Memory-Efficient 3D Scene Understanding via Quantized Multi-View Features

Ashkan Alami[1,2], Fabio Remondino[1]

[1] 3D Optical Metrology (3DOM) unit, Bruno Kessler Foundation (FBK), Trento, Italy – (aalami, remondino)@fbk.eu
[2] University of Trento, Department of Information Engineering and Computer Science (DISI), Trento, Italy

Keywords: 3D scene understanding, foundation models, feature quantization, large-scale point clouds, 2D-to-3D distillation.

Abstract
Projecting high-dimensional representations derived with 2D foundation models onto 3D point clouds via multi-view aggregation has emerged as a powerful paradigm for 3D scene understanding. However, the conventional strategy of storing dense feature vectors - derived by 2D foundation models at each point - introduces a substantial memory overhead that scales linearly with the scene size, thereby limiting scalability and hindering practical deployment in large-scale settings. In this work a memory-efficient framework is proposed to address this limitation through feature quantization. Specifically, projected 2D features are clustered into a compact set of prototypical embeddings, enabling each 3D point to be represented by a single discrete index instead of a full high-dimensional descriptor. This representation drastically reduces memory requirements by orders of magnitude while preserving the semantic richness of the original features. The proposed quantization framework is validated on diverse point clouds, demonstrating that the representations retain strong performance in downstream tasks while significantly improving computational efficiency.

1. Introduction

3D scene understanding has become an essential component in numerous real-world domains, ranging from self-driving cars operating in urban environments to unmanned aerial vehicles conducting large-scale 3D surveys (Geiger et al., 2012; Behley et al., 2019; Caesar et al., 2020; Bayrak et al., 2023). Fuelled by this increasing need, the field has witnessed substantial progresses, with advanced scene understanding approaches delivering strong performance across challenging indoor 3D settings (Dai et al., 2017; Choy et al., 2019; Schult et al., 2022; Wu et al., 2024). Nevertheless, a critical limitation persists: the majority of these approaches fail to generalize to large outdoor environments containing tens of millions of points, such as dense urban landscapes captured via drone/UAV or airborne platforms. This challenge does not stem from algorithmic shortcomings or data scarcity but it is rooted in memory constraints.
2D foundation models such as CLIP (Radford et al., 2021) and the DINO family (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025), pre-trained on web-scale image collections, yield semantically rich visual representations that transfer effectively across diverse tasks and settings, even in the complete absence of labelled test-time data. Beyond being powerful encoders, these representations serve as the foundation for a broad spectrum of downstream applications, including segmentation, detection, retrieval and scene understanding, frequently under zero-shot or few-shot conditions (Peng et al., 2023; Puy et al., 2024; Zhang et al., 2025; Knaebel et al., 2026). This naturally raises a question: is it possible to extend these capabilities to the 3D domain? For small-scale and indoor environments, the answer is yes. Current 3D point clouds are increasingly paired with multi-view RGB imagery and accurate camera calibration data, establishing a natural correspondence between the 2D and 3D modalities. Several recent approaches leverage this link by transferring rich 2D foundation model features onto 3D point clouds through multi-view projection (Rozenberszki et al., 2022; Sautier et al., 2022; Chen et al., 2023; Peng et al., 2023; Alami and Remondino, 2026). The outcomes are convincing, with 3D representations that inherit strong visual and semantic priors, enabling zero-shot inference as well as the development of more broadly capable 3D models (Liu et al., 2023; Peng et al., 2023; Takmaz et al., 2023). Such methods demonstrate impressive performance primarily on indoor benchmarks such as ScanNet (Dai et al., 2017) and driving-oriented datasets like KITTI (Geiger et al., 2012). However, processing scenes that contain millions of points, such as a city block, a large heritage site or a dense aerial point cloud, where each point must carry a high-dimensional feature vector, quickly exhausts the memory resources of any conventional GPU or workstation. As a concrete example, some 10 million 3D points with 1024-dimensional feature vectors alone demand tens of gigabytes of memory. Partitioning the scene into smaller chunks does not offer a viable solution: either the complete set of per-point features must still be maintained in a global store, yielding no reduction in memory usage; or each partition must undergo the full processing pipeline independently, from image retrieval and feature extraction through to projection, only for the intermediate results to be discarded, rendering the overall workflow impractically inefficient. Unlocking the potential of 2D foundation model features for large-scale outdoor scenarios raises a fundamental question: is a unique feature vector truly necessary for every individual 3D point? 3D scenes, and in particular urban environments, exhibit inherent semantic redundancy by nature. Throughout millions of points, identical visual patterns reappear consistently. This observation implies that the continuous, high-dimensional feature space is manageable with strong compression procedures, not through information loss, but by acknowledging that the vast majority of 3D points share common underlying feature representations.

1.1 Aim of the work

This work introduces a quantization-based framework (Section 4) that directly exploits the inherent redundancy present in geospatial 2D-3D datasets. Rather than assigning a unique high-dimensional feature vector to each point, a compact feature bank of K prototypical embeddings is constructed by clustering 2D features across the entire scene. Every point is subsequently mapped to a discrete index within this bank, shrinking the memory footprint of a 10 million 3D points from tens of gigabytes to less than 100 MB, making previously unmanageable large-scale scenes processable on commodity hardware.
Validation results (Section 5) demonstrate that the proposed approach successfully handles datasets that were previously beyond the reach of existing methods, without compromising feature quality and the successive semantic segmentation.
Therefore, the core contributions of this work centres on dismantling the memory barrier that has thus far prevented rich 2D foundation model features from being effectively distilled into large-scale 3D scene understanding and they are summarized as follows:

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-9-2026

1. A memory-efficient pipeline for transferring 2D foundation model features to large-scale 3D point clouds, achieved through the construction of a discrete feature bank and the assignment of compressed index representations to individual points.
2. A thorough evaluation spanning a diverse range of indoor and outdoor 3D scenes, demonstrating that large-scale point clouds can now be processed within practical memory constraints while retaining feature quality sufficient for downstream tasks.

