ETH zürich

OpenMask3D: Open Vocabulary Instance Understanding

Adam Klebus, Jan Ackermann, Ke Li, Ying Xue
Department of Computer Science, ETH Zurich

1 Introduction

Recently developed methods for scene understanding [3] assign CLIP features [2] to points in a scene, allowing for open vocabulary queries. What has yet to be developed are methods for assigning CLIP features to instances in a scene, which may be useful for other kinds of scene understanding tasks.

2 Background

OpenScene [1]:
- State-of-the-art scene understanding method.
- Assigns a CLIP feature to each point in 3D.
- Supports open vocabulary text queries by using a CLIP text encoder and computing similarity to points.

CLIP [2]:
- Learns common features for both text and image data.
- Allows for similarity comparisons between a text query and image data.
- Extended in OpenScene [1] to assign features to points in a point cloud.

Mask3D [3]:
- State-of-the-art instance segmentation on scenes.
- Uses a transformer based architecture to obtain:
  - Instance masks
  - Instance labels
- For our purposes, we modify the model to additionally output instance heatmaps for feature fusion.

ScanNet [4]:
- Large dataset of indoor scene scans.

3 Method

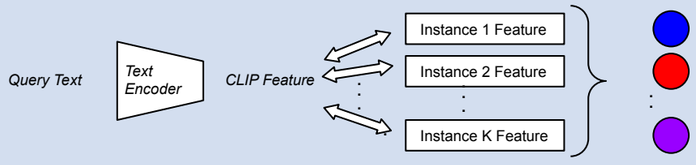
1: Extracting per-instance CLIP features

Algorithm 1 Extracting Open Vocabulary Scene Instance Features

Input: Scene S
Output: Per-instance CLIP features $\{\hat{\mathbf{f}}_i\}_{i \in I}$

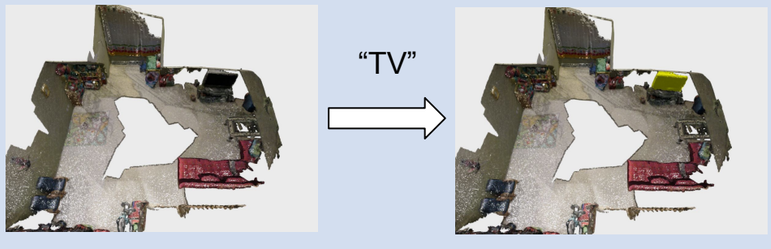
$\{\mathbf{f}_p\}_{p \in P} \leftarrow \text{OpenScene}(S)$ ▷ Extract per-point CLIP features
$\{\mathbf{h}_i\}_{i \in I} \leftarrow \text{Mask3D}(S)$ ▷ Extract instance heatmaps
for $i \in I$ **do**
  $\mathbf{w} \leftarrow \max\{0, \mathbf{h}_i - 0.5\}$
  $\hat{\mathbf{f}}_i \leftarrow \frac{\sum_{p \in P} \mathbf{w}_p \mathbf{f}_p}{\sum_{p \in P} \mathbf{w}_p}$ ▷ Aggragate to per-instance features
end for
return $\{\hat{\mathbf{f}}_i\}_{i \in I}$

2: Processing text queries

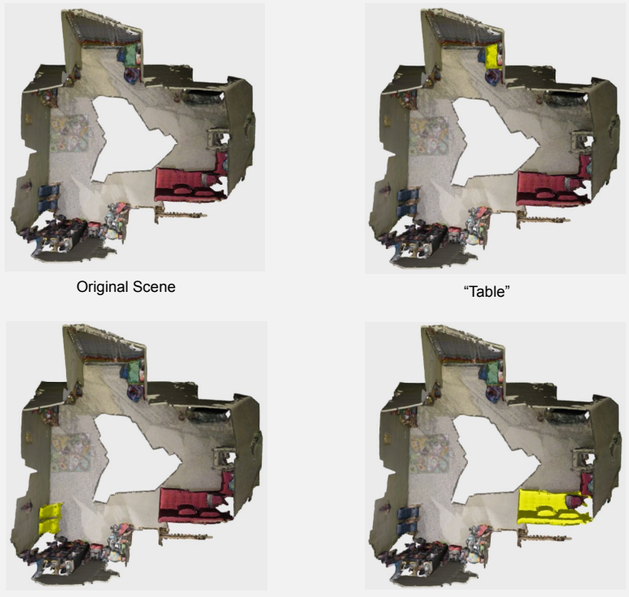


The following is a sketch of the processing procedure for text queries. After a text query is entered, a text to CLIP encoder is used to extract a CLIP feature from the text query. We then compare the text feature to each of the instance features previously extracted to compute a relevancy score for each instance in relation to the query.

3. Visualizing results



4 Results



Example visualizations on the same scene with different queries. We visualize the instances with a relevancy score above a certain threshold. While these here in particular were object word queries, any text query can be input, such as object properties or materials. We welcome you to try out your own queries at our live demo.

5 Future Work

- End-to-end script
- Compatibility with phone taken scans
- SOTA benchmarks
- Web based visualization

References

1. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., & Funkhouser, T. (2023). Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 815-824).

2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

3. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., & Leibe, B. (2022). Mask3D for 3D Semantic Instance Segmentation. *arXiv preprint arXiv:2210.03105*.

4. Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5828-5839).