Diffusion-Based Contextual Reconstruction for Point Cloud Segmentation with Limited Annotations

Jiawei Lian[1], **Zhengxue Wang** [1], **Wentao Qu** [1], **Haobo Jiang**[3], **Le Hui**[2*], **Jian Yang**[1*]

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education
School of Computer Science and Engineering, Nanjing University of Science and Technology
[2] Shaanxi Key Laboratory of Information Acquisition and Processing
School of Electronics and Information, Northwestern Polytechnical University
[3]Nanyang Technological University
{lianjw,zxwang,quwentao,csjyang}@njust.edu.cn, haobo.jiang@ntu.edu.sg, huile@nwpu.edu.cn

Abstract

Point cloud semantic segmentation is fundamental to 3D scene understanding, but dense annotation requirements limit scalability. Although recent label propagation and contrastive learning methods enhance local consistency, the incomplete object coverage caused by sparse annotations hinders global context modeling, ultimately limiting overall performance. To this end, we propose a diffusion-based contextual reconstruction framework for point cloud semantic segmentation with limited annotations. At its core, our framework guides denoising with semantic predictions, using better context reconstruction to enhance the conditional model for better segmentation. Specifically, our contributions include: (1) Diffusion-based segmentation framework: reconstructs contextual semantics from noise under conditional guidance, sharing the decoder with the segmentation module for robust contextual semantic learning. (2) Dynamically aggregates local context from segmentation features and guides denoising with global spatial structure, significantly enhancing denoising quality and contextual awareness. Notably, we pioneer diffusion models for 3D semantic segmentation with limited annotations, enabling efficient single-step inference. Experiments show robustness across varying annotation ratios and state-of-the-art performance on benchmarks.

Code — https://github.com/ljwwwiop/DiCoSeg

Introduction

Point cloud semantic segmentation serves as a fundamental technology for 3D scene understanding (Qian et al. 2022; Wu et al. 2024; Wang et al. 2025c; Hui et al. 2021), with critical applications in scene reconstruction (Hane et al. 2013; Jiang et al. 2021), autonomous driving (Hu et al. 2023), and embodied AI (Gupta et al. 2021; Wang et al. 2024b; Lian et al. 2025). However, this task confronts two fundamental obstacles: the unavoidable dependence on fine-grained point-wise annotations and the prohibitive labeling costs imposed by 3D geometric complexity. Consequently, weakly supervised learning has emerged as a promising direction to overcome these practical bottlenecks. Nonetheless, learning segmentation under sparse annotations remains highly challenging due to limited supervision signals.

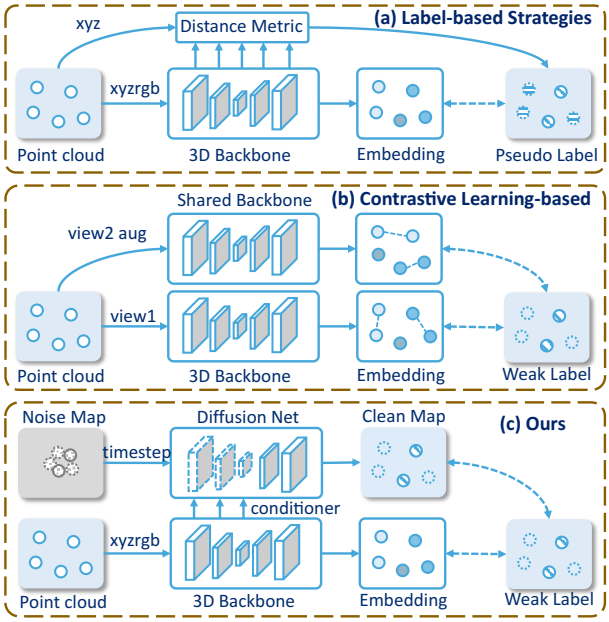


Figure 1: A comparison of existing weakly-supervised point cloud segmentation frameworks: (a) label-based strategy, (b) contrastive learning approach, and (c) our proposed method.

Recent years have witnessed significant progress in weakly supervised learning under limited annotation conditions. As illustrated in Figure 1, existing approaches primarily follow two paradigms: (1) label-based strategies (Pan et al. 2025; Tang et al. 2023; Wang, Yan, and Yang 2024; Wang et al. 2025b) and (2) self-enhanced contrastive learning (Hou et al. 2021; Li et al. 2022; Liu et al. 2023). Among these, label propagation stands as one of the earliest and most prevalent weakly supervised methods. This approach aggregates and propagates features among neighboring points through local geometric distance metrics, effectively enhancing local feature representation.

