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

Zhenzhen Weng, Mehmet Giray Ogut, Shai Limonchik, Serena Yeung
Stanford University
{zzweng, giray98, shailk, syyeung}@stanford.edu

1. Additional Experiment on VOC to non-VOC Generalization

To show that our model is not biased towards the long-tail categories in LVIS, we conduct additional quantitative experiment for PASCAL VOC [1] to non-VOC generalization. In this experiment, we show that our model is able to discover the non-VOC categories in COCO even though the class-agnostic region proposal network was pre-trained on only the categories in VOC. Since the 80 categories in COCO [5] also include the 20 categories in PASCAL VOC, we pre-train our mask proposal network on the 20 categories in PASCAL VOC, and show that our method is able to discover the objects that belong to the rest of the 60 categories in COCO.

Quantitative Results In Table 2 we compare the mean average precision (mAP) of the fully-supervised Mask R-CNN model, the partially-supervised methods (ShapeMask and Mask^X R-CNN) and our method. Although our method uses less amount of supervision than partially-supervised methods, we are able to outperform both methods. What is more, the improvement over semi-supervised models is even stronger given the fact that detecting 60 non-VOC categories in COCO is an easier task compared to the 1200 long-tail categories in LVIS. Our model overall demonstrates performance comparable to that of the fully-supervised Mask R-CNN model.

Cluster Analysis Following the same protocol of the "COCO to LVIS" experiments in the main paper, we try different number of clusters in the hyperbolic clustering algorithm, and then report cluster purity scores on the final clusters (Table 1). The number of clusters that are matched to original COCO categories (excluding the 20 Pascal VOC classes) increases as we increase the number of clusters k. The highest purity scores are obtained at the optimal k determined by the elbow method .

Qualitative Results In Figure 2 we show qualitative examples of the new categories (i.e. non-VOC categories in COCO) discovered using our method. Each row shows the segmentation results on an image in the COCO dataset. In the first example, our model is able to segment non-VOC categories, such as traffic lights, trucks and stop signs. In the second example, our model is able to discover and segment novel categories such as glasses, knives, plates and even hot dogs and slices of bread. In the third example, our model successfully finds new classes such as bed, bag, lamp and paintings. Notice how the frame of the painting is separately detected from its canvas. In the fourth example, television is a PASCAL category. We observe that the poster, laptop, mouse, keyboard, essence bottle books as well as eyes of teddy bears could also be segmented using our method.

No. of Clusters	COCO	$\text{Purity}_{\text{Avg}}$	Purity_{s}	Purity_{m}	Purity_{l}
k=80	41	0.483	0.413	0.478	0.652
k=90	44	0.532	0.468	0.553	0.681
k*=108 (Elbow)	51	0.622	0.524	0.637	0.744
k=200	60	0.520	0.436	0.535	0.669

Table 1. Cluster purity analysis with different number of clusters. The number of discovered clusters determined to correspond to COCO classes excluding PASCAL VOC classes (and used to compute the purity scores) are denoted under the COCO column.

2. Additional Qualitative Examples

In Figure 1 we show additional qualitative examples of model ablations. We show that the designed loss terms are essential in discovering and segmenting fine-grained objects.

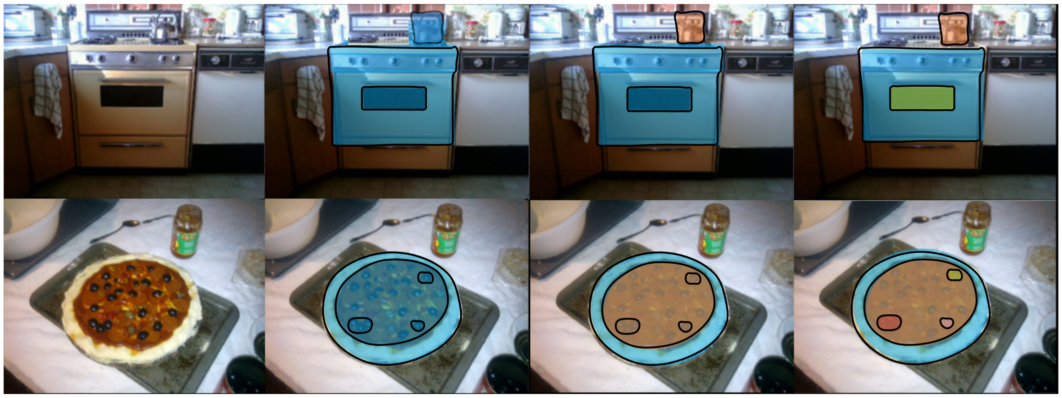

Figure 1. Additional qualitative example showing model ablations. **From left to right**: Original image; segmentation masks obtained using only mask loss term; with mask loss and object loss; with all three loss terms included.

Model	Supervision	mAP	mAP_{50}	mAP_{75}	mAP_s	mAP_m	mAP_l
Mask R-CNN [2]	Full supervision	0.344	0.552	0.363	0.186	0.391	0.479
ShapeMask [4]	VOC Masks + non-VOC Boxes	0.302	0.493	0.315	**0.161**	0.382	0.384
Mask^X R-CNN [3]	VOC Masks + non-VOC Boxes	0.238	0.429	0.235	0.127	0.281	0.335
Ours	**VOC Masks**	**0.327**	**0.525**	**0.331**	0.159	**0.385**	**0.413**

Table 2. Quantitative results on VOC to non-VOC. The fully-supervised Mask R-CNN is trained with the masks and boxes for all categories in COCO (i.e. including VOC and non-VOC categories). The partially-supervised methods (ShapeMask and Mask^X R-CNN) are trained using the masks of the categories that are in VOC and the bounding boxes of the categories that are not in VOC. Our model consumes only VOC masks in pre-training the region proposal network. Each model was evaluated on the non-VOC categories. Our method outperforms the partially-supervised methods in terms of mAP.

References

[1] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[3] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. 2

[4] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9207–9216, 2019. 2

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

Figure 2. Qualitative examples of new object discovery. The region proposal network was pre-trained with VOC categories in the COCO dataset. Each column shows the segmentation results on an image in the COCO dataset.