DyG-Seg: Unsupervised 3D Point Cloud Segmentation via Geometric Manifold Rectification

Benyu Wu[1], Kun Zhou[2*], Xulun Ye[1*]
[1]Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China
[2]School of Architecture & Urban Planning, Shenzhen University, Shenzhen, China
wubenyu02@gmail.com, zhoukun@szu.edu.cn, yexulun@nbu.edu.cn

Abstract

Unsupervised 3D semantic segmentation is vital for label-free open-world perception. However, current methods typically struggle with two core limitations: fixed category assumptions that restrict adaptability in complex scenes, and geometric incompleteness caused by irregular point sampling, which degrade the integrity of local geometric descriptors for fine-grained or long-tail objects. To address these, we propose DyG-Seg, a Dynamic Geometric-Semantic Alignment framework. First, targeting the challenge of incomplete geometry, the Geometric Manifold Rectification (GMR) leverages a diffusion model to reconstruct integral object geometry and recover the underlying manifold topology, thereby enhancing feature discriminability. Utilizing this geometric prior, the Dynamic Prototype Contrastive Clustering (DPCC) applies non-parametric Bayesian inference to automatically estimate the optimal number of clusters and generate pseudo-labels. Finally, our Iterative Optimization with Dynamic Constraints integrates Dynamic Class Balancing (DCB) to mitigate long-tail bias. Capitalizing on the recovered manifold topology, we enforce Spatial Geometric Consistency to rectify structural incompleteness and ensure spatially coherent semantic predictions. Extensive experiments on ScanNet and S3DIS demonstrate that DyG-Seg discovers latent semantic structures and significantly outperforms state-of-the-art unsupervised methods.

1 Introduction

Semantic segmentation of 3D point clouds serves as a cornerstone capability for autonomous driving, embodied AI, and robotic navigation [Dai *et al.*, 2017; Armeni *et al.*, 2017]. While fully supervised paradigms have yielded remarkable progress on specific benchmarks [Qi *et al.*, 2017a; Qi *et al.*, 2017b; Graham *et al.*, 2018], their heavy reliance on extensive point-wise human annotations acts as a severe bottleneck to scalability in open-world scenarios. To alleviate the dependency on expensive labeling, unsupervised 3D segmentation has emerged as a compelling research direction [Zhang *et al.*, 2023; Chen *et al.*, 2023b; Zhang *et al.*, 2025]. However, prevalent methodologies typically operate under a strong assumption: the number of semantic categories K in a scene is a priori known and static. In realistic open environments, scene complexity and object variety are inherently dynamic. Prescriptively fixing the cluster number is not only impractical but often induces semantic ambiguity, leading to erroneous merging or over-segmentation. Consequently, this paper addresses a more challenging yet practical task: achieving automatic semantic discovery and segmentation of 3D scenes without pre-specifying the category count K and devoid of human supervision.

Despite recent strides in leveraging self-supervised features or multi-modal distillation [Caron *et al.*, 2021; Oquab *et al.*, 2023; Kirillov *et al.*, 2023; Peng *et al.*, 2023], substantial hurdles remain when applying existing unsupervised methods to complex indoor and outdoor environments. The foremost bottleneck stems from the rigidity of a fixed K. Most approaches depend on standard clustering algorithms (e.g., K-Means) that force data into a predetermined number of partitions [Caron *et al.*, 2018; Ji *et al.*, 2019; Zhang *et al.*, 2023]. When the preset K mismatches the true semantic structure, the model fails to adapt, causing distinct objects to fuse or single entities to fragment. The second critical limitation is the geometric incompleteness inherent in raw point clouds. Unlike dense and regular 2D images, 3D scans often suffer from occlusion, sensor noise, and varying density, resulting in fragmented object structures. Conventional feature extractors struggle to infer the integral geometric manifold from these incomplete observations, leading to indistinct feature representations that are easily overwhelmed by dominant background patterns. Fundamentally, the inability to dynamically adapt to unknown cluster counts and the failure to rectify incomplete geometric structures represent two pivotal constraints restricting current unsupervised performance.

To bridge these gaps, we propose DyG-Seg, a novel unsupervised framework designed to achieve robust scene segmentation via geometric-semantic alignment. Unlike traditional methods that perform static clustering directly on raw features, we reformulate the task as a progressive process of geometric recovery, structural inference, and semantic learn-

*Corresponding authors.

ing. First, to address the issue of geometric incompleteness, we introduce the Geometric Manifold Rectification (GMR) module. This module leverages a generative reconstruction task to recover latent manifold structures from incomplete inputs, thereby enhancing the integrity of geometric feature representations. Second, to tackle the unknown K problem, we propose the Dynamic Prototype Contrastive Clustering (DPCC) strategy. By integrating non-parametric Bayesian inference [Ferguson, 1973] with contrastive learning, DPCC dynamically infers the optimal number of clusters and learns compact class prototypes from the data distribution. Finally, we implement an Iterative Optimization with Dynamic Constraints mechanism, which cyclically refines pseudo-labels and feature embeddings to ensure robust convergence in complex scenes.

Our main contributions are summarized as follows:

- We propose DyG-Seg, a unified unsupervised framework capable of automatically discovering the number of categories and performing accurate segmentation without human labels.
- We design the Geometric Manifold Rectification (GMR) module, which effectively recovers latent manifolds from incomplete point clouds, significantly improving the geometric completeness of learned features.
- We introduce the Dynamic Prototype Contrastive Clustering (DPCC) strategy integrated with an Iterative Optimization with Dynamic Constraints mechanism. This synergy allows the model to adaptively discover semantic categories while enforcing spatial consistency and class balance.
- Extensive experiments on the challenging ScanNet and S3DIS datasets demonstrate that our method significantly outperforms existing state-of-the-art unsupervised approaches.