*Corresponding authors.

(a) View occlusion **(b) Spatial isolation**

Figure 2: Visualization based on S3DIS point cloud showing (a) viewpoint occlusion and (b) spatial isolation.

However, its heavy reliance on geometric priors often induces pseudo-label propagation errors, constituting a critical bottleneck for performance improvement.

Recognizing the limitations of local metric-based methods, contrastive learning has emerged to capture cross-view semantic similarity and enhance contextual understanding. Notable examples include HybridCR (Li et al. 2022), which employs a dual-branch dynamic view augmentation with global-local contrastive regularization, and CPCM (Liu et al. 2023), which introduces a cross-view masked contrastive framework to better handle occlusions. However, existing methods still face two major challenges (see Figure 2): (1) object truncation under heavy occlusion disrupts cross-view semantic consistency; and (2) spatial isolation weakens augmentation strategies (e.g., masking) in modeling isolated objects effectively. These issues significantly hinder robust context modeling in real-world scenarios.

Inspired by diffusion models' success in image and video reconstruction (Ho, Jain, and Abbeel 2020; Kawar et al. 2022; Jiang et al. 2023b), where progressively better restoration from noise reflects *deeper contextual understanding* (Sohl-Dickstein et al. 2015; Qu et al. 2024; Jiang et al. 2023a), we incorporate this capability into weakly supervised point cloud segmentation. Moreover, diffusion models inherently complement the unstructured nature of point clouds: (i) reconstruction requires no fixed augmentations, and (ii) spatial structure recovery implicitly models occlusions, and geometric relationships. Building on these insights, the challenge is to develop a diffusion-based weakly supervised segmentation framework that effectively leverages diffusion models' strengths.

To address this challenge, we propose DiCoSeg, a **Di**ffusion-based **Co**ntextual reconstruction framework for point cloud **Seg**mentation under limited annotations. The framework comprises two key components: a diffusion-based denoising network (DDN) and a context-aware guidance learning (CGL) module. Specifically, DDN first encodes sparse labels into a semantic map, injects Gaussian noise as the diffusion network's initial input, and then progressively denoises to recover global contextual semantics. Concurrently, CGL uses segmentation semantics as a conditional guide to dynamically aggregate local context and integrate global spatial structure, further enhancing fine-grained semantic reconstruction. Through this synergistic design, DiCoSeg achieves refined contextual modeling, significantly advancing weakly-supervised segmentation performance. Moreover, a shared prediction head enables single-step inference without sacrificing accuracy. Extensive experiments on three public benchmarks (indoor and outdoor) validate the effectiveness of DiCoSeg, achieving state-of-the-art performance. Our contributions are summarized as follows:

1. We propose DiCoSeg, a diffusion-based framework that reconstructs contextual semantics for point cloud segmentation with limited annotations.
2. Two core components: a diffusion-based denoising network for context reconstruction, and a context-aware guidance learning module for conditional feature fusion.
3. To our knowledge, this is the first work to explore diffusion models for 3D point cloud segmentation with limited annotations. Extensive experiments on benchmarks demonstrate DiCoSeg's superior performance.

Related Work

To mitigate the high annotation cost in point cloud semantic segmentation, weakly supervised learning has emerged as a viable alternative. Current approaches predominantly adapt techniques from the image domain (Sun et al. 2024; Qu et al. 2025), including label propagation for annotation completion (Tang et al. 2023), contrastive learning for enhanced feature discriminability (Li et al. 2022), and other modeling frameworks (Hu et al. 2022; Wang et al. 2024a, 2025a). Most studies tend to integrate multiple strategies to construct more robust weakly supervised frameworks.

Label Propagation. Under weakly supervised settings (Sun et al. 2025a,b; Wang, Fang, and Tiwari 2025; Cheng et al. 2021), an increasing number of methods focus on improving label quality and propagation effectiveness. Zhang et al. (2021) leverage transfer learning to introduce sparse pseudo-labels for regularizing network training; OTOC++ (Liu, Qi, and Fu 2021) employs RelationNets to precisely measure 3D graph node similarity for label propagation; while AADNet (Pan et al. 2025) proposes an adaptive label distribution regularization to effectively mitigate distribution imbalance in weakly supervised learning. However, they merely treat pseudo-labels as auxiliary supervision while ignoring occlusion-induced geometric context variations.