2. Related works

2.1 Evolution of 3D scene understanding

The domain of 3D scene understanding has experienced multiple fundamental shifts in methodology. Early approaches depended on manually engineered geometric features (Ni et al., 2017; Czerniawski et al., 2018; Grilli and Remondino, 2020), while the emergence of deep learning subsequently enabled end-to-end processing of raw point clouds (Qi et al., 2017a; Qi et al., 2017b; Thomas et al., 2019; Choy et al., 2019; Lai et al., 2022; Wang, 2023; Wu et al., 2024; Yang et al., 2025). Yet despite considerable progress in architectural innovations, supervised approaches for 3D classification remain inherently bounded by the availability of annotated data. Producing such annotations is resource-intensive across all tasks and this burden is especially pronounced in large 3D settings such as STPLS3D (Chen et al., 2022), SensatUrban (Hu et al., 2022) or ESTATE (Bayrak et al., 2024). The 2D vision community encountered this same constraint and responded not by expanding annotation pipelines, but by pivoting toward self-supervised learning over unlabelled data.

2.2 Foundation models: representation learning at scale

Self-supervised learning redefines how the annotation bottleneck is addressed: rather than depending on human-labelled data, models derive transferable representations autonomously from raw inputs. Among the earliest formulations, contrastive frameworks such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) acquire representations by encouraging consistency across differently augmented views of the same sample. Masked image modeling methods like MAE (He et al., 2022) take a complementary route, employing generative reconstruction as a pretext task, the model is trained to recover obscured regions, compelling it to internalize the structural regularities present in the data.

Modern foundation models integrate these principles at a scale previously unseen. The DINO family (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) pairs self-distillation with masked modeling objectives. The latest 7-billion parameter DINOv3 yields high-dimensional 4096-dimensional visual descriptors trained across more than 1.7 billion images. In parallel, vision-language models such as CLIP (Radford et al., 2021) bridge visual and linguistic modalities via contrastive training on 400 million image-text pairs, facilitating semantically rich, language-driven perception. Collectively, these models have reshaped 2D visual understanding by delivering robust, generalizable feature representations that transfer effectively across a wide spectrum of downstream applications and tasks.

2.3 The evolution of 2D-to-3D distillation

The achievements of self-supervised 2D representation learning have prompted a fundamental shift in how 3D perception is approached. Given the scarcity of large-scale labelled 3D data, modern research has largely abandoned training 3D networks from scratch, concentrating instead on transferring the dense semantic knowledge embedded in 2D foundation models into the 3D domain. This trajectory began with early cross-modal alignment work. Liu et al. (2021a) identified that reconciling the inherent disparity between sparse 3D geometry and dense 2D pixel representations demands explicit statistical calibration, introducing dimension normalization as a mechanism to harmonize features across modalities. Building upon this foundation, contrastive frameworks such as PPKT (Liu et al., 2021b) and SLidR (Sautier et al., 2022) pioneered what can be described as "pixel-to-point" knowledge transfer, showing that 3D backbones can absorb 2D semantic structure simply through shared observation of the same scene. ScaLR (Puy et al., 2024) subsequently demonstrated that simultaneously expanding dataset diversity and upgrading the 2D backbone capacity, for instance by incorporating DINOv2 (Oquab et al., 2023), produces markedly more robust 3D representations suited to demanding environments such as autonomous driving.

As these representations matured, research attention shifted toward deeper architectural integration. Dino in the Room (DITR) (Knaebel et al., 2026) reflects the current state of the art within this line of work, directly injecting DINOv2 features into the skip-connections of a Point Transformer v3 (PTv3) (Wu et al., 2024) architecture to achieve segmentation performance. Complementary efforts such as OpenScene (Peng et al., 2023) and ConceptFusion (Jatavallabhula et al., 2023) have harnessed the language-aligned embedding space of CLIP to support zero-shot, open-vocabulary comprehension of 3D indoor environments. To prevent models from exploiting superficial geometric cues, methods such as Sonata (Wu et al., 2025) and Concerto (Zhang et al., 2025) have introduced intra-modal 3D self-distillation, establishing that 3D networks perform optimally when geometric self-consistency is balanced against external 2D visual supervision.

2.4 Large-scale bottleneck

Although 2D-to-3D distillation methods have shown significant promise on indoor benchmarks and autonomous driving datasets, their memory demands become prohibitive when applied at urban scale. Recent work has extended CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023) features to outdoor datasets such as SemanticKITTI and nuScenes (Behley et al., 2019; Caesar et al., 2020; Chen et al., 2023; Peng et al., 2023; Wang et al., 2023), producing competitive zero-shot segmentation results. Nevertheless, these datasets consist of sparse LiDAR profiles with tens of thousands of 3D points per scan, a scale that remains tractable in memory even when features are stored at full dimensionality.

Urban-scale photogrammetric or LiDAR point clouds introduce a qualitatively distinct set of challenges. Dense reconstructions of city environments can span vast areas, with datasets reaching billions of points. Just for a scene with 10 millions 3D points, retaining 1024-dimensional DINO descriptors at full floating-point precision (float32) incurs a storage cost of 10^7 points $\times$ 1024 dim $\times$ 4 bytes $=$ 40.9 GB. This yields existing distillation pipelines computationally infeasible on conventional hardware.

2.5 Feature quantization in 3D

Vector quantization has found broad adoption in 3D processing, spanning applications from geometry compression (Huang et al., 2022) to shape generation (Van Den Oord et al., 2017). More recently, VQ-FF (Tang et al., 2025) extended this paradigm by quantizing CLIP features rendered from neural radiance fields, targeting memory-efficient semantic querying within indoor environments. However, these prior approaches share a common limitation: they either operate on geometric rather than semantic

representations, function at the level of individual objects or rely on implicit scene encodings such as NeRFs, a formulation that does not scale to large outdoor point clouds. By contrast, the method proposed in this paper directly quantizes dense 2D foundation model features projected onto explicit 3D point cloud structures, offering a scalable pathway to scene-level understanding that bypasses the need for implicit scene reconstruction entirely.

