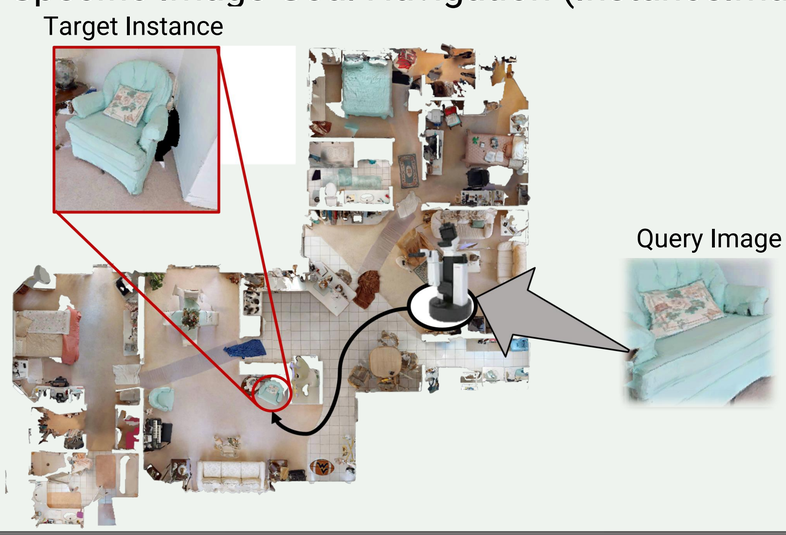
Object Instance Retrieval in Assistive Robotics: Leveraging Fine-Tuned SimSiam with Multi-View Images Based on 3D Semantic Map

Taichi Sakguchi, Akira Taniguchi, Yoshinobu Hagiwara, Lotfi El Hafi, Shoichi Hasegawa, Tadahiro Taniguchi
Ritumeikan University

Task & Motivation

Task

Instance Specific Image Goal Navigation (InstanceImageNav)[1]



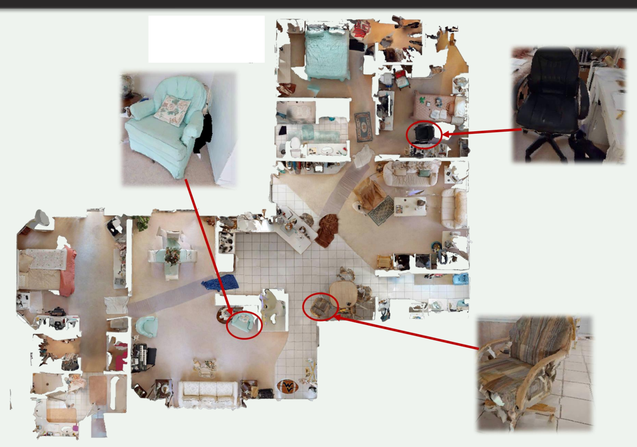
Motibation

Searching Object is fundamental capability for human support robots
In an environment, there are multiple instances of same class

Important

In such environment,
Robot must classify between instances of the same class of objects

Challenges



In domestic environment, there are multiple instances of same class

Challenge1

In such scenario, the robot needs to classify objects fine-grained

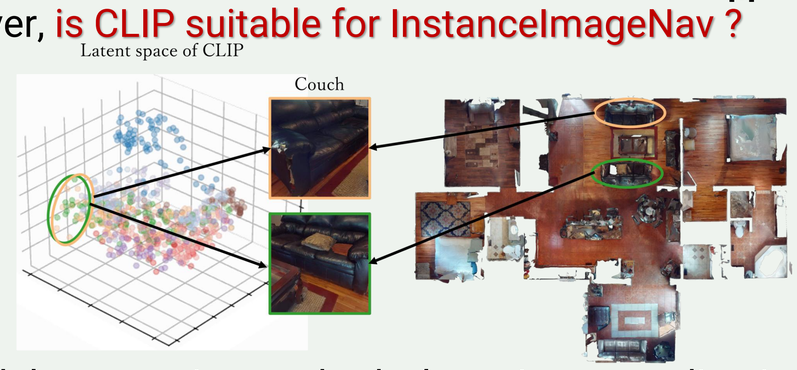
Given query image may be taken from a different viewpoint than when the robot observes the target object

