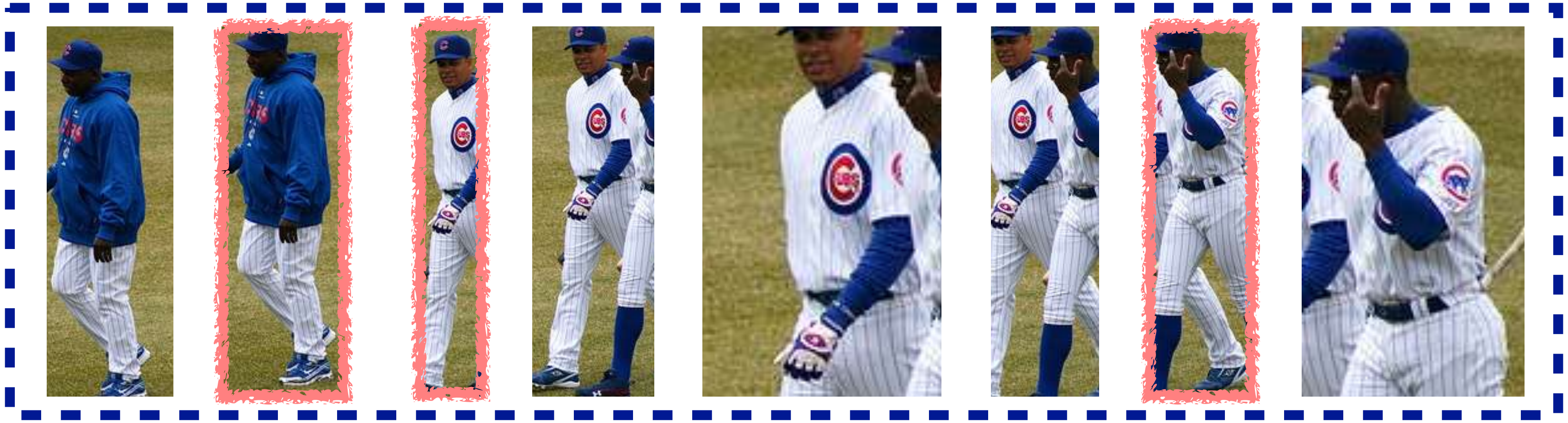
where $\mathbf{L}_Y$ is a submatrix of $\mathbf{L}$ indexed by elements in Y.

Kulesza, Alex, and Ben Taskar. "Determinantal point processes for machine learning." Now Publishers Inc *(2012)*.

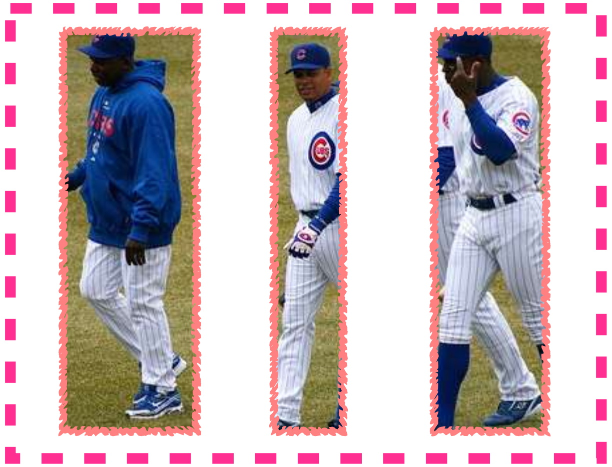
Learning Identity

RILAB
Robot Learning Laboratory

Object Candidates ($\mathscr{Y}$)



Y_{rep}



DPP

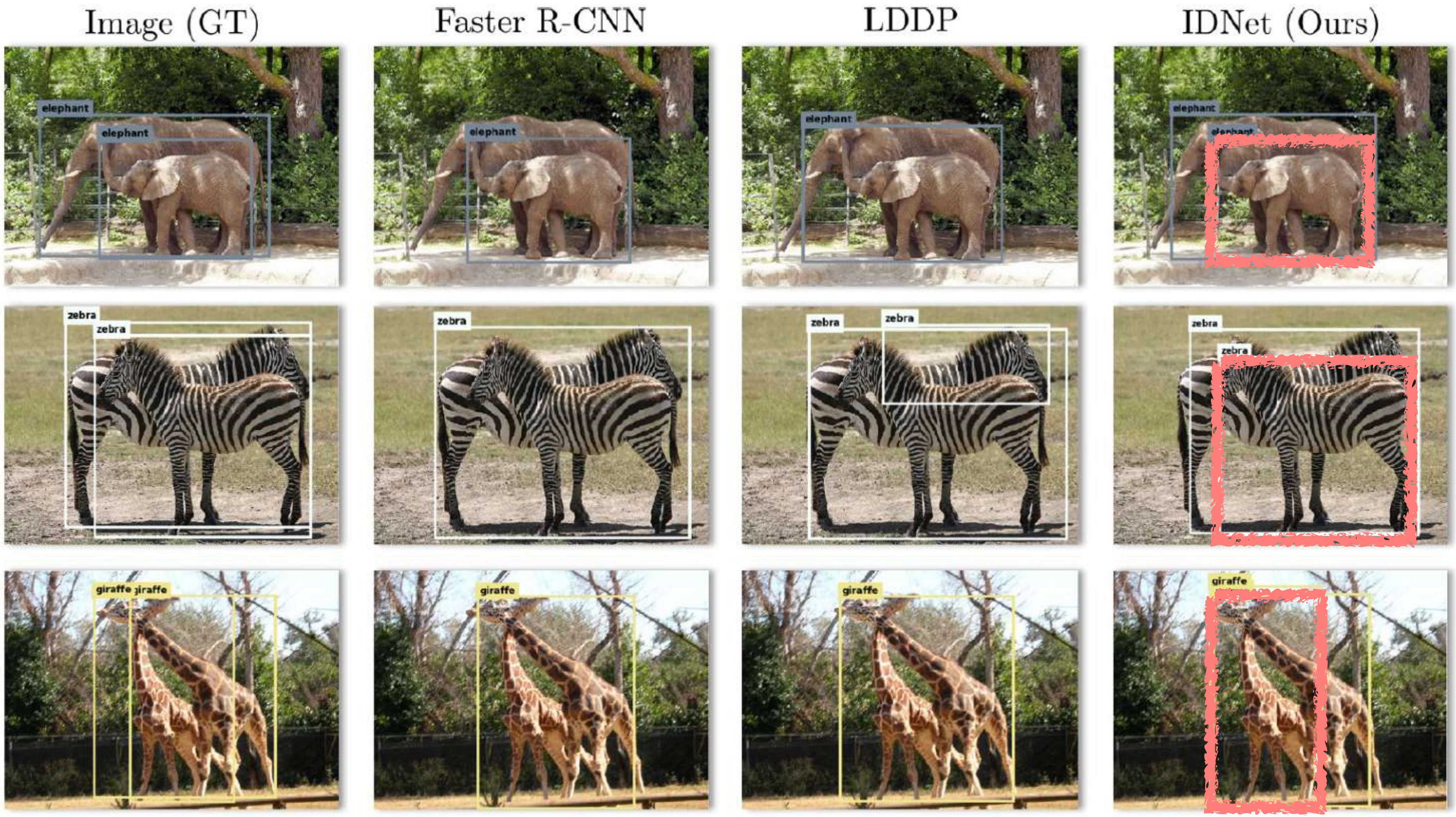
$$\mathscr{P}(Y) = \frac{\det(\mathbf{L}_Y)}{\sum_{A \subseteq S} \det(\mathbf{L}_A)} = \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{L} + \mathbf{I})}$$

$$Loss_{ID}(Y_{rep}, \mathscr{Y}) = -\log(\mathscr{P}_{\mathbf{L}_{\mathscr{Y}}}(Y_{rep})) = -\log\left(\frac{\det(\mathbf{L}_{Y_{rep}})}{\det(\mathbf{L}_{\mathscr{Y}} + \mathbf{I}_{\mathscr{Y}})}\right)$$

$$= -\operatorname{logdet}(\mathbf{L}_{Y_{rep}}) + \operatorname{logdet}(\mathbf{L}_{\mathscr{Y}} + \mathbf{I}_{\mathscr{Y}})$$

Results of Learning Identity

RLLAB
Robot Learning Laboratory



Successfully detected overlapped instances

Learning Correct Category

Robot Learning Laboratory

*Decision: Dog **and** horse*

$\mathcal{Y}_m$

Y_{pos}

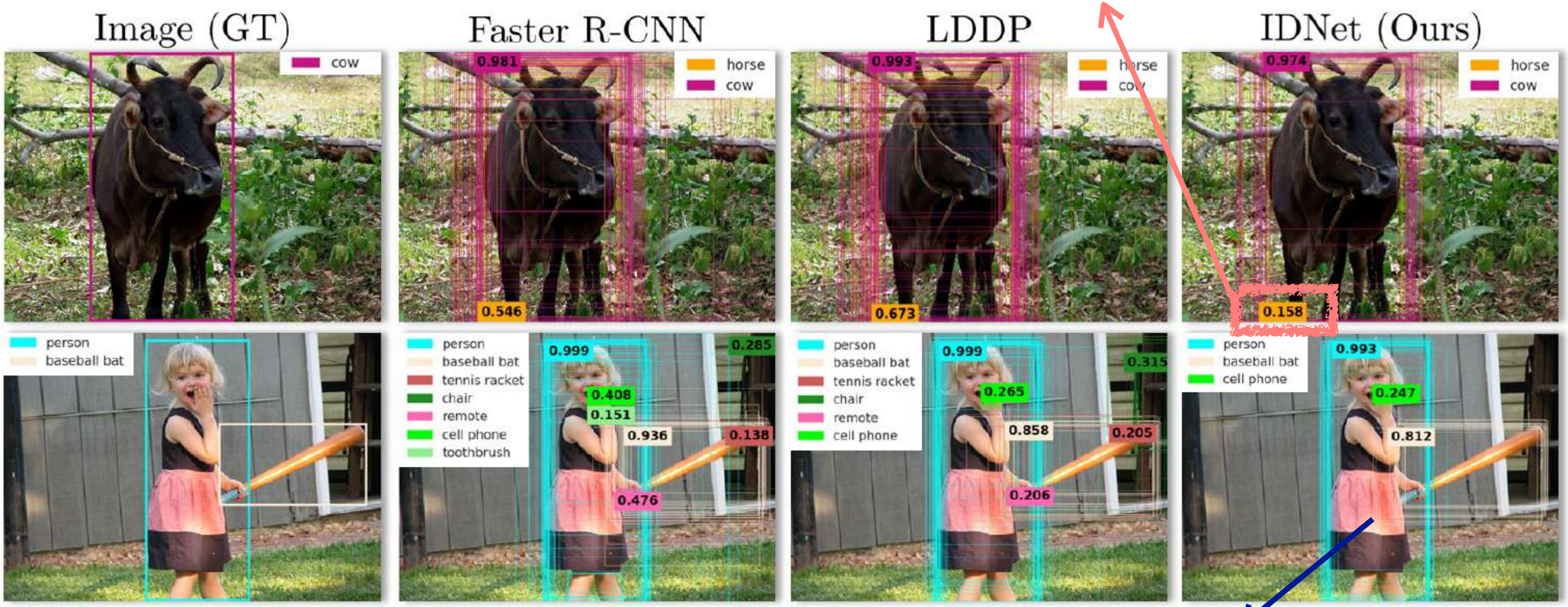
$$\mathcal{L}_{ss}(Y_{pos}, \mathcal{Y}_m) = -\log \sum_{Y\subseteq Y_{pos}} \mathcal{P}_{\mathbf{L}_{\mathcal{Y}_m}}(Y) = -\log \sum_{Y\subseteq Y_{pos}} \frac{\det(\mathbf{L}_Y)}{\det(\mathbf{L}_{\mathcal{Y}_m}+\mathbf{I}_{\mathcal{Y}_m})}$$

$$= -\operatorname{logdet}(\mathbf{L}_{Y_{pos}}+\mathbf{I}_{Y_{pos}}) + \operatorname{logdet}(\mathbf{L}_{\mathcal{Y}_m}+\mathbf{I}_{\mathcal{Y}_m})$$

Learning to **reduce** scores of **wrong categories**

Results of Sparse Score

Successfully reduced the score of wrong categories



Removed wrong categories

DPP Inference

RLLAB
Robot Learning Laboratory

Algorithm 1 Instance-Aware DPP Inference (IDPP).