3. Datasets and metrics

Three complementary datasets are chosen to include a variety of scene scales and data acquisition modalities:

- **ScanNet** (Dai et al., 2017): The complete set of 312 validation scenes, all captured indoors, is used for evaluation. Their relatively modest size allows full per-point features to be retained in memory, which enables a direct quantitative comparison between the compressed and uncompressed representations. Notably, a single shared feature bank is built jointly across all 312 scenes, meaning that only K prototype vectors must collectively encode the visual vocabulary of hundreds of distinct indoor environments.
- **STPLS3D** (Chen et al., 2022): This photogrammetric dataset is based on UAV imagery (4864 × 3648 pixels). Two urban scenes are selected: the Residential Area (RA, approximately 7 million points reconstructed from 1,970 images) and the University of Southern California campus (USC, approximately 29 million points from 4,781 images). Both scenes have ground-truth annotations with eight semantic categories. In the conducted experiments, the grass and dirt classes are combined into a unified category, as their occurrence is heavily skewed across scenes, with some scenes dominated by one class and other largely or entirely absent.
- **Graz** (Farella et al., 2025): This dataset consists of large-scale aerial photogrammetry dataset covering a densely built urban environment. We use a subset comprising 9 high-resolution images (14,144 × 10,560 pixels) and roughly 107.5 million points, a scale at which storing uncompressed per-point features on standard hardware becomes infeasible. This subset therefore acts as a real-world scalability stress test for the proposed framework.

In term of metrics, the proposed method (Section 4) is evaluated along three axes*: (i) Reconstruction quality*, measured by cosine similarity between L2-normalized original and quantized features (Section 5.1), *(ii) Memory efficiency*, reported as storage in MB/GB (Section 5.2) and *(iii) Downstream semantic segmentation,* assessed via per-class IoU and mean IoU (mIoU) under both linear probe and zero-shot settings (Sections 5.3-5.4).

4. Feature quantization framework

The proposed framework takes as input a collection of oriented images $\{I_j\}_{j=1}^{M}$ of a 3D scene alongside its corresponding 3D point cloud $P = \{p_i\}_{i=1}^{N}$. The objective is to assign to each point p_i a semantically rich visual descriptor obtained from a 2D foundation model, while avoiding the need to maintain N full-dimensional feature vectors in memory. This is made possible by the observation that large-scale scenes contain a limited range of distinct semantic patterns, enabling aggressive compression with negligible loss of information. The pipeline (Figure 1) consists of two sequential stages: (1) building a compact feature bank of $K \ll N$ representative embeddings by clustering 2D features extracted across the scene and (2) matching per-pixel features to this bank and lifting the resulting discrete assignments onto the 3D point cloud via known camera parameters.

4.1 Feature extraction and bank construction

Using a pretrained vision encoder ϕ, dense feature maps $\phi(I_j) \in \mathbb{R}^{\mathbb{H}\times\mathbb{W}\times\mathbb{D}}$ are computed for each image I_j. All pixel-wise feature vectors undergo L_2 normalization:

$$\hat{f} = \frac{f}{|f|_2}, \quad f \in \mathbb{R}^{\mathbb{D}} \tag{1}$$

To form the feature bank, mini-batch K-means is run over the normalized features pooled from all scene images. Upon convergence, the K resulting cluster centers are re-normalized to unit norm, yielding the feature bank $\mathcal{B} = \{c_0, \ldots, c_{K-1}\}$. Each centroid in this bank represents a prototypical visual pattern observed across the scene.

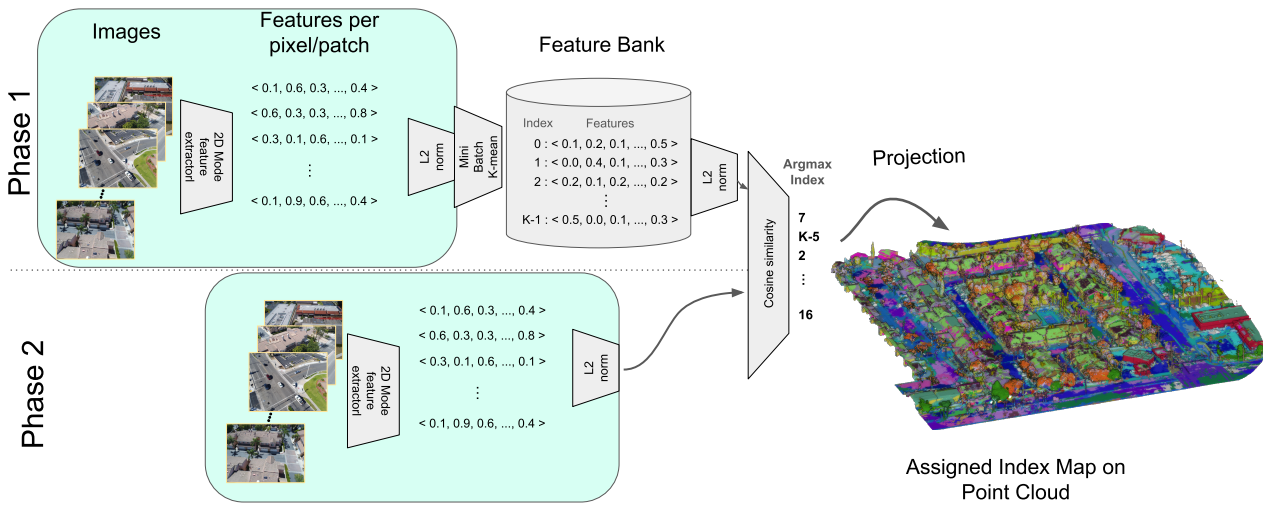

Figure 1. Overview of the proposed pipeline. **Phase 1 - Feature Bank Construction:** Dense per-image features are extracted with a 2D foundation model (e.g., DINOv3), *L2*-normalized and clustered via mini-batch *K*-means to yield *K* centroids forming the feature bank. **Phase 2 - Quantization and Projection:** Patch features are extracted and normalized as in Phase 1, then assigned to their nearest centroid by cosine similarity. The resulting index maps are projected onto the 3D point cloud using the available camera intrinsic and extrinsic parameters.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-9-2026 | © Author(s) 2026. CC BY 4.0 License.