2 Related Work

2.1 Unsupervised 3D Semantic Segmentation

Early approaches primarily relied on fully supervised architectures like point-based MLPs [Hu *et al.*, 2020; Qi *et al.*, 2017a; Qi *et al.*, 2017b] or sparse convolutions [Graham *et al.*, 2018; Zhu *et al.*, 2021]. Despite their effectiveness, prohibitive annotation costs accelerated the shift towards weakly supervised paradigms. While methods utilizing sparse annotations [Hu *et al.*, 2022; Liu *et al.*, 2021; Unal *et al.*, 2022] or 2D tags [Chibane *et al.*, 2022; Wei *et al.*, 2020] mitigate labeling efforts, they still necessitate human intervention. Recently, fully unsupervised segmentation has advanced significantly. For instance, GrowSP [Zhang *et al.*, 2023] proposes progressive superpoint growing, PointDC [Chen *et al.*, 2023b] leverages self-supervised distillation with soft clustering, and LogoSP [Zhang *et al.*, 2025] utilizes global graph grouping in the frequency domain. However, these methods typically operate under the assumption of a fixed category count. Moreover, they cluster raw sparse features directly, neglecting the geometric manifold discontinuity inherent in point clouds, which degrades performance on fine-grained or long-tail categories.

2.2 Leveraging 2D Priors for 3D Segmentation

Distilling knowledge from 2D foundation models, such as SAM [Kirillov *et al.*, 2023; Ravi *et al.*, 2024] and CLIP [Radford *et al.*, 2021], has emerged as a dominant paradigm. By projecting rich semantic priors onto 3D points, numerous approaches have achieved impressive results in open-vocabulary scene understanding [Chen *et al.*, 2023a; Guo *et al.*, 2024; Kim *et al.*, 2024; Li *et al.*, 2024; Liu *et al.*, 2023; Peng *et al.*, 2024; Xiao *et al.*, 2024; Yin *et al.*, 2024]. For instance, OpenScene [Peng *et al.*, 2023] and PLA [Ding *et al.*, 2023] demonstrate effective zero-shot transfer by aligning 3D features with 2D embeddings. However, their reliance on external human supervision (text or image labels) violates the strictly unsupervised premise. Consequently, recent research has pivoted towards distilling *self-supervised* features, notably from DINO [Caron *et al.*, 2021] and DINOv2 [Oquab *et al.*, 2023]. While providing valuable semantic initialization, direct distillation often overlooks the domain gap between dense images and sparse point clouds. Existing works like PointDC utilize these features but struggle to maintain geometric consistency. We argue that 2D priors are beneficial but must be complemented by robust 3D geometric regularization to align semantic features with the underlying manifold.

2.3 Clustering and Category Discovery

Traditional deep clustering methods, such as DeepCluster [Caron *et al.*, 2018] and IIC [Ji *et al.*, 2019], typically rely on K-Means with a pre-specified cluster count K. This rigidity remains a critical bottleneck in 3D segmentation. While recent contrastive approaches [Cho *et al.*, 2021; Xie *et al.*, 2020] enhance feature discriminability, they generally suffer from the same limitation due to the use of static prototypes. Attempts using over-clustering strategies often require manual upper bounds and struggle to converge to true semantic entities. In contrast, Non-parametric Bayesian methods [Ferguson, 1973], particularly the Dirichlet Process Mixture Model (DPMM), offer a theoretical framework for dynamic cluster inference. However, the efficient integration of DPMM into end-to-end deep learning for large-scale 3D semantic discovery remains an underexplored frontier.

2.4 Diffusion Models for Geometric Representation Learning

While diffusion models have achieved remarkable success in 3D shape generation and completion [Ho *et al.*, 2020; Zhou *et al.*, 2021], most approaches prioritize generation fidelity over feature discriminability. Conversely, recent 2D studies [Baranchuk *et al.*, 2021] reveal that the intermediate representations within the denoising process encapsulate rich semantic information. Inspired by this, we repurpose diffusion from a generative tool to a *self-supervised pretext task* for geometric rectification. Instead of synthesizing new objects, we leverage the denoising objective to force the backbone to recover latent manifolds from sparse inputs. This strategy explicitly enhances the model's sensitivity to local geometric structures, effectively mitigating the geometric incompleteness inherent in raw point clouds.