Contrastive Learning. Xie et al. (2020); Zheng et al. (2023) pioneered the first contrastive pre-training framework for point cloud scenes through view perspective augmentation, initiating subsequent research. HybridCR (Li et al. 2022) establishes a multi-level contrastive mechanism across transformed point cloud pairs, local geometric pairs, and category prototype pairs. MILTrans (Yang et al. 2022) introduces category-wise contrastive loss at the scene level. CPCM (Liu et al. 2023) is the first to integrate masked modeling for weakly supervised point cloud segmentation, enhancing the backbone network's reasoning capability for masked contexts through stochastic contrastive learning. However, they focus solely on point- or view-level self-augmentation contrasts, ignoring the inherent rich contextual semantic relationships within the complete scene.

Others. In weakly supervised point cloud semantic segmentation, local partitioning strategies such as KNN sampling

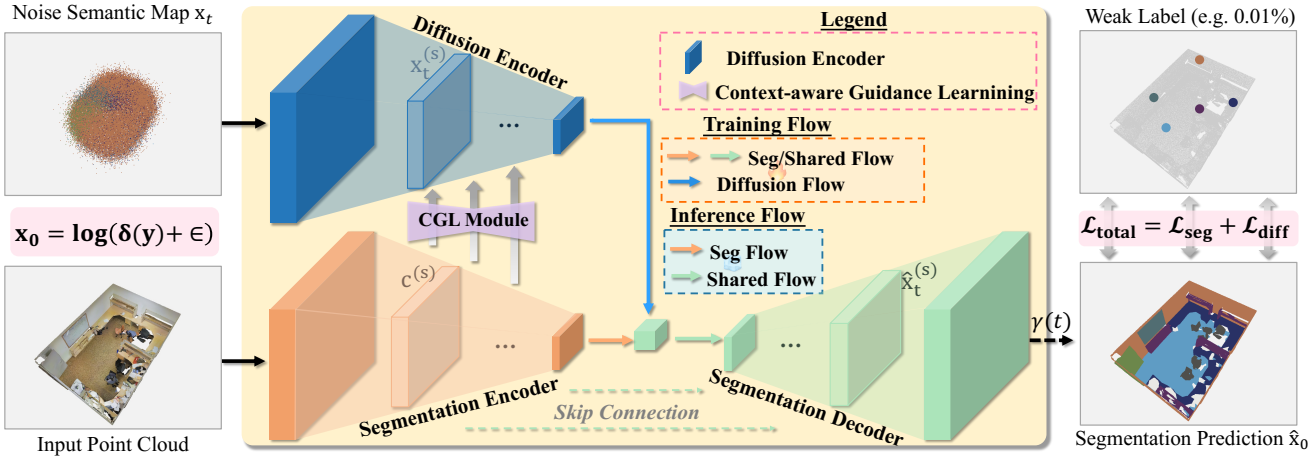


Figure 3: The framework of DiCoSeg consists of two modules: (1) A Diffusion Denoising Network (DDN) that leverages a diffusion encoder and a shared semantic decoder to recover context by reconstructing semantic maps from noise; (2) A Context-aware Guidance Learning (CGL) module, which steers the denoising in the DDN using scene features as a conditioning signal. While both modules are jointly optimized during training, only the semantic branch is needed for efficient single-step inference.

and Class Activation Maps (CAM) were among the earliest explored approaches. Xu and Lee (2020) constructed KNN graphs based on point cloud geometry to facilitate feature embedding learning. SQN (Hu et al. 2022) aggregates hierarchical representations via interpolation within locally labeled neighborhoods. With the advent of more powerful backbones like Transformers, PointCT (Tran et al. 2024) introduces a center-point attention mechanism within neighborhoods to enhance semantic representation. Moreover, MILTrans (Yang et al. 2022) incorporates adaptive global weighted pooling on top of CAM to effectively suppress irrelevant classes and noise points. Despite benefiting from neighborhood consistency in point clouds, they fail to effectively mitigate the adverse impact of outliers on semantics.