4.2 Feature quantization and multi-view projection

With the feature bank $\mathcal{B}$ in place, all scene images are re-processed through the same encoder ϕ. For each image I_j, dense features are extracted and normalized following the procedure in Section 4.1 (Eq. 1), after which every pixel at location u is assigned to its closest bank entry:

$$\mathrm{idx}(u) = \arg\max_{k \in \{0,\dots,K-1\}} \langle \hat{f}_u, c_k \rangle \quad (2)$$

This yields a discrete index map $\mathcal{I}_j \in \{0, \dots, K-1\}^{H \times W}$ for each image. These index maps are subsequently back-projected onto the 3D point cloud using the known camera intrinsics and extrinsics. Each point p_i visible in a given view receives the index of its corresponding pixel. As most points appear in multiple views, each accumulates a set of index candidates, which are then resolved via majority voting to produce a single, consistent assignment per point.

4.3 Memory complexity

Standard 2D-to-3D feature distillation requires storing an $N \times D$ feature matrix, implying $\mathcal{O}(N \cdot D)$ memory. The proposed representation replaces this with a per-point index array $\mathrm{idx} \in \{0, \dots, K-1\}^N$ together with the shared bank $\mathcal{B} \in \mathbb{R}^{\mathbb{K} \times \mathbb{D}}$, bringing total storage down to $\mathcal{O}(N + K \cdot D)$. Given that $K \ll N$ and each index occupies only $\lceil \log_2 K \rceil$ bits, the resulting memory reduction is substantial (see results in Section 5.2).

4.4 Implementation details

Feature extraction: The pipeline is validated using two 2D foundation models, demonstrating its independence from any specific feature extraction backbone. The principal model employed is DINOv3 (Siméoni et al., 2025), selected for the richness of its visual representations and its established role in recent 2D-to-3D feature distillation frameworks (Knaebel et al., 2026, Puy et al., 2024). More precisely, the ViT-H+/16 variant ($D = 1280$) is adopted for the indoor and UAV acquisition settings (ScanNet, STPLS3D), while the satellite-specific ViT-L/16 variant ($D = 1024$) is applied to the aerial dataset (Graz). The second considered backbone is OpenSeg (Ghiasi et al., 2022) ($D = 768$), which supports zero-shot segmentation by producing dense features aligned with the CLIP embedding space. Across both backbones, the quantization stages, comprising bank construction, index assignment and 2D-to-3D projection, remain unchanged.

ScanNet frames are fed at their original resolution (1296 × 968 pixels). STPLS3D drone imagery is downscaled to 1024 × 768 pixels prior to processing. Graz aerial images, given their larger spatial extent, undergo a tiling strategy using 1024 × 1024 pixels patches with a 25% overlap; only centrally cropped patch features are retained to suppress artifacts near tile boundaries.

For OpenSeg, inputs are either resized or tiled to match its native 640×640 px resolution, after which the resulting features are upsampled back to the original image dimensions to align with the projection grid. Feature banks are built across a range of codebook sizes $K \in \{512, 2048, 8192, 16384\}$ via mini-batch K-means clustering.

2D-to-3D projection: Depth-guided back-projection from posed RGB-D frames is applied for ScanNet. For the other outdoor datasets (STPLS3D and Graz), the voxelization and ray-casting approach introduced in Alami and Remondino (2026) is instead adopted.

Linear probe protocol: Each 3D point is represented by the centroid feature retrieved via a frozen embedding lookup, using its assigned bank index. A single linear classifier is then trained on top of these fixed representations to produce per-point semantic label predictions, incorporating inverse-frequency class weighting and early stopping governed by validation mean intersection over union (mIoU).

Zero-shot segmentation: For zero-shot evaluation, a dedicated feature bank is built from OpenSeg features through the same quantization pipeline. At inference time, each bank centroid is compared against CLIP text embeddings of candidate class names following the evaluation protocol of OpenScene (Peng et al., 2023) and every point inherits the relevance score of its corresponding centroid.

5. Experiments

To validate the proposed framework, experiments are conducted on indoor and outdoor scenes (Section 3), covering four aspects: fidelity of scene reconstruction (Section 5.1), efficiency in memory usage (Section 5.2), performance on downstream segmentation tasks (Section 5.3) and zero-shot segmentation capability (Section 5.4). Results are examined through both quantitative metrics (Section 3) and qualitative observations.

5.1 Reconstruction quality

While the proposed feature quantization method yields notable memory savings (Section 5.2), it may potentially degrade the semantic structure inherent in the original features. To examine this, cosine similarity is computed between the raw DINOv3 features and their quantized equivalents. Table 1 summarizes the results across different datasets and bank sizes. For ScanNet and STPLS3D-RA, where the available memory allows retaining the complete set of 3D features (albeit at a considerable cost for RA), the comparison is carried out directly on the point cloud following projection. For the remaining outdoor datasets (STPLS3D-USC, Graz), memory limitations restrict the comparison to 2D image space, where original features are measured against bank centroids prior to projection. Across all datasets, the semantic structure is consistently well-preserved. Reconstruction fidelity increases with larger values of K, indicating that a moderate number of prototypes is sufficient to capture the bounded semantic diversity present in real-world environments.