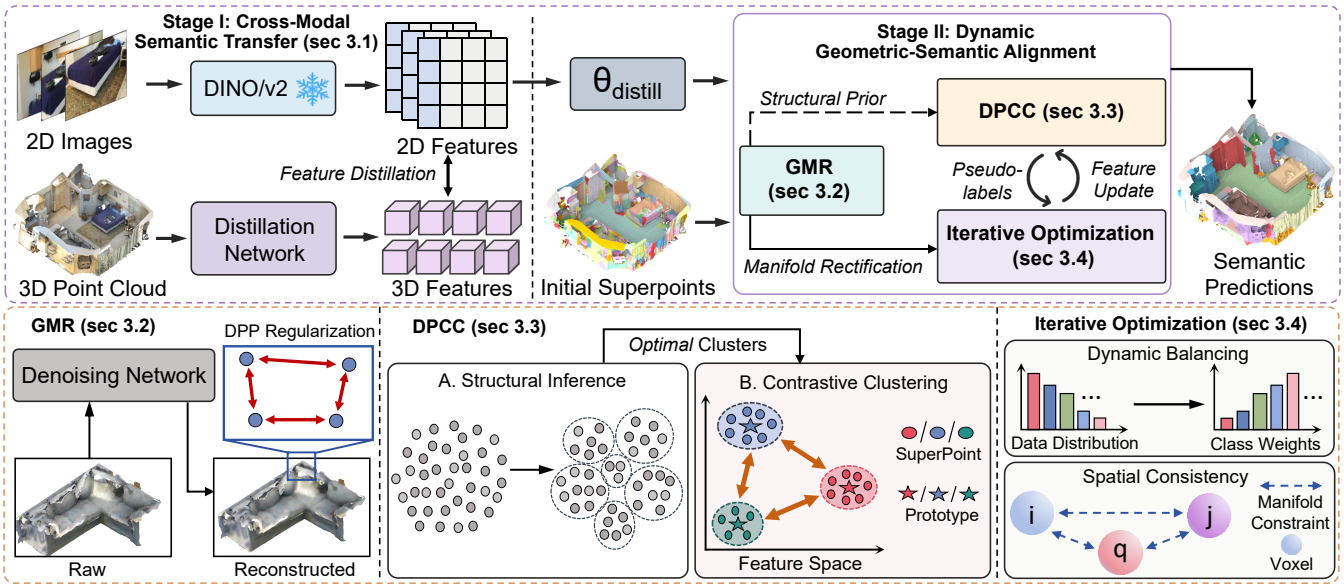

Figure 1. Overview of the proposed DyG-Seg framework. The top row illustrates the two-stage pipeline, progressing from cross-modal semantic initialization to dynamic geometric-semantic alignment. The bottom row details the three core components: Geometric Manifold Rectification (GMR) leverages a denoising network to recover latent structural topology; Dynamic Prototype Contrastive Clustering (DPCC) integrates structural inference with prototype learning for category discovery; and Iterative Optimization enforces dynamic class balancing and spatial consistency constraints.

3 Methodology

We present DyG-Seg, a framework designed to address the challenges of unsupervised point cloud segmentation by seamlessly integrating geometry, structure, and semantics. The pipeline begins with a 2D-to-3D distillation module that transfers semantic knowledge from foundation models [Caron *et al.*, 2021; Oquab *et al.*, 2023; Kirillov *et al.*, 2023] to initialize superpoint features [Chen *et al.*, 2023b; Zhang *et al.*, 2025]. To mitigate geometric sparsity, the Geometric Manifold Rectification (GMR) module serves as a structural prior, refining the latent manifold of the input data. Subsequently, the core Dynamic Prototype Contrastive Clustering (DPCC) module adaptively infers the optimal cluster number and learns discriminative prototypes. Finally, the segmentation network is optimized through dynamic class balancing and spatial consistency constraints to ensure robust predictions.

3.1 Cross-Modal Semantic Transfer

To leverage the rich semantic priors encapsulated in large-scale vision foundation models, we implement a multi-view 2D-to-3D distillation strategy. Following established paradigms [Chen *et al.*, 2023b; Peng *et al.*, 2023], this process distills pixel-level semantics to the 3D domain through a projection-then-regression pipeline.

Feature Projection. Given a point cloud $\mathcal{P} = \{p_i\}_{i=1}^{N}$ and its associated multi-view images $\mathcal{I} = \{I_k\}_{k=1}^{K}$, we first extract pixel-level features using a frozen 2D encoder Φ_{2D} (e.g., DINOv2 [Oquab *et al.*, 2023]). By leveraging camera parameters and depth maps, these 2D features are unprojected into 3D space. For points visible across multiple views, we average their projected features to obtain the target semantic embedding $\mathbf{f}_i^{\text{target}}$. Points lacking valid projections are excluded from the supervision signal.

3D Distillation. We employ a learnable 3D backbone Φ_{3D} (e.g., SparseConv [Graham *et al.*, 2018]), which takes the raw point cloud as input to regress these projected 2D features. The network predicts 3D features $\mathbf{f}_i \in \mathbb{R}^C$ supervised by maximizing the cosine similarity with the target embeddings:

$$\mathcal{L}_{\text{distill}} = \frac{1}{N_v} \sum_{i=1}^{N_v} \left(1 - \frac{\mathbf{f}_i \cdot \mathbf{f}_i^{\text{target}}}{\|\mathbf{f}_i\|_2 \|\mathbf{f}_i^{\text{target}}\|_2} \right), \tag{1}$$

where N_v denotes the number of points with valid 2D projections.

Superpoint Partitioning. Upon convergence, the frozen distillation network Φ_{3D} generates geometry-aware semantic features for the entire dataset. To enforce local consistency and reduce computational complexity for the subsequent stage, we partition the point cloud into disjoint superpoints $\mathcal{S} = \{s_j\}_{j=1}^{M}$ using an unsupervised region growing algorithm based on geometric curvature and color similarity. Finally, we compute the unified embedding $\mathbf{z}_j$ for each superpoint s_j by average pooling the distilled features of its constituent points:

$$\mathbf{z}_j = \frac{1}{|\mathcal{N}_j|} \sum_{p_k \in \mathcal{N}_j} \Phi_{\text{3D}}(p_k), \tag{2}$$

where $\mathcal{N}_j$ represents the index set of points in s_j. The resulting matrix $\mathbf{Z} \in \mathbb{R}^{M \times C}$ provides the initialized geometric-semantic representation for the subsequent stage.

3.2 Geometric Manifold Rectification

In unsupervised settings, sparse and noisy 3D acquisitions often degrade the extraction of consistent semantic features [Chen *et al.*, 2021; Wang *et al.*, 2021]. To address this, we introduce the Geometric Manifold Rectification (GMR) module. Unlike standard denoising approaches [Pang *et al.*, 2023] that merely maximize the likelihood of observed coordinates, GMR operates as a *structure-guided generative process*. We formulate the rectification as learning a Density-Rectified Vector Field, where the denoising trajectory is actively modulated by a repulsive topological prior to ensuring both geometric fidelity and structural uniformity.