$Y^* = \emptyset$
while $\mathcal{Y} \neq \emptyset$ **do**
  $j^* = \arg\max_{j \in \mathcal{Y}} \log(\prod_{i \in Y^* \cup \{j\}} \mathbf{q}_i^2 \cdot \det(\mathbf{S}_{Y^* \cup \{j\}}))$
  $Y = Y^* \cup \{j^*\}$
  if $\mathrm{Cost}(Y) > \mathrm{Cost}(Y^*)$ **then**
    $Y^* \leftarrow Y$
    delete j^* from $\mathcal{Y}$
  else
    return Y^*
  end if
end while
return Y^*

- where $\mathrm{Cost}(Y) = \log(\prod_{i \in Y} \mathbf{q}_i^2 \cdot \det(\mathbf{S}_Y))$

Results on CrowdHuman Dataset

RLLAB
Robot Learning Laboratory

Method	Inference	mAP				
		$\mathtt{crowd}_3$	$\mathtt{crowd}_4$	$\mathtt{crowd}_5$	$\mathtt{crowd}_6$	$\mathtt{crowd}_7$
# of images		4,370	3,879	3,143	2,087	1,052
Faster R-CNN [62]	NMS	52.0	51.8	51.1	44.4	44.2
RepLoss [73]	NMS	52.2	52.0	51.5	48.4	44.2
LDDP [4]	LDPP	52.9	52.8	52.5	52.0	51.4
IDNet	IDPP	**58.9**	**56.3**	**55.8**	**54.9**	**54.2**

Baseline

Crowd Detection Methods

Results on COCO Dataset

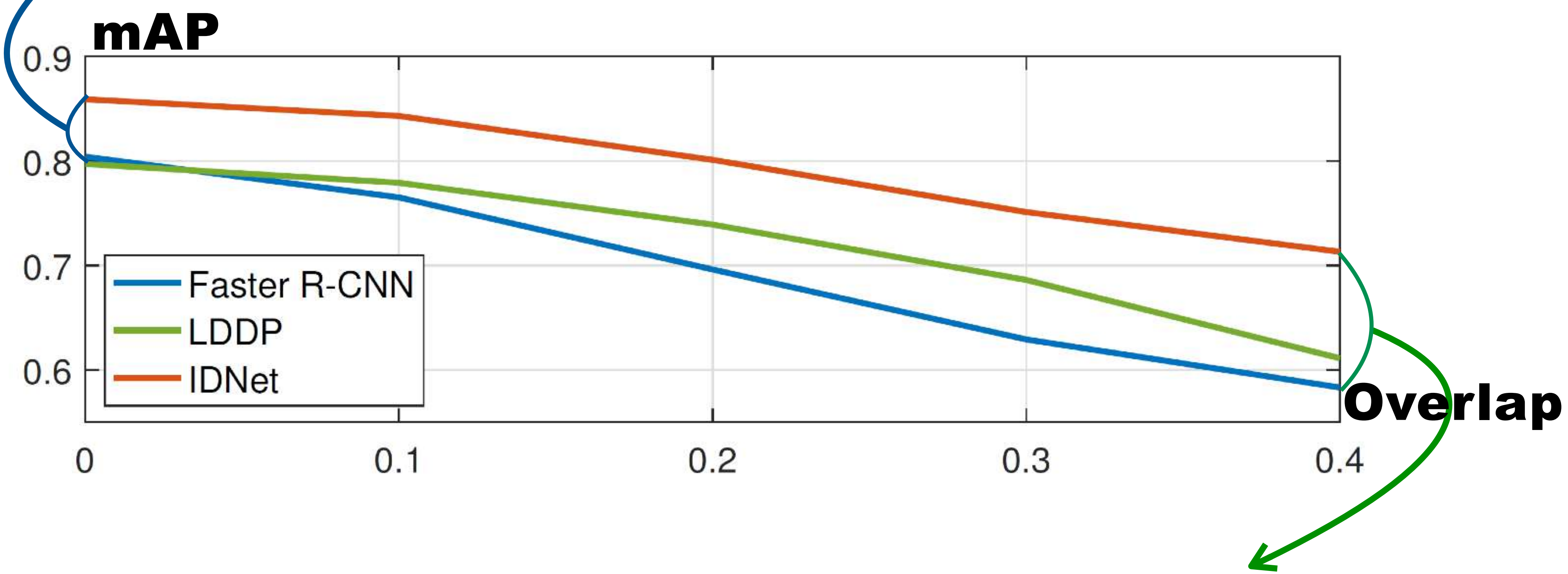
RLLAB
Robot Learning Laboratory

Method	Inference	Backbone	AP		AP_{50}		AP_{75}	
			test	crowd	test	crowd	test	crowd
Faster R-CNN [62]	NMS	VGG-16	26.2	19.2	46.6	36.9	26.9	18.4
LDDP [4]	LDPP	VGG-16	26.4	19.6	46.7	37.9	26.8	18.6
IDNet	IDPP	VGG-16	**27.3**	**20.5**	**47.6**	**38.2**	**28.2**	**20.0**
Faster R-CNN [62]	NMS	ResNet-101	31.5	23.5	52.0	42.5	33.5	23.0
LDDP [4]	LDPP	ResNet-101	31.4	23.8	51.7	43.0	33.4	23.4
IDNet	IDPP	ResNet-101	**32.7**	**24.4**	**53.1**	**43.4**	**34.8**	**24.4**

Results on MS COCO

Ablation Study

Outperforms baseline methods



Effective as overlap gets harder

Summary

- Proposes an end-to-end object detection framework for ***crowded*** situation using ***object relationship.***
- Proposes ***two losses*** using Determinantal Point Processes
  - ID (Instance identity) loss, which learns the identity of objects.
  - SS (Sparse score) loss, which removes confusing categories.

Roadmap

Passive Learning | **Interactive Learning**

Object Detection

DOG, DOG, CAT

CVIU 2020

Image Goal

ICCV 2021

CoRL 2022 (oral)

Object Goal

Chair

TV

Sofa

CVPR 2023 (submitted)

- Obin Kwon, **Nuri Kim**, Yunho Choi, Hwiyeon Yoo, Jeongho Park, and Songhwai Oh., “**Visual Graph Memory with Unsupervised Representation for Visual Navigation**,” International Conference on Computer Vision (ICCV-21)
- **Nuri Kim**, Obin Kwon, Hwiyeon Yoo, Yunho Choi, Jeongho Park, and Songhwai Oh., “**Topological Semantic Graph Memory for Image-Goal Navigation**,” Conference on Robot Learning (CoRL-22), ***oral presentation***

Navigation

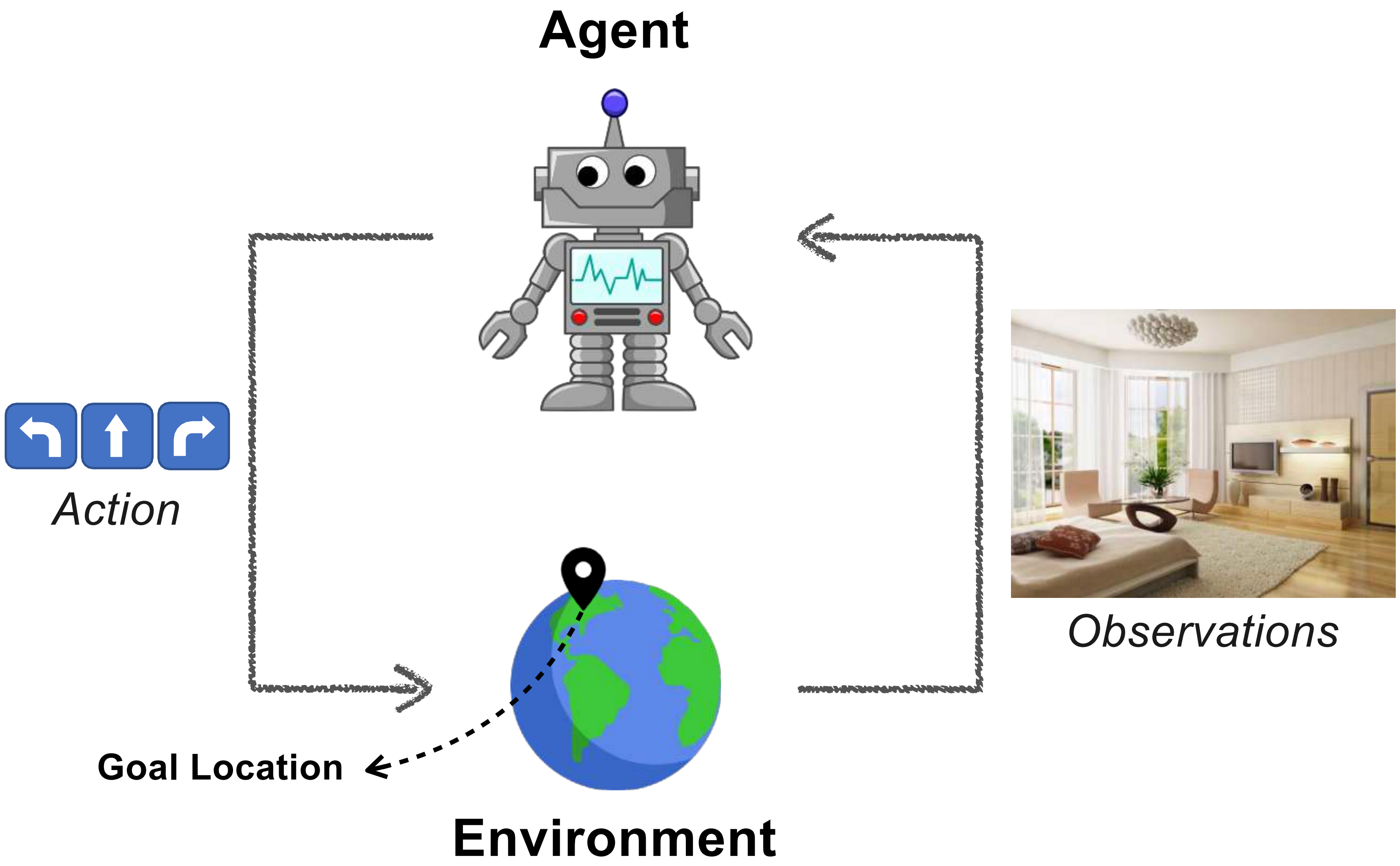
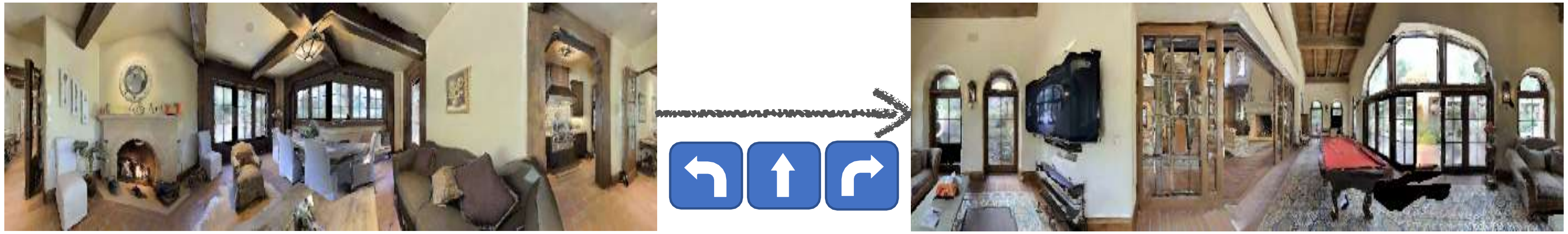


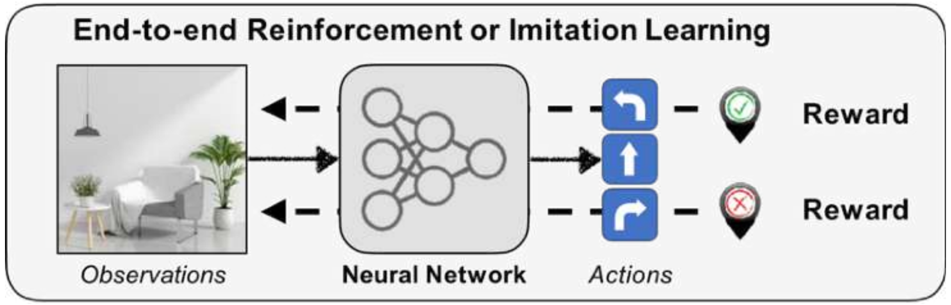
Image Goal Navigation

Source Image **Goal Image**



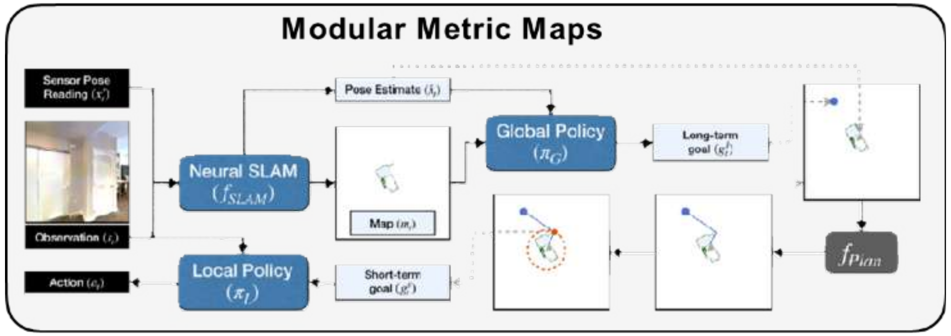
- Agent observations are panoramic images
- Take actions to navigate to the goal location
- Take the **stop** action at the goal location

