Unsupervised Discovery of the Long-Tail in Instance Segmentation Using Hierarchical Self-Supervision

Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, Serena Yeung
Stanford University

Motivation

- Instance segmentation is an active topic in computer vision that is usually solved by using supervised learning approaches over very large datasets composed of object level masks.
- This work proposes a method that can perform unsupervised discovery of long tail categories in instance segmentation, through **self-supervised learning** of instance embeddings of masked regions.
- We use hyperbolic space (Poincare ball) to embed the mask features, because it is able to efficiently embed hierarchical features with arbitrarily low distortion.

Overview

Our proposed method consists of 3 steps:

(1) class-agnostic **mask proposal generation** using a region proposal network (pre-trained on common categories in COCO)

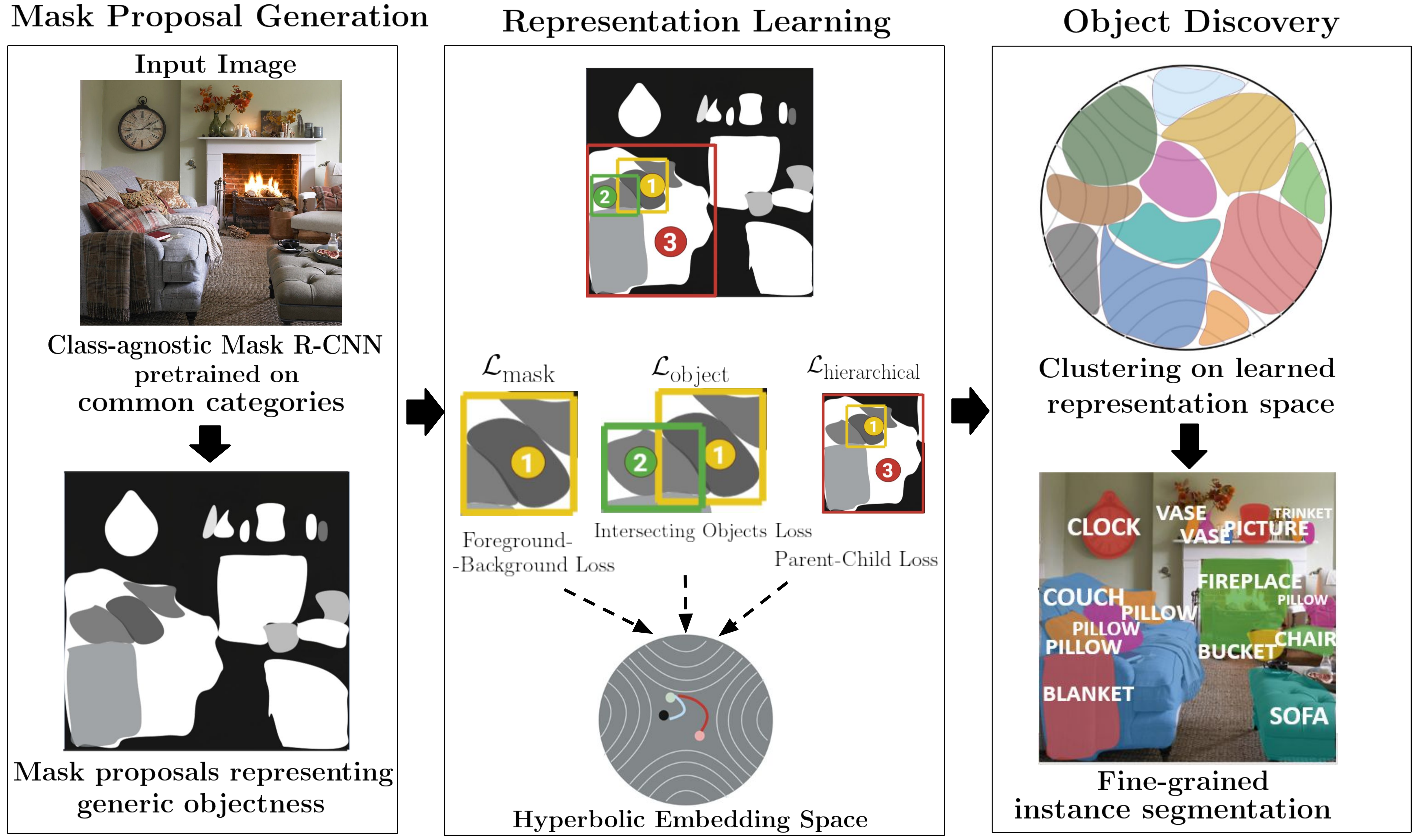
(2) sampling of the masks using sampling rules that exploits the relationship and hierarchical structure within the mask proposals, and **representation learning** of the sampled mask features using triplet losses with a hyperbolic (Poincare ball) embedding space.

