Breaking the 2D Dependency: What Limits 3D-Only Open-Vocabulary Scene Understanding

Domenico D'Orsi
Dept. of Computer Engineering
University of Pisa, Italy
d.dorsi@studenti.unipi.it

Fabio Carrara
CNR-ISTI
Pisa, Italy
fabio.carrara@isti.cnr.it

Fabrizio Falchi
CNR-ISTI
Pisa, Italy
fabrizio.falchi@isti.cnr.it

Nicola Tonellotto
Dept. of Computer Engineering
University of Pisa, Italy
nicola.tonellotto@unipi.it

***Abstract*—Open-vocabulary 3D scene understanding, i.e., recognizing and classifying objects in 3D scenes without being limited to a predefined set of classes, is a foundational task for robotics and extended reality applications. Current leading methods often rely on 2D foundation models to extract semantics, then projected in 3D. This paper investigates the viability of a purely 3D-native pipeline, thereby eliminating dependencies on 2D models and reprojections. We systematically explored various architectural combinations using established 3D components. However, our extensive experiments on benchmark datasets reveal significant performance limitations with this direct 3D-native approach, with performance metrics falling short of expectations. Rather than a simple failure, these outcomes provide critical insights into the current deficiencies of existing 3D models when cascaded for complex open-vocabulary tasks. We highlight the lessons learned, identify the pipeline's limitations (e.g., segmenter-encoder domain gap, robustness to imperfect segmentations), and posit future research directions. We argue that a fundamental rethinking of model design and interplay is necessary to realize the potential of truly 3D-native open-vocabulary understanding.**

***Index Terms*—Open-vocabulary 3D scene understanding, 3D Scene Segmentation, Multimodal Point Cloud Encoder, 3D-Only Pipeline.**

I. Introduction

The pursuit of robust 3D scene understanding remains a cornerstone of computer vision, fueled by its potential to revolutionize human-computer interaction, robotics, and immersive experiences. Traditionally, scene understanding has relied on predefined taxonomies. However, the ability to comprehend scenes in an open-vocabulary manner — recognizing and classifying objects without closed-set class constraints — represents a more flexible and scalable paradigm.

Current state-of-the-art open-vocabulary 3D scene understanding methods [1]–[7] predominantly extract semantic information from registered 2D images using 2D vision language models and map it into 3D representations. Some methods attach semantic features to individual points of 3D point clouds, either learned [1] or obtained via aggregation strategies [3]. Other methods, such as [2], [5], [6], map 2D-derived features to groups of points provided by agnostic 3D segmentation. LERF [4] and LangSplat [2] instead distill the 2D semantic features into a fitted parametric radiance field, respectively a NeRF [8] or Gaussian Splatting [9] representation, and are then able to render semantic features. While effective, pipelines based on 2D multi-modal models are usually complex due to the necessity of accessing and processing a large collection of 2D-3D registered data, which is not independent from the scene reconstruction process. This raises the question: can we achieve effective open-vocabulary 3D scene understanding using a purely 3D-native pipeline, operating solely on 3D data (e.g., colored point clouds)? Such an approach could offer streamlined processing, reduced data/storage requirements, and applicability in scenarios where 2D data might not be available.

In this paper, we present an investigation into the design and implementation of such a 3D-native pipeline. We set up a simple two-stage pipeline that sequences a 3D object segmenter directly with a multimodal (text-3D) point cloud encoder for open-vocabulary classification.

We explored combinations of state-of-the-art components for these tasks. Section III details our experimental evaluation on benchmark datasets, revealing significant performance limitations with this direct 3D-native approach. These outcomes, rather than indicating a simple failure, provide critical insights into the current deficiencies of existing 3D models when cascaded for complex open-vocabulary tasks. Through targeted experiments and analysis, we articulate the key lessons learned regarding these limitations (e.g., segmenter-encoder domain gap, robustness to imperfect segmentations) and identify what is critically missing in the current state-of-the-art to enable effective purely 3D open-vocabulary understanding. Building on these findings, we propose future research directions that we believe are essential for progress in realizing the potential of truly 3D-native open-vocabulary understanding.