Memory



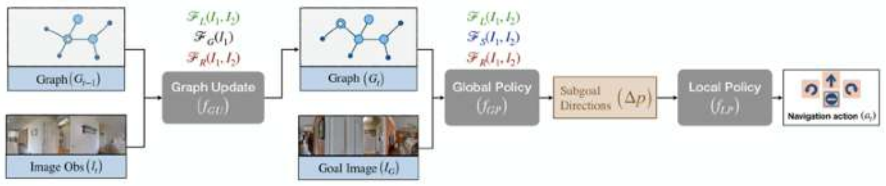
Implicit Memory

- High sample complexity
- Ineffective in large environments



Metric Map Memory

- Can not learn semantic priors
- Pose error accumulation

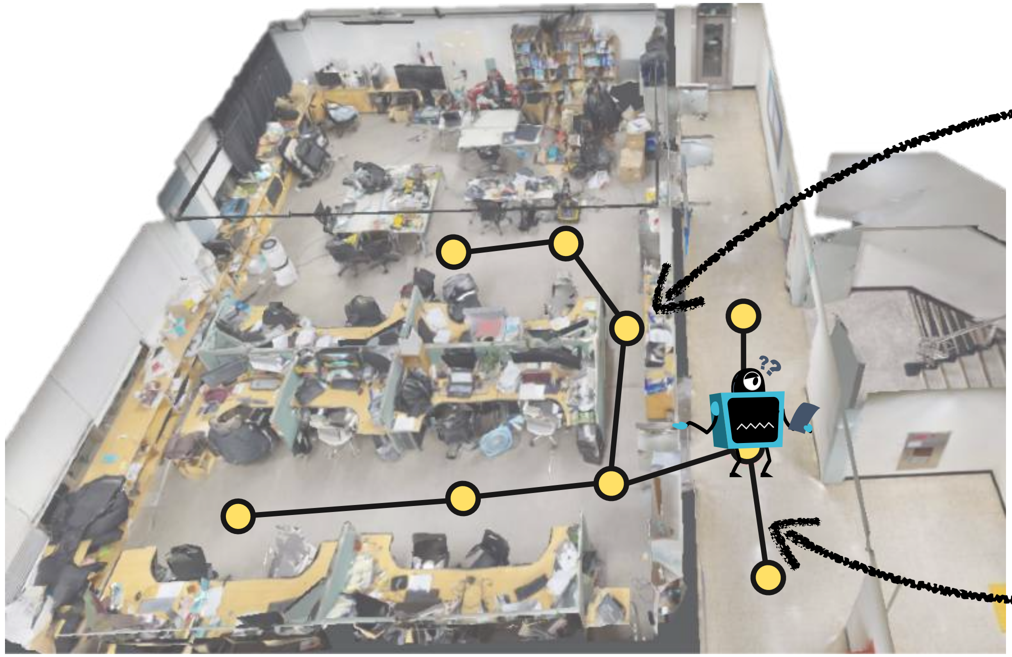


Topological Graph Memory

- Concise and precise
- Accurate pose sensor is not required

Topological Graph Memory

RLLAB
Robot Learning Laboratory



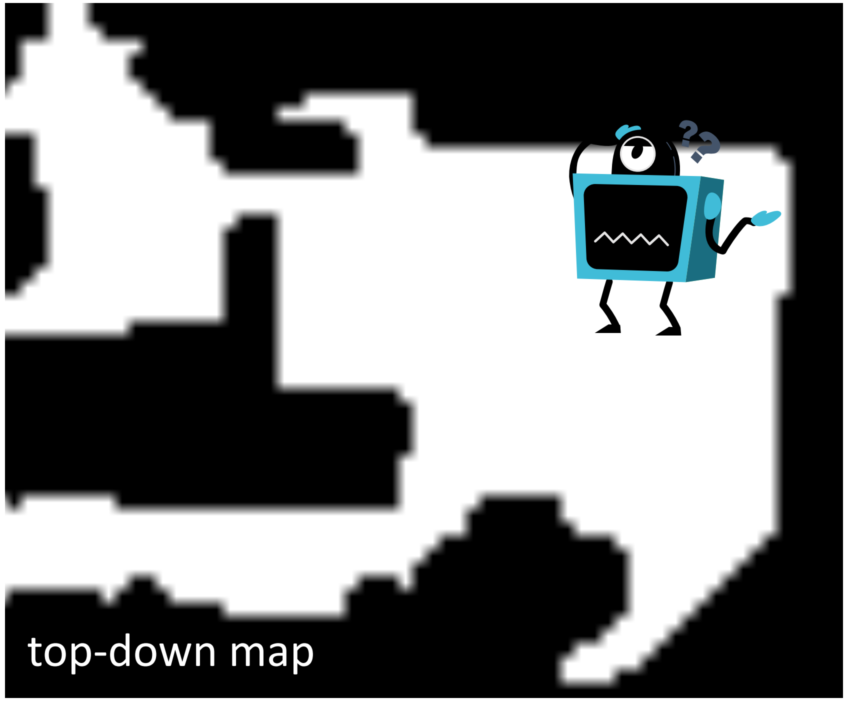
A vertex represents an area in the environment

An edge represents the relationship between two vertices, such as reachability and proximity

Topological Graph Memory

RLLAB
Robot Learning Laboratory

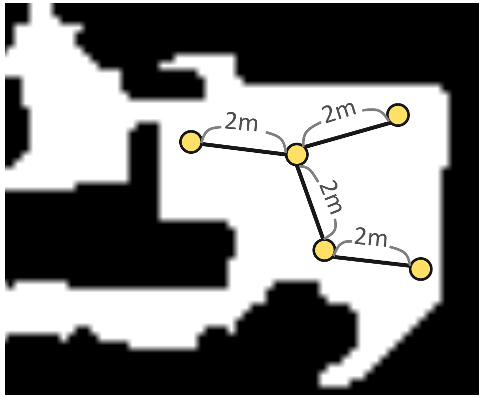
How to build a graph map?



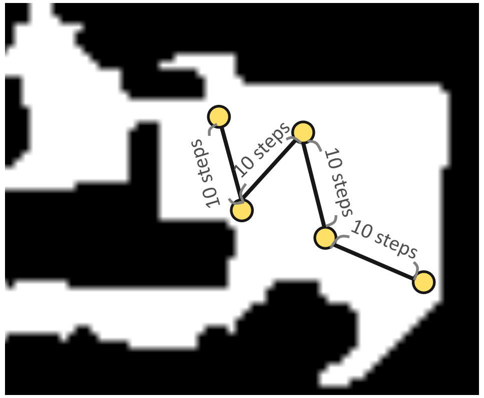
Visual Graph Memory

Previous graph-based navigation methods usually select the vertices and edges based on the following standards :

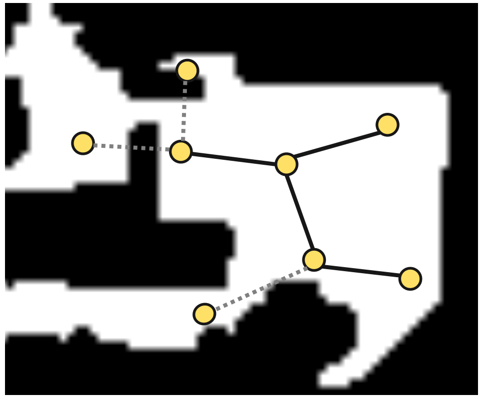
Spatial Distance



Temporal Distance



Visibility / Occlusion



Several learning-based methods build a graph map using a pretrained classifier network, based on images.

It is trained to determine whether the two image observations are close or not, based on the predefined rules.

Elaborately designed annotation rules based on accurate geometric information are required for preparing datasets.

- What is the adequate distance between each node?
- How can we determine the two nodes are visible from each other?

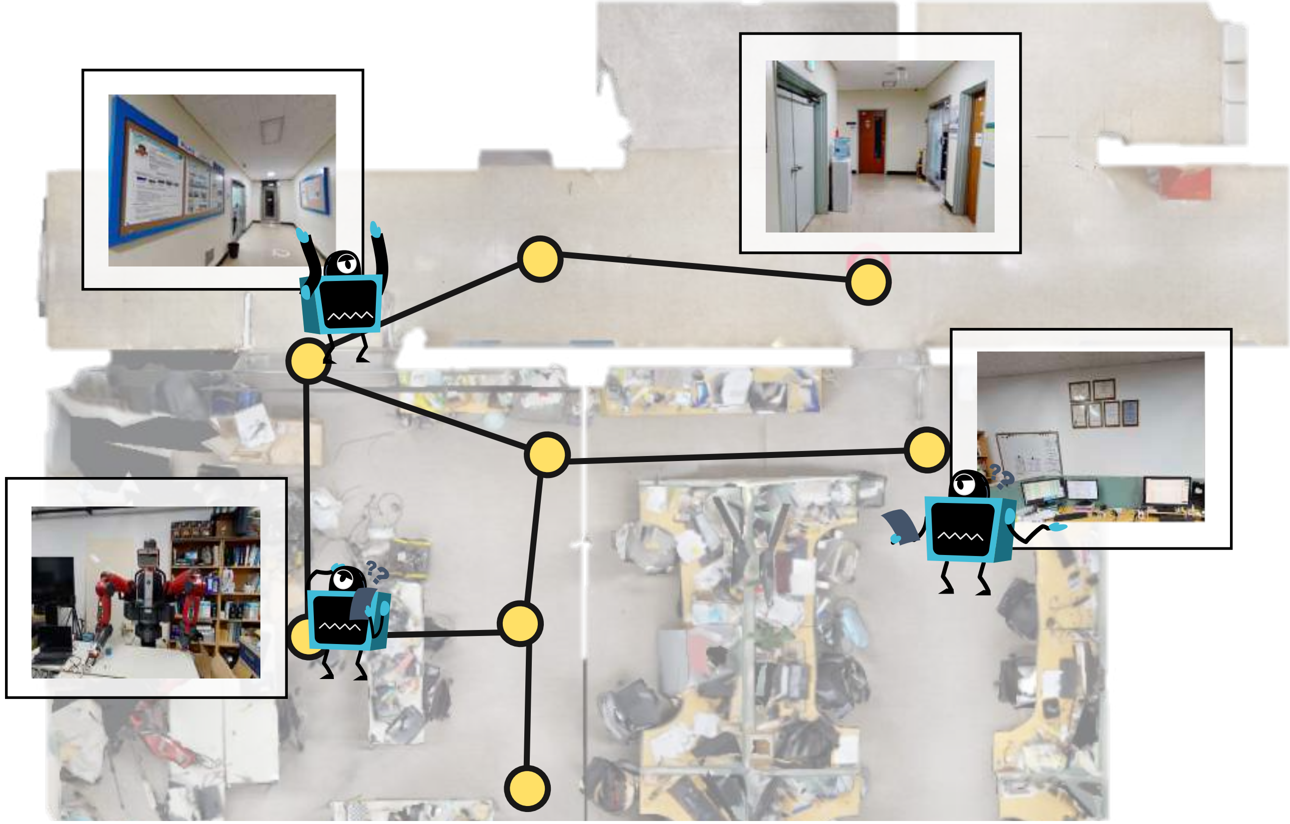
Visual Graph Memory

Furthermore**, the perception about relative distance can be vary** depending on the appearance of the environment. For example, the image pairs below are 1.5m apart from each other.

We can recognize that the camera position has certainly moved from the original position in the first pair. However, in the second pair, the translation is not visually significant as much as the first pair.

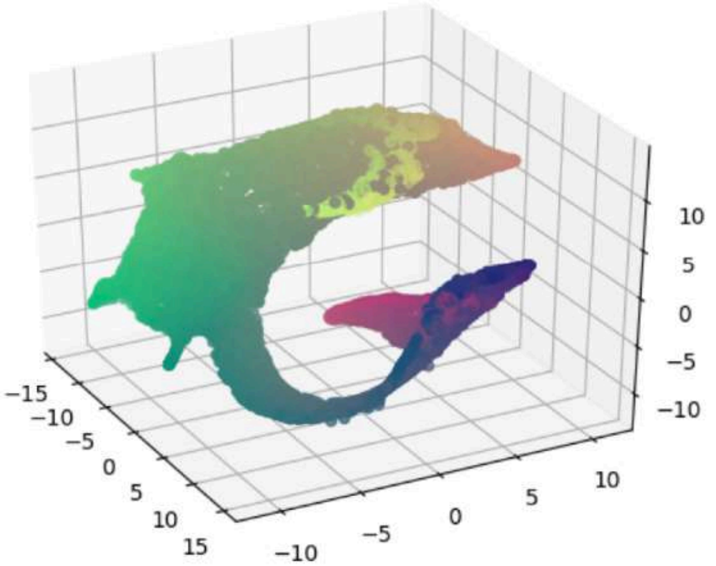
Visual Graph Memory

Human remembers the novel landmarks rather than equally-spaced distances.
Human subconsciously knows which places are good to be the landmarks.
We aimed to inject this characteristic into the graph-based navigation system.



Visual Graph Memory

We hypothesized that using unsupervised image representation is sufficient to build a graph map.
We have collected 10000 images from each (training) environment in habitat simulator and trained an image encoder.
The image encoder is trained using unsupervised contrastive learning, without any annotation labels.
This image encoder transforms image observations to feature embeddings.
The more the images have a similar appearance, the closer the distance between the encoded features.



Visual Graph Memory

RLLAB
Robot Learning Laboratory

Comparing to other previous methods, ours can build a sufficient graph map during the navigation.

Agent’s Trajectory | Distance-based | Supervised Localization | VGM (unsupervised)

Observation

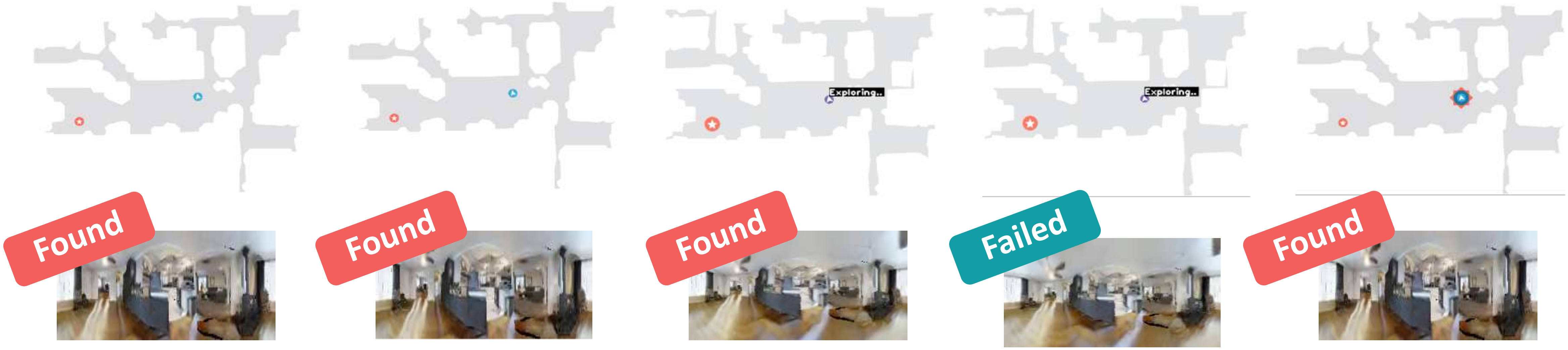
Visual Graph Memory

Target Image: 

Agent

Target

Exp4nav	SMT	SPTM	Neural Planner	VGM (ours)



Visual Graph Memory

Target Image: 

Agent

Target

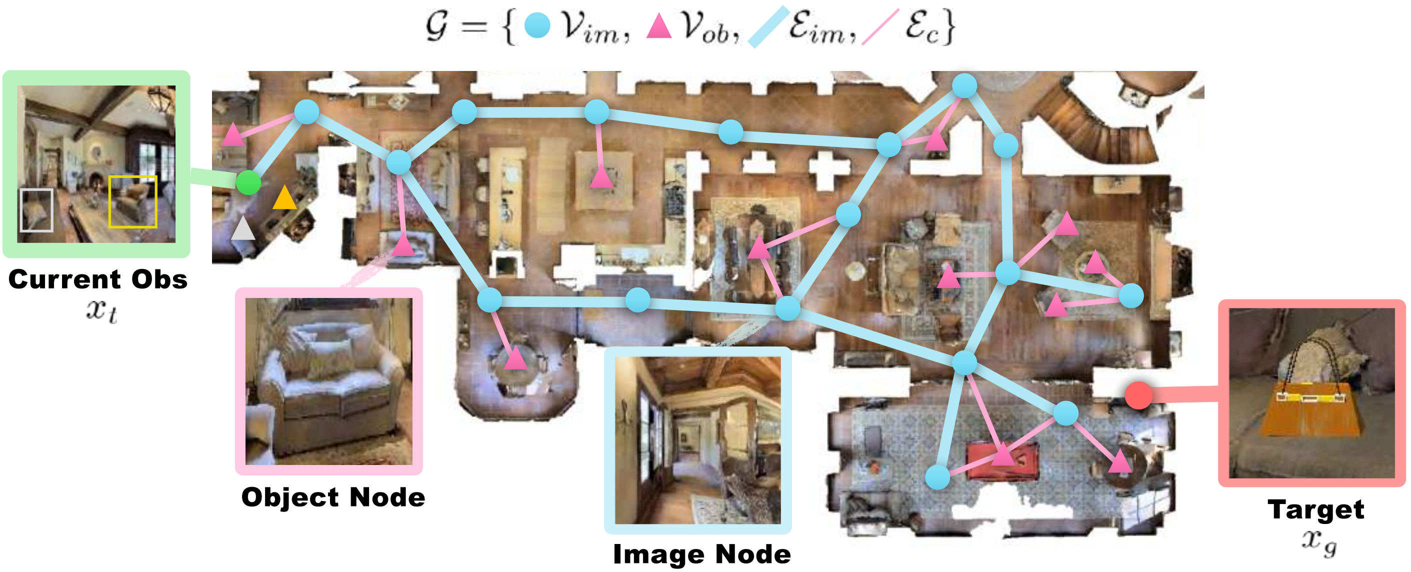
Exp4nav | SMT | SPTM | Neural Planner | VGM (ours)

Visual Graph Memory

The proposed navigation framework brings performance improvement over other types of memory models.

