Supplementary Materials:
Open-Vocabulary Instance Segmentation via Robust Cross-Modal Pseudo-Labeling

Dat Huynh[1]* Jason Kuen[2] Zhe Lin[2] Jiuxiang Gu[2] Ehsan Elhamifar[1]
[1]Northeastern University [2]Adobe Research
[1]{huynh.dat,e.elhamifar}@northeastern.edu [2]{kuen, zlin, jigu}@adobe.com

1. Effect of Pseudo-Mask Quantity

Figure 1 show the segmentation improvement of the student model compared to the teacher with respect to different amounts of pseudo masks on target classes in Conceptual Captions. We partition target classes into groups of different ranges of pseudo masks and report the average improvement per group. We observe significant improvements often correlated with classes having many pseudo masks, which shows the effectiveness of pseudo masks generated by our cross-modal pseudo-labeling framework. On the other hand, the classes with the fewest number of pseudo masks (less than 1000) degrade their performances. This is because the student model overfits to a few pseudo masks and cannot generalize to new samples in these classes.

2. Pseudo-Mask Visualization

Figure 2 visualizes the pseudo masks predicted by the teacher model and the noise levels estimated by the student model. Here, we notice that objects of novel classes, which are distinct from their background such as “bus”, “rib” or “umbrella”, can be easily segmented with high reliability scores since their pixel colors are clearly different from background colors. Our cross-modal pseudo labeling can also segment overlapping objects such as “cat” on “bed”.

On the other hand, partially-segmented pseudo masks such as “dog” have irregular boundaries with no clear distinction between pixels in foreground and background. As the student has difficulty learning to classify between background and foreground pixels of the pseudo masks, our training process drives the noise levels high and reduces the reliability scores of these masks. Similarly, mis-localized pseudo masks such as “tie” also result in low reliability scores due to indistinguishable object boundaries.

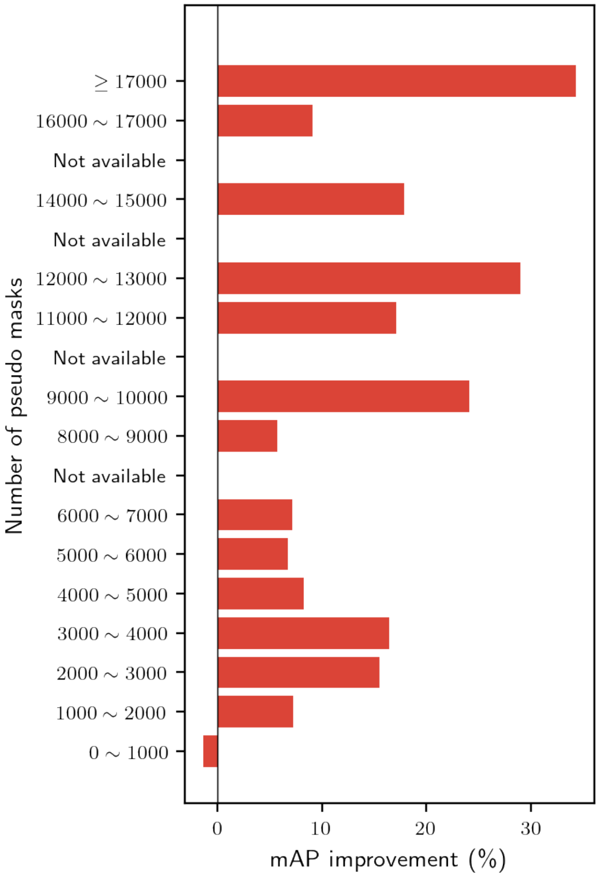

Figure 1. Segmentation improvement of the student model compared to the teacher in our method w.r.t. different amounts of pseudo masks for novel classes in Conceptual Captions.

3. Additional Qualitative Results

Figure 3 shows additional qualitative results of our methods compared to OVR on MS-COCO datasets. Thanks to additional pseudo masks, our method learns to segment the entire ”dog” from target classes while OVR mis-classifies the

*This work was done during Dat Huynh’s internship at Adobe Research.

