FOSSIL: Free Open-Vocabulary Semantic Segmentation through Synthetic References Retrieval Supplementary Material

In this supplementary material we:

- provide additional implementation details on the proposed method;
- provide a qualitative evaluation of the Reference Collection Generation step, focusing on the heatmaps and the binary masks extracted for each generated concept;
- extend our qualitative evaluation of predictions made on the benchmarks considered, providing a more comprehensive view of the performance of our model.

A. Additional Implementation Details

A.1. Efficient Retrieval and Clustering

To optimize the execution of retrieval and clustering procedures, we harness the capabilities of the `faiss` library [3]. Specifically, our approach leverages a retrieval index based on the Hierarchical Navigable Small World graph exploration (HNSW) [4] technique. This method employs an approximation of the nearest neighbor search, thereby enhancing the efficiency of retrieving Textual Retrieval Embeddings and Visual Reference Embeddings.

A.2. Textual Prompt Templates

During both the Reference Collection Generation and Prototype Creation phases, we encapsulate arbitrary textual concepts using pre-defined prompt templates. These templates are the ones introduced in CLIP [6], as follows:

- `itap of a {}.`
- `a bad photo of the {}.`
- `a origami {}.`
- `a photo of the large {}.`
- `a {} in a video game.`
- `art of the {}.`
- `a photo of the small {}.`

B. More Qualitatives

B.1. Reference Collection Generation

In Figure 1, we present a visual representation of the qualitative results obtained during the Reference Collection Generation step. In particular, we illustrate: 1) the caption used to condition Stable Diffusion [7], 2) the resulting generated image, 3) the heatmaps corresponding to nouns, extracted through DAAM [8], and 4) the binarized heatmaps used to perform region pooling on the dense features of the visual backbone. These qualitatives show the effectiveness of DAAM in localizing words in the generated image. The resulting heatmaps can be thresholded to produce approximate binary masks for the identified concept. It is worth noting that minor inaccuracies in border delineation do not significantly impact the resulting feature vector, as these masks are employed for feature averaging.

B.2. Prediction Qualitatives

In Figure 2, we present qualitative results depicting the predictions made by our method on three benchmark datasets: PASCAL Context [5], Cityscapes [2] and COCOStuff [1]. Additionally, we provide a comparison by displaying the same predictions without the utilization of OpenCut, emphasizing the impact of OpenCut in refining masks. Specifically, the showcased qualitative results underscore the robust recognition capabilities of FOSSIL in identifying semantic elements within the scenes. However, it is noteworthy that the integration of OpenCut is essential for refining the resulting segments, particularly in areas adjacent to borders.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 3