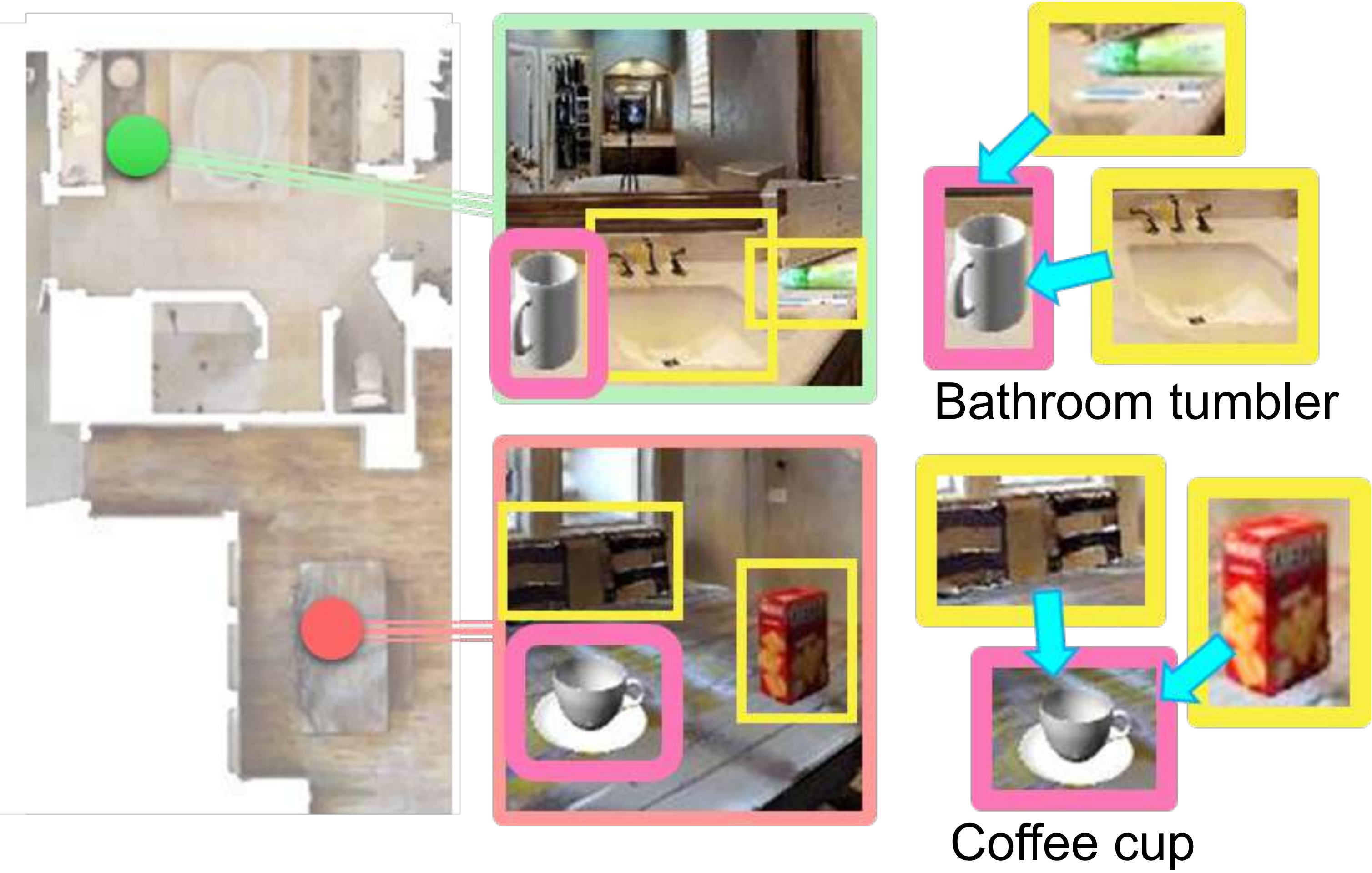
Methods	Memory Type	Need Pose Information	Navigation Results	
			Success Rate	SPL
CNN + LSTM	hidden vector	yes	0.49	0.45
ANS + predicted target pose	metric map	yes	0.58	0.18
Exp4nav	metric map	yes	0.59	0.51
SMT	stack of image features	yes	0.68	0.56
Neural Planner	graph	yes	0.60	0.36
Exploration + SPTM	graph	no	0.58	0.35
NTS	graph	yes	0.63	0.43
VGM (ours)	**graph**	**no**	**0.76**	**0.64**

Semantic Navigation



Object Context

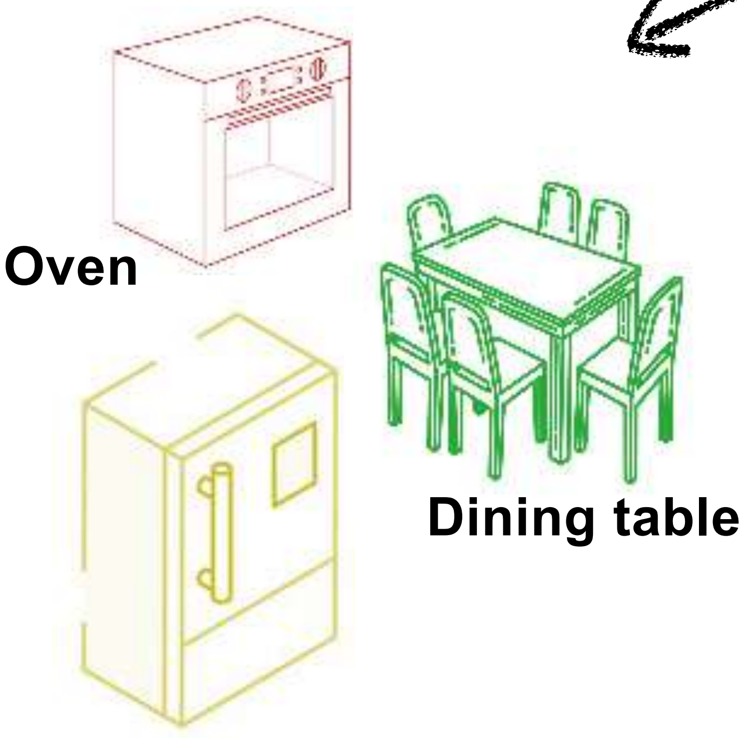
RLLAB
Robot Learning Laboratory

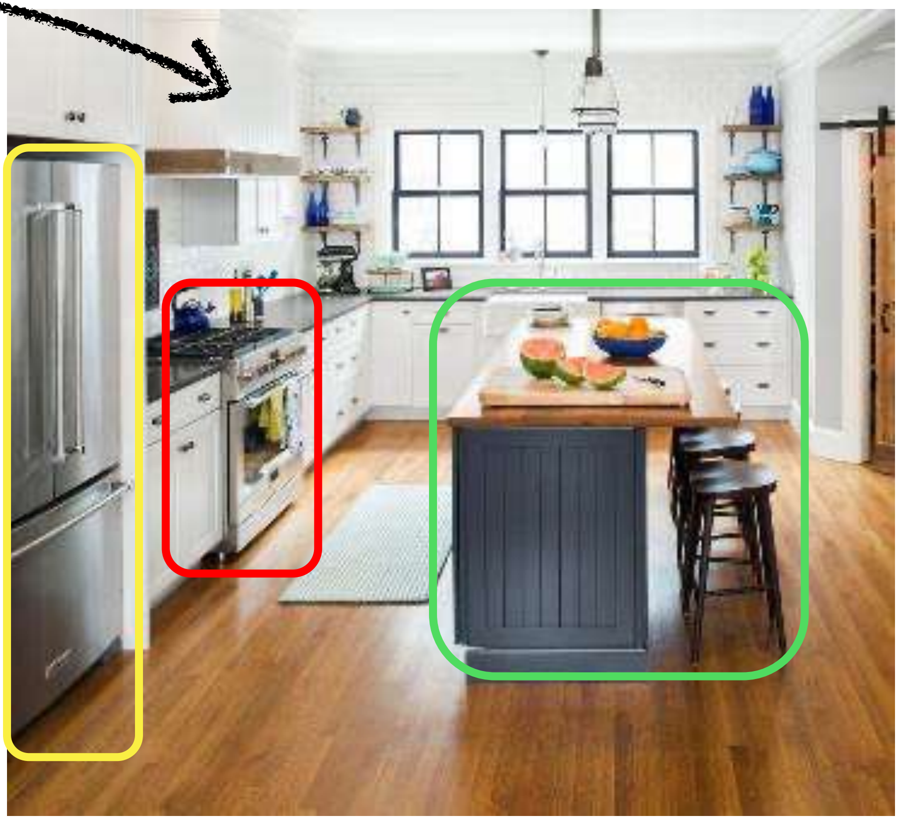


Neighboring objects make an object unique

Place-Object Context

RLLAB
Robot Learning Laboratory





Kitchen

How to embed landmark knowledge into topological graph memory?