Unified Density-Aware Generative Objective. We treat the geometric recovery not as two separate tasks, but as optimizing a unified target potential. Let $\mathbf{p}_t$ be the noisy coordinates at timestep t, and $\Psi_\theta(\mathbf{p}_t, t)$ be the denoising network predicting the clean manifold geometry $\hat{\mathbf{p}}_0$. Standard diffusion models [Ho *et al.*, 2020] approximate the score function $\nabla_{\mathbf{p}} \log q(\mathbf{p})$ to map noise back to data. However, this solely captures spatial location, ignoring the density distribution on the manifold. To unify geometry and topology, we define a Target Potential $\mathcal{E}(\hat{\mathbf{p}}_0)$ that integrates geometric reconstruction with structural constraints:

$$\mathcal{E}(\hat{\mathbf{p}}_0) = \underbrace{\|\hat{\mathbf{p}}_0 - \mathbf{p}_{\text{GT}}\|^2}_{\text{Geometric Fidelity}} + \lambda_{dpp} \underbrace{(-\log\det(\mathbf{L}_{\hat{\mathbf{p}}_0} + \epsilon\mathbf{I}))}_{\text{Repulsive Potential}}, \quad (3)$$

where $\mathbf{p}_{\text{GT}}$ denotes the underlying clean surface (approximated by the reconstruction target). The second term is derived from the Determinantal Point Process (DPP) [Kulesza and Taskar, 2012], where $\mathbf{L}_{\hat{\mathbf{p}}_0}$ is the Gaussian similarity kernel of the predicted points. Geometrically, minimizing this potential induces a Repulsive Gradient Field $\nabla_{\hat{\mathbf{p}}}\mathcal{L}_{\text{DPP}}$ that counteracts the tendency of diffusion models to collapse points into high-density modes.

Gradient-Guided Optimization. During training, we compel the network to minimize the expected energy of the rectified coordinates. The unified optimization objective is formulated as:

$$\mathcal{L}_{\text{geo}} = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[\mathcal{E}(\Psi_\theta(\mathbf{p}_t, t))\right]. \quad (4)$$

Remark: This formulation implicitly performs *structural guidance* during training. Through back-propagation, the gradients from the Repulsive Potential flow through the denoising network Ψ_θ, updating the backbone parameters. This compels the network to learn a rectified vector field that pushes points towards the manifold surface while simultaneously spreading them uniformly, effectively creating a "cleaner" geometric support for the subsequent segmentation task.

3.3 Dynamic Prototype Contrastive Clustering

Following the geometric rectification, the framework proceeds to semantic discovery. A critical limitation in existing unsupervised segmentation methods is the reliance on a pre-specified number of clusters K [Caron *et al.*, 2018; Ji *et al.*, 2019; Zhang *et al.*, 2023]. In open-world scenarios, the semantic structure is typically unknown and varies across scenes. To overcome this, we propose Dynamic Prototype Contrastive Clustering (DPCC), a unified non-parametric framework that orchestrates semantic structure discovery and representation learning via an iterative Expectation-Maximization (EM) paradigm. DPCC treats the superpoint embeddings $\mathbf{Z}$ as observations and models the semantic discovery as a Bayesian inference process. We maintain a dynamic set of prototypes whose existence probability is governed by a Dirichlet Process, while their feature distribution is optimized to guide the backbone training.

E-Step: Bayesian Structural Inference. In the inference step, we formulate the discovery of semantic categories as a Variational Inference (VI) problem within a Dirichlet Process Mixture Model (DPMM) [Ferguson, 1973]. To model the potentially infinite number of clusters, we adopt the Stick-Breaking construction [Sethuraman, 1994]. Let $\mathbf{V} = \{v_k\}_{k=1}^{K_{\max}}$ be a set of latent variables drawn from a Beta distribution $v_k \sim \text{Beta}(1, \alpha)$, where α controls the sparsity. The existence probability π_k for the k-th prototype is defined as:

$$\pi_k = v_k \prod_{l=1}^{k-1}(1 - v_l). \quad (5)$$

To infer the optimal structure (pseudo-labels $\hat{\mathbf{y}}$), we approximate the posterior by maximizing the Evidence Lower Bound (ELBO) [Blei and Jordan, 2006]:

$$\mathcal{L}_{\text{ELBO}} = \underbrace{\mathbb{E}_q[\log p(\mathbf{Z}|\hat{\mathbf{y}})]}_{\text{Data Fitting}} - \underbrace{\text{KL}(q\|p)}_{\text{Complexity Penalty}}. \quad (6)$$

This objective creates a dynamic competition: the *Data Fitting Term* encourages activating more prototypes to capture fine-grained details, while the *Complexity Penalty Term* imposes a sparsity constraint. Through this optimization, the model automatically eliminates redundant prototypes, converging to the optimal cluster number K_{opt} and generating high-quality pseudo-labels $\hat{\mathbf{y}}$ for the subsequent training phase.

M-Step: Discriminative Representation Learning. In the learning step, guided by the structurally inferred priors, we employ Prototype Contrastive Clustering [Caron *et al.*, 2021; Xie *et al.*, 2020] to shape the semantic space. We define the set of active prototypes as $\mathbf{C} = \{\mathbf{c}_k\}_{k=1}^{K_{opt}}$. These prototypes serve as the anchors for metric learning. For each superpoint feature $\mathbf{z}_j$, the assignment probability to the k-th prototype is modeled as a softmax distribution:

$$P(y_j = k|\mathbf{z}_j) = \frac{\exp(\mathbf{z}_j \cdot \mathbf{c}_k/\tau)}{\sum_{l=1}^{K_{opt}} \exp(\mathbf{z}_j \cdot \mathbf{c}_l/\tau)}, \quad (7)$$

where τ is a temperature hyperparameter. The backbone network is then supervised using the generated pseudo-labels $\hat{y}_j$ from the E-step. By minimizing the cross-entropy loss between the predicted distribution and the pseudo-labels, this contrastive objective enforces intra-class compactness and inter-class separability, ensuring that the dynamically discovered clusters are semantically distinct.

