SYNCERE AI

UNIVERSITY OF TORONTO | Robotics Institute

LangRob @ CoRL 2024

Find Everything: A General Vision Language Model Approach to Multi-Object Search

Daniel Choi*, Angus Fung, Haitong Wang, Aaron Hao Tan

Background

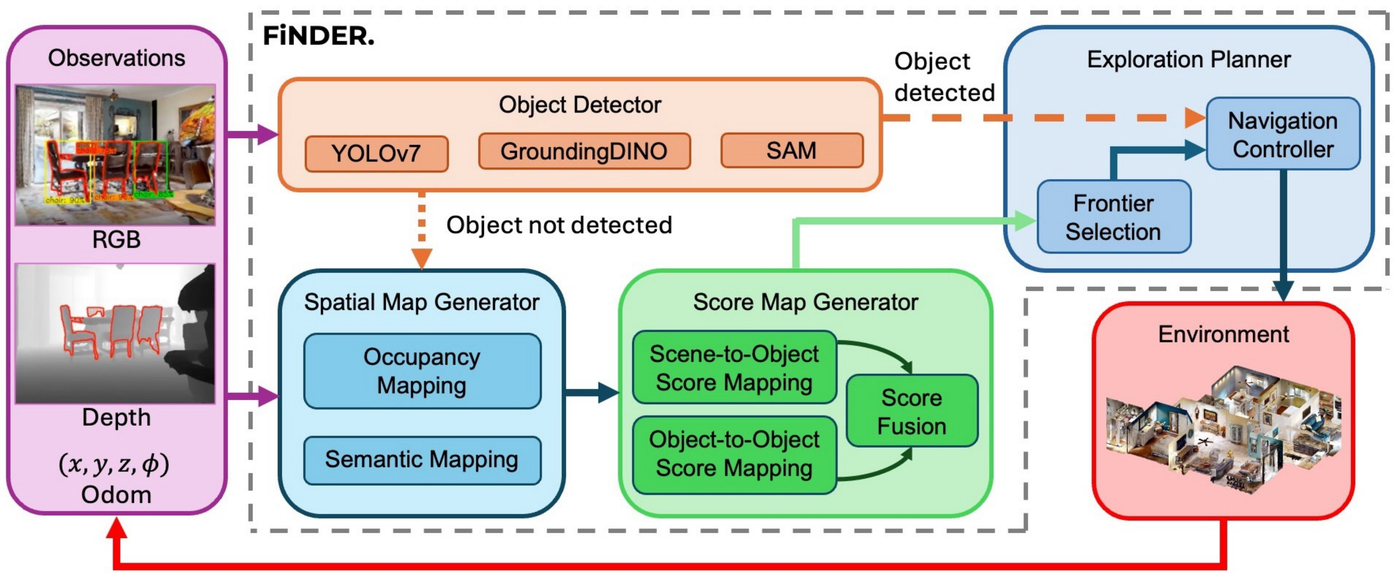
- **Multi-Object Search** Problem involves navigating to a sequence of locations to **maximize the likelihood of finding target objects while minimizing travel costs.**
- Existing approaches face challenges such as inefficient exploration due to **limited semantic modeling between objects and scenes, and poor generalization** caused by the **sim-to-real gap.**
- Single-Object Search methods cannot handle **multiple objects simultaneously** and rely on **coarse, noisy embeddings unsuitable for dense environments.**

Contributions

- **Multi-Channel Score Maps:** Introduced to simultaneously capture and track the semantic correlation between multiple target objects, the environment, and objects within the environment.
- **Fusion Technique:** Combines scene-level correlations with object-level correlations, to overcome the limitations of coarse scene-level embeddings.
- **Extensive Simulation and Real-World Validation:** Demonstrated extensive simulation and real-world experiments to validate Finder's performance. The code is also available upon request.