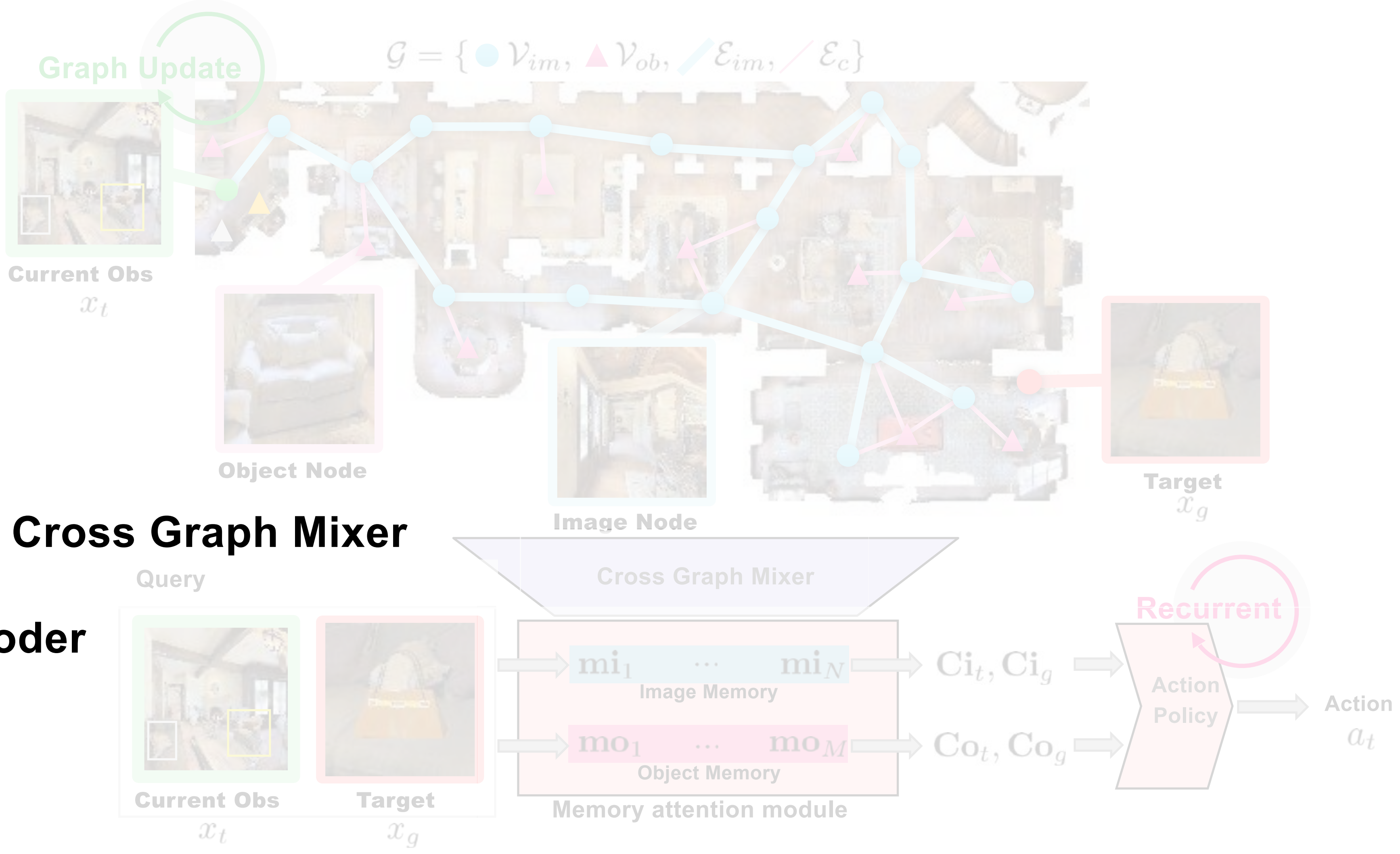
Topological Semantic Graph Memory

RLLAB
Robot Learning Laboratory

Graph Builder

Cross Graph Mixer

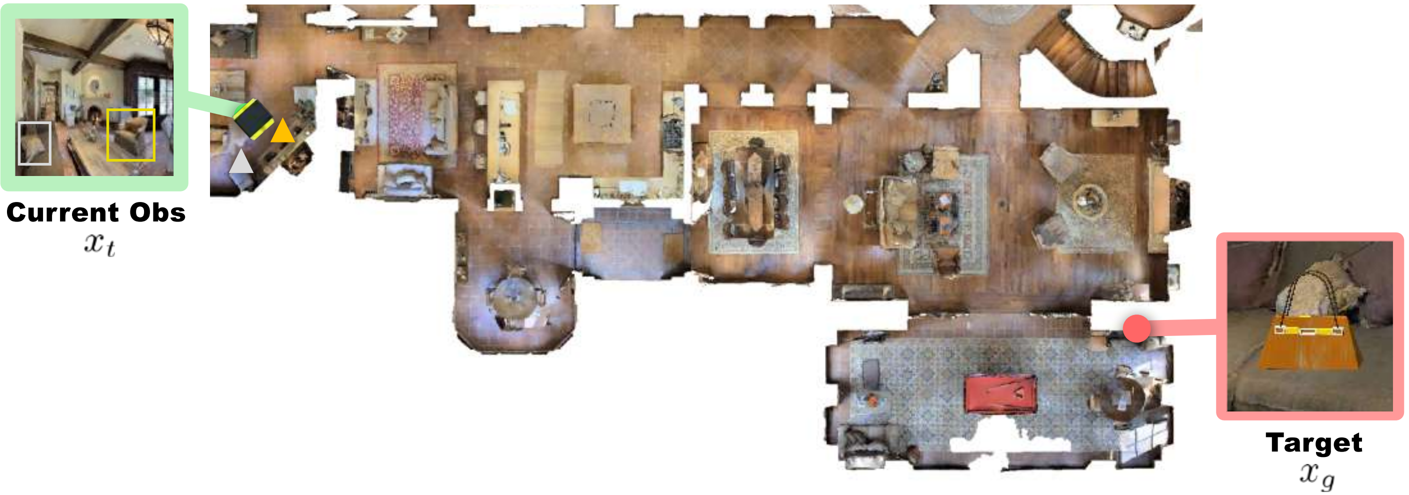
Memory Decoder



Graph Builder: Overview

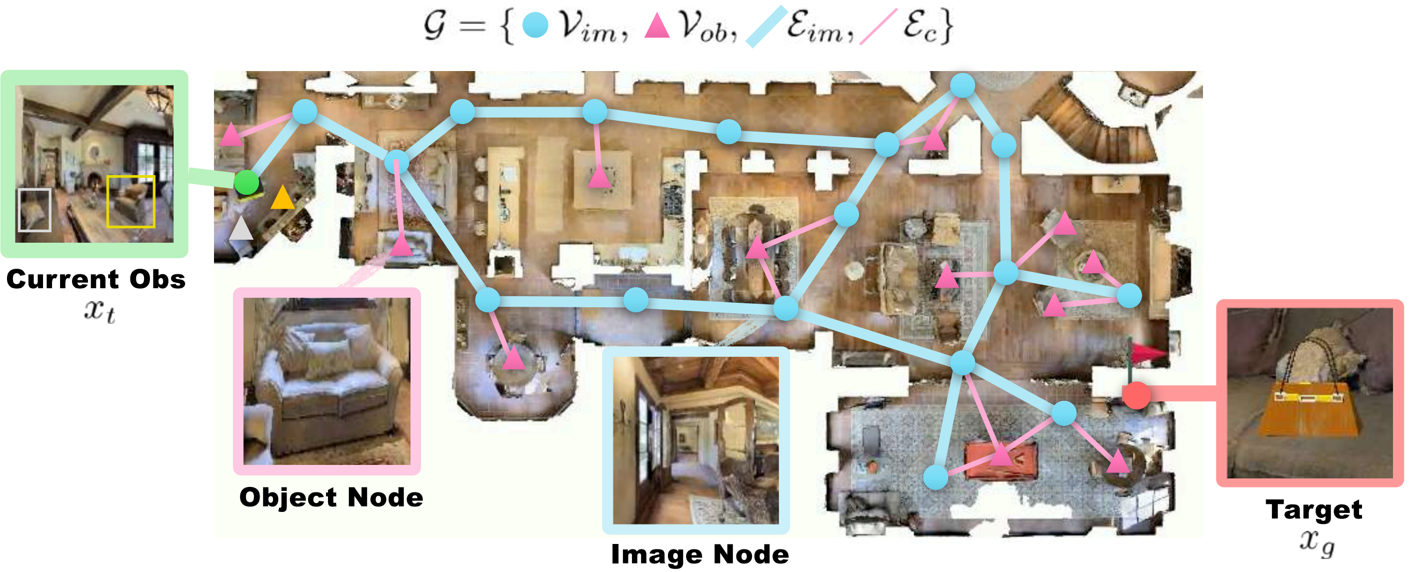
RLLAB
Robot Learning Laboratory

$\mathcal{G} = \{ \mathcal{V}_{im}, \mathcal{V}_{ob}, \mathcal{E}_{im}, \mathcal{E}_c \}$



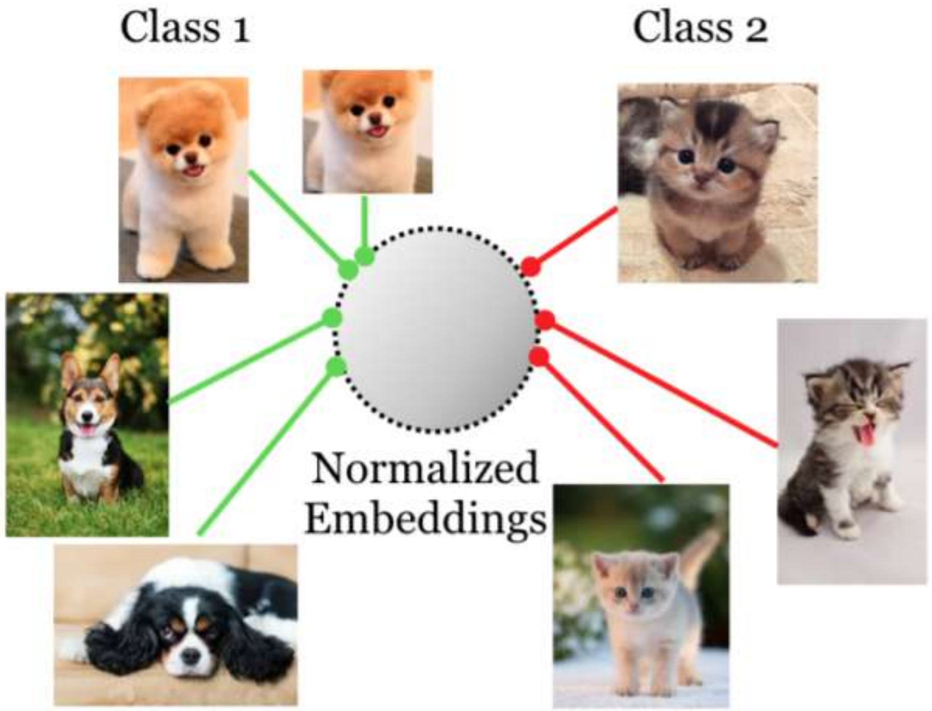
✱ Note that floorplan and node positions are only used for illustration and not given as input to agent

Graph Builder: Overview

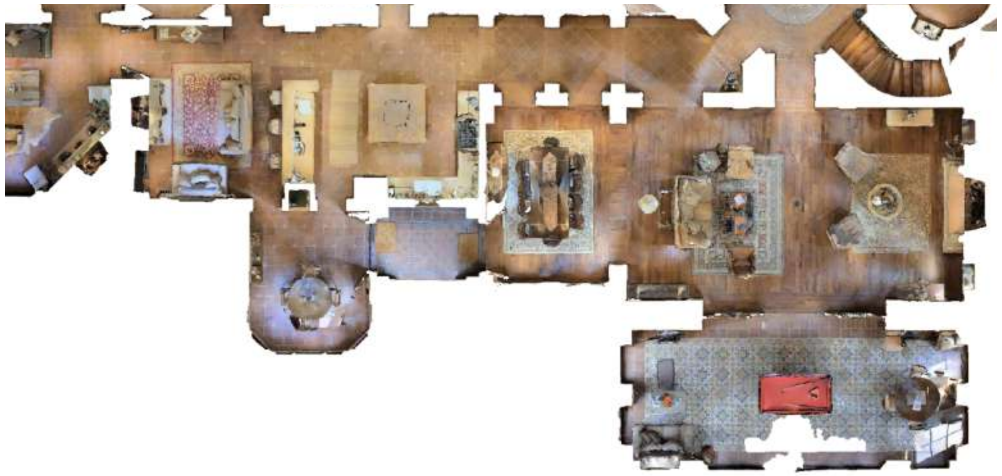


✱ Note that floorplan and node positions are only used for illustration and not given as input to agent

Graph Builder: Object Graph



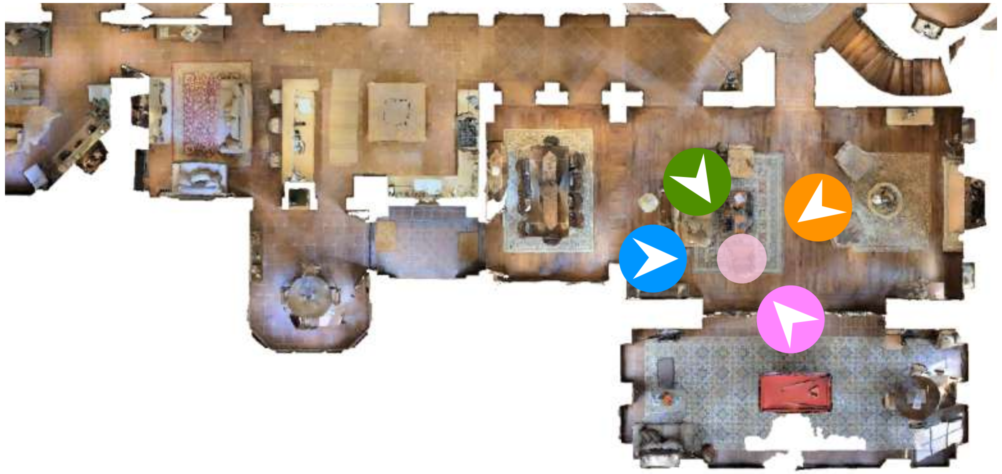
Contrastive Learning



Graph Builder: Object Graph

RLLAB
Robot Learning Laboratory

Collect an object from different viewpoints



Graph Builder: Object Graph

RLLAB
Robot Learning Laboratory

Query

Top 5 objects in the environment (among ~7000 candidates)

The object encoder successfully find a query object from different viewpoints

Graph Builder: Object Graph

RLLAB
Robot Learning Laboratory

Observation

Object Nodes: Individual objects

Detected objects are connected to the current node

Object Memory

* Color represents the 3-dim tsne feature of the place

Image Nodes

Agent's Current Image Node

Object Nodes

Graph Builder: Object Graph

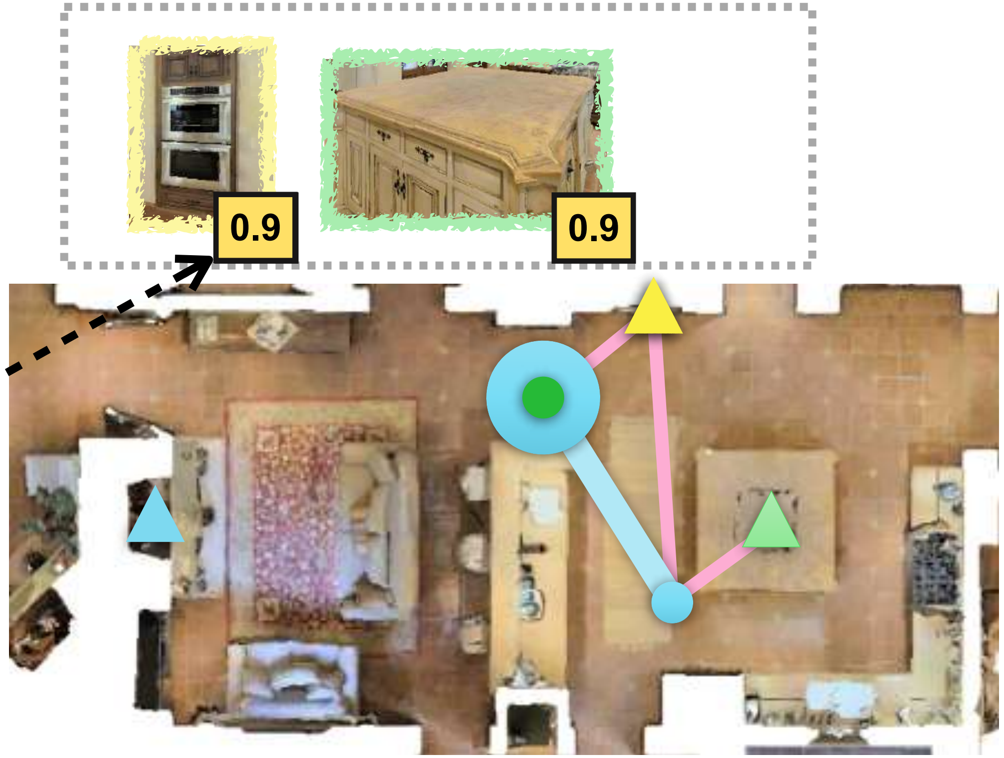
RLLAB
Robot Learning Laboratory

Similarity is **high** and the category is the same.
It indicates that the object is **already in the memory.**

Since **detection score is higher** than the memory node,
It is used to update the memory node.
The node is connected to the lastly localized image node.