Diffusion Modeling for Point Cloud. Luo and Hu (2021) pioneered the application of diffusion models to 3D point cloud object reconstruction and generation, establishing a new research paradigm. PointDif (Zheng et al. 2024) introduces a diffusion-based pre-training approach for object-level point cloud understanding. For scene-level tasks, DifFUSER (Le et al. 2024) incorporates diffusion models in BEV space to simulate robust components for enhanced 3D detection stability, while CDSegNet (Qu et al. 2025) presents a novel end-to-end single-stage fully-supervised diffusion framework for segmentation. Nevertheless, diffusion-based approaches remain unexplored for 3D semantic segmentation under limited annotations. Consequently, we propose a diffusion-based contextual reconstruction framework that achieves efficient scene segmentation from sparse annotations via semantic completion.

Method

In this part, Section 3.1 introduces the problem and preliminaries. Sections 3.2 and 3.3 detail the DiCoSeg, including a diffusion denoising network and context-aware guidance module. Section 3.4 presents the overall training objective.

Preliminaries

Problem Setup. We define the point cloud dataset as $\mathcal{D} = \{\mathbf{P}, \mathbf{Y}\}$, where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N]$, $\mathbf{p}_i \in \mathbb{R}^{1\times F}$ denotes a 3D point set with F-dimensional features, and N total points. The label set is defined as $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_M]$, $\mathbf{y}_i \in \{0,1\}^{1\times C}$, where $M \ll N$ is the number of labeled points, and C denotes the number of semantic categories. When $M \ll N$, the task is considered weakly supervised point cloud semantic segmentation, and the labeling ratio is defined as $\frac{M}{N}$. Specifically, under the 1% label setting, only $M = 1\% \times N$ points are randomly selected and annotated. Note that all labeled points are sampled uniformly at random.

Diffusion Model. Given a target data sample $\mathbf{x}_0 \sim \mathcal{D}$, a conditioning signal $\mathbf{c} \sim \mathcal{D}_c$, and a latent variable $\mathbf{z} \sim \mathcal{D}_{\text{noise}}$, the conditional denoising diffusion probabilistic model (DDPM) follows an autoregressive generation process (Ho, Jain, and Abbeel 2020). A predefined forward process q gradually corrupts $\mathbf{x}_0$ into pure noise $\mathbf{z}$, while a learnable reverse process p_θ, guided by condition $\mathbf{c}$, reconstructs the target sample through denoising. Let $\mathbf{x}_0$ represent the target semantic labels and $\mathbf{c}$ the segmented point cloud of the corresponding scene. Since semantic labels are inherently discrete and incompatible with diffusion modeling, a continuous relaxation is applied. Specifically, the labels $y \in \{0, 1, \ldots, C-1\}^N$ are first one-hot encoded using a delta function $\delta(y)$, then smoothed with a small constant $\varepsilon > 0$, resulting in:

$$\mathbf{x}_0 = \log(\delta(y) + \varepsilon), \quad \mathbf{x}_0 \in \mathbb{R}^{N\times C}, \tag{1}$$

where N is the number of points and C the number of semantic classes. The smoothing factor ε mitigates numerical instability during training.

To enable conditional generation, noisy samples are constructed as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \cdot \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t} \cdot \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(0, I), \tag{2}$$

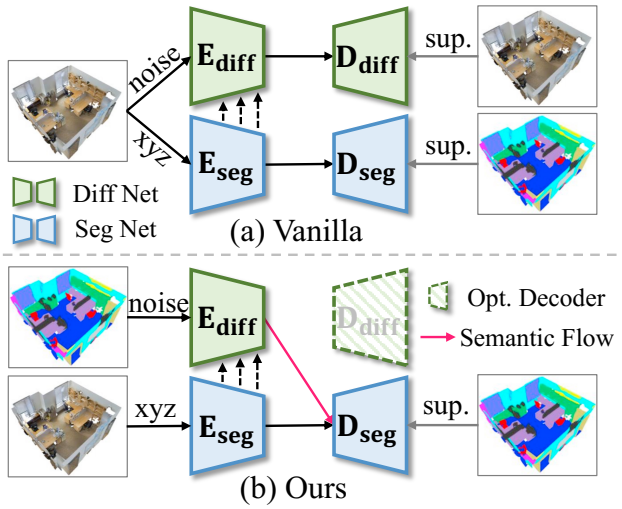


Figure 4: (a) Point cloud scene reconstruction (vanilla diffusion), (b) Ours: Semantic-aware reconstruction (diffusion decoder-free DiCoSeg).

where $\bar{\alpha}_t$ is the cumulative noise schedule from the forward diffusion process, and $\mathbf{z}$ denotes Gaussian noise.