In summary, DPCC bridges structural inference and representation learning: the Bayesian prior determines the *cardinality* of the prototypes, while the contrastive objective governs the *semantic content* these prototypes represent, yielding

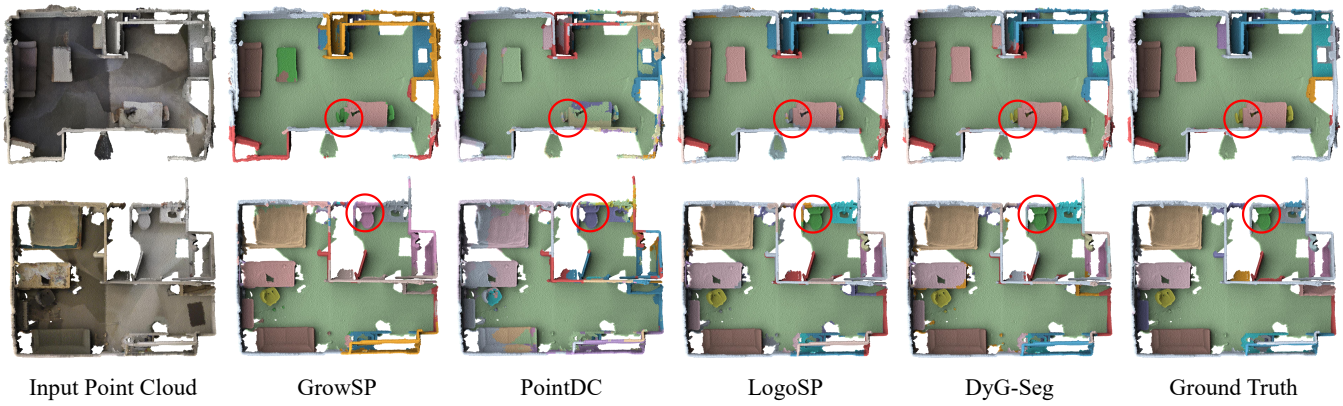

Figure 2. Qualitative results on ScanNet dataset. Red circles highlight the differences.

robust pseudo-labels to supervise the iterative optimization in the subsequent stage.

3.4 Iterative Optimization with Dynamic Constraints

In the M-Step of our framework, the backbone network is updated using the pseudo-labels $\hat{\mathbf{y}}$ generated by DPCC. To address the inherent challenges of unsupervised point cloud segmentation—specifically long-tail class distributions and spatial discontinuity—we incorporate dynamic constraints into the optimization landscape.

Dynamic Class Balancing. Real-world 3D scenes exhibit severe class imbalance, where dominant categories (e.g., walls, floors) can overshadow rare objects during training. Directly supervising with raw pseudo-labels risks model collapse into trivial solutions [Caron *et al.*, 2018; Ji *et al.*, 2019]. To mitigate this, we implement an online frequency estimation mechanism (illustrated as "Dynamic Balancing" in Fig. 1). We maintain a global class counter $\mathbf{v} \in \mathbb{R}^{K_{opt}}$ updated via Exponential Moving Average (EMA) at each iteration t:

$$\mathbf{v}^{(t)} = m \cdot \mathbf{v}^{(t-1)} + (1-m) \cdot \mathbf{v}_{\text{batch}}^{(t)}, \tag{8}$$

where m is the momentum coefficient. We then derive an inverted re-weighting vector $\mathbf{w}$ to modulate the loss contribution of each class k:

$$w_k = \frac{1}{\ln(\lambda + v_k^{(t)})}, \quad \tilde{\mathbf{w}} = \text{Normalize}(\mathbf{w}). \tag{9}$$

This ensures that tail classes with lower frequencies receive larger gradients, effectively flattening the optimization landscape.

Spatial Geometric Consistency. To enforce the spatial smoothness of predictions, we construct a spatial affinity graph where edges $\mathcal{E}$ connect adjacent superpoints. Crucially, this adjacency is defined on the *rectified manifold* recovered by GMR (Sec. 3.2), ensuring that topological constraints are applied to geometrically valid neighbors rather than noise. We minimize the Kullback-Leibler (KL) divergence between the predictive distributions of neighboring superpoints:

$$\mathcal{L}_{\text{spatial}} = \sum_{(i,j)\in\mathcal{E}} \text{KL}(P(\cdot|\mathbf{z}_i) \| P(\cdot|\mathbf{z}_j)). \tag{10}$$

This term propagates high-confidence predictions to spatially connected regions, smoothing out isolated semantic noise.

Unified Objective and Iterative Refinement. The DyG-Seg framework unifies geometric recovery and semantic learning into a single optimization objective. The total loss $\mathcal{L}_{\text{total}}$ combines the segmentation-specific task loss with the geometric rectification objective defined in Eq. (4):

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{ce}}(\hat{\mathbf{y}}, \tilde{\mathbf{w}}) + \lambda_{\text{s}}\mathcal{L}_{\text{spatial}}}_{\text{Segmentation Task } (\mathcal{L}_{\text{seg}})} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}, \tag{11}$$

where $\mathcal{L}_{\text{ce}}$ denotes the weighted cross-entropy loss against pseudo-labels. This joint optimization creates a positive feedback loop (depicted as "Feature Update" in Fig. 1): $\mathcal{L}_{\text{geo}}$ provides a cleaner geometric support for feature extraction [Chen *et al.*, 2021], which in turn improves the clustering quality in DPCC; simultaneously, $\mathcal{L}_{\text{seg}}$ injects semantic discriminability back into the backbone. Through this iterative self-training, the framework progressively aligns geometric structures with semantic categories, converging to robust segmentation performance.