Observation

Object Memory



* Color represents the 3-dim tsne feature of the place

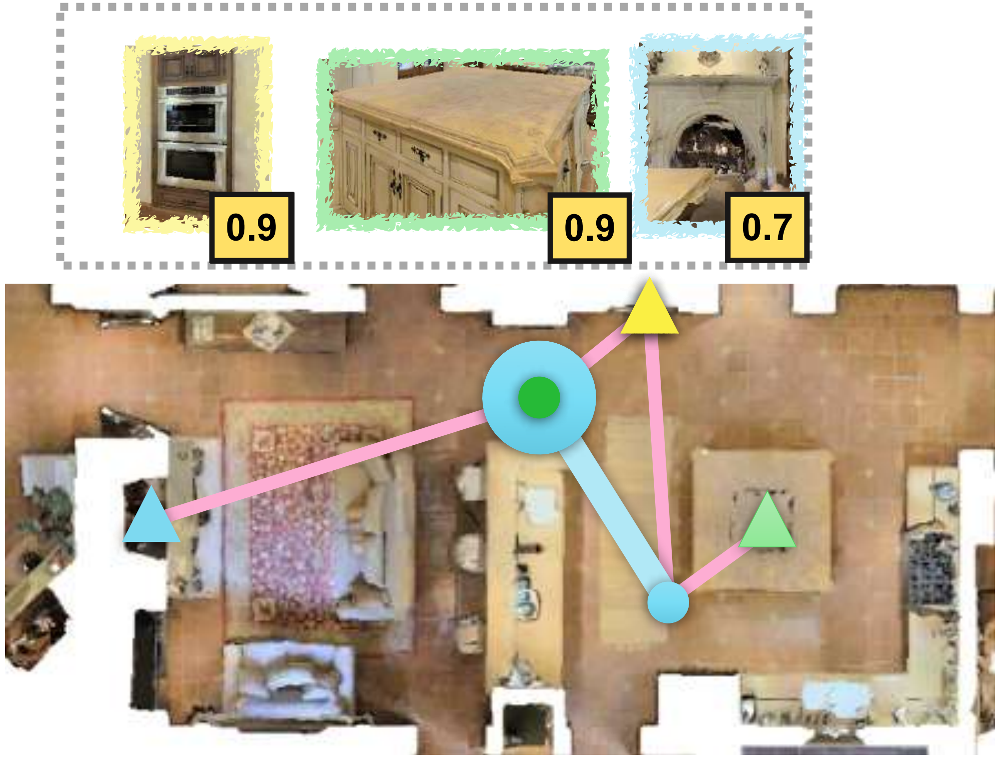
- **Image Nodes**
- **Agent's Current Image Node**
- **Object Nodes**

Graph Builder: Object Graph

Observation

Similarity with memory is low.
It is added to a memory as a new node
and connected to the lastly localized image node.