II. An Exploratory 3D-Native Framework

Our proposed framework consists of a sequence of distinct processing modules, each transforming the input 3D data. We begin with a colored point cloud, formally defined as $\mathcal{P} = \{(p_i, c_i)\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^3$ represents the spatial coordinates of the i-th point and $c_i \in \mathbb{R}^3$ its corresponding color information. We then perform the following steps:

a) Binary Mask Proposer: The initial stage involves class-agnostic 3D instance segmentation, which aims to identify and delineate individual objects within the scene. Starting from the colored point cloud $\mathcal{P}$, this step generates a set of binary masks, $\mathcal{M} = \{M_j\}_{j=1}^{K}$. Each mask M_j provides a

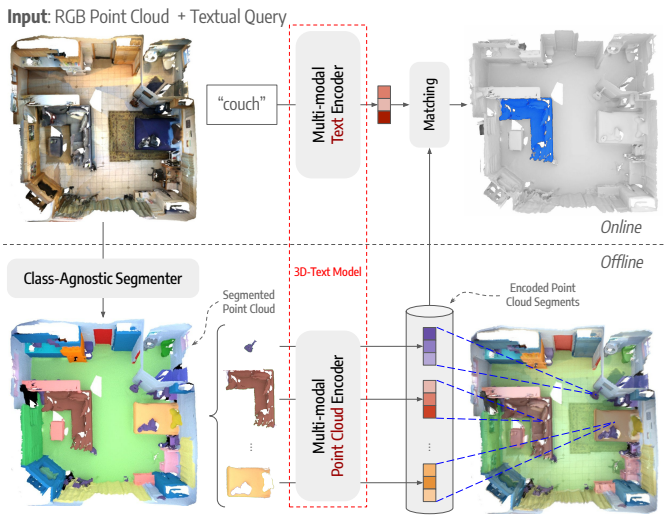

Fig. 1. Schematic of the explored purely 3D-native open-vocabulary scene understanding pipeline. While conceptually straightforward, this study reveals significant challenges in its practical realization with current SOTA components.

binary label $m_{i,j} \in \{0, 1\}$ for every point p_i in the original point cloud, indicating its belonging to the j-th segmented instance.

b) Point Cloud Extraction from Masks: We then extract the point cloud corresponding to each segmented object. Given the original colored point cloud $\mathcal{P}$ and a specific mask M_j, we derive the object-specific point cloud $\mathcal{P}_j$ by selecting only those points (p_i, c_i) from $\mathcal{P}$ for which the corresponding mask value $m_{i,j}$ is 1.

c) Multimodal Point Cloud Encoder: Each extracted object point cloud $\mathcal{P}_j$ is then fed into a multimodal (text-3D) point cloud encoder E_{PC}. Depending on the model used, the input point clouds may be preprocessed, including coordinate normalization, point re-densification, and/or downsampling, resulting in $\mathcal{P}'_j$. The encoder's role is to process $\mathcal{P}'_j$ and produce a dense embedding vector $e_j \in \mathbb{R}^d$. This embedding is designed to encapsulate the key geometric and semantic characteristics of the represented object.

d) Query Embedding and Matching: For the final open-vocabulary classification task, textual queries t, such as "chair" or "table", are encoded into embedding vectors $e_t \in \mathbb{R}^d$ using the text encoder component of the multimodal model E_{T}. The semantic relationship between an object embedding e_j and a text query embedding e_t is then quantified using the cosine similarity: $s(e_j, e_t) = \frac{e_j \cdot e_t}{\|e_j\| \|e_t\|}$. The classification of the object is determined by identifying the textual query that yields the highest similarity score with its corresponding embedding.

III. Experiments and Analysis of Limitations

We evaluated our proposed 3D-native framework for open-vocabulary scene understanding on the ScanNet200 benchmark [10], [11]. This challenging dataset comprises approximately 1500 real-world scenes and features 200 object classes with a long-tailed distribution.

TABLE I
ZERO-SHOT 3D SEGMENTATION AP METRICS ON SCANNET200. GRAY ROWS ARE SUPERVISED OR MULTI-MODAL METHODS NOT DIRECTLY COMPARABLE TO 3D-ONLY ZERO-SHOT SETTINGS.