4 Experiments

4.1 Experimental Setup

Datasets. To validate the effectiveness of our proposed framework, we conduct extensive experiments on two large-scale public indoor 3D benchmarks. ScanNetV2 consists of 1,201 scenes for training and 312 scenes for validation, where points are annotated into 20 semantic categories [Dai *et al.*, 2017]. S3DIS contains 271 rooms distributed across 6 areas [Armeni *et al.*, 2017],where points are annotated into 13 semantic categories. Following standard protocols [Zhang *et al.*, 2023; Chen *et al.*, 2023b; Zhang *et al.*, 2025], we evaluate our method using 6-fold cross-validation, reporting the average performance across all folds to ensure a comprehensive assessment.

Baselines. We compare our proposed method with comprehensive state-of-the-art unsupervised 3D semantic segmentation methods, including: 1) K-means: a classical clustering algorithm applied directly to per-point $xyzrgb$ features. 2) IIC [Ji *et al.*, 2019] and 3) PiCIE [Cho *et al.*, 2021]:

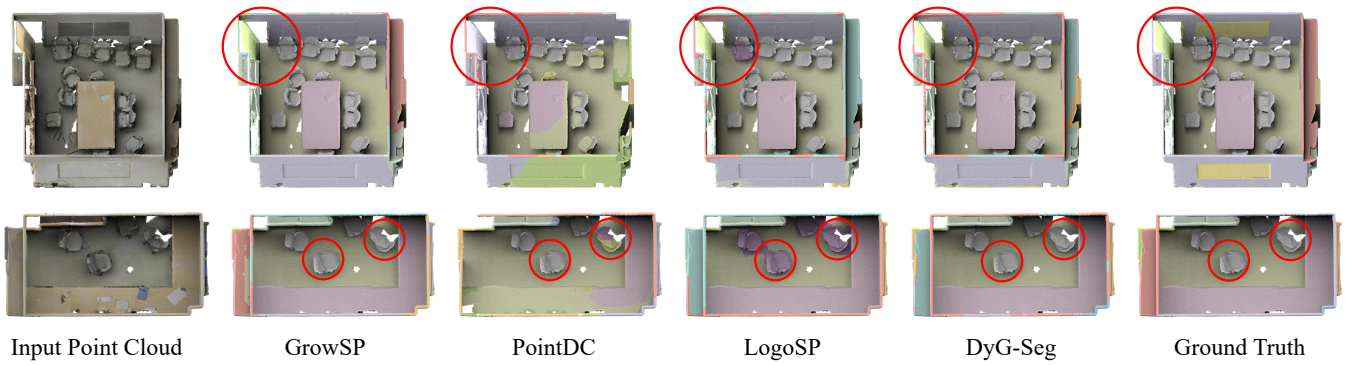

Figure 3. Qualitative results on S3DIS dataset. Red circles highlight the differences.

	Methods	OA(%)	mAcc(%)	mIoU(%)
Unsupervised	K-means	10.1	10.0	3.4
	IIC	27.7	6.1	2.9
	PiCIE	20.4	16.5	7.6
	GrowSP	57.3	44.2	25.4
	PointDC	63.7	-	25.7
	PointDC-DINOv2	64.7	45.0	29.6
	LogoSP	64.7	50.8	35.8
	DyG-Seg (Ours)	**70.6**	**53.0**	**38.4**

Table 1: Quantitative results on the validation split of ScanNet dataset. All 20 categories are evaluated.

		mIoU(%)	curtain.	shower.	bathtub.
Supervised	PointNet++	33.9	24.7	14.5	58.4
	DGCNN	44.6	38.9	22.4	47.4
	PointCNN	45.8	37.6	22.9	57.7
	SparseConv	72.5	75.4	87.0	64.7
Unsupervised	GrowSP	26.9	6.1	**9.3**	0
	PointDC	22.9	-	-	-
	LogoSP	32.7	33.4	0	21.1
	DyG-Seg (Ours)	**37.1**	**39.8**	3.0	**55.5**

Table 2: Quantitative results on the **online hidden split** of ScanNet dataset. All 20 categories are evaluated. We also report the metrics for some tail classes.

	ScanNet	**S3DIS**
Predicted Cluster Number (K')	20.3	12.4
Ground Truth Number (K)	20	12

Table 3: Comparison of predicted (K') and ground truth (K) cluster numbers.

methods adapted from the unsupervised 2D segmentation domain, modified to utilize the same 3D backbone as ours. 4) GrowSP [Zhang *et al.*, 2023]: a pioneering work that employs a superpoint growing strategy for semantic segmentation. 5) PointDC [Chen *et al.*, 2023b]: a feature distillation method utilizing the 2D pre-trained STEGO [Hamilton *et al.*, 2022] model. For a fair comparison, we also implement 6) PointDC-DINOv2, which replaces the teacher model with DINOv2 [Oquab *et al.*, 2023], consistent with our setting. 7) LogoSP [Zhang *et al.*, 2025]: a recent framework that incorporates local-to-global graph learning for unsupervised segmentation.

Evaluation Metrics. The standard metrics, including mean Intersection-over-Union (mIoU), Overall Accuracy (OA), and mean Accuracy (mAcc), are reported across all classes.

