Mitigating Objectness Bias and Region-to-Text Misalignment for Open-Vocabulary Panoptic Segmentation

Supplementary Material

A. Per-Image Inference Latency

We measured the per-image inference time on the ADE20k validation set to assess computational efficiency. All three methods run at comparable speeds: 0.31 s for FC-CLIP [7], 0.30 s for MAFT+ [3], and 0.32 s for OVRCOAT. The maximum difference of 0.02 s indicates that the proposed method delivers competitive performance without introducing significant computational overhead.

B. Extended Qualitative Analysis

We provide five example images from the internet [1, 2, 4–6] to highlight model performance. In three cases, the ADE20K [8] vocabulary is extended with musical instruments, workshop tools, or safari animals (Fig. 1). OVRCOAT outperforms both FC-CLIP [7] and MAFT+ [3], leveraging COAT to recover masks and OVR to classify them more accurately. Nevertheless, certain instruments, such as flutes and the baroque guitar, are not detected, likely due to their unconventional appearance. Tool masks are also suboptimal, with oversegmentation (e.g., the drill) and undersegmentation (e.g., the wrench and chisel) in the unseen setting. All models miss distant animals on the safari scene, likely due to their small size. These examples demonstrate the superior generalisation capabilities of OVRCOAT, while exposing common limitations.

The museum scene evaluates robustness to visually ambiguous content: blank paintings matching the wall colour. OVRCOAT detects the most paintings, while MAFT+ [3] fails to detect any. One of the paintings is incorrectly included in the wall segment by OVRCOAT, highlighting remaining segmentation challenges.

Finally, the Shibuya scene illustrates a real-world scenario where all models perform similarly. OVRCOAT correctly classifies buildings and segments more civilians, but fails to capture five signboards in ambiguous regions. This underscores that such areas remain a significant challenge for open-vocabulary models and represent a key area for future investigation into segmentation robustness.

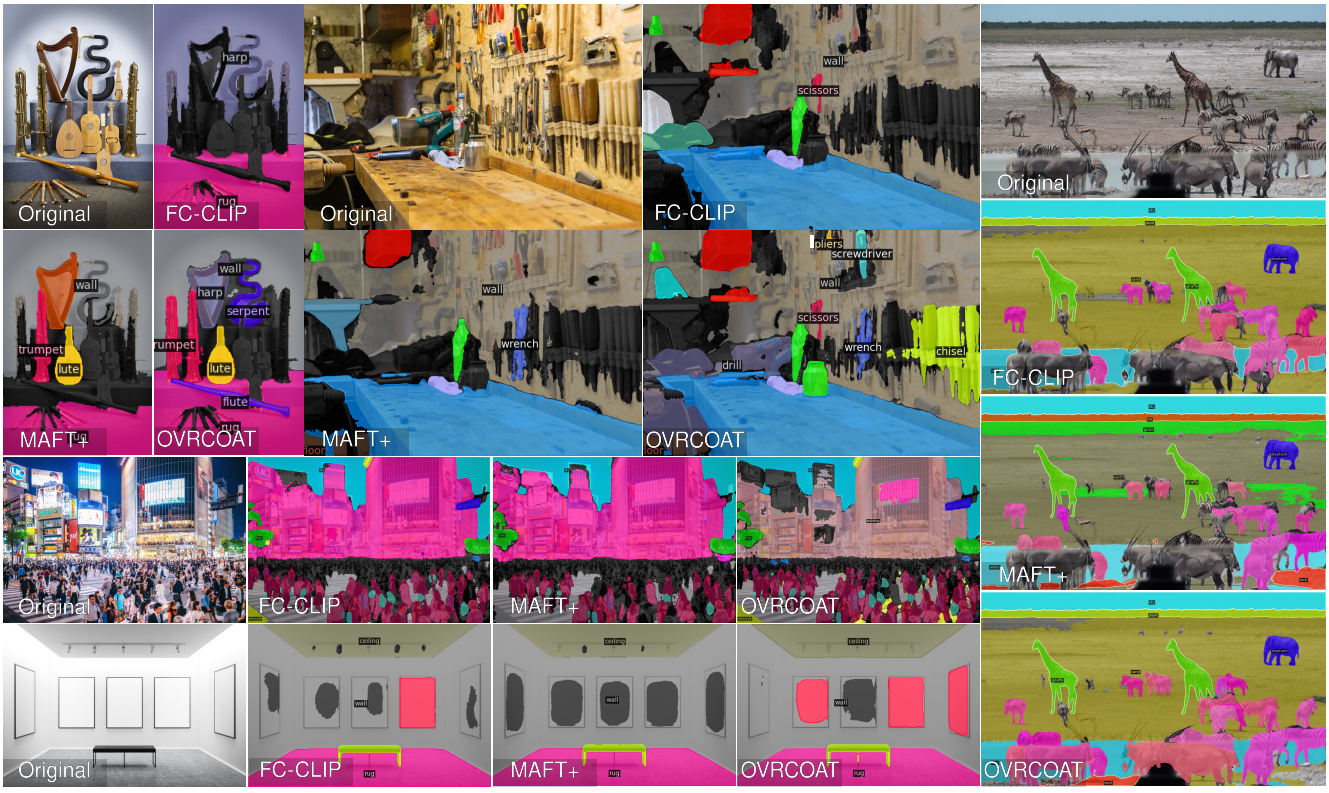

Figure 1. Qualitative comparison on unconventional scenes using an extended ADE20K vocabulary.

References

[1] A. E. Booth. Own work. https://commons.wikimedia.org/w/index.php?curid=73612035, 2015. 1

[2] iStock / Nikada. Shibuya scramble crossing, tokyo, japan, 2019. 1

[3] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 399–416. Springer, 2024. 1

[4] Reno Liquidators. Creating a workshop at home – must-haves. 1

[5] Tips.at Redaktion. Italienischer künstler verkauft unsichtbare kunst für preise zwischen 14.000 und 30.000 euro.

[6] YouTube / i.ytimg.com. Thumbnail for youtube video (id: C-4x4qpmtms), 2026. 1

[7] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1