Challenge2

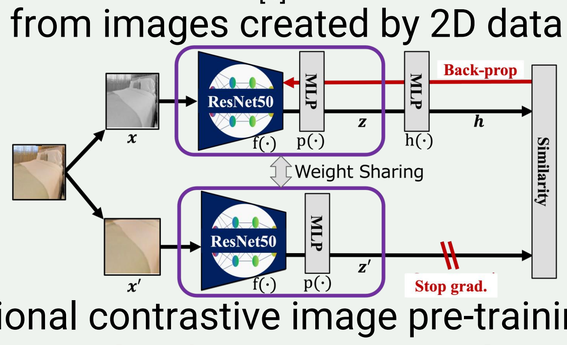
It is important to learn viewpoint invariance of instance images

Previous works

CLIP[2] attracts attention in object searching task[3]
However, is CLIP suitable for InstanceImageNav ?



Unimodal contrastive method[4] learn instance discriminative representation from images created by 2D data augmentation



Dose conventional contrastive image pre-training methods learn the similarity of multi-view images of an instance ?

Proposed Method

Main Idea

- Learning instance discriminative representation by leveraging contrastive learning
- Learning viewpoint invariant instance representations utilizing contrastive learning with multi-view images

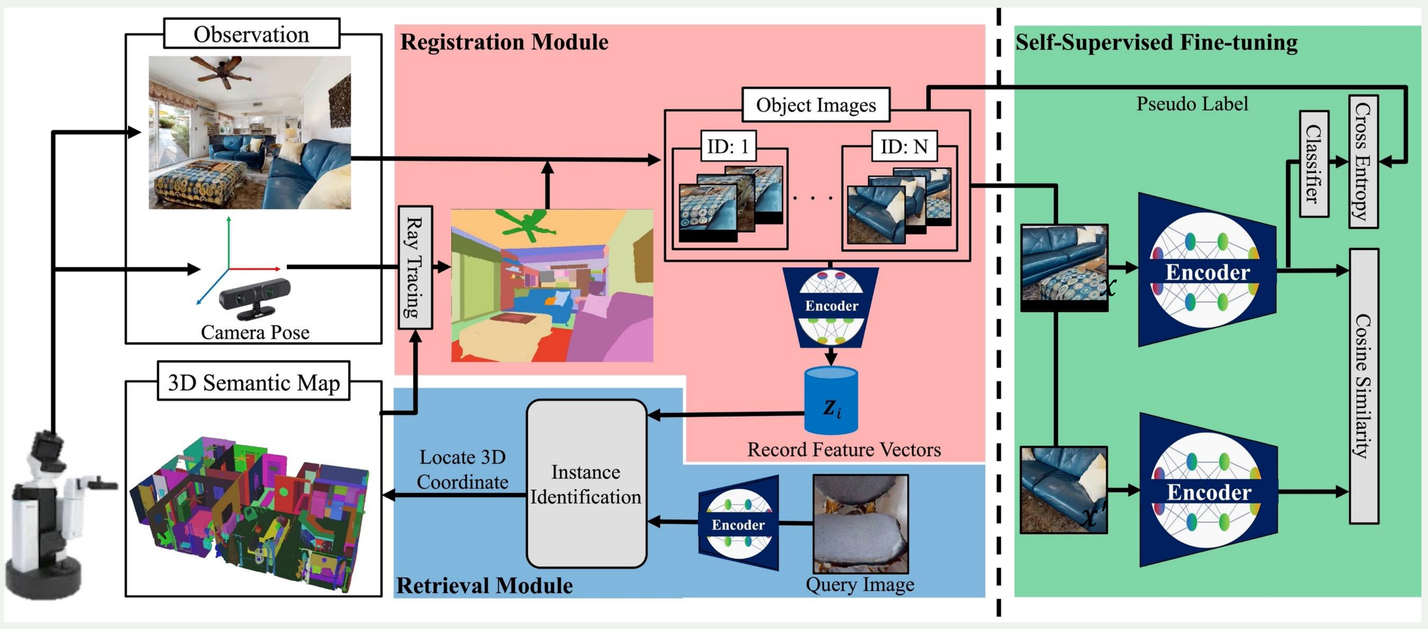
x, x' : Same instance images observed by robot from different view-points

y, y' : Instance ID of x, x' labeled by robot utilizing 3D Semantic Map

$\hat{y}, \hat{y'}$: Instance ID of x, x' labeled by robot utilizing 3D Semantic Map

$$L = \frac{1}{2}\{\mathrm{CosSim}(f(x), h(x')) + \mathrm{CosSim}(f(x), h(x'))\} + \frac{1}{2}\{\mathrm{CE}(\hat{y}, y) + \mathrm{CE}(\hat{y'}, y')\}$$

Proposal: Semantic Instance Multi-view Contrastive Fine-tuning (SimView)



Experiment Result

Purpose

1. Evaluating if contrastive learning of only image pairs is more suitable than CLIP for identifying the instance
2. Evaluating if SimView learn viewpoint invariance of same instance images compared with prior contrastive methods

Dataset : 9 scenes which are included in Habitat Matterport 3D[5]

Task : Image retrieval
Evaluate whether the robot can find instances that are identical to the query image.

