Geometrically-driven Aggregation for Zero-shot 3D Point Cloud Understanding Supplementary Materials

Guofeng Mei Luigi Riz Yiming Wang Fabio Poiesi
Fondazione Bruno Kessler, Via Sommarive, 18, 38123 Trento, Italy
{gmei,luriz, ywang, poiesi}@fbk.eu

In this supplementary material, we provide implementation details in Sec. A and further experimental analyses in Sec. B.

A. Implementation Details

A.1. VLM representation extraction

For classification tasks on both the ModelNet40 [8] and ObjectScanNN [9] datasets, we adhere to the method used in PointCLIPv2 [12], sampling 1,024 points from each point cloud. These points are then projected into depth images from 10 different views for the extraction of VLM representations. As in PointCLIPv2 [12], we use the ViT-B/16 [10] model as the visual encoder within the CLIP framework, which comprises 12 layers of multi-head self-attention (MHSA). We extract the VLM representations corresponding to image patches during the attention process at the final MHSA layer. These VLM representations are then subjected to bilinear interpolation, a technique we utilize to upscale the VLM representations to the original size of the image, specifically to 224 × 224. Given that a single view projection captures only a partial point cloud, we utilize multi-view back projection to ensure thorough predictions for all points in the cloud. For points observable from multiple views, we perform linear interpolation of the VLM representations, fine-tuning this process based on the varying weights of the different views. These VLM representations are then fed into our GeoZe to obtain enhanced point-level representations. Subsequently, max-pooling is applied to generate the global features. Additionally, we extract global features using the ViT-B/16 model from each view, consistent with PointCLIPv2.

For the part segmentation task on the ShapeNetPart [11] dataset, we sample 2,048 points per point cloud. These points are then projected onto depth images from 10 different viewpoints for VLM representation extraction using the ViT-B/16 model. After extraction, these VLM representations undergo bilinear interpolation for upsampling to their original image size. Similar to the classification process, we apply multi-view back projection to the point cloud using the VLM representations. These VLM representations are subsequently processed through our GeoZe to achieve enhanced point-level representations.

Table 1. Hyper-parameters configurations for different datasets.

Dataset	Network	Γ	$\bar{N}$	K_1	K_2
ModelNet40	PointCLIPv2[12]	16	256	32	24
ObjectScanNN	PointCLIPv2[12]	16	256	32	24
ShapeNetPart	PointCLIPv2[12]	16	256	32	24
ScanNet	OpenSeg [3]	8	3000	48	32
	LSeg [5]	8	3000	48	32
	ConceptFusion[4]	8	3000	48	32
nuScenes	LSeg [5]	8	2400	48	32

In assessing the semantic segmentation performance on the ScanNet [2] and nuScenes [1] datasets, we utilize the VLM representations that are provided by OpenScene [6]. To ensure consistency and facilitate a fair comparison, we adhere to the standard voxel size of 0.02m as in OpenScene. These VLM representations are then processed using GeoZe, which facilitates the enhancement of point-level representations.

A.2. Parameters

Tab. 1 presents the dataset-specific hyperparameters including the number of iterations (Γ), the number of superpoints ($\bar{N}$), the number of points (K_1) used for computing similarity, and the number of neighboring points (K_2) for local aggregation. Specifically, increasing the number of superpoints from $\frac{N}{8}$ to a maximum of $\bar{N} \leq \frac{N}{4}$ can slightly improve results, as denoted by $\frac{N}{8} \leq \bar{N} \leq \frac{N}{4}$. N is the number points of a point cloud. Our experiments suggest that optimal accuracy is achieved when the number of superpoints is maintained between one-eighth and one-third of the original point cloud size. Balancing both accuracy and time complexity, the parameters we adopted in our experiments are as listed in Tab. 1.

A.3. Geometric representation (FPFH) extraction

In our experiments, we compute FPFH [7] features for all points. To improve the computation time, we first downsample M reference points. Then, we sample K_3 neighboring points for each point from the M reference points within a radius r_1 for estimating the normals and K_4 neighboring points for each point from M reference points within a radius r_2 for estimating FHFH. The details are reported in Tab. 2.

Table 2. Hyper-parameters configurations for FPFH computation on different datasets.