(3) unsupervised **clustering** to identify the distinct object categories of the embedded masks.

Intuition of the Loss terms in Representation Learning

- $\mathcal{L}_{\text{mask}}$: the foreground (masked region) feature of each region proposal is closer to the bounding box feature than to the background (non-masked region) feature.
- $\mathcal{L}_{\text{object}}$: mask proposals that are overlapping are likely to be about the same object, so their feature should be close.
- $\mathcal{L}_{\text{hierarchical}}$: smaller masks that are part of the larger masks have hierarchical relationship in their visual features.

Method



Example clusters discovered through clustering

“L” means categories in LVIS, without “L” means novel categories that are not in LVIS

Experiments

We conduct experiments on LVIS dataset.

Training: The mask proposal generation network is trained on the 80 common categories in COCO without consuming any annotations on the long-tail categories in LVIS.

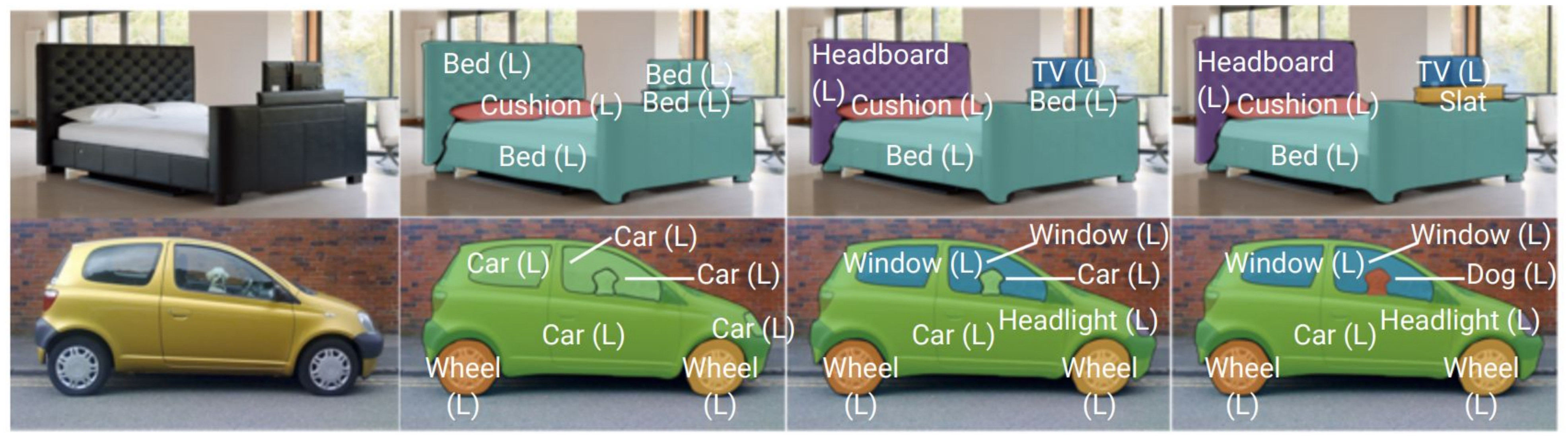
Evaluation: Hyperbolic K-Means clustering is run with 1462 number of clusters (chosen by Elbow method).

Model	Supervision	mAP	mAP_{50}	mAP_{75}	mAP_r	mAP_c	mAP_f	mAP_s	mAP_m	mAP_l
Mask R-CNN	Fully Supervised	0.201	0.327	0.212	0.072	0.199	0.284	0.106	0.214	0.325
ShapeMask [31]	COCO masks+LVIS boxes	0.084	0.137	0.089	0.056	0.084	0.102	0.062	0.088	0.103
$Mask^X$ R-CNN [25]	COCO masks+LVIS boxes	0.056	0.095	0.058	0.024	0.051	0.079	0.031	0.056	0.078
Ours (rand. init. backbone)	COCO masks	0.096	0.139	0.104	0.051	0.092	0.168	0.075	0.107	0.139
Ours	**COCO masks**	**0.109**	**0.160**	**0.113**	**0.087**	**0.105**	**0.174**	**0.092**	**0.129**	**0.147**

Ablation study to test the effectiveness of each triplet loss term

	mAP	mAP50	mAP75
w/o $\mathcal{L}_{\text{mask}}$	0.0689	0.0842	0.0707
w/o $\mathcal{L}_{\text{object}}$	0.0374	0.0455	0.0396
w/o $\mathcal{L}_{\text{hierarchical}}$	0.0846	0.1082	0.0921
Full model	**0.1086**	**0.1597**	**0.1125**

Qualitative results showing model ablations.



Input image | Only mask loss | Mask loss and object loss | All three losses