Method	AP	AP_{50}	AP_{25}
Closed-set Fully Supervised			
Full Mask3D	26.9	36.2	41.4
Zero-shot with 2D data			
OpenScene (LSeg)	6.0	7.7	8.5
OpenScene (OpenSeg)	11.7	15.2	17.8
OpenMask3D (CLIP)	15.4	19.9	23.1
Zero-shot 3D-only data			
Mask3D + ULIP	0.2	0.5	0.6
Mask3D + ULIPv2	**0.4**	**0.6**	**0.8**
Segment3D + ULIP	0.2	0.4	0.5
Segment3D + ULIPv2	0.3	0.5	0.6
Segment3D + ReCon	0.1	0.2	0.3
Segment3D + ReCon++	0.2	0.3	0.3

A. Tested Models

For our experiments, we selected off-the-shelf models for the key components of our pipeline. Specifically, we employed the most performant state-of-the-art models with publicly available code and checkpoints.

For the object masks proposal, we evaluated a) Mask3D [5], a transformer-based architecture for 3D point cloud instance segmentation that provides class-agnostic binary object masks as intermediate output, and b) Segment3D [6], which shares the same architecture of Mask3D but is trained with weakly-labeled data produced by applying 2D segmentation models to posed RGBD images.

For point cloud encoding, we evaluated a) ULIP [12], a CLIP-like multimodal contrastive model that embeds point clouds with a PointBERT [13] encoder into the frozen CLIP space, b) ULIPv2 [14], an improved version of ULIP with a scaled-up point cloud encoder [15] and trained on a larger automatically annotated dataset, c) ReCon [16], a contrastive 3D representation learning model that integrates generative and contrastive objectives to enhance the transferability and generalization of learned embeddings, d) ReCon++ [17], an extension of ReCon trained within the ShapeLLM framework using a large-scale dataset of image-text-shape triplets, which further boosts the quality of the learned point cloud embeddings by leveraging richer cross-modal supervision. ULIP and ReCon encode only the point cloud's geometry, while ULIPv2 and ReCon++ also incorporate point color information.

B. Performance of Pipeline Combinations

Table I presents the zero-shot 3D instance segmentation performance of various combinations of the selected segmenters and encoders on ScanNet200. The results are notably poor across all configurations. The best-performing combination, Mask3D with ULIPv2, achieves an Average Precision (AP) of only 0.4%. Even Segment3D and ReCon(++) — proposed as improvements to their competitors (Mask3D and ULIP, respectively) — consistently achieve low performance, indicating a fundamental limitation in the effectiveness of a

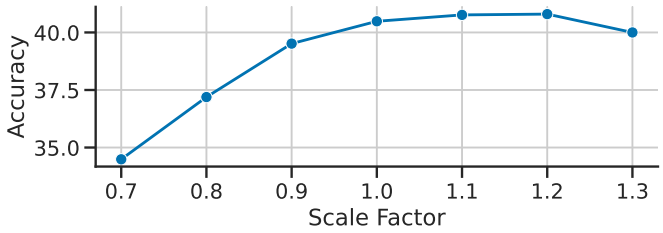

Fig. 2. Accuracy of ReCon against the point cloud coordinate scale factor on the zero-shot classification of the ScanObjectNN dataset.

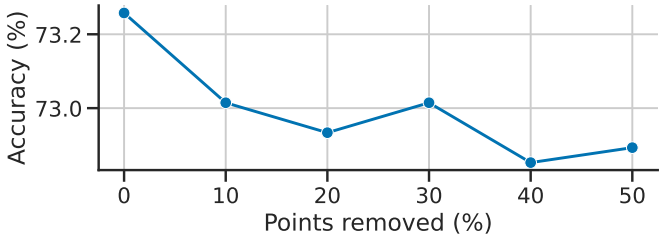

Fig. 3. Accuracy of ULIP against point cloud density on the zero-shot classification of the ModelNet40 dataset. On the x-axis, we report the percentage of points randomly removed from the input point cloud.

straightforward 3D-native pipeline. This is probably due to gaps between the specific domains on which each model was initially pretrained. We investigate the limitations of this seemingly logical pipeline more deeply.