Dataset	$\bar{M}$	K_3	K_4	r_1	r_2
ModelNet40	512	32	100	0.04	0.08
ObjectScanNN	512	32	100	0.04	0.08
ShapeNetPart	512	32	100	0.04	0.08
ScanNet	4800	32	100	0.05	0.10
nuScenes	5200	32	100	0.05	0.10

B. Additional Experimental Analyses

B.1. VLM representation anchors

VLM representation anchor is designed to find important VLM representations that are more suitable for semantic alignment. In our effort to cover a broad spectrum of categories and improve the anchors' general applicability, we conducte experiments using VLM representations derived from 32 point clouds. This idea is inspired by the concept of a memory bank for constrastive learning. We also produce geometric representation anchors, employing the same weight parameters as those used for calculating the VLM representation anchors. The purpose of these geometric representation anchors is to aid the anchor projection process, thereby mitigating issues of semantic misalignment. This is achieved by searching for the closest VLM representation anchor for each point, taking into account both the similarities of VLM representations with VLM representation anchors and geometric representations with geometric representation anchors. Fig. 1 displays the results of zero-shot semantic segmentation on ScanNet [2], using the OpenSeg as the feature extractor. This figure clearly shows the improved performance in zero-shot semantic segmentation achieved by incorporating VLM representation anchors (presented in the bottom row), especially when compared to the method that does not utilize VLM representation anchors, as seen in the third row of the figure.

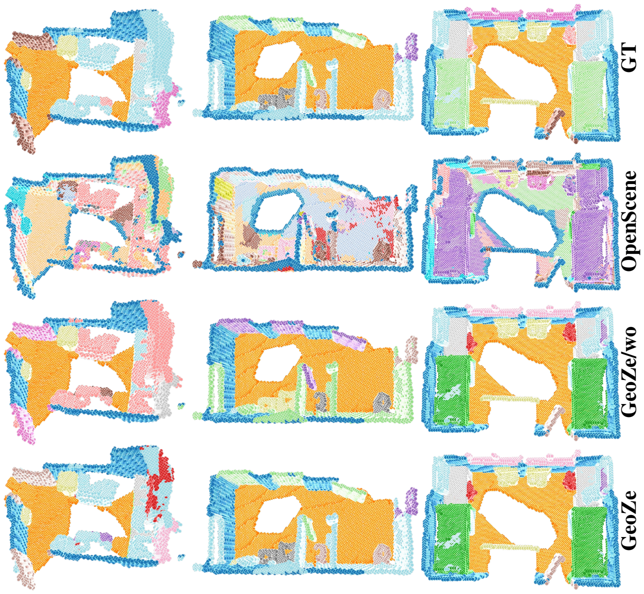
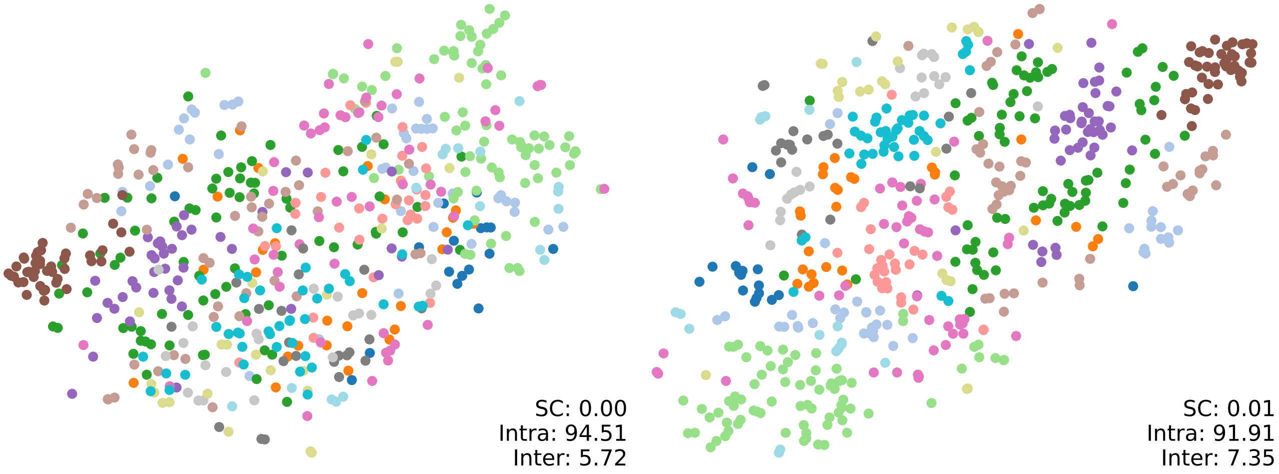


Figure 1. Zero-shot semantic segmentation results on ScanNet using OpenSeg feature extraction. (top row) ground-truth annotations, (second row) OpenScene (OpenSeg) [6], (third row) GeoZe/wo (OpenSeg) without using VLM representation anchors, and (bottom row) GeoZe (OpenSeg). 'wo' indicates the absence of VLM representation anchors.