For semantic segmentation, the model is trained to maximize the log-likelihood of the predicted class corresponding to the ground-truth label. The training objective is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{(\mathbf{x}_0,t,\mathbf{z})}\left[-\log\left(\text{Dec}\left(\text{Enc}(\mathbf{x}_t,t,\mathbf{c})\right)^{(y)}\right)\right], \quad (3)$$

where $\text{Enc}(\cdot)$ and $\text{Dec}(\cdot)$ denote the diffusion encoder and decoder of the conditional diffusion model, and $(\cdot)^{(y)}$ extracts the predicted logit for the ground-truth class.

Diffusion-Based Denoising Network

As shown in Figure 3, we propose a diffusion-based framework for weakly supervised semantic segmentation of large-scale point clouds. The framework's core is a diffusion-based denoising network (DDN) consisting of a diffusion encoder that captures contextual dependencies under noise, and a shared semantic decoder that steers denoising toward semantically consistent, discriminative predictions. This design enables the denoising trajectory to progressively reconstruct coherent contextual semantics.

Network Architecture Analysis. Unlike existing diffusion-based point cloud methods that focus on low-level geometric reconstruction (e.g., scene/object generation), our framework specializes in high-level semantic context modeling (Figure 4). While prior works often overlook semantically consistent representations critical for scene understanding, our method progressively restores complete semantics from noisy inputs under weak supervision. This process inherently requires point-wise reasoning, creating a natural alignment between diffusion-based generation and point cloud segmentation.

Most segmentation networks employ U-Net architectures, where encoders extract geometric-semantic features and decoders reconstruct point-wise predictions through cascaded upsampling. Capitalizing on the structural similarity between the diffusion model's reverse denoising (which progressively recovers semantics) and segmentation tasks, we share the semantic decoder between both branches in the DDN module. This unified design ensures semantic consistency while reducing parameters and improving training efficiency. Thus, our DDN mainly consists of a dedicated denoising encoder and a shared semantic decoder.

Denoising Feedforward. Based on the above architectural design and analysis, we feed the continuous semantic feature $\mathbf{x}_t$ into the denoising diffusion network for forward prediction. Specifically, this corresponds to modeling the reverse of a Markovian noising process, where the forward process is defined as:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\,\mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \quad (4)$$

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right). \quad (5)$$

where $t \in \{1,\ldots,T\}$ denotes the diffusion timestep and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is the cumulative noise coefficient. A noisy feature $\mathbf{x}_t$ sampled from this distribution serves as the input to the network, guiding the subsequent denoising-based semantic modeling.

Next, in the specific process, the noisy feature $\mathbf{x}_t$ is formed as a sparse tensor and fed into the diffusion encoder for multi-stage feature extraction. At each stage, a time-step embedding $\gamma(t)$ explicitly conditions the diffusion step. Meanwhile, point cloud segmentation features are encoded via the CGL module into a condition vector $\mathbf{c}$, guiding the encoder to learn denoised representations. At stage s, the encoder computation is:

$$\mathbf{x}_t^{(s+1)} = \text{Encoder}^{(s)}\left(\text{CGL}\left([\mathbf{x}_t^{(s)}, \gamma(t)], \mathbf{c}^{(s)}\right)\right), \quad (6)$$

where $\mathbf{x}_t^{(s)}$ is the input from the initial or previous downsampling stage, and $\mathbf{x}_t^{(s+1)}$ is the resulting hidden state of the next layer. Outputs at each stage are forwarded via skip connections to the corresponding decoder blocks for contextual semantic reconstruction.

