Instance-Guided Unsupervised Domain Adaptation for Robotic Semantic Segmentation

MICHELE ANTONAZZI micant@kth.se

LORENZO SIGNORELLI lorenzo.signorelli@student.unimi.it

MATTEO LUPERTO matteo.luperto@unimi.it

NICOLA BASILICO nicola.basilico@unimi.it

ICRA

Context

Domain shift problem in semantic segmentation for indoor robotics

- Training on a source domain: general-purpose datasets, often in controlled conditions
- Testing on a target domain: unseen environments, often with novel visual aspects

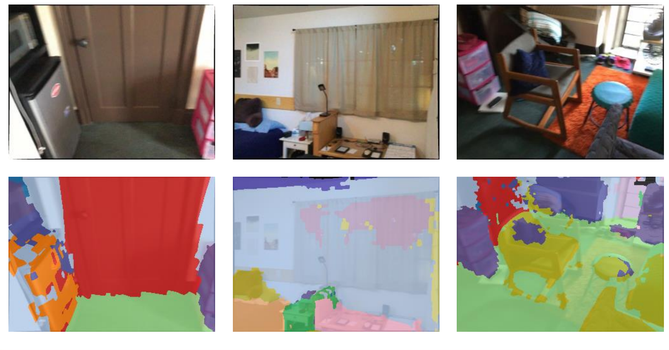
} Performance degradation

- Fine-tuning with manual annotations is not practical in real-world robot deployments
- Self-supervision with pseudo-labels is a common solution for unsupervised adaptation[1]

Motivation

Multi-view consistent self-supervision

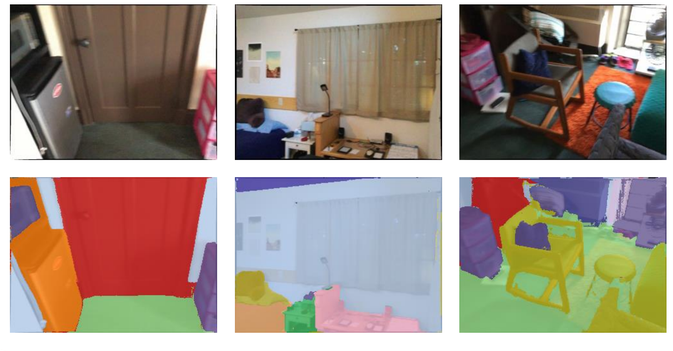
- Aggregating pseudo-labels from different point of views into the same 3D region (voxels in a 3D map) to force spatial consistency
- Obtaining spatially consistent pseudo-labels from the map
- **Limitations**:
  - Visual artifacts produced by the rendering process
  - Multiple categories can be assigned to the same object



Solution

Instance-guided unsupervised adaptation

- Revise pseudo-labels exploiting object-instance consistency: propagate pseudo-labels using guidance from an instance map
- We exploit the zero-shot instance segmentation model (Segment Anything v2[3]) devising two complementary prompting strategies