4.2 Implementation Details

Implemented in PyTorch, our framework utilizes SparseConv [Graham *et al.*, 2018] as the backbone, with initial priors distilled from pre-trained DINOv2 (ViT-S/14) [Oquab *et al.*, 2023]. The training pipeline comprises initialization clustering followed by online self-training. The GMR module is trained end-to-end with 50 diffusion timesteps, incorporating a DPP regularization term (Gaussian kernel, $\sigma = 0.05$, weight 0.01) to maintain geometric uniformity. We employ the Adam optimizer with a learning rate of 10^{-3} and train for 200 epochs. For DPCC, the concentration parameter α is set to 0.4. Key hyperparameters include a dynamic class balancing momentum of 0.99, a smoothing term of 1.2, a spatial consistency loss weight of 0.15, a geometric loss weight of 0.8, and a contrastive temperature of 0.1.

4.3 Evaluation on ScanNet

Settings and Protocol. We report results on both the offline validation and online hidden test splits. Consistent with [Zhang *et al.*, 2023; Chen *et al.*, 2023b; Zhang *et al.*, 2025], we employ Hungarian matching to align predicted labels for evaluation. The initial superpoint count is set to 40, with the final category count dynamically inferred (Table 3). Undefined points are excluded from loss computation.

Results and Analysis. As shown in Tables 1 and 2, DyG-Seg significantly surpasses baselines on both splits. A key strength is its long-tail robustness; notably, *bathtub* performance approaches supervised levels. This improvement stems from our dynamic class balancing and diffusion-based manifold recovery, which effectively capture the geometric features of sparse objects. Qualitative results are provided in Figure 2 and the Appendix.

	Methods	OA(%)	mAcc(%)	mIoU(%)
Supervised	PointNet	77.5	59.1	44.6
	PointNet++	77.5	62.6	50.1
	SparseConv	88.4	69.2	60.8
Unsupervised	K-means	21.4	21.2	8.7
	IIC	28.5	12.5	6.4
	PiCIE	61.6	25.8	17.9
	GrowSP	78.4	57.2	44.5
	PointDC	54.1	-	22.6
	PointDC-DINOv2	75.7	48.7	40.2
	LogoSP	**82.8**	55.9	46.5
	DyG-Seg (Ours)	82.0	**58.9**	**48.5**

Table 4: Quantitative results on Area-5 of S3DIS dataset. Only 12 classes are evaluated.

		mIoU(%)	beam	col.	wind.
Supervised	PointNet	44.6	0.1	10.6	54.9
	PointNet++	50.1	0	8.1	55.6
	SparseConv	60.8	0.1	36.7	37.6
Unsupervised	K-means	8.7	0.2	2.5	12.0
	IIC	6.4	0	2.1	0.1
	PiCIE	17.9	0	0.3	2.2
	GrowSP	44.5	0	14.8	27.6
	PointDC	22.6	0	1.8	12.2
	PointDC-DINOv2	40.2	0	0.8	25.8
	LogoSP	46.5	0	3.3	57.8
	DyG-Seg (Ours)	**48.5**	**0.2**	**18.9**	**60.8**

Table 5: Performance comparison on S3DIS Area-5, highlighting improvements on challenging tail categories.

4.4 Evaluation on S3DIS

Settings and Protocol. Following standard practice [Zhang *et al.*, 2023; Zhang *et al.*, 2025], we exclude the *clutter* category during training and evaluation due to its semantic ambiguity. We adopt standard 6-fold cross-validation, maintaining hyperparameters consistent with the ScanNet experiments.

Results and Analysis. As reported in Tables 4 to 6, DyG-Seg consistently outperforms baselines. Crucially, our method achieves state-of-the-art mIoU performance. Its distinct advantage lies in its robustness towards long-tail distributions, securing the best results on tail categories. This confirms that our dynamic graph structure and manifold recovery provide substantial benefits beyond the quality of initial feature priors. Additional results are in the Appendix.

4.5 Ablation Study

To validate the contribution of each component, we conduct ablation studies on the ScanNet validation set, investigating the impact of Geometric Manifold Rectification (GMR), Dynamic Prototype Contrastive Clustering (DPCC), and Iterative Optimization. Results in Table 7 show that removing GMR significantly degrades performance, confirming the criticality of geometric manifold recovery for sparse 3D data. Furthermore, Figure 4 demonstrates that the model achieves optimal performance with 40 initial superpoints and 50 diffusion timesteps, effectively balancing semantic granularity with geometric smoothing. Comprehensive ablation analysis is provided in the Appendix.

	Methods	OA(%)	mAcc(%)	mIoU(%)
Supervised	PointNet	75.9	67.1	49.4
	PointNet++	77.1	74.1	55.1
	SparseConv	89.4	78.1	69.2
Unsupervised	K-means	20.0	21.5	8.8
	IIC	32.8	14.7	8.5
	PiCIE	46.4	28.1	17.8
	GrowSP	76.0	59.4	44.6
	PointDC	55.7	37.7	26.0
	PointDC-DINOv2	74.4	51.5	41.3
	LogoSP	79.2	58.0	46.3
	DyG-Seg(Ours)	**79.9**	**61.8**	**47.5**

Table 6: Quantitative results of 6-fold cross validation on S3DIS.Only 12 classes excluding *clutter* are evaluated.

DPCC	GMR	Spatial Consistency	Dynamic Balancing	mIoU(%)
✓				32.4
✓	✓			35.3
✓	✓	✓		37.5
✓		✓	✓	34.9
✓	✓	✓	✓	**38.4**

Table 7: Ablation study of different components on the ScanNet dataset. The vertical line separates the module configurations from the performance metric.