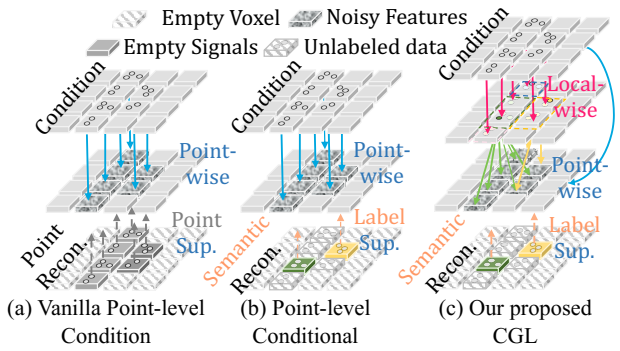
During decoding, the network adopts a symmetric architecture that progressively upsamples and fuses features from corresponding encoder stages via skip connections, thereby reconstructing denoised semantic representations layer by layer. At stage s, the decoding process is formulated as:

$$\hat{\mathbf{x}}_t^{(s)} = \text{Decoder}^{(s)}\left([\text{Up}(\mathbf{x}_t^{(s+1)}), \mathbf{x}_t^{(s)}, \mathbf{c}^{(s)}]\right), \quad (7)$$

where $\hat{\mathbf{x}}_t^{(s)}$ denotes the decoded feature at stage s, and $\mathbf{x}_t^{(s)}$ is the feature from the encoder at the same stage passed via skip connection. Finally, a semantic head is applied to the decoded features to perform point-wise classification, yielding the denoised semantic output $\hat{\mathbf{x}}_0$.

Context-aware Guidance Learning

Under weak supervision, sparse annotations often result in incomplete object surface coverage (Figure 5), making traditional point-to-point geometric reconstruction strategies ineffective for semantic recovery. This is primarily because



(a) Vanilla Point-level Condition (b) Point-level Conditional (c) Our proposed CGL

Figure 5: (a) Scene reconstruction with point-wise supervision, (b) Transferred point supervision for semantic reconstruction, (c) Proposed CGL: Fuses local patches before point relation modeling.

isolated point labels lack sufficient geometric and semantic context, hindering accurate feature modeling. Therefore, precise semantic reconstruction requires incorporating local geometric structures and contextual information to compensate for the missing supervision. We propose a context-aware guidance learning module that aggregates local patch-wise context and key semantic cues, then integrates global spatial structure to guide the diffusion process in recovering fine-grained semantics (Figure 6). It consists of two components: a Local Context Aggregation (LCA) block that captures informative local features, and a Global Geometric Fusion (GGF) block that encodes holistic spatial structure for semantic guidance.

Local Context Aggregation. To enable effective conditioning under sparse supervision, we design a local context aggregation module that distills informative geometric priors and injects them into the diffusion stream. Given an intermediate condition semantic feature representation $\mathbf{c}^{(s)}$, we voxelize the point cloud by computing:

$$\mathbf{v}_i = \left\lfloor \frac{\mathbf{p}_i}{v} \right\rfloor, \quad \mathbf{p}_i \in \mathbb{R}^3, \tag{8}$$

where v denotes the voxel size. Points in the same voxel are aggregated to produce initial voxel-level features. We then compute an importance score for each voxel—based on feature variance—and retain the top-K voxels to form context-rich patches:

$$s_j = \left\| \texttt{scatter\_std}(\mathbf{c}^{(s)}, \mathbf{v})_j \right\|_2, \tag{9}$$

$$\mathbf{c}^{(s)}_{\text{top-}K} = \texttt{TopK}_j \left(\texttt{scatter\_mean}(\mathbf{c}^{(s)}, \mathbf{v})_j,\ s_j \right). \tag{10}$$

These aggregated semantic anchors provide explicit guidance and enhance contextual awareness, effectively guiding the diffusion model to recover fine-grained semantic structures. The selected top-K features are fused into the noisy representation via a residual dense convolution:

$$\mathbf{x}_t^{(s)\prime} = \mathbf{x}_t^{(s)} + \Psi^{(s)} \left[\mathbf{x}_t^{(s)}, \mathbf{c}^{(s)}_{\text{top-}K} \right], \tag{11}$$

where $\Psi^{(s)}$ denotes a dense feed-forward network based on 1D dense convolution, applied across all points to inject local semantic priors.

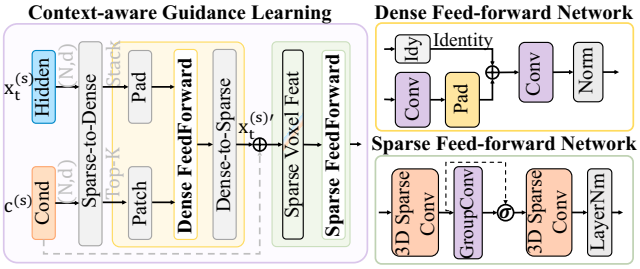


Figure 6: Detail of the CGL module: Processed intermediate and conditional features from LCA and GGF are fed into the next denoising encoder. (N: points, B: batch, d: dim).