Dataset	Comparison	K=512	K=2048	K=8192	K=16384
ScanNet	3D: Centroid vs. Projected	0.67	0.71	0.74	0.76
STPLS3D-RA	3D: Centroid vs. Projected	0.79	0.80	0.80	0.80
STPLS3D-RA	2D: Centroid vs. Extracted	0.80	0.84	0.86	0.87
STPLS3D-USC	2D: Centroid vs. Extracted	0.80	0.83	0.85	0.86
Graz	2D: Centroid vs. Extracted	0.84	0.86	0.87	0.87

Table 1. Feature reconstruction quality evaluated through cosine similarity scores. For ScanNet and RA datasets, assessment is performed in 3D by comparing quantized against full features directly on the point cloud; for all remaining datasets, evaluation is carried out in 2D by comparing quantized against original features in image space. A single unified bank, shared across all 312 ScanNet scenes, is employed throughout. As shown by the results, similarity to the original 2D model features decreases with smaller bank sizes and improves with larger ones, indicating better feature preservation as bank size increases.

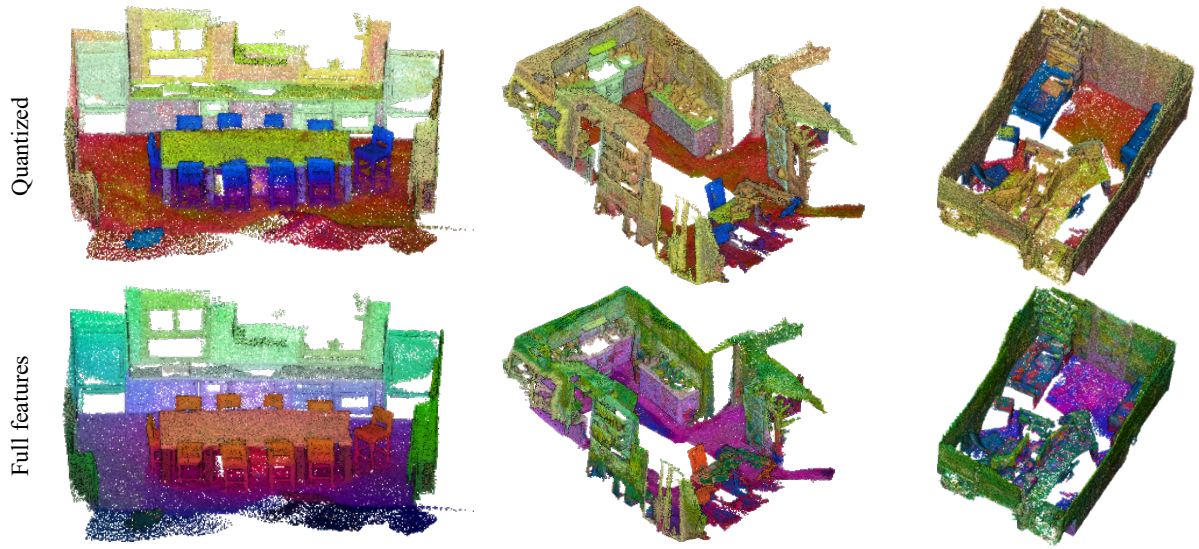

Figure 2. UMAP projections illustrating feature distributions across three indoor ScanNet scenes. *Top*: quantized features retrieved from the bank ($K = 8192$). *Bottom*: uncompressed DINOv3 features in their original form.

The comparatively lower similarity observed on ScanNet is explained by the use of a single shared bank across all 312 scenes, which requires the K centroids to span a considerably broader spectrum of indoor visual appearances, ranging from kitchen and bathroom to office and corridor scenes. Figure 2 and Figure 3 illustrate the UMAP (McInnes et al., 2018) projections of the feature space for indoor and outdoor scenes, respectively. UMAP is a dimensionality reduction method that projects high-dimensional features into a lower-dimensional space while preserving local neighbourhood structure. Here, the features are projected onto 3 components, which are visualized as RGB colours. The quantized bank features closely mirror the cluster organization of the original uncompressed representations, providing visual confirmation that the discrete encoding preserves the underlying semantic layout of the feature space.

Dataset	K=512	K=2048	K=8192	K=16384
STPLS3D-RA	0.95	0.96	0.96	0.96
STPLS3D-USC	0.95	0.96	0.96	0.96
Graz	0.95	0.95	0.96	0.96

Table 2. Reconstruction fidelity of OpenSeg features, assessed with cosine similarity between the original 2D features and their nearest corresponding bank centroids.

Table 2 presents an equivalent analysis for the same datasets using OpenSeg features. Cosine similarity surpasses 0.96 at K=8192 across all scenes, markedly higher than the 0.85–0.87 range recorded for DINOv3 on the corresponding datasets. This discrepancy aligns with the fundamental differences between the two feature spaces: OpenSeg features are CLIP-aligned, meaning their variability is inherently constrained by the set of textual concepts encountered during training. DINO features, by contrast, are learned through self-supervision on purely visual data, enabling them to encode a far richer variety of visual patterns than any text-based vocabulary could encompass, naturally leading to greater feature diversity across identical scenes.

5.2 Memory Efficiency

Table 3 quantifies the memory reduction achieved through the proposed quantization scheme. Drawing on the findings of Table 1, K=8192 is selected for all subsequent experiments, as gains beyond this threshold are marginal while the bank size doubles at each increment. For datasets where storing the full feature set is infeasible, the theoretical memory footprint is reported assuming float32 precision. The quantization delivers over 99.9% memory reduction on outdoor scenes. For the complete ScanNet validation set spanning 312 scenes, storage requirements drop from 127 GB to just 537 MB. For STPLS3D-USC, comprising approximately 29 million points, the theoretical full-feature cost of 150 GB is reduced to roughly 101 MB. Most strikingly, the Graz dataset — with 107.5 million points — would demand over 500 GB for uncompressed features, well beyond the capacity of standard hardware, yet the entire scene is handled using only 257 MB. It is also worth highlighting that conventional dimensionality reduction techniques such as PCA do not address this challenge: compressing feature dimensionality from 1024 to 128 still necessitates storing a per-point vector, amounting to approximately 5.1 GB for 10 million points, two orders of magnitude larger than what the proposed method requires.

