Incremental Class Discovery for Semantic Segmentation with RGBD Sensing

Yoshikatsu Nakajima[1,2], Byeongkeun Kang[1], Hideo Saito[2], Kris Kitani[1]

[1]Carnegie Mellon University, [2]Keio University

ICCV 2019 Seoul, Korea

← Video Results!

Motivation

Semantic Scene Reconstruction

A task of incrementally building a dense, semantically annotated 3D map in real-time

Issue

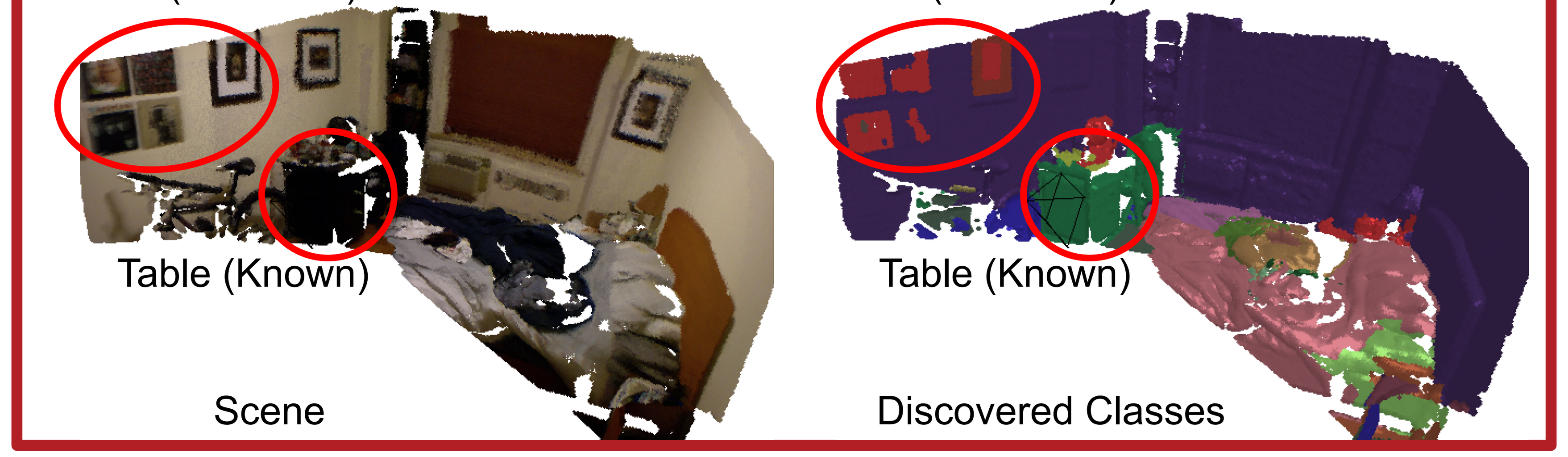
In real-world, many types of objects exist.
However, most approaches [1, 2] assume closed world.

Our Approach

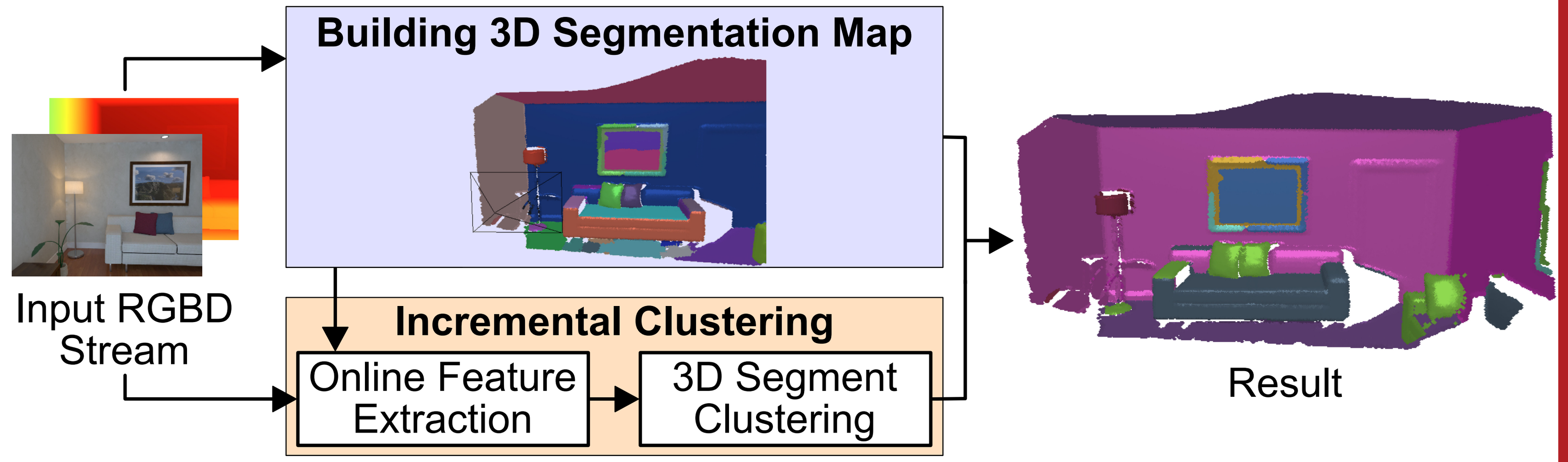
Incrementally segment both learned and unseen classes

