Supplementary material for Mask-free OVIS: Open-Vocabulary Instance Segmentation without Manual Mask Annotations

Vibashan VS*, Ning Yu†, Chen Xing†, Can Qin‡, Mingfei Gao†,
Juan Carlos Niebles†, Vishal M. Patel*, Ran Xu†
*Johns Hopkins University, ‡Northeastern University, †Salesforce Research
{vvishnu2, vpatel36}@jhu.edu, qin.ca@northeastern.edu,
{ning.yu, cxing, jniebles, ran.xu}@salesforce.com

COCO Caption vs Image-label pseudo-caption:
Pseudo-caption generation: Since the pre-trained vision-language models are trained on full sentences, we need to feed the image-labels into a prompt template first, and use them to generate a pseudo-captions. Specifically, given image-labels [category-1,category-2...,category-n], we randomly sample a prompt from 63 prompt templates [1,6] and the pseudo-caption are generated as "{Prompt-x} + {category-1 and category-2 and ... category-n}". For example, as shown in Fig. 1 bottom row - the sampled prompts are *"A black and white photo of the {category}."* and *"A photo of {category} in the scene."* and the image-labels are "zebra" and "giraffe". Thus, the generated pseudo-captions are *"A black and white photo of the zebra and giraffe."* and *"A photo of one zebra and giraffe in the scene."*

COCO Caption and Pseduo-caption activation map: Given a caption and object of interest c_t ("zebra","giraffe"), corresponding activation map is generated using GradCAM [7] and pre-trained vision-language model [4]. As we can observe from Fig. 1, both human-provided COCO captions and image-labels based pseudo-captions produce similar activations for "zebra" and "giraffe". Even for the same image with different human-provided COCO captions, the generated activation

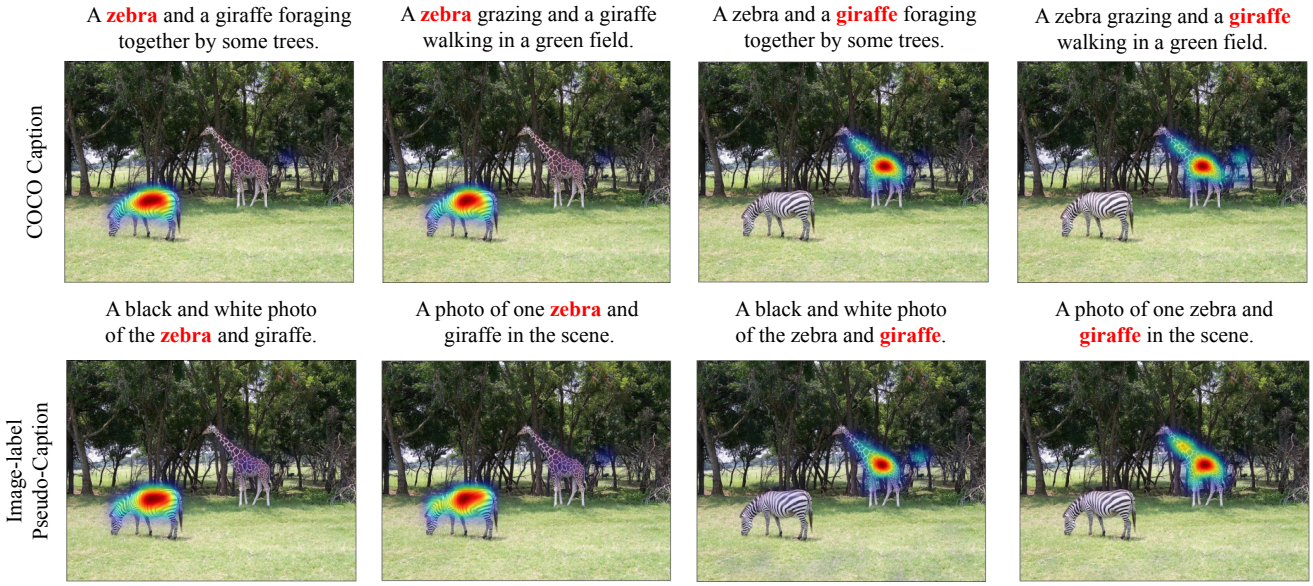

Figure 1. **Top row:** Given COCO caption and object of interest c_t ("zebra","giraffe"), corresponding activation map is generated using GradCAM. **Bottom row:** Given pseudo-caption generated from image-labels and object of interest c_t ("zebra","giraffe"), corresponding activation map is generated using GradCAM. The original activation map are of 1/16'th of the image size and we perform bilinear interpolation to obtain an activation map of image size.

maps are similar for "zebra" and "giraffe". This shows that irrespective of caption type, the GradCAM generates similar activation map for the object of interest resulting in similar pseudo-masks.