Segmentation Quality. Segment3D provided better masks qualitatively than Mask3D, as demonstrated quantitatively in [6]. Still, it reaches an AP of 27.7 on class-agnostic segmentation on the ScanNet200 validation set, leaving room for improvement on the segmentation part of the pipeline. We note that Segment3D often provided more mask proposals with respect to Mask3D, but in the evaluation pipeline, this penalized performance as most of those valid masks end up being erroneously classified in the ScanNet200 taxonomy, lowering the AP metric overall.

Point Cloud Quality. We investigated whether some mismatch in point cloud properties, i.e., scale and density, between the ones used to train the encoders and those produced by the segmenter might be a significant factor in performance degradation. We started investigating scale. After matching the segmented point cloud scale with the one required by the point cloud encoder (via coordinate centering, normalization, and/or scaling), we applied a variable scale factor to coordinates to test the robustness of encoders to scale differences. Figure 2 shows the zero-shot classification accuracy of ReCon when varying the point cloud coordinate scale factor on ScanObjectNN [18], its training dataset. The performance degradation is only marginal when the scale is perturbed. Indeed, the same test on ScanNet200 revealed no significant changes in the performance of our pipeline for the 3D instance segmentation task. We also investigated point cloud density, as point cloud embedders are usually trained on complete, high-quality, high-density clouds with respect to the ones segmented from scans,

TABLE II
ZERO-SHOT CLASSIFICATION PERFORMANCE (PERCENTAGES) OF A LINEAR PROBE ON SEGMENT3D + ULIPV2 EMBEDDINGS ON SCANNET200.

	Micro-averaged			Macro-averaged		
Accuracy	Precision	Recall	F1-score	Precision	Recall	F1-score
51.0	51.0	51.0	50.0	20.0	15.0	16.0

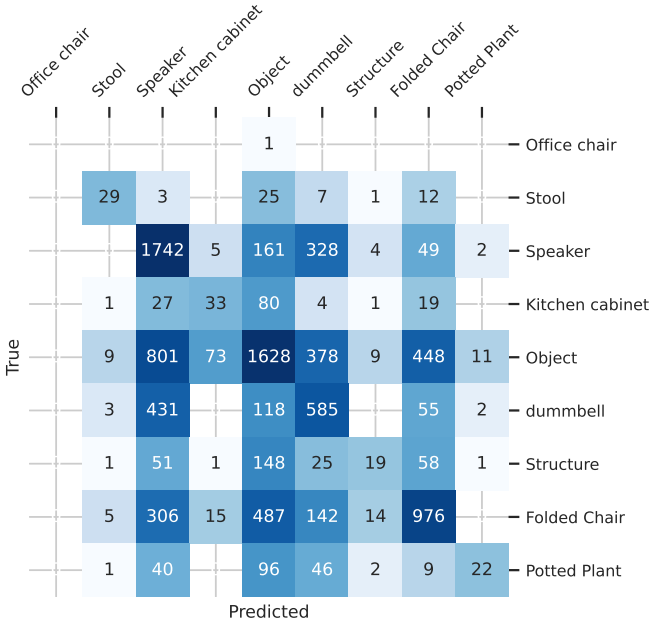

Fig. 4. Confusion matrix (most frequent classes only) of linear probing over Segment3D + ULIPv2 embeddings on ScanNet200. Predominant classes show moderate recognition, suggesting embeddings are not entirely uninformative.

which can vary from very few points to more than 10k points, depending on the dimensions of the objects. ULIP has the most stringent density requirement, as it has been trained with 8192 points per cloud. We tested the robustness of ULIP to point cloud density by performing zero-shot classification of ModelNet40 clouds, randomly removing increasing percentages of points up to 50% (the common lowest percentage of objects seen in raw scans to be recognizable), observing no strong degradation (see Figure 3).

Embeddings Quality. We also investigated whether the point cloud encoders failed to extract meaningful semantic information from the segmented objects. To do this, we trained a simple linear classifier on top of the frozen embeddings generated by Segment3D + ULIPv2 on ScanNet200 and attempted to predict the ground truth labels. Table II shows the zero-shot classification performance, and Figure 4 shows the confusion matrix of the most frequent predicted classes in ScanNet200. The linear classifier achieved a weighted F1 Score of 50.0%, and the confusion matrix indicates some ability to recognize frequent and geometrically well-defined classes. This suggests that the point cloud encoders are not producing entirely uninformative embeddings. Note, however,

that the representation gap is more prominent in less frequent classes, as testified by the micro-averaged metrics.