B.2. Classification visualization

In Fig. 2, we report a t-SNE comparison between the class representations extracted with PointCLIPv2 and GeoZe on ScanObjectNN [9] (S-PB-T50-RS). We quantify t-SNE clusters with three clustering metrics: Silhouette Coefficient, Inter-cluster Distance, and Intra-cluster Distance. From these metrics, we can confirm the efficacy of GeoZe in better separating point features of diverse categories compared to PointCLIPv2.

Fig. 3 compares the same global features using traditional clustering metrics for a more detailed assessment of the differences between GeoZe and PointCLIPv2. For a comprehensive analysis of intra- and inter-cluster statistics, we consider six extrinsic clustering measures, that explicitly compare classification predictions with ground-truth annotations. The Adjusted Rand Index (ARI) evaluates the similarity of cluster assignments through pairwise comparisons. The Adjusted Mutual Information (AMI) assesses the agreement of cluster assignments. Homogeneity (H) gauges the proportion of instances from a single class in a cluster, akin to Precision. Completeness (C) measures the proportion of a given class's instances assigned to the same cluster, similar to Recall. The V-measure (V) quantifies clustering correctness using conditional entropy analysis. The Fowlkes-Mallows score (FM) evaluates clustering accuracy through the geometric mean of pairwise Precision and Recall. Higher scores in all these metrics indicate better performance. In Fig. 3, the histogram values are normalized, with the maximum value for each score set to 1. GeoZe outperforms PointCLIPv2 in all metrics.



(a) PointCLIPv2 [12]

(b) GeoZe

Figure 2. T-SNE embeddings of (a) PointCLIPv2 [12] and (b) GeoZe on ScanObjectNN [9] (S-PB-T50-RS). GeoZe produces better separated and grouped clusters for different categories, as evidenced by the superior silhouette coefficient (SC) and greater inter-cluster distance (inter), alongside a smaller intra-cluster distance (intra).

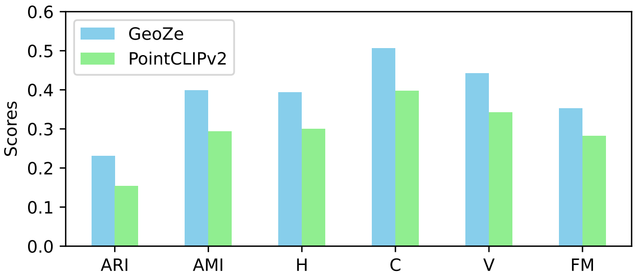


Figure 3. Comparison of clustering metrics (the higher the better) on ModelNet40 [8]. Metrics key: ARI: adjusted rand index, AMI: adjusted mutual information, H: homogeneity score, C: completeness score, V: V-measure, FM: Fowlkes-Mallows score.

B.3. Segmentation results on nuScenes

Fig. 4 showcases a range of qualitative results obtained using our method, GeoZe, on the nuScenes outdoor dataset [1]. The effectiveness of GeoZe is underscored by its ability to produce more semantically coherent segmented regions, a property primarily attributed to our novel clustering technique. Additionally, the synergy of local and global aggregation techniques bolsters the VLM representation, making it more geometrically aware. Another reason is the integration of geometric representation assignment, which plays a pivotal role in reducing semantic misalignment during anchor projection. The positive effects of our method are particularly visible at the boundaries of point clouds, where GeoZe substantially lowers noise levels, outperforming the compared method OpenScene. This noise reduction is especially effective owing to the distinct geometric structures commonly found at these boundary zones.

B.4. VLM representation guided clustering

In this section, we demonstrate the enhancement in semantic clustering performance achieved by integrating geometric information into VLM representation. We use the clustering scores to compute each cluster's prototypical representation. Specifically, these prototypes are weighted averages of VLM representations, based on the clustering scores. Subsequently, each point is assigned to the prototype of its corresponding cluster. PCA projection is then applied to visualize the clustering results. Fig. 5 showcases some qualitative clustering results on ShapeNetPart [11] using different point-level coordinates (Coord.) and representations. The top row demonstrates results using only coordinates for clustering, while the second row combines the coordinates with VLM representations. The third row integrates coordinates and geometric representations (FPFH), offering more meaningful partitioning compared to using just coordinates or coordinates with VLM representations. The bottom row features GeoZe, which considers both coordinates and geometric representations (FPFH) for clustering, guided by VLM representation similarity, achieving the best clustering results (same parts tend to share colors).