Key ideas to meke clusters

- Utilize deep features for grouping learned object classes
- Utilize geometric features for grouping unseen object classes



Overview



Building 3D Segmentation Map

- To identify object regions in the scene
- Use for aggregating information from 2D image segmentation

Incremental Clustering

- Associate objects of the same class and discover new classes

Method

Building 3D Segmentation Map

Dense SLAM

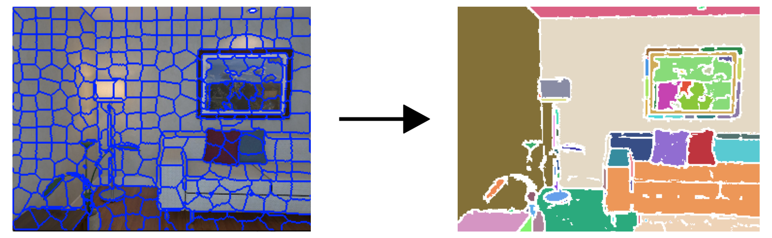
- To estimate camera poses and build a 3D map

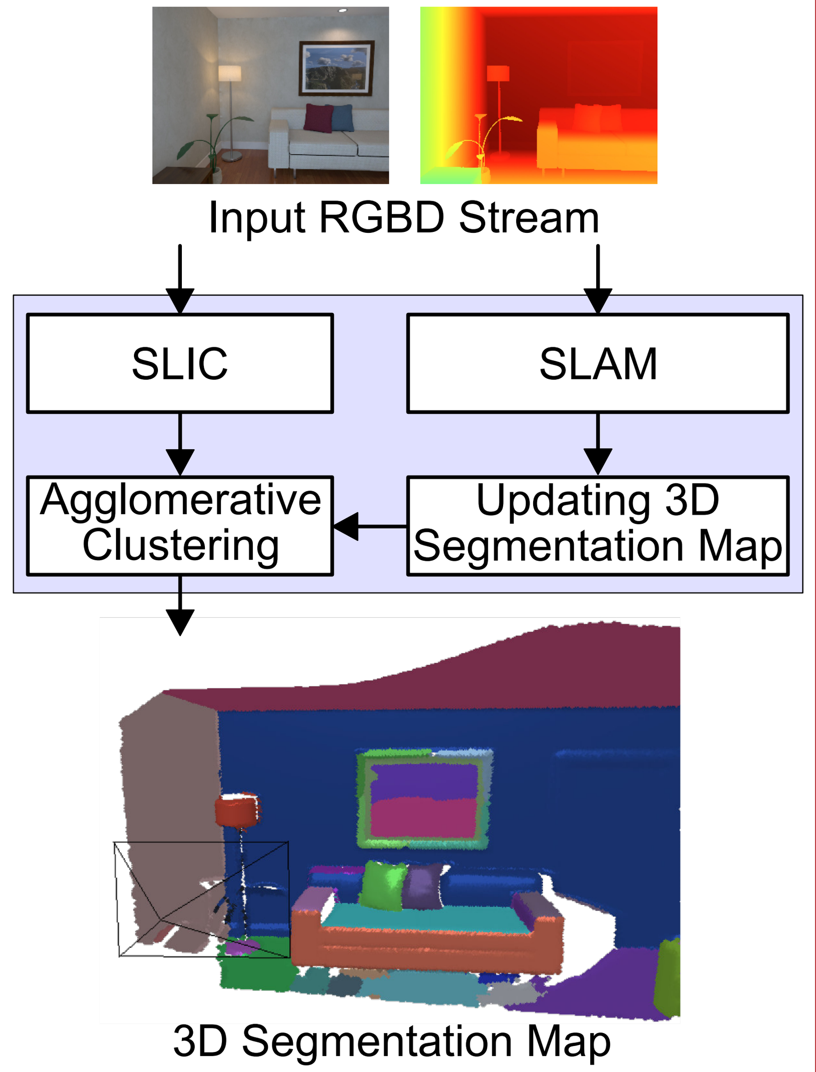
RGBD SLIC Superpixel Segmentation

- Distance metric using CIELAB color, normal, and image coordinates

Agglomerative Clustering

- Merge superpixels to produce object-level segments
- Based on similarity in color, geometric distance in 3D, and convexity in shape





3D Segmentation Map Update

- Update the 3D map using the 2D segmentation result of current frame

Incremental Clustering

Online Feature Extraction

Assign and update the following features to each region in the 3D segmentation map

- Deep features - To recognize learned object classes
- Geometric features - To recognize unseen object classes
- Entropy - To estimate the reliability of the deep features

$$f^{\text{CNN}}_{l_i=\mathcal{R}(u)} \leftarrow \frac{\Gamma f^{\text{CNN}}_{l_i=\mathcal{R}(u)} + \mathcal{F}^{\text{CNN}}_t(u)}{\Gamma+1}, \ \Gamma \leftarrow \Gamma+1$$

$f^{\text{CNN}}_{l_i}$: Deep feature assigned to region l_i of the 3D segmentation map
$\mathcal{F}^{\text{CNN}}_t$: Feature map of U-Net, $\mathcal{R}$: Camera view of the 3D segm. map