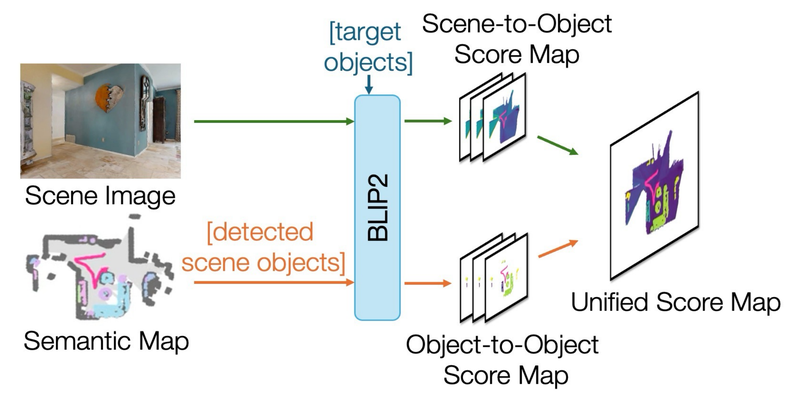
FiNDER

Architecture



- **Object Detector:** Detects both target and scene objects using YOLOv7 and G-DINO for Open Vocabulary Detection.
- **Spatial Map Generator:** Builds occupancy and semantic maps from RGB-D data.
- **Exploration Planner:** Selects navigation waypoints based on unified score maps.
- **Score Map Generator:** Two-part score mapping (StO and OtO) and fusion for guiding navigation.

Unified Score Map



- **Scene-to-Object (StO) Map** is the product between the confidence mask for field-of-view with cosine similarity score between scene and target embeddings.
- **Object-to-Object (OtO) Map** weighs each scene object by their semantic similarity to target objects, creating a map of relevant object correlations.
- **Unified Score Map** is the element-wise addition of the OtO and StO map.

Experiments

Habitat Simulator

Score Map Channel 1 (Chair)

Score Map Channel 2 (TV)

Score Map Channel 3 (Bed)

Unified Score Map

Description: Demo of multi-channel score maps for multi-object search.
Environment: HM3D Dataset
3 Target Objects: Chair, TV, Bed

Sim-to-Real

- **Environment areas:** Study, Fireplace and Lounge.
- **Equipment:** TurtleBot with Kinect camera capturing RGB-D.
- **Target Objects:** Garbage Bin, Fireplace, Laptop, Shoes, etc.
- **To assess:**
  - Finder's **generalizability** in real-world environments.
  - Finder's **scalability** in handling increasing # of targets

Results

Comparison | Ablation | Scalability

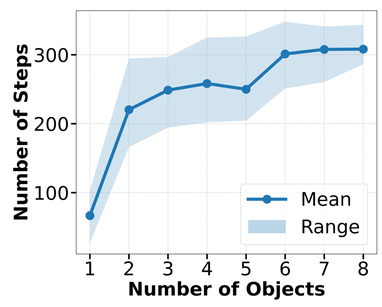


Table 1: Comparison between Finder and SOTA methods

Methods	HM3D SR↑	HM3D MSPL↑	MP3D SR↑	MP3D MSPL↑
Random Walk	0.5%	0.0043	0.0%	0.0
MultiON	–	–	23.9%	0.159
CoW	14.2%	0.113	1.9%	0.059
L3MVN (Zero-Shot)	27.2%	0.187	6.6%	0.043
L3MVN (Feed-Forward)	28.1%	0.188	7.3%	0.051
VLFM	32.4%	0.155	12.6%	0.104
Oracle	100.0%	1.0	100.0%	1.0
Finder (ours)	**63.4%**	**0.389**	**55.4%**	**0.344**

Table 2: Ablation Results

Variants	SR↑	MSPL↑
Finder w/o StO	61.5%	0.364
Finder w/o OtO	58.3%	0.337
Finder (ours)	**63.4%**	**0.389**

- **Comparison Study** (SOTA methods)
- **Ablation Study** (StO and OtO)
- **Scalability Study** (# of Objects)

Takeaways

- **FiNDER:** A VLM-based approach solving the multi-object search problem across diverse environments.
- **Superior Performance:** Outperformed SOTA methods in simulated and real-world tests in success rate (SR) and path efficiency (MSPL).
- **Future Goals:** Expand to handle dynamic objects and interactive search scenarios.

Links

FiNDER — Video / Website

Author's Website

SYNCERE AI — SYNCERE AI