Global Geometric Fusion. After local aggregation, we further inject global structural priors into the 3D voxel space. At stage s, the noisy voxel features $\mathbf{x}_t^{(s)\prime}$ are fused with the voxel-wise semantic features $\mathbf{c}^{(s)}$, which encode 3D spatial structure, through a multi-layer sparse convolution:

$$\mathbf{x}_t^{(s)\prime\prime} = \mathbf{x}_t^{(s)\prime} + \Phi^{(s)} \left[\mathbf{x}_t^{(s)\prime} \oplus \mathbf{c}^{(s)} \right], \tag{12}$$

where $\Phi^{(s)}$ is a sparse feed-forward network based on multi-group 3D sparse convolutions, designed to propagate spatial information across the entire scene. This fusion injects global structural context into the diffusion process, reinforcing semantic consistency at a coarse scale.

Training and Inference

During training, we jointly optimize the semantic reconstruction and segmentation branches end-to-end, sharing data preprocessing and hyperparameter settings. The training objective is: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{seg}}$. During inference, only the segmentation branch is retained for efficient prediction, as denoising depends on its conditional features.

Experiments

Experiment Setting

Datasets. We evaluate our method on three widely used benchmarks: S3DIS, ScanNetV2, and SemanticKITTI. S3DIS (Armeni et al. 2016) consists of 271 rooms from 6 indoor areas, covering 13 semantic categories. Following common practice, we train our model on Areas 1, 2, 3, 4, and 6, and evaluate on Area 5. ScanNetV2 (Dai et al. 2017) provides 1,513 RGB-D scans collected from 707 diverse indoor scenes, with 21 annotated semantic classes per point. We use the official split with 1,201 scenes for training and 312 for validation. SemanticKITTI (Behley et al. 2019) is a large-scale outdoor LiDAR dataset annotated with 19 semantic classes. Following standard protocol, we train on sequences 00–10, using sequence 08 for validation.

Annotation setting. For fair comparison, we report performance under varying label rates: 1%, 0.1%, and 0.01% for S3DIS; 0.1%, 0.01%, and 1% for SemanticKITTI; and 20 points per scene (approximately 0.014%), 1%, and 0.1% for ScanNetV2. All annotations are randomly sampled.

Settings	Method	mIoU	Where
100%	PointNet++ (Qi et al. 2017)	50.0	NeurIPS
	SparseUNet (Base)	65.4	CVPR
	PointNext (Qian et al. 2022)	**69.2**	NeurIPS
100%	SQN (Hu et al. 2022)	63.7	ECCV
	RandLA-Net (Hu et al. 2020)	64.6	CVPR
	HybridCR (Li et al. 2022)	65.8	CVPR
	ERDA (Tang et al. 2023)	68.3	NeurIPS
	PointCT (Tran et al. 2024)	67.9	WACV
	DiCoSeg (Ours)	68.6	
1%	RandLA-Net (Hu et al. 2020)	59.8	CVPR
	SQN (Hu et al. 2022)	63.6	ECCV
	PointCT (Tran et al. 2024)	67.6	WACV
	DiCoSeg (Ours)	**68.3**	
0.1%	SQN (Hu et al. 2022)	61.4	ECCV
	CPCM (Liu et al. 2023)	66.3	ICCV
	PointCT (Tran et al. 2024)	68.3	WACV
	AADNet (Pan et al. 2025)	67.2	AAAI
	DiCoSeg (Ours)	**68.3**	
0.01%	PointNext (Qian et al. 2022)	58.4	NeurIPS
	SQN (Hu et al. 2022)	45.3	ECCV
	CPCM (Liu et al. 2023)	59.3	ICCV
	AADNet (Pan et al. 2025)	60.8	AAAI
	DiCoSeg (Ours)	**61.1**	

Table 1: Comparison on S3DIS Area5 under various weakly-supervised settings. **Bold** indicates the best performance.