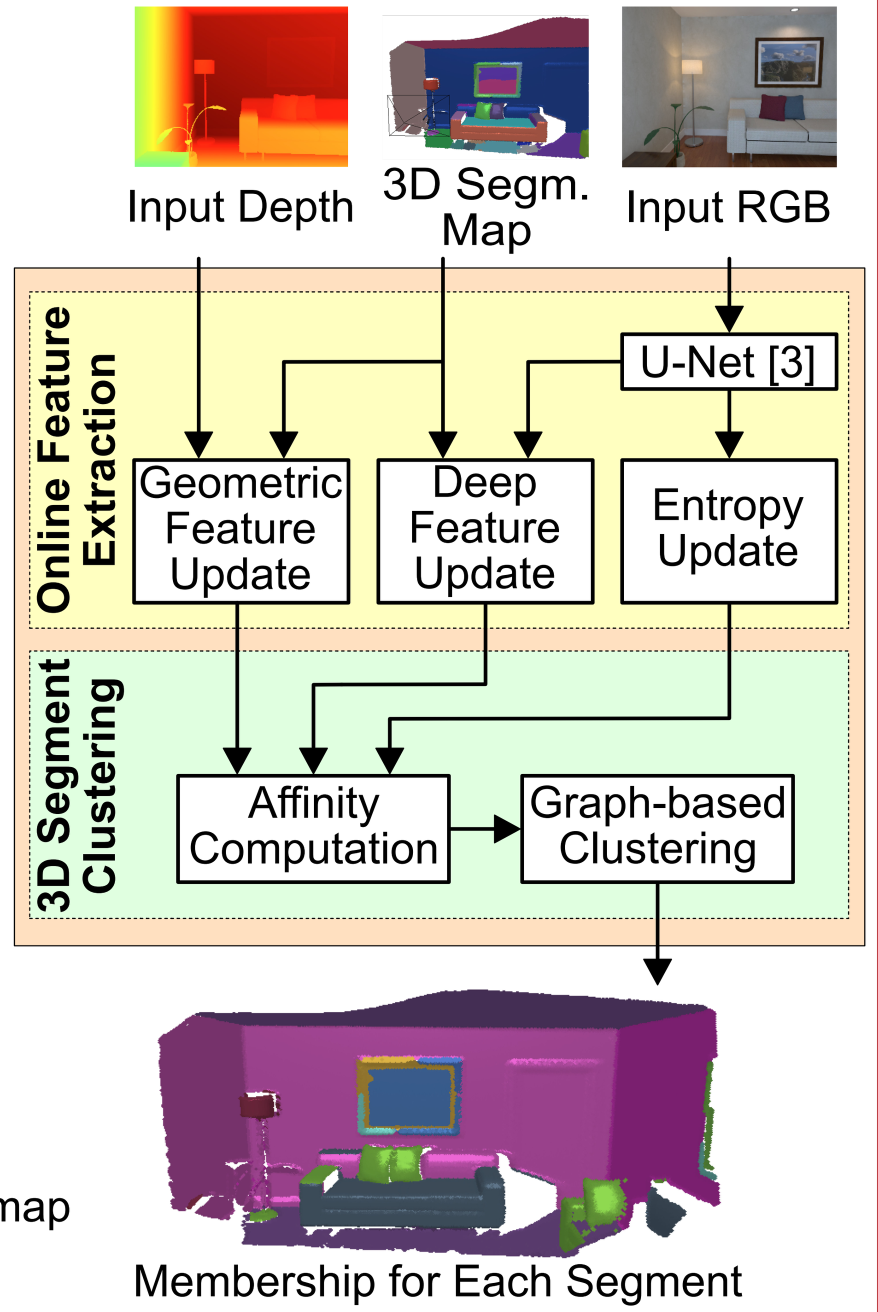
3D Segment Clustering

- Compute affinities between each region using the assigned features with weighting based on entropy
- Feed the affinity matrix to Markov clustering

$$w_i = \frac{e_{l_i}}{\log N}, \ w_j = \frac{e_{l_j}}{\log N}$$

$$\text{distance}(i,j) = \|(1-w_i)f^{\text{CNN}}_{l_i} - (1-w_j)f^{\text{CNN}}_{l_j}\|_2 + \|w_i f^{\text{GEO}}_{l_i} - w_j f^{\text{GEO}}_{l_j}\|_2$$

$f^{\text{GEO}}_{l_i}$: Geometric feature assigned to region l_i of the 3D segmentation map
e_{l_i}: Entropy assigned to region l_i, N: Number of learned classes



Results

Verified our method on the NYUDv2 dataset [4].
Trained U-Net excluding Ceiling, Picture, TV, and Window.

Accuracy

Table: **Quantitative comparison.** Supervised methods vs open set method (ours)

Method	Classes in training data									Novel classes				Mean IoU
	Bed	Book	Chair	Floor	Furn.	Obj.	Sofa	Table	Wall	Ceil.	Pict.	TV	Wind.	
U-Net[3]	50.3	22.4	36.6	55.6	36.9	27.3	48.4	33.8	55.1	-	-	-	-	-
[2]	62.8	**27.3**	**42.6**	**68.4**	44.6	24.6	45.0	**42.3**	26.8	-	-	-	-	-
Ours	**64.2**	22.3	41.8	67.4	**56.2**	**28.6**	**49.3**	41.0	**63.2**	**29.3**	**28.7**	**52.2**	**53.9**	**46.1**