Object Memory

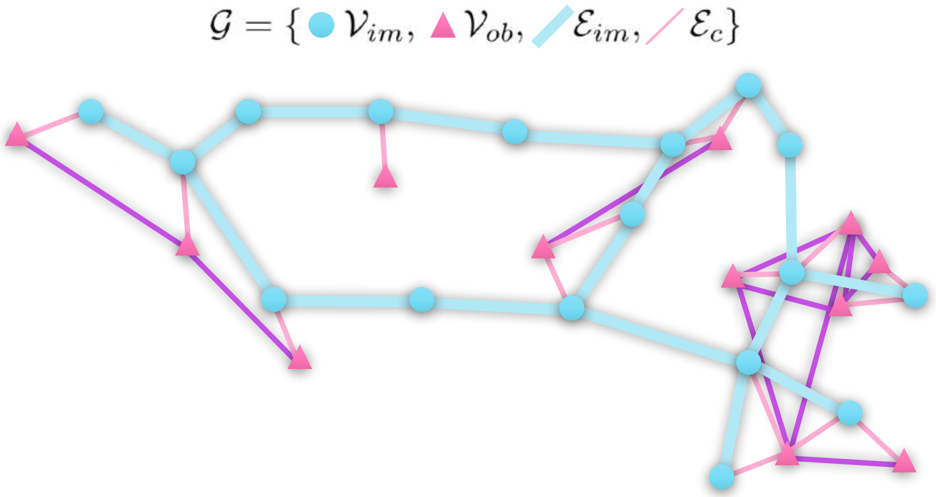


* Color represents the 3-dim tsne feature of the place

- Image Nodes
- Agent's Current Image Node
- Object Nodes

Graph Builder: Object Graph

RLLAB
Robot Learning Laboratory



$$A_{ob} = A_c^T (A_{im} + \mathrm{I}) A_c$$

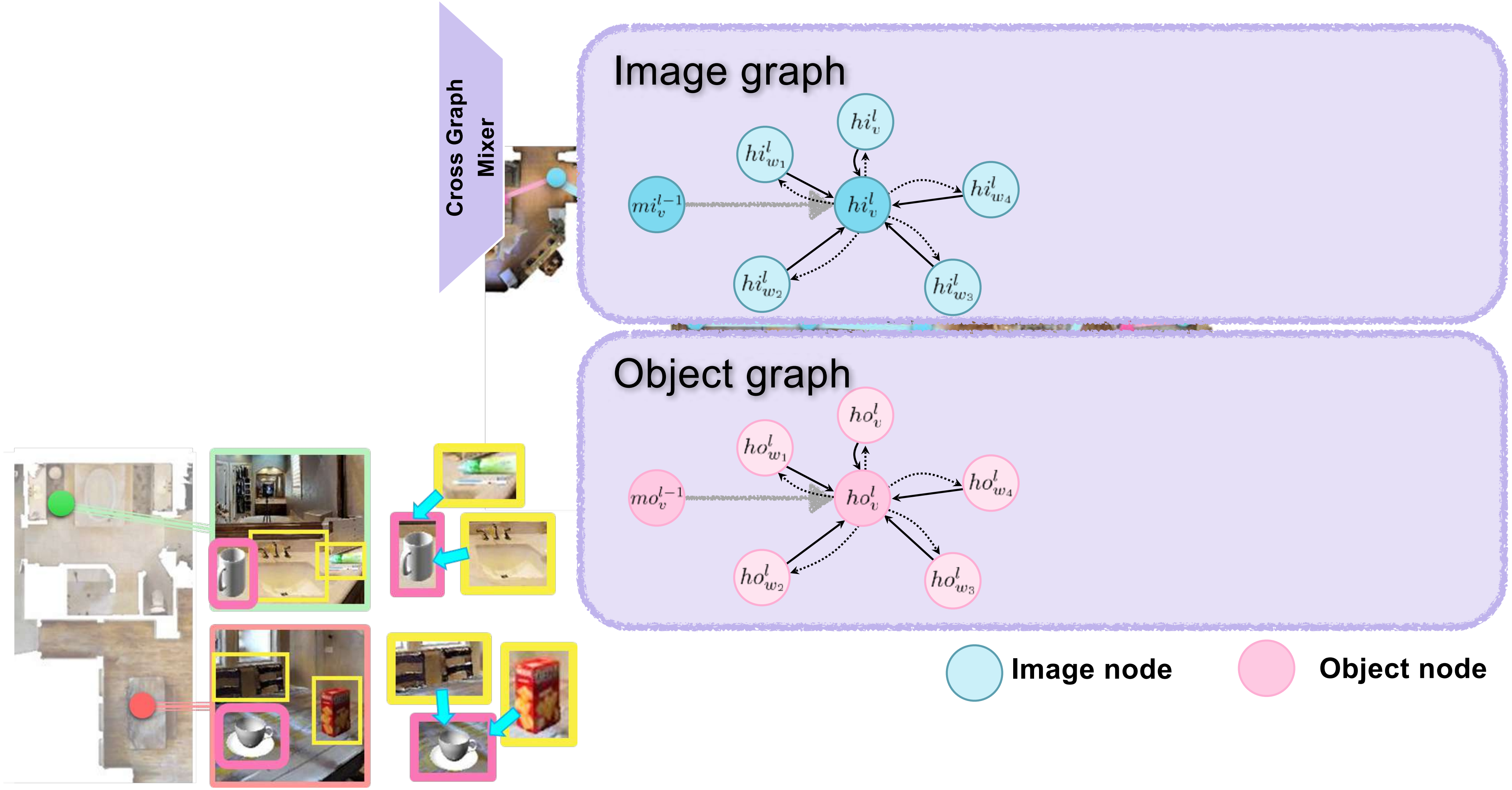
A_{im} : image affinity matrix

A_{ob} : object affinity matrix

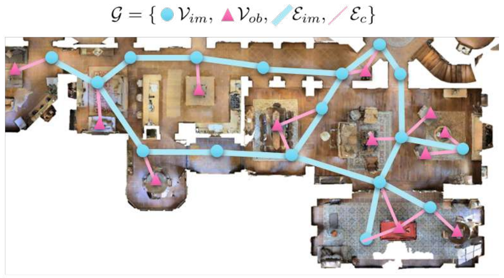
A_c : image-object affinity matrix

Cross Graph Mixer: Self Update

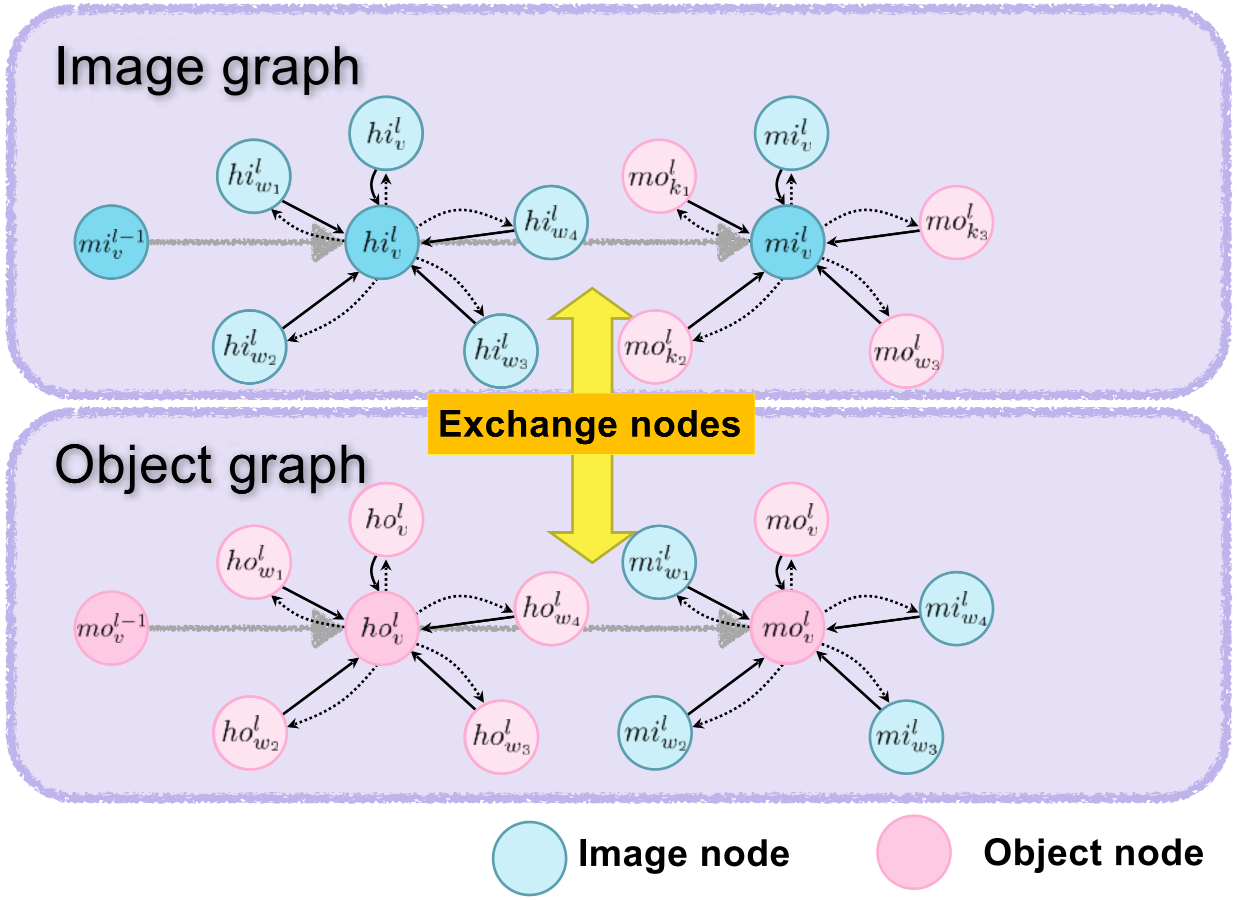
RILAB
Robot Learning Laboratory

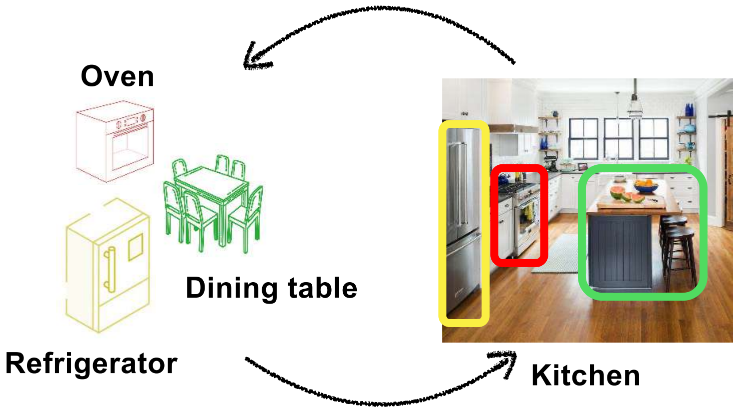


Cross Graph Mixer: Cross Update

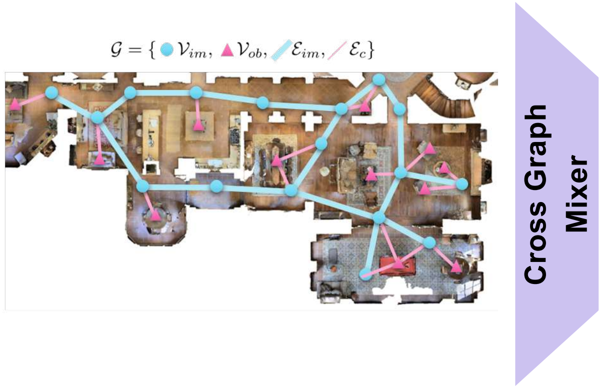


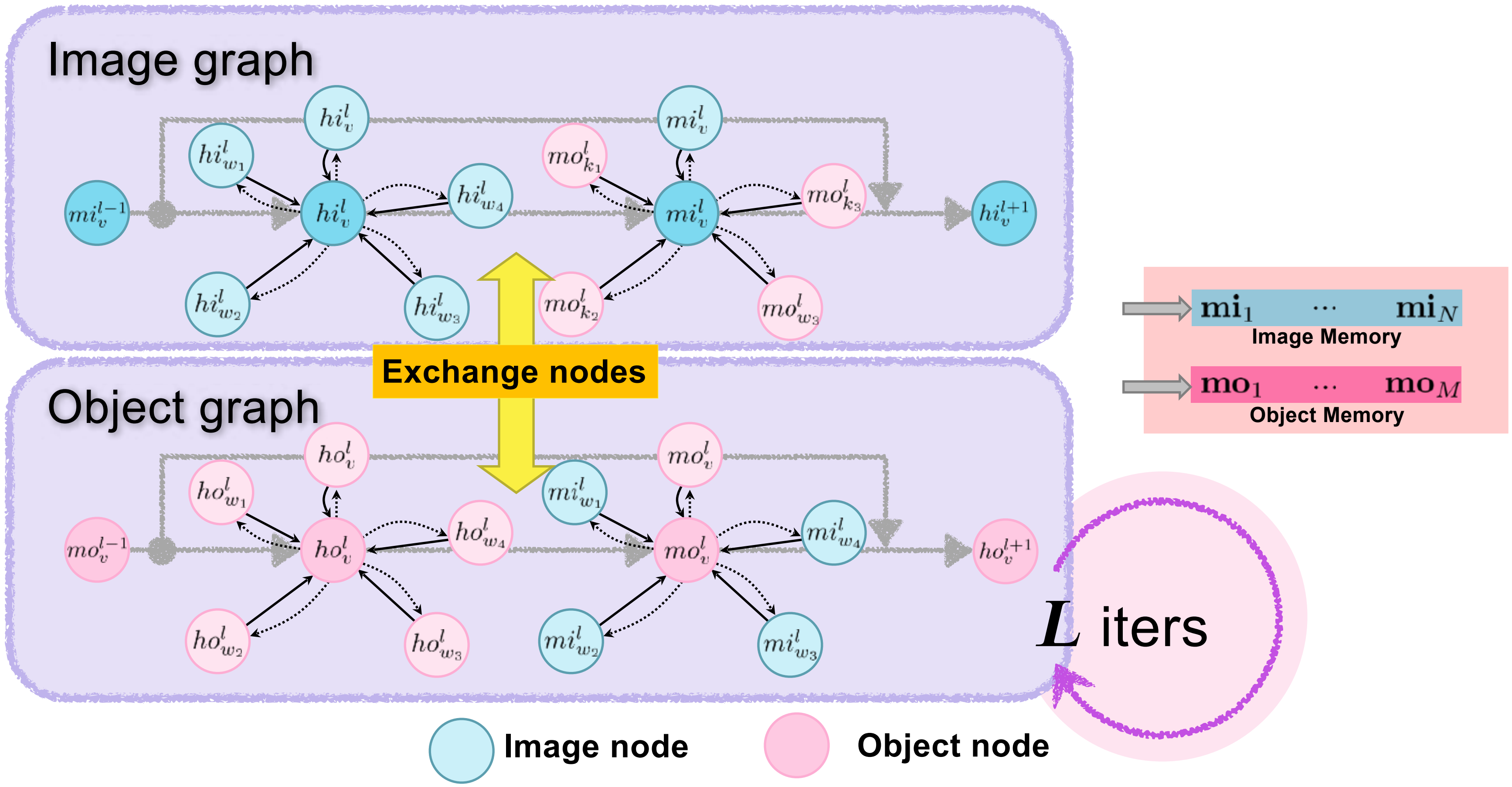
Cross Graph Mixer



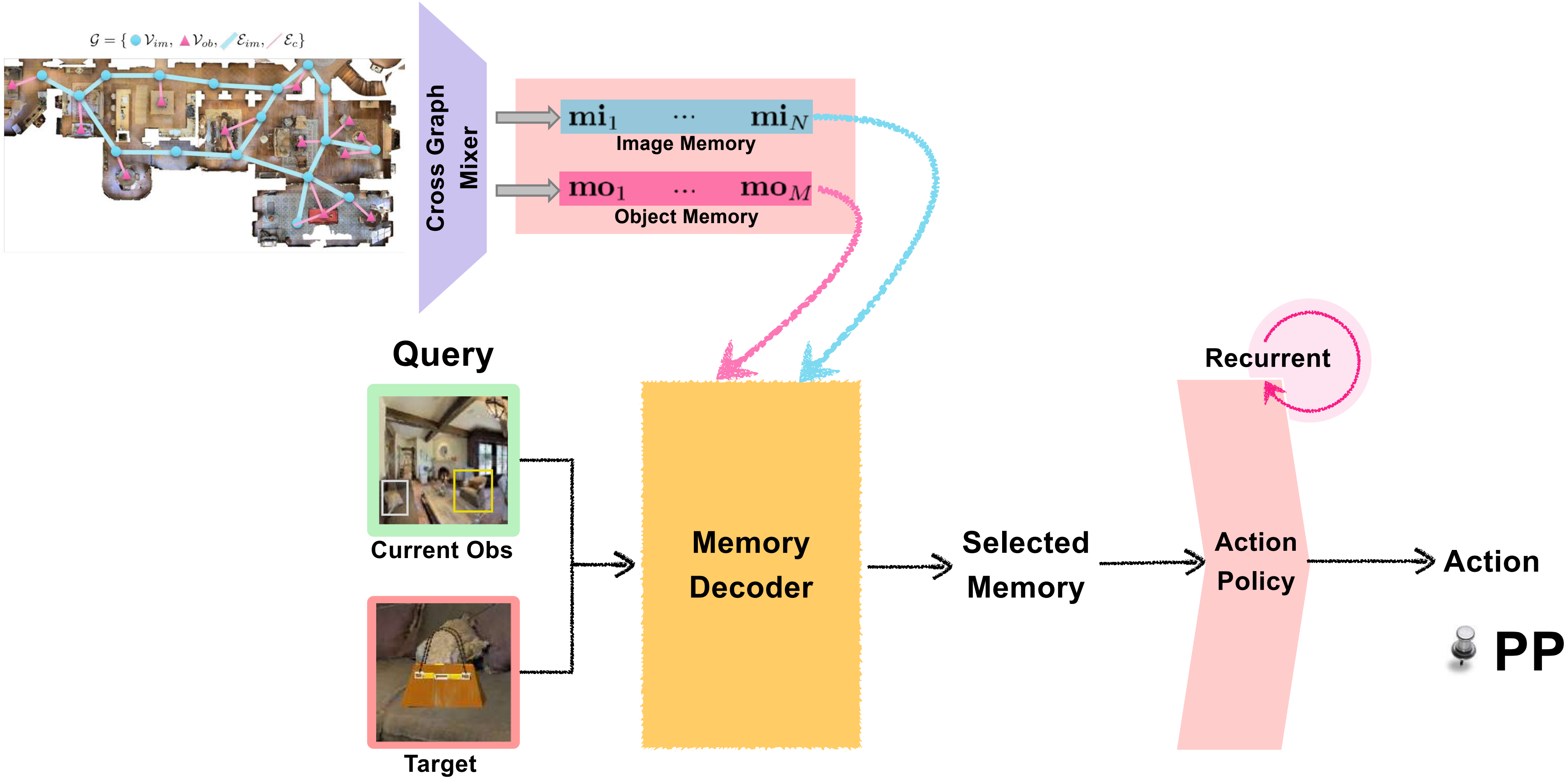


Cross Graph Mixer: Cross Update





Memory Decoder



PPO algorithm

Demo Video

RLLAB
Robot Learning Laboratory

TSGM (Ours)

VGM [2]

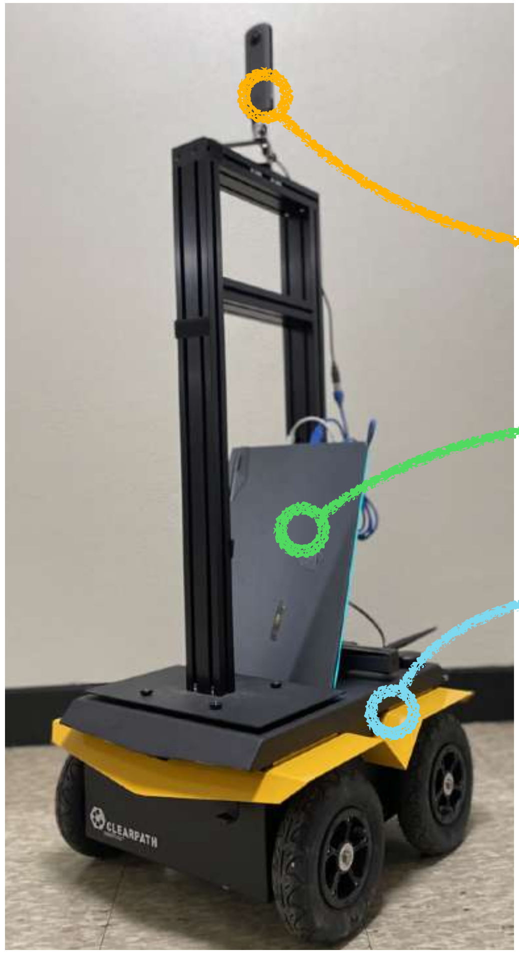
[2] Obin Kwon, et al. "Visual graph memory with unsupervised representation for visual navigation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

Real-World Demonstration

RLLAB
Robot Learning Laboratory

Robot specification

Height: 1.2m



Richo Theta 360° Camera (RGB sensor)

Intel Core i7 and GeForce RTX 2080

Jackal UGV from Clearpath

Real-World Demonstration

RLLAB
Robot Learning Laboratory

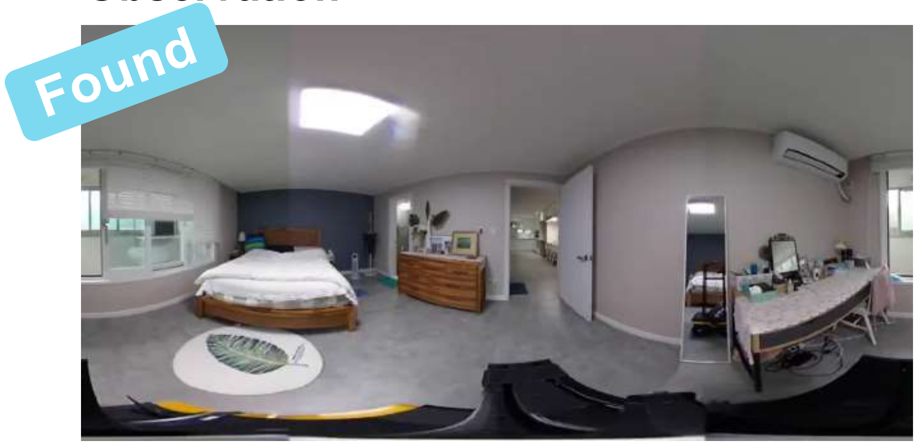
Start Position

Goal Position

✱ we estimated the robot and object locations to draw graphs on the map

Goal *7.55m

Observation



Results

RILAB
Robot Learning Laboratory

Method	Memory	No Pose	Object	Easy		Medium		Hard		Overall	
				Success	SPL	Success	SPL	Success	SPL	Success	SPL
RGBD + RL [26]	implicit	✗	✗	72.5	69.5	53.1	48.6	22.3	17.7	49.3	45.3
ANS [17]	metric	✗	✗	74.2	20.5	68.4	22.9	29.9	11.0	57.5	18.1
Exp4nav [5]	metric	✗	✗	70.2	61.8	60.6	52.4	46.9	38.5	59.2	50.9
SMT [8]	graph	✗	✗	81.9	77.4	65.6	52.2	55.6	39.7	67.7	56.4
Neural Planner [20]	graph	✗	✗	71.7	41.3	64.7	38.5	42.0	27.0	59.5	35.6
SPTM [9]	graph	✔	✗	66.5	40.6	64.2	38.5	42.1	25.4	57.6	34.8
VGM [18]	graph	✔	✗	86.1	79.6	81.2	**68.2**	60.9	45.6	76.1	64.5
TSGM (Ours)	graph	✔	✔	**91.1**	**83.5**	**82.0**	68.1	**70.3**	**50.0**	**81.1**	**67.2**

Implicit memory — Metric-map memory — Topological Memory

Results

Path Type	Method	Easy		Medium		Hard		Overall	
		Success	SPL	Success	SPL	Success	SPL	Success	SPL
Straight	NRNS [27]	67.1	57.8	52.4	41.2	32.6	22.4	50.7	40.5
	VGM [18]	81.0	54.4	82.0	69.9	67.3	54.4	76.7	59.6
	TSGM (Ours)	**94.4**	**92.1**	**92.6**	**84.3**	**70.3**	**62.8**	**85.7**	**79.7**
Curved	NRNS [27]	31.7	13.0	29.0	13.6	19.2	10.4	26.6	12.3
	VGM [18]	81.0	45.5	78.8	59.5	62.2	46.9	74.0	50.6
	TSGM (Ours)	**93.6**	**91.0**	**89.7**	**77.8**	**64.2**	**55.0**	**82.5**	**74.1**

SPL: Success weighted by normalized inverse Path Length

$$\frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(p_i, l_i)}$$

Ablation Study on Cross Graph Mixer

RLLAB
Robot Learning Laboratory

Update	Success	SPL
No	0.533	0.393
Visual	0.578	0.446
Object	0.613	0.458
Cross	**0.627**	**0.471**

Ablation study on Cross graph mixer updates

Summary

RLLAB
Robot Learning Laboratory

- **Integrated semantic information** to topological graph memory
  - To the best of our knowledge, we firstly constructed object graph on the topological graph.
- TSGM can connect objects in proximity even though the adjacent objects are not in the same view, which makes a **spatially meaningful** graph memory.
- TSGM gives **object connection**s and **object-place connections** to the agent, and outperforms SOTA methods on image goal navigation.