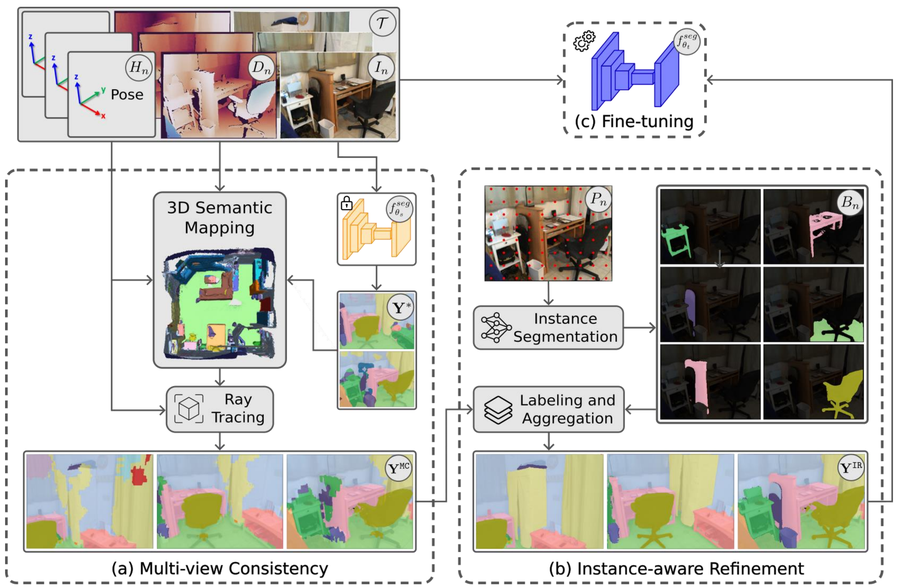
Method

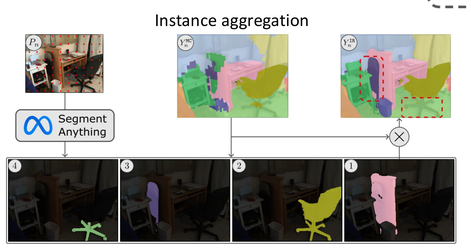
a) Multi-view consistency

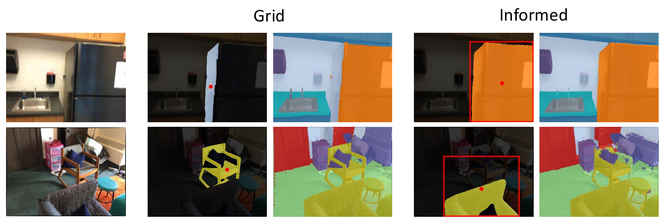
- Spatially-consistent pseudo-labels $\mathbf{Y}^{\mathrm{MC}}$ are ray-traced from the map

b) Instance-aware refinement

- Object instances B are obtained using SAM v2
  - **Grid:** the prompts are a grid of points uniformly distributed in the image
  - **Informed:** point and bounding box pairs derived from the pseudo-labels
  - **Combined:** pseudo-labels obtained with both strategies are used during training
- Object instances are labeled using the object categories in Y_n^{MC}
- Labeled instances are aggregated in an instance-refined pseud-label Y_n^{IR}







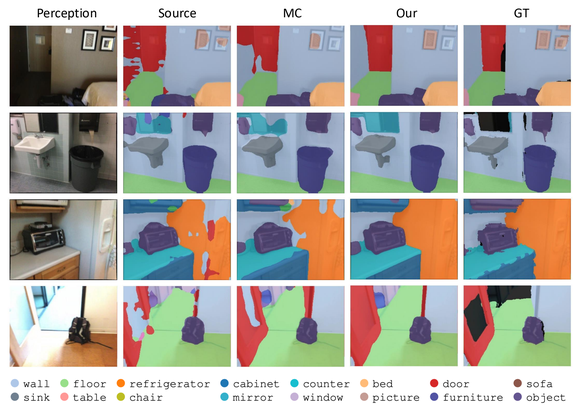
Results

Evaluation protocol

- **Dataset:** Scannet[4], environments 1-10 for adaptation/testing, 11-707 for initializing the source model
- **Model:** DeepLab v3
- **Mapping framework:** Kimera semantics
- **Baseline:** multi-view consistency[2] (MC)

Results

- Our instance refinement (IR) step improves pseudo-labels quality
- Fine-tuning with our improved pseudo-labels improve model performance
- Prompting strategies performs similarly but Combined wins



Env	Source	Baseline MC	Baseline Δ_{PT}	Our method IR_G	Our method Δ_{MC}	Our method IR_I	Our method Δ_{MC}	Our method IR_{GI}	Our method Δ_{MC}
1	47,1	49,6	5,3%	**55**	10,9%	51,2	3,2%	54,3	9,5%
2	**44,1**	35,2	-20,2%	39,6	12,5%	39,6	12,5%	39,3	11,6%
3	30,6	30,5	-0,3%	32,8	7,5%	34,4	12,8%	**34,7**	13,8%
4	55,3	55,8	0,9%	57,2	2,5%	56,2	0,7%	**57,7**	3,4%
6	49,2	50,5	2,6%	52,8	4,6%	**54,7**	8,3%	53	5,0%
7	52	54,4	4,6%	55,5	2%	**57,2**	5,1%	56,2	3,3%
10	57,5	63,2	9,9%	68,7	8,7%	67,6	7,0%	**68,7**	8,7%
Avg	48	48,5	0,4%	51,7	7%	51,6	7,1%	**52,0**	7,9%

References

[1] M. Schwonberg, et al., "Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving," IEEE Access, 2023

[2] J. Frey, et al., "Continual adaptation of semantic segmentation using complementary 2d-3d data representations," IEEE RAL, 2022

[3] N. Ravi, et al., "Sam 2: Segment anything in images and videos," 2024.

[4] A. Dai, et al., "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in Proc. of CVPR, 2017.