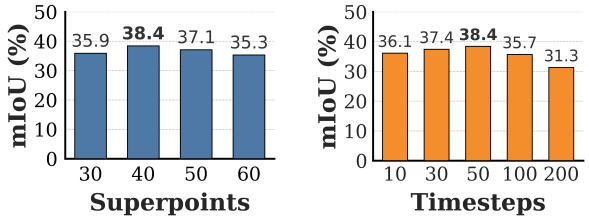

Figure 4. Hyperparameter sensitivity on ScanNet. Impact of the initial superpoint count (Left) and diffusion timesteps (Right) on segmentation performance.

5 Conclusion

In this paper, we presented DyG-Seg, a novel unsupervised framework for 3D point cloud segmentation. By introducing a diffusion-based Geometric Manifold Rectification (GMR) mechanism, our method goes beyond simple denoising to actively recover latent topology from geometrically incomplete observations, providing a robust geometric foundation. We further proposed the Dynamic Prototype Contrastive Clustering (DPCC) and Iterative Optimization to enforce spatial consistency and feature distinctiveness. Extensive experiments on ScanNet and S3DIS benchmarks demonstrate that DyG-Seg significantly outperforms existing state-of-the-art unsupervised methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62471266, 62006131, 62071260; the National Natural Science Foundation of Zhejiang Province under Grant LQ24F020001; and the Ningbo Major Research and Development Plan Project (No.2023Z224).

References

[Armeni *et al.*, 2017] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

[Baranchuk *et al.*, 2021] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[Blei and Jordan, 2006] David M Blei and Michael I Jordan. Variational inference for dirichlet process mixtures. 2006.

[Caron *et al.*, 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[Chen *et al.*, 2021] Ye Chen, Jinxian Liu, Bingbing Ni, Hang Wang, Jiancheng Yang, Ning Liu, Teng Li, and Qi Tian. Shape self-correction for unsupervised point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8382–8391, 2021.

[Chen *et al.*, 2023a] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[Chen *et al.*, 2023b] Zisheng Chen, Hongbin Xu, Weitao Chen, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, et al. Pointdc: Unsupervised semantic segmentation of 3d point clouds via cross-modal distillation and super-voxel clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14290–14299, 2023.

[Chibane *et al.*, 2022] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *European conference on computer vision*, pages 681–699. Springer, 2022.

[Cho *et al.*, 2021] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16794–16804, 2021.

[Dai *et al.*, 2017] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[Ding *et al.*, 2023] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7010–7019, 2023.

[Ferguson, 1973] Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

[Graham *et al.*, 2018] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.

[Guo *et al.*, 2024] Haoyu Guo, He Zhu, Sida Peng, Yuang Wang, Yujun Shen, Ruizhen Hu, and Xiaowei Zhou. Sam-guided graph cut for 3d instance segmentation. In *European Conference on Computer Vision*, pages 234–251. Springer, 2024.

[Hamilton *et al.*, 2022] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2020] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11108–11117, 2020.

[Hu *et al.*, 2022] Qingyong Hu, Bo Yang, Guangchi Fang, Yulan Guo, Aleš Leonardis, Niki Trigoni, and Andrew Markham. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *European Conference on Computer Vision*, pages 600–619. Springer, 2022.

[Ji *et al.*, 2019] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9865–9874, 2019.

[Kim *et al.*, 2024] Hyeonseong Kim, Sung-Hoon Yoon, Minseok Kim, and Kuk-Jin Yoon. On-the-fly category discovery for lidar semantic segmentation. In *European Conference on Computer Vision*, pages 231–249. Springer, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[Kulesza and Taskar, 2012] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012.

[Li *et al.*, 2024] Xiang Li, Jian Ding, Zhaoyang Chen, and Mohamed Elhoseiny. Uni3dl: A unified model for 3d vision-language understanding. In *European Conference on Computer Vision*, pages 74–92. Springer, 2024.

[Liu *et al.*, 2021] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1726–1736, 2021.

[Liu *et al.*, 2023] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. *Advances in Neural Information Processing Systems*, 36:37193–37229, 2023.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Pang *et al.*, 2023] Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1:2440001, 2023.

[Peng *et al.*, 2023] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023.

[Peng *et al.*, 2024] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Yujing Sun, Tai Wang, Xinge Zhu, and Yuexin Ma. Learning to adapt sam for segmenting cross-domain point clouds. In *ECCV*, pages 54–71. Springer, 2024.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PmLR, 2021.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Sethuraman, 1994] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[Unal *et al.*, 2022] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2697–2707, 2022.

[Wang *et al.*, 2021] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021.

[Wei *et al.*, 2020] Jiacheng Wei, Guosheng Lin, Kim-Hui Yap, Tzu-Yi Hung, and Lihua Xie. Multi-path region mining for weakly supervised 3d semantic segmentation on point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4384–4393, 2020.

[Xiao *et al.*, 2024] Zihao Xiao, Longlong Jing, Shangxuan Wu, Alex Zihao Zhu, Jingwei Ji, Chiyu Max Jiang, Wei-Chih Hung, Thomas Funkhouser, Weicheng Kuo, Anelia Angelova, et al. 3d open-vocabulary panoptic segmentation with 2d-3d vision-language distillation. In *European Conference on Computer Vision*, pages 21–38. Springer, 2024.

[Xie *et al.*, 2020] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, pages 574–591. Springer, 2020.

[Yin *et al.*, 2024] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024.

[Zhang *et al.*, 2023] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 17619–17629, 2023.

[Zhang *et al.*, 2025] Zihui Zhang, Weisheng Dai, Hongtao Wen, and Bo Yang. Logosp: Local-global grouping of superpoints for unsupervised semantic segmentation of 3d point clouds. In *CVPR*, pages 1374–1384, 2025.

[Zhou *et al.*, 2021] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, pages 5826–5835, 2021.

[Zhu *et al.*, 2021] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, pages 9939–9948, 2021.