LEARNING RGB-D FEATURE EMBEDDINGS FOR UNSEEN OBJECT INSTANCE SEGMENTATION

Yu Xiang, 10/12/2020
How can a robot manipulate objects in this cluttered kitchen?
MODEL-BASED OBJECT RECOGNITION

3D models

Not scalable

PoseCNN + PoseRBPF

Xiang et al. RSS’18
Deng et al. RSS’19
Can we train a model to segment unseen objects in images?

It is difficult to obtain 3D model for every object.
SEGMENTATION ENABLES GRASPING

Unseen Object Segmentation + GraspNet

Xie et al. CoRL’19
Mousavian et al. ICCV’19
LEARNING THE CONCEPT OF “OBJECT”

- Learning from data

ImageNet: Deng et al. CVPR’09

Internet Images, not suitable for indoor robotic settings

COCO Dataset: Lin et al. ECCV’14
LEARNING FROM SYNTHETIC DATA

RGB

Depth

Instance Label

ShapeNet objects in the PyBullet simulator

Xie et al. CoRL’19

Need to deal with the sim-to-real gap

40,000 scenes
7 RGB-D images per scene
PREVIOUS WORKS: LEARNING FROM DEPTH

- Synthetic depth generalizes better to the real depth images

Dex-Net 2.0
Mahler et al. RSS’17

UOIS-Net
Xie et al. CoRL’19
CAN WE UTILIZE NON-PHOTOREALISTIC SYNTHETIC RGB IMAGES?

- Depth is not good for transparent objects or thin objects

ClearGrasp
Sajjan et al. ICRA’20
OUR WORK: LEARNING RGB-D FEATURE EMBEDDINGS FOR SEGMENTATION

METRIC LEARNING LOSS FUNCTION

- Intra-cluster loss function

\[
\mu^k = \frac{\sum_{i=1}^{N} x_i^k}{\| \sum_{i=1}^{N} x_i^k \|}
\]

Spherical mean

\[
d(\mu^k, x_i^k) = \frac{1}{2} (1 - \mu^k \cdot x_i^k)
\]

Cosine distance

\[
\ell_{\text{intra}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{1}{\sum_{i=1}^{N}} 1 \left\{ d(\mu^k, x_i^k) - \alpha \geq 0 \right\} \frac{d^2(\mu^k, x_i^k)}{\sum_{i=1}^{N} 1 \left\{ d(\mu^k, x_i^k) - \alpha \geq 0 \right\}}
\]

- Inter-cluster loss function

\[
\ell_{\text{inter}} = \frac{2}{K(K - 1)} \sum_{k<k'} \left[ \delta - d(\mu^k, \mu^{k'}) \right]^2
\]
FUSING RGB AND DEPTH

(a) Early Fusion

(b) Late Fusion Addition

(c) Late Fusion Concatenation
MEAN SHIFT CLUSTERING

- von Mises-Fisher (vMF) mean shift for unit length vectors  
- Find local maxima of the von Mises-Fisher distribution

\[ p(x; \mu, \kappa) = C(\kappa) \exp(\kappa x^T \mu) \]

**Algorithm 1:** von Mises-Fisher mean shift clustering

**Input:** Feature embedding matrix \( X \in \mathbb{R}^{n \times C} \), \( \kappa \), \( \epsilon \), number of seed \( m \), number of iteration \( T \) 
Sample \( m \) initial clustering centers from \( X \) as the \( m \) furthest points, denote it as \( \mu^{(0)} \in \mathbb{R}^{m \times C} \); 
for \( t \leftarrow 1 \) to \( T \) do
  - Compute weight matrix \( W \leftarrow \exp(\kappa \mu^{(t-1)} X^T) \) ;
  - Update \( \mu^{(t)} \leftarrow WX \); 
  - Normalize each row vector in \( \mu^{(t)} \) to obtain \( \mu^{(t)} \);
end
Merge cluster centers in \( \mu^{(T)} \) with cosine distance smaller than \( \epsilon \); 
Assign each pixel to the closest cluster center ;

Kobayashi and Otsu. ICPR’10
TWO-STAGE CLUSTERING

RGB

Depth

Feature Map

Initial Label

RoI Feature Map

Initial Label

Segment split

Refined Label

Refined Label
EXPERIMENTS: DATASETS

- Object Cluster Indoor Dataset (OCID), 2,390 RGB-D images  
  Sushi et al. ICRA’19

- Object Segmentation Database (OSD), 111 RGB-D images  
  Richtsfeld et al. IROS’12
EFFECT OF THE INPUT MODE

Mask R-CNN. He et al. CVPR’17
EFFECT OF THE TWO-STAGE CLUSTERING
COMPARISON TO STATE-OF-THE-ARTS

Mask R-CNN. He et al. CVPR’17
UOIS-2D. Xie et al. CoRL’19
ANECDOTAL EXAMPLE ON TRANSPARENT OBJECTS

ClearGrasp
Sajjan et al. ICRA’20
CONCLUSION

- Learning RGB-D feature embeddings from synthetic data with a metric learning loss that transfers well to the real world

- Adding non-photorealistic RGB images to Depth can still improve in our method

- Using RGB images can handle objects with bad or missing depth information such as transparent, flat or thin objects

Questions?