Dataset	Points	Full Features	Ours	Reduction
ScanNet (312 scenes)	--	127 GB	537 MB	99.6%
STPLS3D-RA	7 mil	32.4 GB*	55.6 MB	99.8%
STPLS3D-USC	29 mil	150 GB*	100.8 MB	99.9%
Graz	107.5 mil	512.6 GB*	257 MB	99.9%

Table 3. Memory footprint comparison between uncompressed DINOv3 features and their quantized counterpart at K = 8192. For outdoor datasets, full-feature memory is estimated assuming float32 precision, as these are not physically stored. * Theoretical value; actual storage is impractical on commodity hardware.

5.3 Downstream Semantic Segmentation

To confirm that the quantized representation preserves semantically meaningful information suitable for downstream use, a linear probe evaluation is conducted on STPLS3D. Each point is characterized by its corresponding prototype vector retrieved from the bank and a linear classifier is trained on these vectors to perform per-point semantic label prediction. Table 4 reports per-class IoU at K=8192.

For RA, where retaining the full set of 3D features is memory-feasible albeit costly, the uncompressed baseline yields 54.7% mIoU, whereas the quantized counterpart gets 45.8%, a moderate

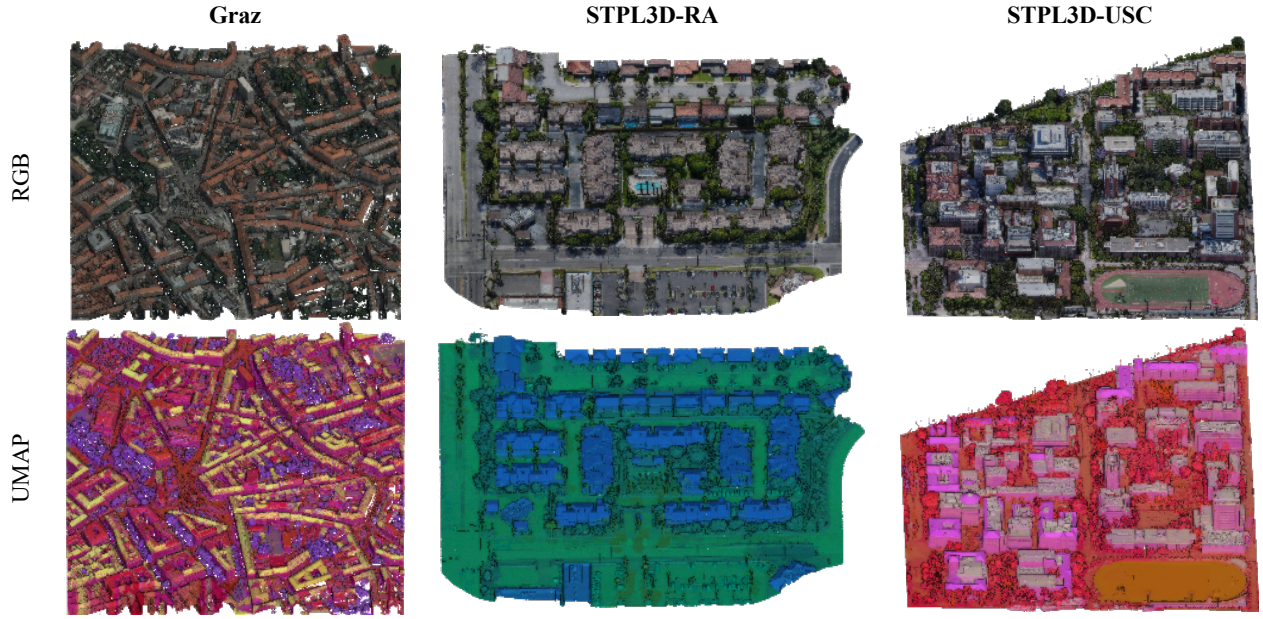

Figure 3. Qualitative results on outdoor scenes. *Top*: Outdoor RGB point clouds. *Bottom*: UMAP projection of quantized features from the computed bank ($K = 8192$). Across each scene, regions of similar semantic contents are consistently highlighted with the same color, validating that the feature bank successfully retains meaningful semantic structure. Note that colors are not consistent across the different scenes.

	Scene	Building	Tree	Street	Grass	Car	Fence	Pole	Clutter	mIoU
Linear prob	RA† (1:5.4)	87.6	71.2	76.6	56.1	58.7	45.8	37.4	4.2	54.7
	RA (1:5.4)	82.2	62.6	69.6	37.9	52.3	37.3	21.0	3.8	45.8
	USC (1:24)	79.2	59.0	40.2	20.1	22.1	5.3	15.1	7.4	31.1
Zero-shot	RA	62.6	53.7	53.7	5.1	20.7	0.5	1.3	-	28.1
	USC	74.6	63.8	31.8	8.2	5.2	1.7	0.6	-	26.6

Table 4. Per-class IoU (%) on STPLS3D at $K = 8192$. *(Top)* Linear probe segmentation results, with train/test split ratios indicated in parentheses. † indicates results without compressions. *(Bottom)* Zero-shot segmentation results using quantized OpenSeg features, with no task-specific training involved. The clutter class is omitted from zero-shot queries as it encompasses miscellaneous objects lacking a meaningful textual description.

decline given the substantial memory savings documented in Table 3. Among quantized configurations, RA reaches competitive mIoU (45.8%). The highly imbalanced train/test splits (reaching up to 1:29) disadvantage underrepresented categories such as fences and poles. Coarse structural elements, including buildings and trees, are reliably identified across configurations, whereas fine-grained categories remain difficult to distinguish, a limitation attributable to the patch-level processing inherent to the DINOv3 backbone.