Generalization of WSPN network trained on VOC (20 categories) and test on COCO (80 categories) dataset:

Iterative masking lead to better guidance function:

Fig. 3 present iterative masking visualization for G going from 0 to 4. From Fig. 3, we can observe that, the initial GradCAM [7] activation for the towel category is less at G=0. Utilizing this activation map as guidance function to generate pseudo-labels will generate less accurate pseudo-labels. However, by performing the proposed iterative masking strategy, we can observe that the activation map is progressively shifting towards a less discriminative part in successive iterations and the combination of all activation's $\mathbf{\Phi}_t$ covers the entire object effectively. Note that even after multiple iterations, there is no GradCAM [7] activation for the baby, which shows the robustness of the pre-trained model towards region-text alignment.

Masking for more iteration might produce redundant activation:

Fig. 3 present iterative masking visualization for G going from 0 to 4. From Fig. 3, we can observe that the initial GradCAM [7] activation for the Jet ski category is towards the most discriminative parts. Also, the activation for the person category is negligible due to the good region-text alignment property of pre-trained vision-language model. However, as the masking iteration increases after G=3, we can observe that the water regions around Jet ski are starting to activate. Hence, more steps might completely mask the object and will start producing redundant activation. Quantitatively we observed G-3 produces optimal pseudo-labels, as discussed in section 4.3.

References

[1] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 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. 6

[3] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 6

[4] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[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. 6

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[7] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1, 2

Figure 2. Qualitative analysis of WSPN network trained on VOC 2007 (20 categories), VOC 2012 (20 categories) and COCO Base 2017 (48 categories) using image-labels in a weakly-supervised manner and tested on COCO 2017 (80 categories). Here, the WSPN trained on VOC 2007 and VOC 2012 has not seen categories such as *"Pizza", "Kite", "Clock"* etc during training. Still, we can observe the VOC trained WSPN network can localize these unseen categories similar when tested on COCO 2017. This is due to weakly-supervised training, where the WSPN network learns to localize the object irrespective of categories it is trained on. As a result, the WSPN network trained on VOC is able to localize COCO object categories which are not seen during training. This show the generalization capability of the WSPN model and can be used to localize any objects for different dataset. Green: Ground truth bounding box, Red: Top 50 WSPN proposals.