Roadmap

Passive Learning

Object Detection

DOG, DOG, CAT

CVIU 2020

Interactive Learning

Image Goal

ICCV 2021

CoRL 2022 (oral)

Object Goal

Chair

TV

Sofa

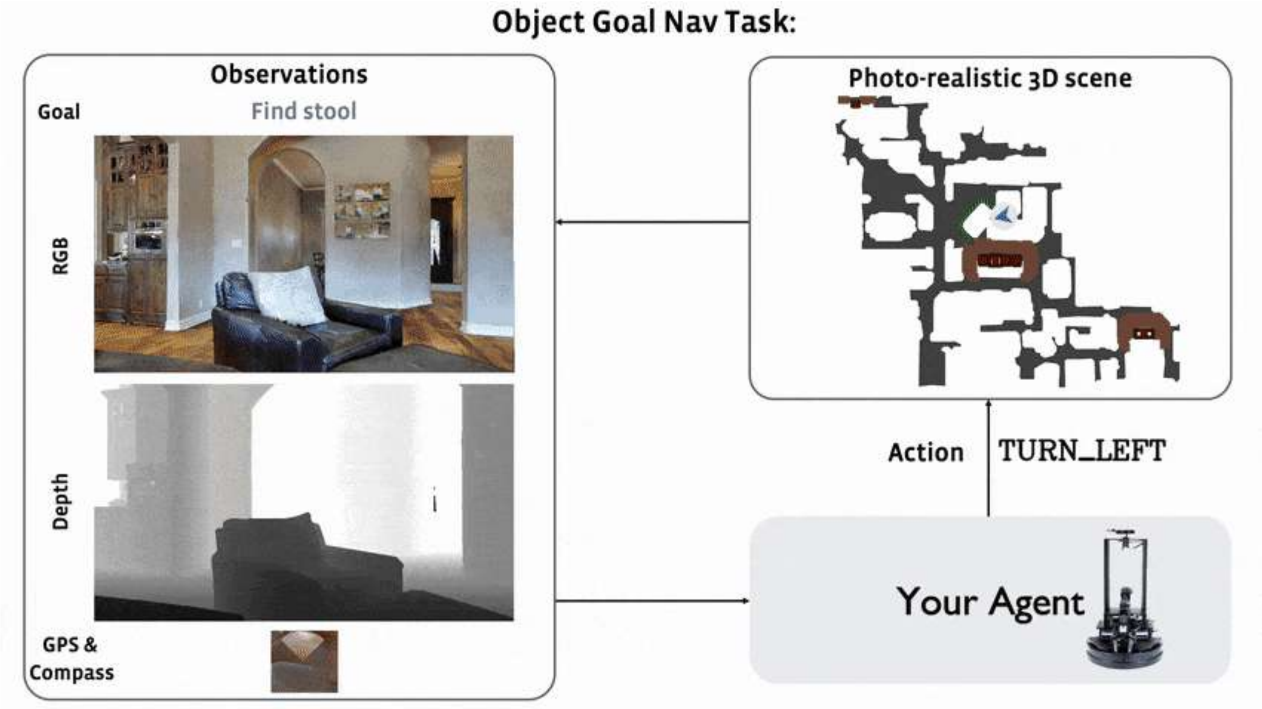
CVPR 2023 (submitted)

Nuri Kim, Jeongho Park, and Songhwai Oh., “**Relational Semantic Visual Graph for Object-Goal Navigation**,” Computer Vision and Pattern Recognition 2023 (CVPR-23, ***submitted***)

Object Goal Navigation

RILAB
Robot Learning Laboratory

- In ObjectNav, an agent is initialized at a random starting position and orientation in an unseen environment and asked to find an instance of an **object category** (*'find a chair'*) by navigating to it. A map of the environment is not provided and the agent must only use its sensory input to navigate.
- The agent is equipped with an **RGB-D camera and a (noiseless) GPS+Compass sensor**. GPS+Compass sensor provides the agent's current location and orientation information relative to the start of the episode. We attempt to match the camera specification (field of view, resolution) in simulation to the Azure Kinect camera, but this task does not involve any injected sensing noise.



Object Goal Navigation

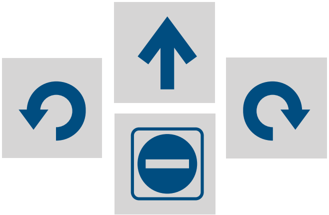
Input

Observation (RGBD)

Global Pose Sensor (x, y, θ)

Goal: Chair

Output



Action

Conclusion

- We propose detection algorithm for building ***semantic knowledge*** in passive learning methods.
- Using the know-how, we build ***navigation agents*** that can utilize semantic knowledge.
- The proposed approaches do not need a pose sensor for long-term planning, which makes the agent ***robust to noises*** and applicable to real-world applications.

Conclusion

Passive Learning

Interactive Learning

Object Detection

DOG, DOG, CAT

CVIU 2020

Image Goal

ICCV 2021
CoRL 2022 (oral)

Object Goal

Chair
TV
Sofa

CVPR 2023 (submitted)

Building navigation agents capable of *semantic understanding* by learning *relationship* of data using *graphs*

Future Directions

You are in a bedroom. Turn around to the left until you see a door leading out into a hallway, go through it. Hang a right and walk between the island and the couch on your left. When you are between the second and third chairs for the island stop.

Future Directions

RLLAB
Robot Learning Laboratory

- Active detectio

Table 1. **Habitat ObjectNav results on MP3D.** We report the results from the top-performing methods. † This is privileged.

Method	No Global Pose	MP3D (val)		
		Success ↑	SPL ↑	DTS ↓
BC	✗	3.8	2.1	7.5
DDPPO [38]	✗	8.0	1.8	6.9
Red-Rabbit [43]	✗	34.6	7.9	-
THDA [24]	✗	28.4	11.0	5.6
FBE [41]	✗	22.7	7.2	6.7
ANS [7]	✗	27.3	9.2	5.8
PONI [29]	✗	31.8	12.1	5.1
ANS + SI [3]	✗	27.9	13.1	6.1
SemExp + SI [3]	✗	34.7	**15.1**	5.8
RSVG (ours)	✔	**39.0**	[illegible]	[illegible]
RSVG − Update	✔	33.3	[illegible]	[illegible]
RSVG + GT†	✔	62.0	[illegible]	[illegible]

23% drop

*Object Goal Navigation Results on **MP3D** dataset*

Future Directions

- Active detection and navigation using Nerf

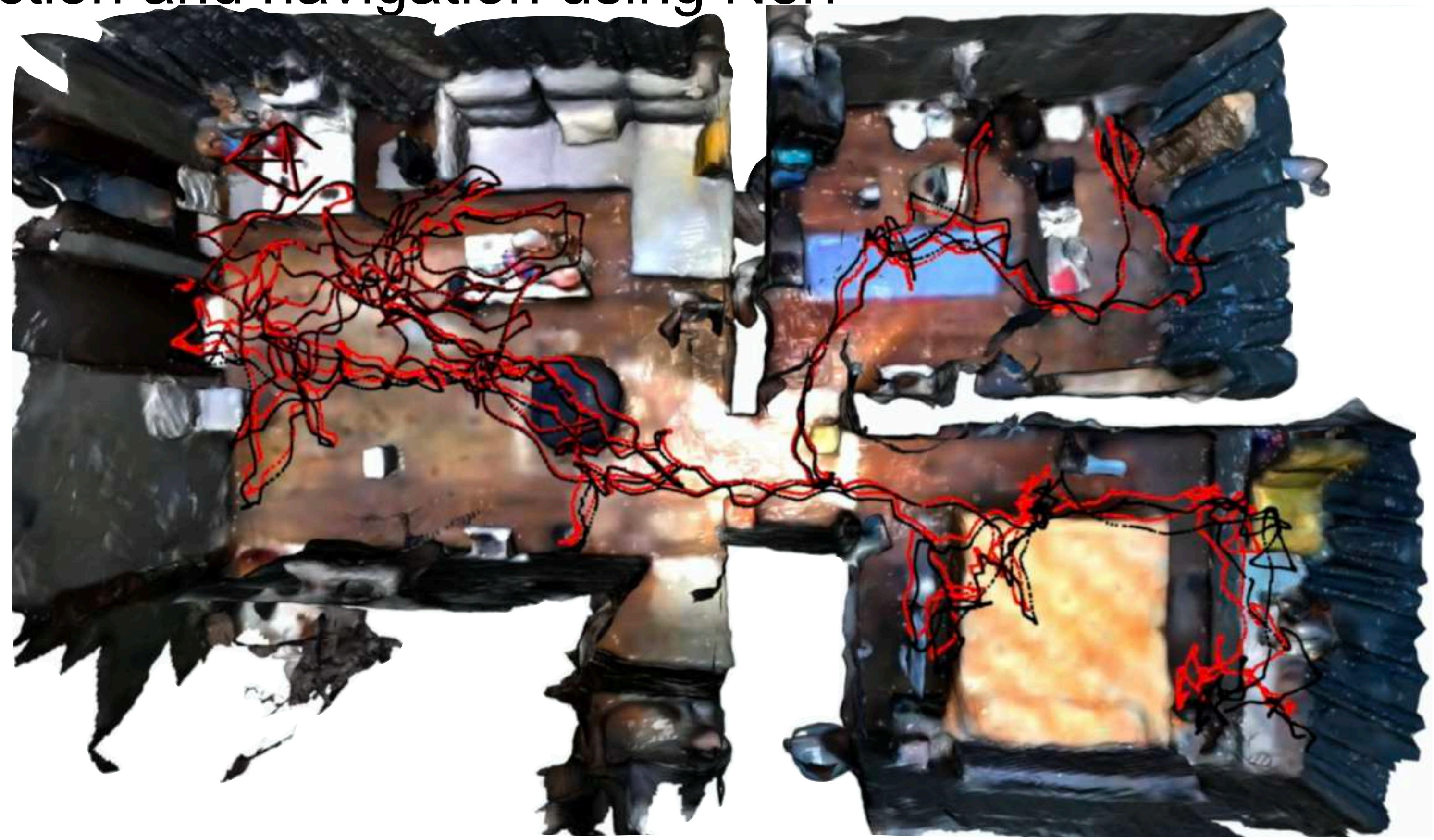
Object detector with passive learning

Object detector with interactive learning

Future Directions

RLLAB
Robot Learning Laboratory

- Active detection and navigation using Nerf



Interesting Papers on Visual Navigation

- Exploration
  - Mid-level visual representations improve generalization and sample efficiency for learning active tasks, CoRL 2019
  - SplitNet: Sim2Sim and Task2Task Transfer for Embodied Visual Navigation, ICCV 2019
  - Learning Exploration Policies for Navigation, ICLR 2019
  - Learning To Explore Using Active Neural SLAM, ICLR 2020
- Active Vision
  - Viewpoint Selection for Visual Failure Detection, IROS 2017
  - A dataset for developing and benchmarking active vision, ICRA 2017
  - Geometry-aware recurrent neural networks for active visual recognition, NIPS 2018
  - Learning to look around: Intelligently exploring unseen environments for unknown tasks, CVPR 2018
  - Embodied Visual Recognition, ICCV 2019
  - SEAL: Self-supervised Embodied Active Learning using Exploration and 3D Consistency, NeurIPS 2021

Interesting Papers on Visual Navigation

RLLAB
Robot Learning Laboratory

- Point Goal Navigation
  - A Behavioral Approach to Visual Navigation with Graph Localization Networks, RSS 2019
  - Learning Exploration Policies for Navigation, ICLR 2019.
  - Sparse Graphical Memory for Robust Planning, arXiv 2020
  - Active Neural Localization, ICLR 2018
  - Active Neural SLAM, ICLR 2020
- Image Goal Navigation
  - Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning, ICRA 2017
  - Semi-Parametric Topological Memory for Navigation, ICLR 2018
  - Sparse Graphical Memory for Robust Planning, arXiv 2020
- Object Goal Navigation
  - Auxiliary Tasks and Exploration Enable ObjectNav, ICCV 2021
  - Treasure Hunt Data Augmentation for Semantic Navigation, ICCV 2021
  - Object Goal Navigation using Goal-Oriented Semantic Exploration, NeurIPS 2020
  - Learning to Map for Active Semantic Goal Navigation, ICLR 2022
  - PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning, CVPR 2022

Interesting Papers on Visual Navigation

- Visual Language Navigation
  - Hierarchical Cross-Modal Agent for Robotics Vision-and-Language Navigation, ICRA 2021
  - Waypoint Models for Instruction-guided Navigation in Continuous Environments, ICCV 2021
  - LM-Nav: Robotic Navigation with Large Pre-Trained Models of Language, Vision, and Action, CoRL 2022

Research Tips

RLLAB
Robot Learning Laboratory

On Richard Feynman's problem solving

- The Feynman problem solving algorithm:

  1. Write down the problem
  2. Think very, very hard
  3. Write down the solution

Research Tips

RLLAB
Robot Learning Laboratory

- Keep up with recent researches
  - Google scholar keyword alerts
  - Paper study with colleagues
- Organize research materials
  - EndNote (paper)
  - Notion (research journal)
  - Slack (experimental results)
  - Github (code)
  - PPT (organize intersting papers in ppt)
  - LaTex (write paper -> experiment -> revise paper -> …, for this, use ChatGPT)
- Visualize your work
  - Wandb / Tensorboard (training)
  - ipython notebook (simple test/visualize)
  - at least plt.show()

Thank you for your attention