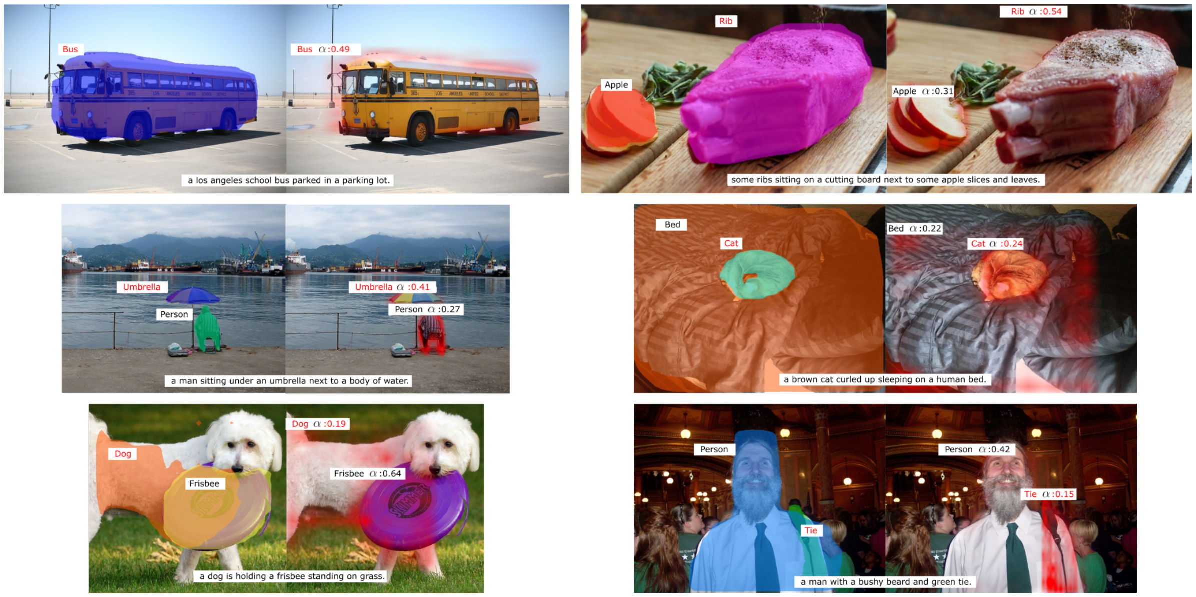
Bus
Bus α :0.49
a los angeles school bus parked in a parking lot.
Rib
Apple
Rib α :0.54
Apple α :0.31
some ribs sitting on a cutting board next to some apple slices and leaves.
Umbrella
Person
Umbrella α :0.41
Person α :0.27
a man sitting under an umbrella next to a body of water.
Bed
Cat
Bed α :0.22
Cat α :0.24
a brown cat curled up sleeping on a human bed.
Dog
Frisbee
Dog α :0.19
Frisbee α :0.64
a dog is holding a frisbee standing on grass.
Person
Tie
Person α :0.42
Tie α :0.15
a man with a bushy beard and green tie.

Figure 2. Visualization of pseudo masks and their noise levels for base classes (in **back**) and novel classes (in **red**) in MS-COCO.

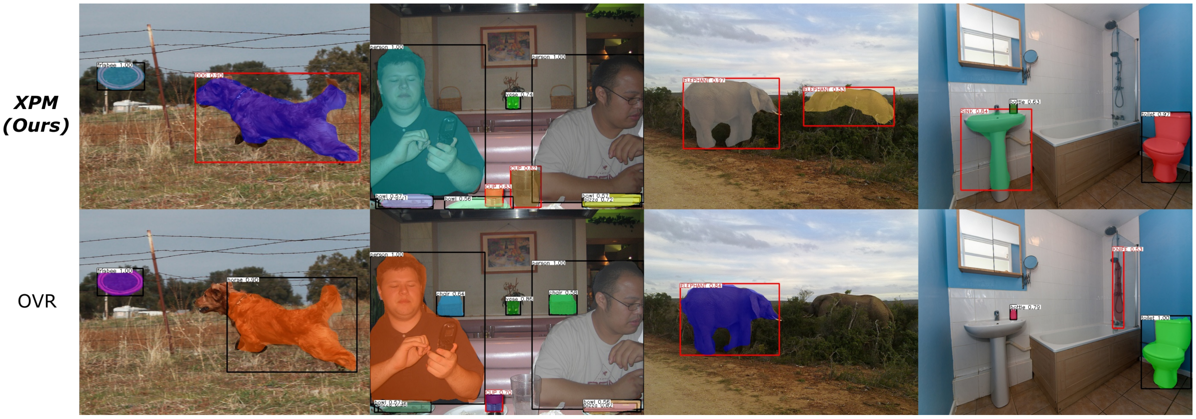

Figure 3. Qualitative comparison between our method and `OVR` for base classes (in **back**) and novel classes (in **red**) in MS-COCO.

object as “horse” and cannot segment its complete shape. Overall, we observe that our method significantly improves the recall of novel object classes in images.