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the*

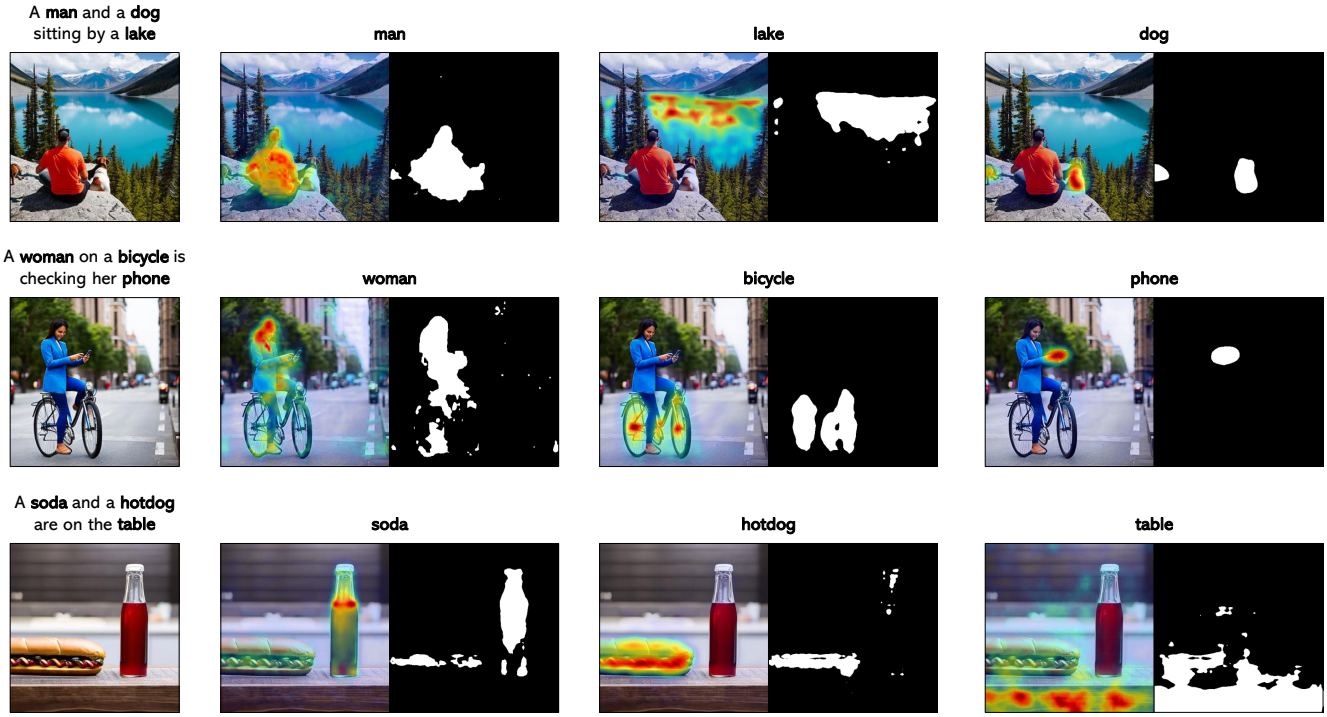

Figure 1. Qualitative results of the Reference Collection step. On the left, we report three captions used to condition Stable Diffusion and generate the corresponding images. On the right, we report the heatmap obtained through DAAM [8] for each noun from the caption and the corresponding binary mask.

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016. 1, 3

[3] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019. 1

[4] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1

[5] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2014. 1, 3

[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 *Proceedings of the International Conference on Machine Learning*, 2021. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[8] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2023. 1, 2

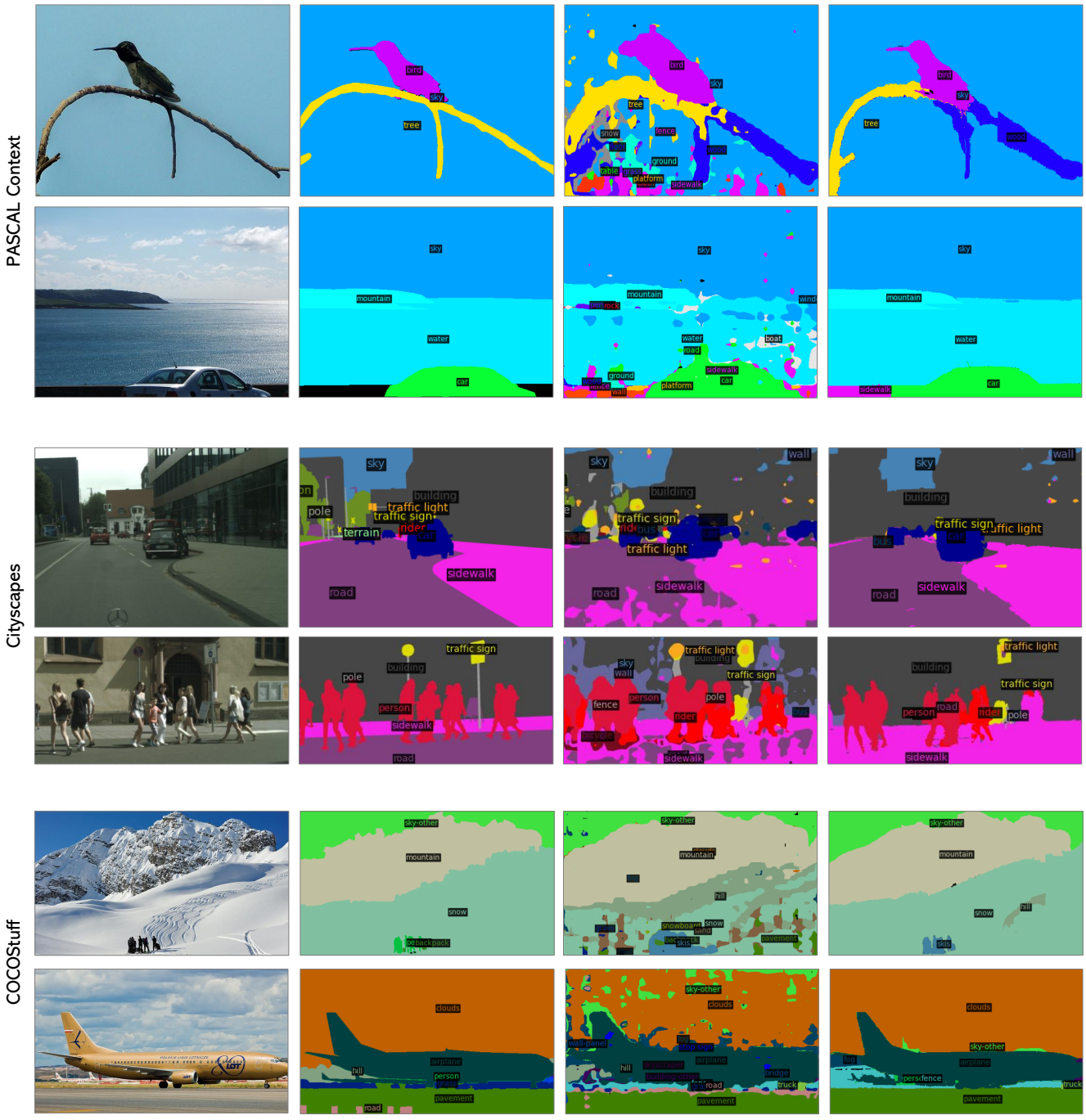

Figure 2. Qualitative results, comparing FOSSIL with and without the OpenCut component on the PASCAL Context [5], Cityscapes [2] and COCOStuff [1].