Acknowledgement. This work was supported by the PNRR project FAIR- Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan,

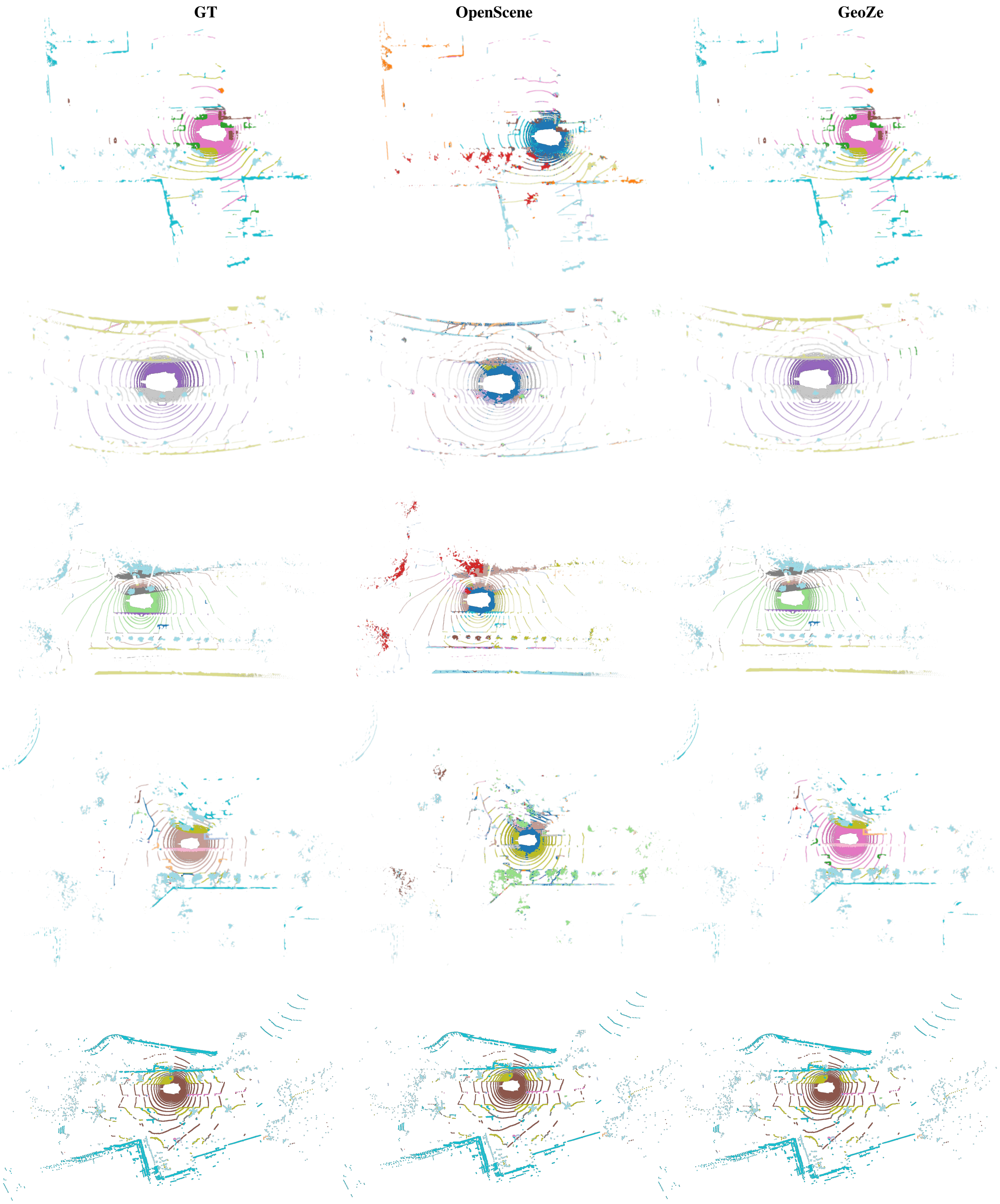


Figure 4. Zero-shot semantic segmentation results on nuScenes [1] using OpenSeg as feature extractor. (left column) ground-truth annotations, (middle column) OpenScene (OpenSeg) [6], and (right column) GeoZe.