Overall, we deem the poor performance of the full zero-shot pipeline likely stems from more complex issues, such as a distribution mismatch between encoder training data (typically clean, isolated objects) and the noisy, partial segments from real-world scenes, and the difficulty of aligning these "raw" 3D segment embeddings with text embeddings in the absence of strong 2D visual cues.

IV. Discussion

We explored a fully 3D-native pipeline for open-vocabulary scene understanding by sequentially combining state-of-the-art 3D segmenters and encoders. The approach, however, performed poorly on the ScanNet200 benchmark (Table I), revealing fundamental limitations of current 3D models when applied to this task.

We deem that the performance gap arises from several compounding factors:

a) Error Propagation and Domain Shift: Errors propagate and amplify in sequential pipelines. Imperfect outputs from the initial segmentation stage — e.g., noisy or incomplete masks from Segment3D (27.7% class-agnostic AP on ScanNet200 val) — directly degrade encoder performance. Existing 3D encoders are trained on clean object datasets (e.g., ScanObjectNN, ModelNet40) and struggle more with noisy, partial, and out-of-distribution segments. While isolated perturbations (Figures 2, 3) show limited impact in controlled experiments, real-world segment imperfections combine to create a significant domain mismatch.

b) 3D Encoder Limitations on Real-World Segments: Despite reasonable performance on frequent and distinct classes (e.g., 50.0% weighted F1-score from linear probing Segment3D+ULIPv2 embeddings, Table II), macro F1 remains low (16.0%). This indicates a lack of robustness in handling less common or geometrically ambiguous objects. Encoders pretrained on clean datasets fail to extract discriminative features from sparse, noisy, and occluded segments typical in scene-level data.

c) Lack of Strong 3D-Text Alignment: Current 3D-native models lack the web-scale semantic grounding that 2D models like CLIP provide. Projecting 2D semantics into 3D has proven effective, but our design intentionally bypassed this. The results suggest that 3D encoders are still insufficiently aligned with textual concepts to support open-vocabulary classification in complex scenes.

Overall, the experiment demonstrates that naively chaining current 3D modules is inadequate for open-vocabulary scene understanding, despite the success of similar modular architectures in the image/2D domain. The modular approach fails to address interoperability issues and the specific challenges of real-world data. This does not invalidate 3D-native methods but highlights the need for better 3D foundational models.

V. Conclusion and Future Directions

Our investigation into a purely 3D-native pipeline for open-vocabulary scene understanding highlights the limitations of current 3D models when applied directly to real-world data. The sequential combination of existing 3D segmentation and encoding modules proved insufficient to match the performance of approaches leveraging 2D vision-language priors. This result is not merely a matter of insufficient model capacity but reflects deeper issues of data distribution, model robustness, and lack of semantic grounding.

Our experiments show that direct 3D-native pipelines, while conceptually straightforward, perform poorly on benchmarks such as ScanNet200. As shown in Table I, all tested configurations yield low Average Precision (AP), highlighting a substantial gap relative to 2D-assisted methods. This underperformance cannot be explained by minor variations in point cloud scale or density alone (Figures 2 and 3). Instead, the primary limitations arise from compounded issues: error propagation from inaccurate segmentations, domain shift between clean training data and noisy real-world inputs, and insufficient alignment between 3D geometry and text semantics. Although object embeddings retain some semantic structure (Table II, Figure 4), the signal is too weak to support reliable open-vocabulary classification in a cascaded setup.

Robustness to real-world data imperfections remains a fundamental challenge. Unlike 2D vision-language models trained on billions of image-text pairs, 3D models lack large-scale, diverse supervision. Existing 3D encoders are generally trained on clean, complete objects from synthetic datasets. Addressing the domain gap requires not only better data augmentation (e.g., simulating occlusions, varying point density, adding noise) but also more realistic pretraining datasets derived from in-the-wild scene scans. Synthetic data generation, e.g., via composition of clusters of objects available in large databases like Objaverse [19], combined with the simulation of scan conditions (scan noise, occlusion, varying density), comprises a viable direction to cope with current data limitations.

