Learning to Explore, Navigate and Interact for Visual Room Rearrangement

Ue-Hwan Kim*, Young-Ho Kim*, Jin-Man Park, Hwan-Soo Choi, and Jong-Hwan Kim
KAIST

CVPR VIRTUAL JUNE 19-25

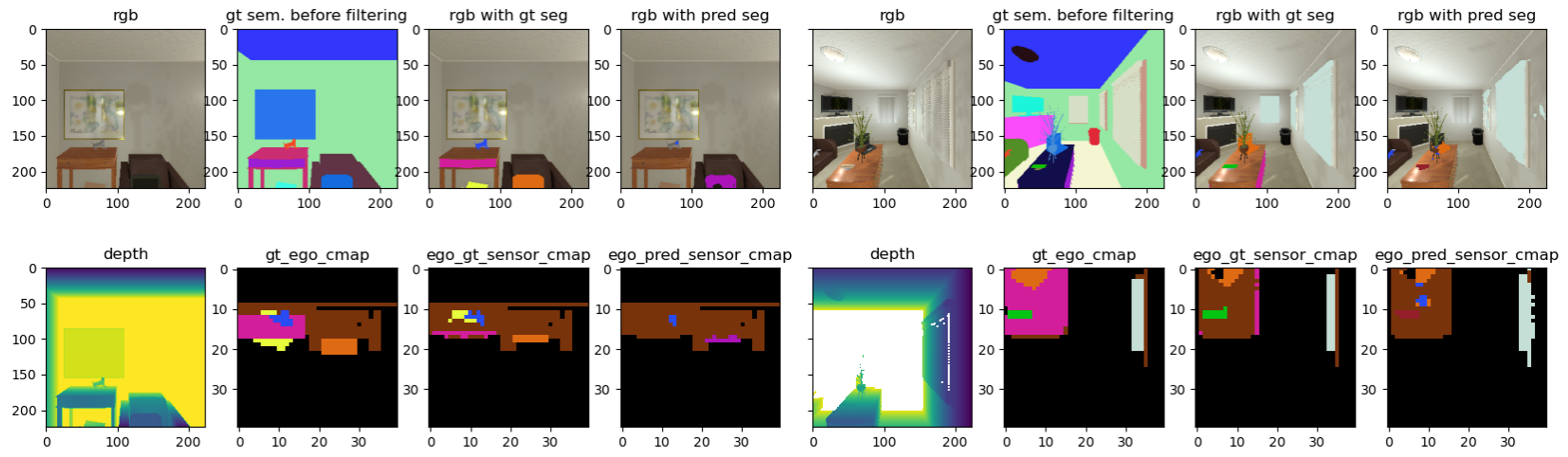
I. Introduction

- Embodied AI
  - Learning of embodied physical interactions with surrounding environments
  - Tasks involving direct physical interaction with objects are drawing increasing attention
  - The visual room rearrangement task

- Motivation
  - An end-to-end formulation
    - Straightforward
    - Expensive cost of pure learning
  - A three-phased modular architecture (TMA)
    - Learning modules along with hand-crafted feature processing modules
    - Advantage of learning + reduced cost of learning

II. Semantic Mapping

- Semantic Map Construction
  - Map representation: K x M x M [1]
    - An obstacle map, the explored area, the current agent location, the past agent locations, and C categories of semantics
  - Semantic segmentation: Swin Transformer [2]



III. TMA

- Phase 1: Exploration
  - Long-term goal
    - Reinforcement learning module [3]
    - (input) current semantic map → (output) long-term goal
    - Reward: newly explored area
  - Short-term goal:
    - Planning module (knowledge-base)
    - (input) long-term goal → (output) sequence of actions
    - The shortest path from the current location to the long-term goal

- Phase 2: Inspection
  - Identical structure as Phase 1
  - Long-term goal
    - Reinforcement learning module [3]
    - (input) current semantic map + map from phase 1

- Phase 3: Rearrangement
  - Change detection
    - Changes: location and state of objects
    - Distance metric: class and size similarities

$$d = w_{\text{class}} \cdot s_{\text{class}} + w_{\text{class}} \cdot s_{\text{size}}$$

  - Planning
    - The order of rearranging each object which is optimal in respect of time complexity
    - Selecting one order from N! permutations of orders (N < 5)
  - Rearrangement
    - Rearrange each object step by step
    - A* planner → sequence of actions

IV. Experiment

- Settings
  - AI2-THOR Rearrangement Challenge
    - 2-Phase track: walkthrough and un-shuffle phases
    - 6,000 unique rearrangement scenarios (4,000/1,000/1,000 for train, validation and test, respectively)
  - Metrics and results
    - Success rate, % Fixed Strict, % Energy Remaining and % Misplaced for each split of dataset

Split	100·Success Rate	100·%Fixed Strict	%E	% Misplaced
Train	0.5	1.0	1.01	1.00
Val.	0.0	0.9	1.00	1.00
Test	0.1	0.6	1.01	1.01

V. Conclusion

- Contribution
  - A three-phased modular architecture (TMA) for visual room rearrangement
    - Taking advantages of deep learning in understanding of room environment
    - Ensuring robustness of long-horizon decision making via planning

References

[1] Chaplot, Devendra Singh, et al. "Semantic curiosity for active visual learning." ECCV, 2020.
[2] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv, 2021.
[3] John Schulman, et al. "Proximal policy optimization algorithms." arXiv, 2017