5.4 Zero-Shot Segmentation

Zero-shot semantic segmentation is assessed using quantized OpenSeg features on STPLS3D, with $K = 8192$ maintained for consistency with preceding experiments. Each semantic class is represented by a text query, which is matched against the bank centroids and all points sharing a given prototype are collectively assigned the corresponding label, following the procedure outlined in Section 4.4. Per-class outcomes are reported in Table 4. In the absence of any labelled supervision, dominant structural categories are reliably detected: buildings surpass 62% IoU across all scenes, while trees exceed 49%. Grass IoU exhibits considerable variation, falling below 10% in RA and USC, where it appears in small, fragmented patches underneath tree canopies. Fine-grained categories such as fences and poles remain difficult to segment, a trend consistent with the linear probe findings in Table 4. Collectively, these results demonstrate that the quantized representation retains adequate semantic structure to support open-vocabulary querying.

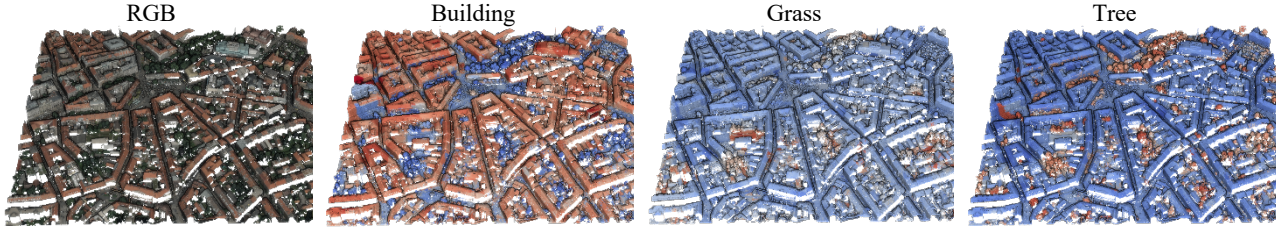

Figure 4. Visual results of the zero-shot segmentation of the Graz dataset. Each figure shows the per-point cosine similarity between the allocated bank centroid and its corresponding query ("building", "trees", "grass"). Higher similarity values are represented by warmer/reddish tones, while cooler/blueish tones reflect lower similarity.

Qualitative zero-shot outcomes on the Graz dataset are visualized in Figure 4. The figure displays the urban point cloud coloured according to the cosine similarity between each point's assigned bank centroid and a specified text query ("building", "trees", "grass"), computed using quantized OpenSeg features ($K = 8192$); warmer colours denoting higher similarity and cooler colours indicating lower values. Despite the large spatial extent of the scene and the aggressive compression applied, the resulting similarity maps remain spatially coherent and align closely with the intended semantic categories.

6. Conclusions

The core contribution of this work addresses the memory scalability challenge that has long constrained the application of 2D foundation model representations to large-scale 3D point clouds. Taking advantage of the natural semantic redundancy present in large scenes, the presented approach distils 2D foundation model representations into a compact set of K prototype vectors; then it maps each 3D point to a single integer index corresponding to its closest prototype, entirely eliminating the need to store dense per-point feature vectors. The resulting reduction in memory footprint exceeds 99%, enabling the processing of scenes with over 100 million points on commodity hardware. Evaluations on both indoor and large-scale outdoor datasets demonstrate that the semantically essential information is well preserved through quantization, with the compressed representations remaining competitive on downstream tasks, including semantic segmentation. While there remains room to explore more advanced clustering mechanisms, the framework establishes a strong and practical foundation for transferring and compressing 2D foundation model knowledge into large-scale 3D environments. Taken together, the findings demonstrate that high-density foundation model features can be efficiently compressed and deployed on widely available consumer hardware, opening-up a viable path toward scalable and accessible 3D scene understanding. For practical deployments, the proposed method enables to train 3D models such as Point Transformers directly on distilled 2D features alongside spatial information or extending recent 3D approaches which are typically limited to small indoor scenes, to much larger outdoor environments. For instance, methods such as OpenScene could be trained in a zero-shot setting similar to ours, using the proposed quantized features as a memory-efficient alternative to dense per-point embeddings. As with any quantization method, some information loss is unavoidable and in our case this can be more noticeable for fine-grained and rare concepts. When necessary, users can manually add such concept features to the bank to preserve them.

Directions for future investigation include coupling quantized representations with 3D backbone networks for large-scale supervised learning, developing scene-complexity-aware strategies for dynamically sizing the feature bank and broadening the framework's applicability to sequential and temporal point cloud data.

References

Alami, A., Remondino, F., 2026. Open-Vocabulary Segmentation of Aerial Point Clouds. *Remote Sensing,* 18(4).

Bayrak, O. C., Remondino, F., and Uzar, M., 2023. A new dataset and methodology for urban-scale 3D point cloud classification. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-1/W3-2023, pp. 1–8.

Bayrak, O. Ma, Z., Farella, E.M., Remondino, F., Uzar, M., 2024. ESTATE: A large dataset of under-represented urban objects for 3D point cloud classification. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci*., XLVIII-2-2024, 25-32.

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. *Proc. ICCV*.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving. *Proc. CVPR*, pp. 11621-11631.

Caron, M., Touvron, H., Misra, I., J´egou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proc. CVPR*, pp. 9650-9660.

Chen, M., Hu, Q., Yu, Z., THOMAS, H., Feng, A., Hou, Y., McCullough, K., Ren, F., Soibelman, L., 2022. Stpls3D: A large-scale synthetic and real aerial photogrammetry 3D point cloud dataset. *Proc. BMCV,* BMVA Press.

Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W., 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. *Proc. CVPR,* pp. 7020-7030.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. *Int. conference on machine learning,* PmLR, 1597-1607.