Recent advances in self-supervised learning for point clouds [20], [21] provide promising tools for robust point cloud encoding, also in the presence of noisy, partial instance segments, as they are not limited to the availability of coupled text-3D data, as current 3D multimodal models do. A promising solution could be mapping the textual space (e.g., the CLIP textual space) to the self-supervised modality (similar to [22]), providing a language-aligned representation without changing the point cloud self-supervised space.

Another interesting direction is the inclusion of *iterative refinement* mechanisms. For instance, initial segmentations could be refined based on recognition feedback, using generative reconstruction models [23], [24]. Such architectures could dynamically revisit segmentation hypotheses when uncertainty in classification is high, mitigating the accumulation of early-stage errors.

In summary, the disappointing results of the pipeline do not undermine the potential of 3D-native scene understanding. Rather, they reveal that achieving competitive performance requires new architectures, training paradigms, and supervision strategies explicitly designed to handle the complexities of real-world 3D data and open-vocabulary semantics.

ACKNOWLEDGMENTS

This work was supported, in part, by the SUN XR project funded by the Horizon Europe Research & Innovation Programme (GA n. 101092612), the Spoke "FutureHPC & BigData" of the ICSC - Centro Nazionale di Ricerca in High-Performance Computing, Big Data and Quantum Computing, the FoReLab project (Departments of Excellence), and the NEREO PRIN project funded by the Italian Ministry of Education and Research Grant no. 2022AEFHAZ.

REFERENCES

[1] S. Peng, K. Genova, A. T. ChiyuMaxJiang, M. Pollefeys, and T. A. Funkhouser, "Openscene: 3d scene understanding with open vocabularies. 2023 ieee," in *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 815–824.

[2] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "Pla: Language-driven open-vocabulary 3d scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[3] K. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, "Conceptfusion: Open-set multimodal 3d mapping," *Robotics: Science and Systems (RSS)*, 2023.

[4] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023.

[5] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, "Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels," in *European Conference on Computer Vision*. Springer, 2024, pp. 278–295.

[7] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "Langsplat: 3d language gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 051–20 060.

[8] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, July 2023. [Online]. Available: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

[10] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[11] D. Rozenberszki, O. Litany, and A. Dai, "Language-grounded indoor 3d semantic segmentation in the wild," in *European Conference on Computer Vision*. Springer, 2022, pp. 125–141.

[12] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, "Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1179–1189.

[13] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 313–19 322.

[14] L. Xue, N. Yu, S. Zhang, A. Panagopoulou, J. Li, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles *et al.*, "Ulip-2: Towards scalable multimodal pre-training for 3d understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 091–27 101.

[15] G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[16] Z. Qi, R. Dong, G. Fan, Z. Ge, X. Zhang, K. Ma, and L. Yi, "Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 223–28 243.

[17] Z. Qi, R. Dong, S. Zhang, H. Geng, C. Han, Z. Ge, L. Yi, and K. Ma, "Shapellm: Universal 3d object understanding for embodied interaction," in *European Conference on Computer Vision*. Springer, 2024, pp. 214–238.

[18] D. Liu, C. Chen, C. Xu, Q. Cai, L. Chu, F. Wen, and R. Qiu, "A robust and reliable point cloud recognition network under rigid transformation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–13, 2022.

[19] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35 799–35 813, 2023.

[20] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li, "Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training," *Advances in neural information processing systems*, vol. 35, pp. 27 061–27 074, 2022.

[21] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 769–21 780.

[22] L. Barsellotti, L. Bianchi, N. Messina, F. Carrara, M. Cornia, L. Baraldi, F. Falchi, and R. Cucchiara, "Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2411.19331

[23] S. Mo, E. Xie, R. Chu, L. Hong, M. Niessner, and Z. Li, "Dit-3d: Exploring plain diffusion transformers for 3d shape generation," *Advances in neural information processing systems*, vol. 36, pp. 67 960–67 971, 2023.

[24] X. Yan, L. Lin, N. J. Mitra, D. Lischinski, D. Cohen-Or, and H. Huang, "Shapeformer: Transformer-based shape completion via sparse representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.