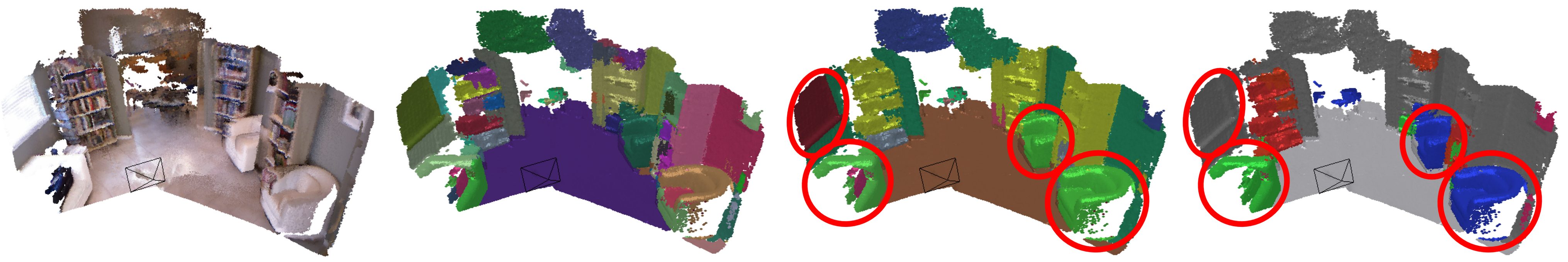


Fig.1: **Qualitative results.** The proposed method discovers various classes including both unseen classes and the classes in the training dataset.

Efficiency

Average processing time: 93.2 ms/frame (10.7 Hz)
Space complexity: # of regions × dimension of features

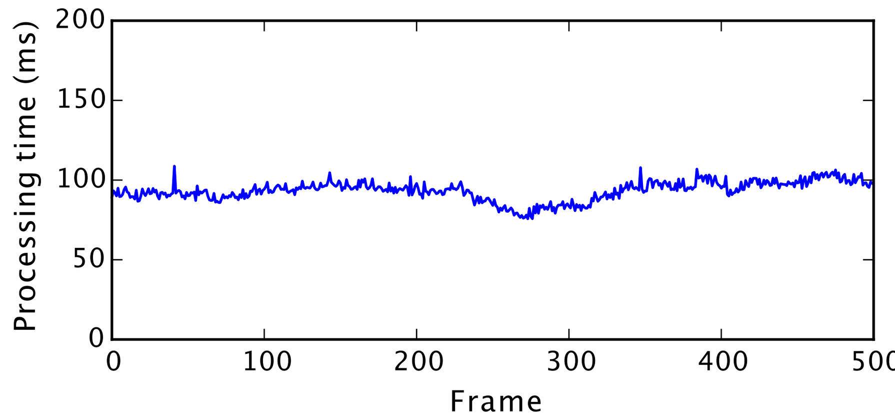


Fig.2: **Runtime analysis.** The processing time is stable even though the 3D map grows larger.

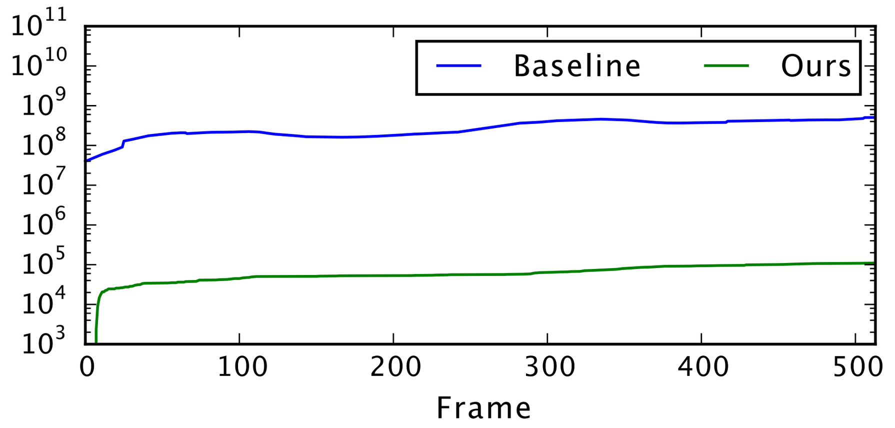


Fig.3: **Memory usage.** The baseline assigns features to each element, *i.e.* 3D point, of the 3D map as in [1].

[1] McCormac *et al.*, "SemanticFusion: Dense 3D Semantic Mapping with Convolutional Neural Networks," ICRA 2017
[2] Nakajima *et al.*, "Fast and Accurate Semantic Mapping through Geometric-based Incremental Segmentation," IROS 2018
[3] Ronneberger *et al.*, "U-Net: Convolutional Networks for Biomedical Image Segmentation," MICCAI 2015
[4] Silberman *et al.*, "Indoor Segmentation and Support Inference from RGBD Images," ECCV 2012