Choy, C., Gwak, J., Savarese, S., 2019. 4D spatio-temporal convnets: Minkowski convolutional neural networks. *Proc. CVPR*, pp. 3075-3084.

Czerniawski, T., Sankaran, B., Nahangi, M., Haas, C., Leite, F., 2018. 6D DBSCAN-based segmentation of building point clouds for planar object classification. *Automation in Construction*, 88, 44-58.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *Proc. CVPR*, pp. 5828-5839.

Farella, E. M., Morelli, L., Remondino, F., Qin, R., Schachinger, B., Legat, K. et al., 2025. Investigating the New Ultracam Dragon Hybrid Aerial Mapping System. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 48, 117–125.

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? The Kitti vision benchmark suite. *Proc. CVPR*, pp. 3354–3361.

Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y., 2022. Scaling open-vocabulary image segmentation with image-level labels. *Proc. ECCV,* Springer, pp. 540–557.

Grilli, E., Remondino, F., 2020. Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, 9(6), 379.

He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *Proc. CVPR*, pp. 16000-16009.

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. *Proc. CVPR*, pp. 9729-9738.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2022. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2), 316-343.

Huang, T., Zhang, J., Chen, J., Ding, Z., Tai, Y., Zhang, Z., Wang, C., Liu, Y., 2022. 3QNet: 3D point cloud geometry quantization compression network. *ACM Transactions on Graphics (TOG)*, 41(6), 1-13.

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

Knaebel, K., Yilmaz, K., de Geus, D., Hermans, A., Adrian, D., Linder, T., Leibe, B., 2026. DINO in the room: Leveraging 2D foundation models for 3D segmentation. *Proc. 3DV*.

Lai, X., Liu, J., Jiang, L., Wang, L., Zhao, H., Liu, S., Qi, X., Jia, J., 2022. Stratified transformer for 3d point cloud segmentation. *Proc. CVPR*, pp. 8500-8509.

Liu, Z., Qi, X., Fu, C.-W., 2021a. 3D-to-2D distillation for indoor scene parsing. *Proc. CVPR*, pp. 4464-4474.

Liu, Y.-C., Huang, Y.-K., Chiang, H.-Y., Su, H.-T., Liu, Z.-Y., Chen, C.-T., Tseng, C.-Y., Hsu, W. H., 2021b. Learning from 2D: Contrastive pixel-to-point knowledge transfer for 3D pre-training. *arXiv preprint arXiv*:2104.04687.

Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z., 2023. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36, 37193-37229.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 3(29), 861.

Ni, H., Lin, X., Zhang, J., 2017. Classification of ALS point cloud with improved point cloud segmentation and random forests. *Remote Sensing*, 9(3), 288.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. et al., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv*:2304.07193.

Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T. et al., 2023. Openscene: 3D scene understanding with open vocabularies. *Proc. CVPR*, pp. 815-824.

Puy, G., Gidaris, S., Boulch, A., Sim´eoni, O., Sautier, C., P´erez, P., Bursuc, A., Marlet, R., 2024. Three pillars improving vision foundation model distillation for lidar. *Proc. CVPR*.

Qi, C. R., Su, H., Mo, K., Guibas, L. J., 2017a. Pointnet: Deep learning on point sets for 3D classification and segmentation. *Proc. CVPR*, pp. 652-660.

Qi, C. R., Yi, L., Su, H., Guibas, L. J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al., 2021. Learning transferable visual models from natural language supervision. *ICML*, 8748-8763.

Rozenberszki, D., Litany, O., Dai, A., 2022. Language- grounded indoor 3d semantic segmentation in the wild. *Proc. ECCV*, Springer, 125-141.

Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R., 2022. Image-to-lidar self-supervised distillation for autonomous driving data. *Proc. CVPR*, 9891-9901.

Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B., 2022. Mask3d: Mask transformer for 3D semantic instance segmentation. Proc. *ICRA*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., et al., 2025. DINOv3. *arXiv:2508.10104.*

Takmaz, A., Fedele, E., Sumner, R. W., Pollefeys, M., Tombari, F., Engelmann, F., 2023. Openmask3d: Open-vocabulary 3D instance segmentation. Proc. *NIPS,* pp. 68367-68390.

Tang, G., Agarwal, A., Han, W., Darrell, T., Bai, Y., 2025. Vector Quantized feature fields for fast 3D semantic lifting. *arXiv e-prints arXiv*:2503.

Thomas, H., Qi, C. R., Deschaud, J.-E., Marcotegui, B., Goulette, F., Guibas, L. J., 2019. Kpconv: Flexible and deformable convolution for point clouds. *Proc. ICCV*, pp. 6411-6420.

Van Den Oord, A., Vinyals, O. et al., 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, P.-S., 2023. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42(4),1-11.

Wang, Y., Huang, S., Gao, Y., Wang, Z., Wang, R., Sheng, K., Zhang, B., Liu, S., 2023. Transferring clip's knowledge into zero-shot point cloud semantic segmentation. *Proc. 31st ACM International Conference on Multimedia*, 3745-3754.

Wu, X., DeTone, D., Frost, D., Shen, T., Xie, C., Yang, N., Engel, J., Newcombe, R., Zhao, H., Straub, J., 2025. Sonata: Self-supervised learning of reliable point representations. *Proc. CVPR*.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point transformer v3: Simpler faster stronger. *Proc. CVPR*, pp. 4840-4851.

Yang, Y.-Q., Guo, Y.-X., Xiong, J.-Y., Liu, Y., Pan, H., Wang, P.-S., Tong, X., Guo, B., 2025. Swin3D: A pretrained transformer backbone for 3d indoor scene understanding. *Computational Visual Media*, 11(1), 83-101.

Zhang, Y., Wu, X., Lao, Y., Wang, C., Tian, Z., Wang, N., Zhao, H., 2025. Concerto: Joint 2D-3D self-supervised learning emerges spatial representations. *Proc. NeurIPS*.