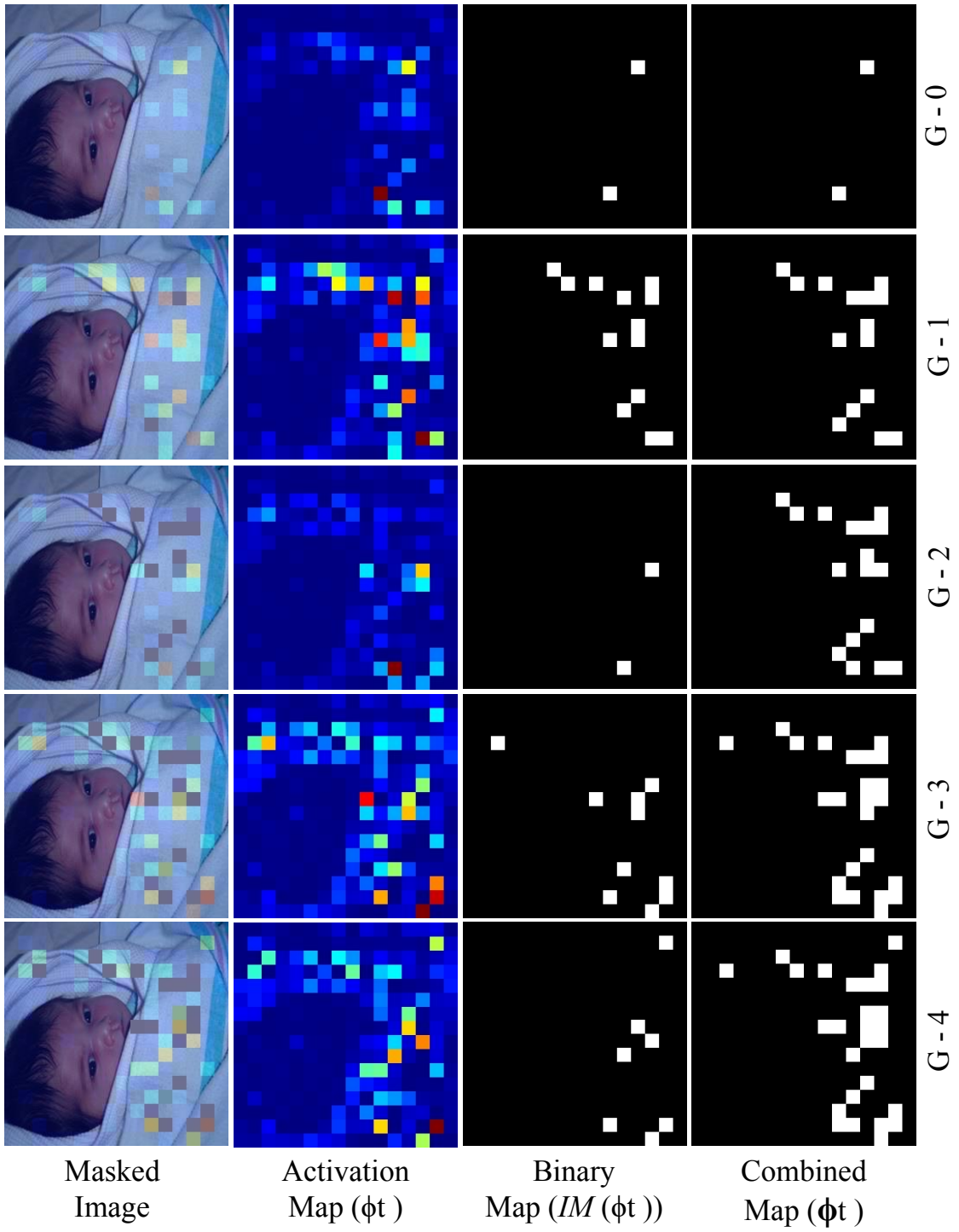

Figure 3. Iterative masking visualization for ”Towel” category.