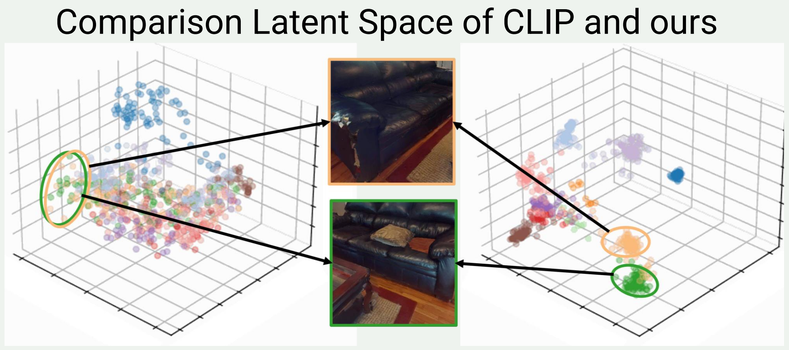
Metrics : mean Average Precision

$$mAP = \frac{1}{\mathrm{K}}\sum_{k=1}^{\mathrm{K}} \mathrm{AP}_k$$

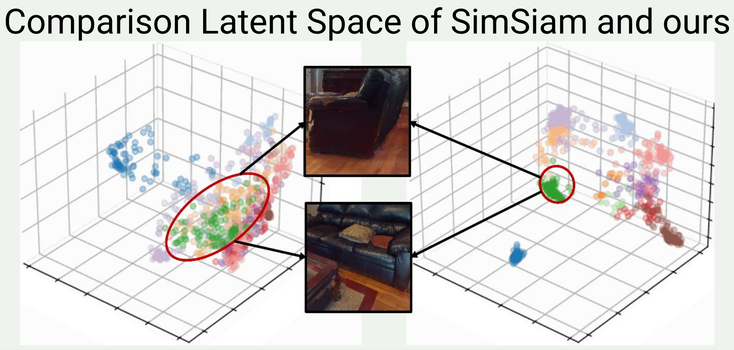
K : Number of trials
AP_k : Average Precision of k-th trial

Method	Arch.	Env. 1	Env. 2	Env. 3	Env. 4	Env. 5	Env. 6	Env. 7	Env. 8	Env. 9	Avg.
SimView (ours)	ResNet50	**0.79**	**0.67**	**0.68**	**0.91**	0.68	**0.45**	**0.72**	**0.74**	**0.8**	**0.72**
SimSiam	ResNet50	**0.70**	0.58	0.56	0.80	0.65	0.41	0.63	0.68	0.75	0.64
DINOv2	ViT-B/14	0.51	0.48	0.45	0.66	**0.71**	0.41	0.55	0.44	0.61	0.54
SimCLR	ResNet50	0.65	0.56	0.52	0.83	0.64	**0.45**	0.66	0.62	0.75	0.63
CLIP	ResNet50	0.51	0.36	0.35	0.51	0.48	0.44	0.42	0.44	0.59	0.46
CLIP	ViT-B/16	0.56	0.38	0.37	0.47	0.48	0.37	0.38	0.35	0.48	0.43

Comparison Latent Space of CLIP and ours



Comparison Latent Space of SimSiam and ours



Suggestion

1. Unimodal contrastive learning methods better at identifying instances than CLIP
2. SimView learn more invariant feature representations than SimSiam

[1] Krantz, Jacob, et al. "Navigating to objects specified by images." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.
[3] Chen, Boyuan, et al. "Open-vocabulary queryable scene representations for real world planning." *IEEE International conference on Robotics and Automation*. 2023.
[4] Chen, Xinlei,et al. "Exploring simple siamese representation learning." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
[5] Yadav, Karmesh, et al. "Habitat-matterport 3d semantics dataset." *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2023.