Navigating to Objects in the Real World

Theophile Gervet ¹
Soumith Chintala ⁴
Dhruv Batra ³,⁴
Jitendra Malik ²,⁴
Devendra Chaplot ⁴

¹ Carnegie Mellon University
² University of California Berkeley
³ Georgia Tech
⁴ Meta AI
Unseen environment: No experience, No map

Inputs
- Toilet
- Goal Category
- Observation (RGBD)
- Pose Sensor

Output

Goal Category: Toilet

Action
Spatial Scene Understanding
Navigable Space Detection
Spatial Scene Understanding
Navigable Space Detection

Semantic Scene Understanding
Object Detection

Chair
Oven
Spatial Scene Understanding
Navigable Space Detection

Semantic Scene Understanding
Object Detection

Semantic Exploration Priors
Where is a toilet more likely to be found?
Spatial Scene Understanding
Navigable Space Detection

Semantic Scene Understanding
Object Detection

Semantic Exploration Priors
Where is a toilet more likely to be found?

Episodic Memory
Keep track of explored and unexplored areas

Unexplored

Explored
Classical Navigation

([A Frontier-based Approach for Autonomous Exploration, Yamauchi, CIRA 1997])
End-to-end Learning

[Image of a diagram showing the flow of data from input modalities like RGB and depth, through segmentation and ResNet18 features, to a GRU network that processes previous action and hidden state to output action.]

[Habitat-Web, Learning Embodied Object-Search Strategies from Human Demonstrations at Scale. Ramrakhya et al., CVPR 2022]
Modular Learning

Classical

- Modular
- Explicit Memory/Maps & Planning
- Heuristic Policy
- Long-term Memory and Planning
- Semantic Exploration Priors

End-to-end Learning

- End-to-end
- Implicit Memory & Planning
- Learned Policy
- Long-term Memory and Planning
- Semantic Exploration Priors
Modular Learning

(Point Cloud) $X \quad Y \quad Z$

Semantic Scores $C_1 \quad C_2 \quad C_3$

Depth

RGB

Semantics

Pose

3D Semantic Map

2D Semantic Map with Semantic Goal

Goal-Oriented Semantic Policy

Project

Sum Height

Pose

Plant

Goal Category

Depth

RGB

Semantics

Action

Planner

Methods

Classical
- Geometric Map
- Heuristic Exploration
- No Training

End-to-end Learning
- End-to-end
- Large-scale IL + RL fine-tuning
  - 77,000 human trajectories
  - 200M frames of RL

Modular Learning
- Semantic Map
- Goal-Oriented Exploration
  - 10M frames of RL
Empirical Evaluation
3 Approaches
6 Unseen Homes
6 Goal Object Categories
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Sim</th>
<th>Real World</th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular Learning</td>
<td>0.81</td>
<td>0.90</td>
<td>0.47</td>
</tr>
<tr>
<td>Classical</td>
<td>0.78</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.23</td>
<td>0.77</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Classical Learning

End-to-end Learning
Goal: couch

SPL: 0.74, 78 steps

Modular

Third-person view

Success

Observation

Predicted Semantic Map

End-to-End

Third-person view

Failure

Observation

Predicted Semantic Map

Classical

Third-person view

Success

Observation

Predicted Semantic Map
Modular Learning is Reliable

Success Rate

<table>
<thead>
<tr>
<th></th>
<th>Sim</th>
<th>Real World</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular Learning</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td>Classical</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.77</td>
<td>0.23</td>
</tr>
</tbody>
</table>

SPL

<table>
<thead>
<tr>
<th></th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular Learning</td>
<td>0.47</td>
</tr>
<tr>
<td>Classical</td>
<td>0.64</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Classical vs Modular Learning

**Success Rate**

- **Modular Learning**
  - Sim: 0.81
  - Real World: 0.90
- **Classical**
  - Sim: 0.78
  - Real World: 0.80
- **End-to-end Learning**
  - Sim: 0.23
  - Real World: 0.77

**SPL**

- Modular Learning: 0.47
- Classical: 0.42
- End-to-end Learning: 0.16
Classical vs Modular Learning

**Goal:** bed

SPL: 0.90, 98 steps

**Semantic Exploration**

SPL: 0.52, 152 steps

**Frontier Exploration**
End-to-end fails to Transfer

Success Rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Sim</th>
<th>Real World</th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular Learning</td>
<td>0.81</td>
<td>0.90</td>
<td>0.47</td>
</tr>
<tr>
<td>Classical</td>
<td>0.78</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.77</td>
<td>0.23</td>
<td>0.42</td>
</tr>
</tbody>
</table>

SPL stands for Signal-to-Noise Ratio.
End-to-end Failures

Goal: Toilet

Goal: TV

Goal: Plant
Modular vs End-to-end Transfer

<table>
<thead>
<tr>
<th></th>
<th>Success Rate</th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sim</td>
<td>Real World</td>
</tr>
<tr>
<td>Modular Learning</td>
<td>0.81</td>
<td>0.90</td>
</tr>
<tr>
<td>Classical</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.23</td>
<td>0.78</td>
</tr>
</tbody>
</table>

SPL: Simultaneous Performance Level
Simulation

Real World

Predicted Semantic Map

Segmentation Model
Trained in Real World

mAP@0.5 = 0.50

Segmentation Model
Trained in Simulation

mAP@0.5 = 0.45

Domain Gap

Domain Gap

Domain Invariance

mAP@0.5 = 0.35

Bed false positive

Toilet false negative

Chair false negative

TV false negative

Plant false negative

0: chair
1: couch
2: potted plant
3: bed
4: toilet
5: tv
Modular Learning Sim vs Real

**Success Rate**

<table>
<thead>
<tr>
<th>Method</th>
<th>Sim</th>
<th>Real World</th>
<th>SPL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modular Learning</td>
<td>0.81</td>
<td>0.90</td>
<td>0.47</td>
</tr>
<tr>
<td>Classical</td>
<td>0.78</td>
<td>0.80</td>
<td>0.64</td>
</tr>
<tr>
<td>End-to-end Learning</td>
<td>0.77</td>
<td>0.23</td>
<td>0.58</td>
</tr>
</tbody>
</table>

- Modular Learning:
  - Sim: 0.81
  - Real World: 0.90
  - SPL: 0.47

- Classical:
  - Sim: 0.78
  - Real World: 0.80
  - SPL: 0.64

- End-to-end Learning:
  - Sim: 0.77
  - Real World: 0.23
  - SPL: 0.58
Modular Learning Sim vs Real

Success Rate

<table>
<thead>
<tr>
<th></th>
<th>Sim</th>
<th>Real World</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Visual Reconstruction Errors** (including Segmentation Errors)
- **Physical Reconstruction Errors**
- **Exploration Failures**
- **Depth Sensor Errors**
Real-world Depth Sensor Errors

- Door approach at an angle
- Noisy depth
- Closed door
Real-world Depth Sensor Errors

Mirror reflection

Reflected depth

Hallucinated bed mapped

Collisions in mirror
Takeaways

For practitioners:
• Modular learning can reliably navigate to objects with 90% success

For researchers:
• Models relying on RGB images are hard to transfer from sim to real ➡ leverage modularity and abstraction in policies
• Disconnect between sim and real error modes ➡ evaluate semantic navigation on real robots
Thank you!

Webpage: https://theophilegervet.github.io/projects/real-world-object-navigation