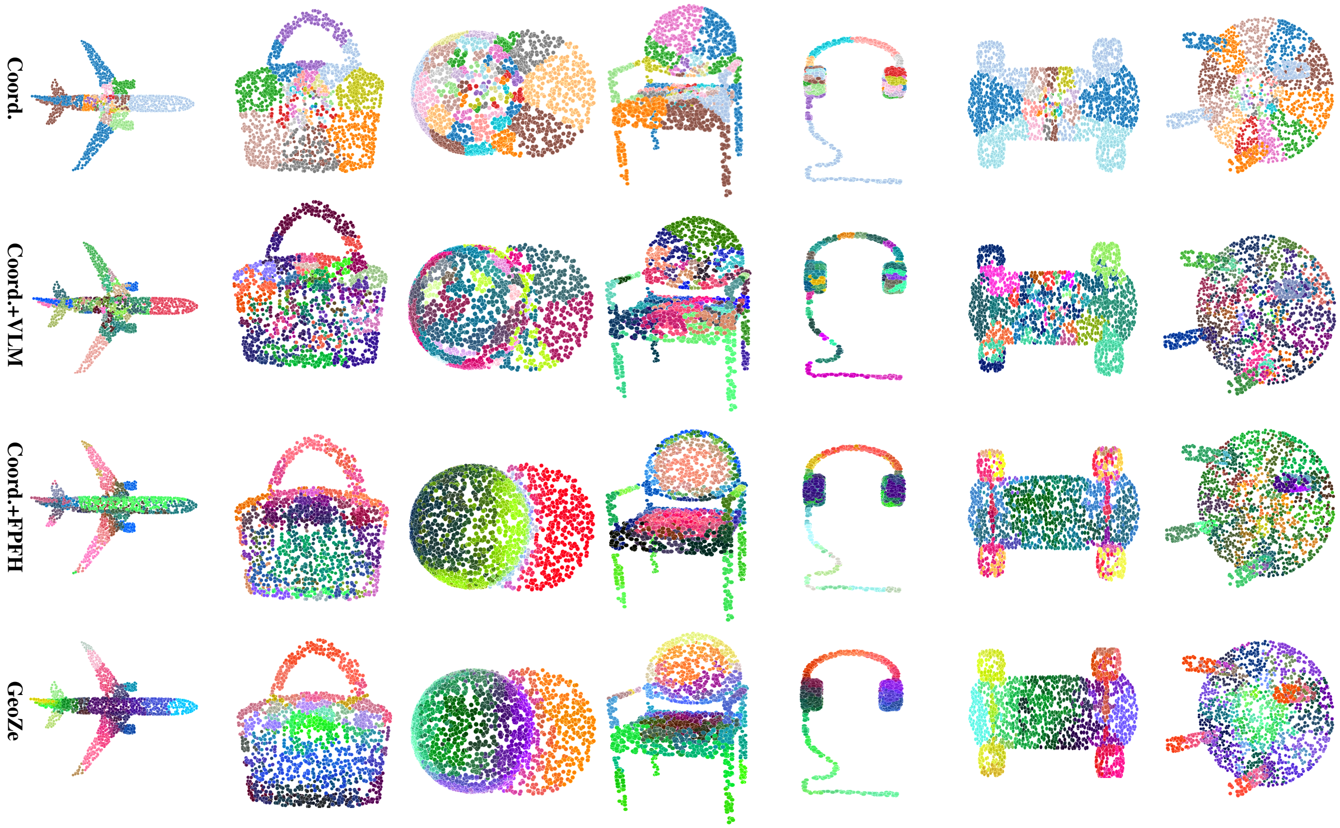


Figure 5. Visualization of clustering results on ShapeNetPart [11] using various sources: (top row) coordinates only; (second row) coordinates with VLM representations; (third row) coordinates and geometric representations (FPFH); (fourth row) coordinates and geometric representations (FPFH), guided by VLM representation similarity. Coord. represents Coordinates.

Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 1, 3, 4

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1, 2

[3] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, pages 540–557. Springer, 2022. 1

[4] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. ConceptFusion: Open-set Multimodal 3D Mapping. In *RSS*, 2023. 1

[5] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1

[6] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. 1, 2, 4

[7] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *ICRA*, 2009. 2

[8] Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconvdae: Deep volumetric shape learning without object labels. In *ECCV*, 2016. 1, 3

[9] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 1, 2, 3

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 1

[11] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM TOG*, 2016. 1, 3, 5

[12] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. PointCLIPv2: Prompting CLIP and GPT for Powerful 3D Open-world Learning. In *ICCV*, 2023. 1, 3