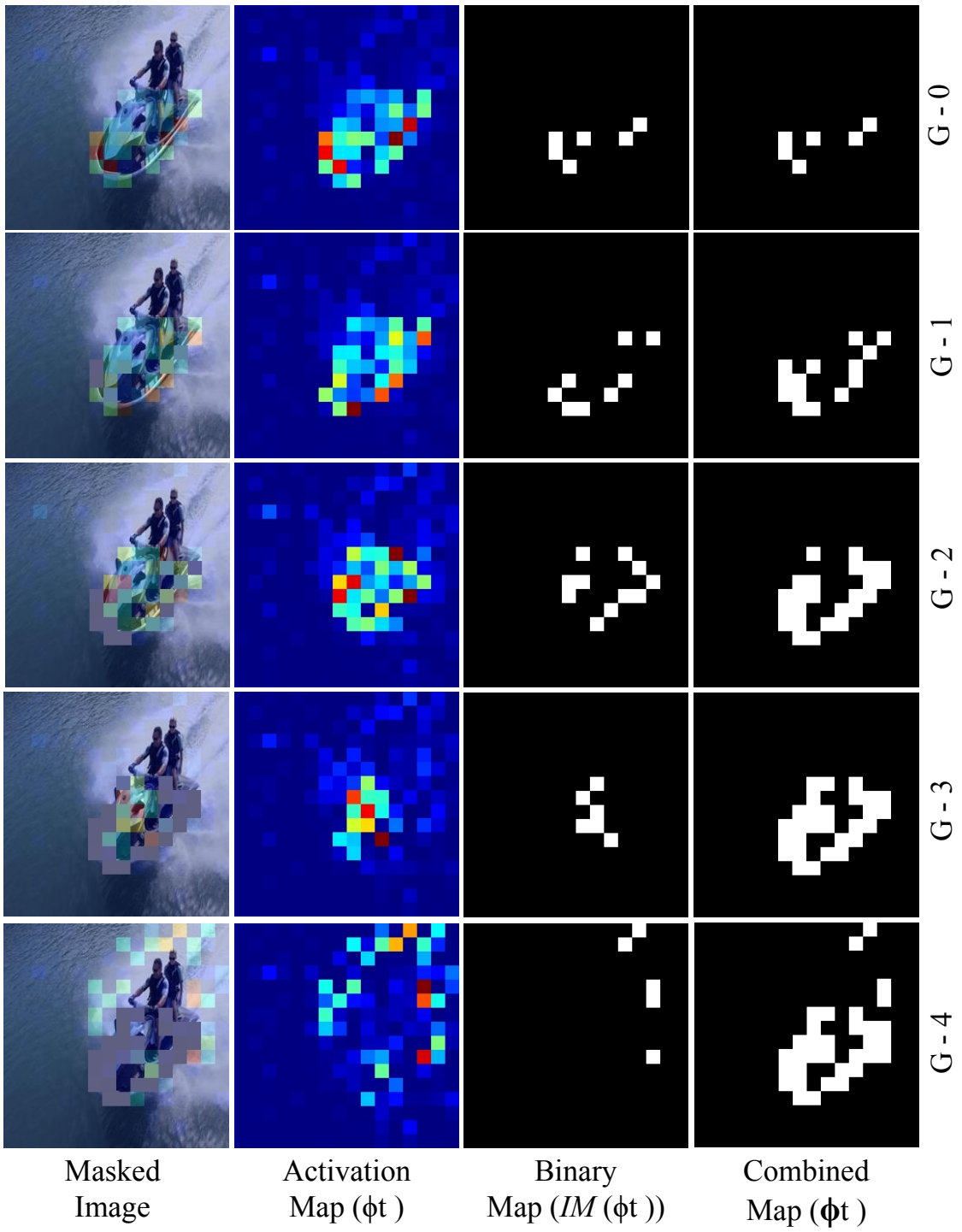

Figure 4. Iterative masking visualization for ”Jet ski” category.

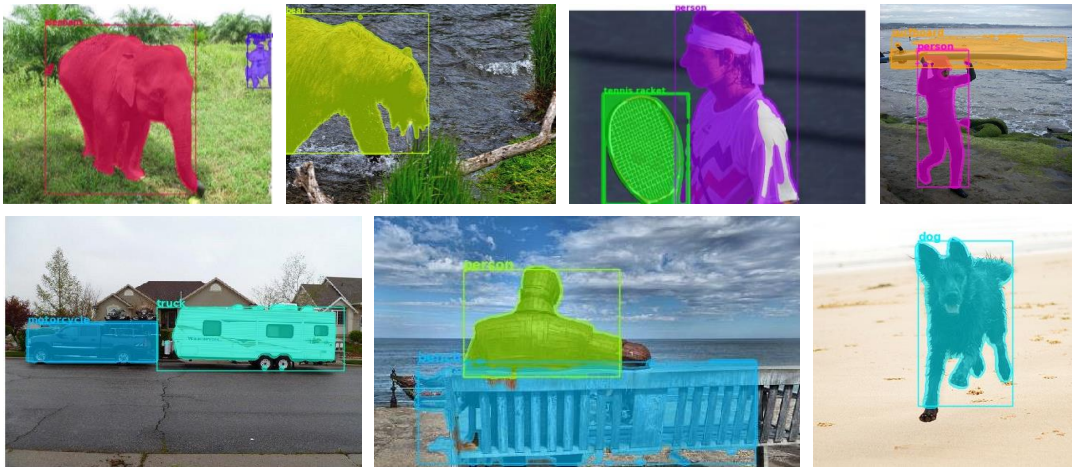

Figure 5. Visualization of pseudo-mask generated for COCO dataset [5] using our pipeline. Note that, the generated box-level and pixel-level annotations are noisy (incomplete mask and less accurate bounding box).

Figure 6. Visualization of pseudo-mask generated for Open Images [3] dataset using our pipeline. Note that, the generated box-level and pixel-level annotations are noisy (incomplete mask and less accurate bounding box).

Figure 7. Visualization of Mask-RCNN [2] predictions trained on pseudo-masks generated on COCO and Open Images - top and bottom row, respectively. Mask-RCNN training helps the model learn to filter the noise present in the pseudo-mask producing better-quality (complete mask and tight bounding box) predictions.