Implementation. We adopt SparseUNet (Choy, Gwak, and Savarese 2019) as the backbone segmentation architecture, leveraging sparse convolution as the core computation. The voxel size v is set to $[0.05, 0.05, 0.05]$ for S3DIS and SemanticKITTI, and $[0.02, 0.02, 0.02]$ for ScanNetV2. The diffusion module is configured with $T = 1000$ time steps using a linear scheduler, and the feature dimension per step is set to 128. `TopK` aggregated patch sizes per stage: $[2048, 2048, 1024, 512, 256]$. **Further experiment and setup details are in the supplementary material.**

Evaluation Protocols. We adopt mean Intersection-over-Union (mIoU) as the primary evaluation metric, with mean class accuracy (mAcc) used as a supplementary measure in selected experiments.

Comparison Results

Results on S3DIS. We conduct systematic evaluations of DiCoSeg on S3DIS Area 5 under varying annotation ratios (100%, 1%, 0.1%, and 0.01%). As shown in Table 1, with SparseUNet as the baseline, DiCoSeg demonstrates consistent and significant performance gains across all supervision levels. In the 100% fully supervised setting, DiCoSeg achieves a 3.2% mIoU improvement over the SparseUNet baseline and performs comparably to the strong fully supervised method PointNext. Under sparse annotation, DiCoSeg surpasses the weakly supervised PointCT at the 1% level and significantly outperforms the recent AADNet at both 0.1% and 0.01% levels, demonstrating robustness under extremely low-label regimes. Notably, at the 0.01% setting, PointNext suffers a severe performance drop to 58.4% mIoU, while DiCoSeg maintains a strong 61.1%, highlighting its superior label efficiency. Please refer to the supplementary material for detailed per-class results on the S3DIS dataset.

Settings	Method	mIoU	Where
100%	PointNet++ (Qi et al. 2017)	53.5	NeurIPS
	SparseUNet (Base)	70.2	CVPR
	PointNext (Qian et al. 2022)	**71.2**	NeurIPS
1%	HybridCR (Li et al. 2022)	56.8	CVPR
	ERDA (Tang et al. 2023)	63.0	NeurIPS
	PointCT (Tran et al. 2024)	65.6	WACV
	DGNet (Pan et al. 2024)	67.4	NeurIPS
	AADNet (Pan et al. 2025)	66.8	AAAI
	DiCoSeg (Ours)	**70.7**	
0.1%	SQN (Hu et al. 2022)	58.4	ECCV
	PointCT (Tran et al. 2024)	63.7	WACV
	DiCoSeg (Ours)	**70.3**	
20pts	PointNext (Qian et al. 2022)	54.6	NeurIPS
	EDRA (Liu et al. 2023)	57.0	NeurIPS
	DGNet (Pan et al. 2024)	62.9	NeurIPS
	AADNet (Pan et al. 2025)	62.5	AAAI
	DiCoSeg (Ours)	**68.4**	

Table 2: Quantitative comparisons on ScanNetV2.

Settings	Method	mIoU	Where
0.1%	SQN (Hu et al. 2022)	50.8	ECCV
	DGNet (Pan et al. 2024)	51.8	NeurIPS
	AADNet (Pan et al. 2025)	53.3	AAAI
	DiCoSeg (Ours)	**62.6**	
0.01%	CPCM (Liu et al. 2023)	34.7	ICCV
	SQN (Hu et al. 2022)	39.1	ECCV
	DGNet (Pan et al. 2024)	41.2	NeurIPS
	DiCoSeg (Ours)	**49.6**	

Table 3: Quantitative comparisons on SemanticKITTI.

Results on ScanNetV2. Compared to S3DIS, ScanNetV2 presents greater challenges with its diverse and complex indoor scenes. We evaluate DiCoSeg under 1%, 0.1%, and 20 points (pts) annotation, with results summarized in Table 2. Despite the limited supervision, DiCoSeg surpasses DGNet by 3.3% at 1%, outperforms PointCT by 6.6% at 0.1%, and exceeds DGNet by 5.5% under the 20 pts setting. These results demonstrate the strong generalization and label efficiency of DiCoSeg under sparse supervision.

Results on SemanticKITTI. In contrast to the dense indoor datasets S3DIS and ScanNetV2, SemanticKITTI is a sparse

Settings	CGL Module		S3DIS		ScanNetV2	
	LCA	GGF	mIoU	mAcc	mIoU	mAcc
Base	–	–	65.1	70.6	66.2	72.4
w/o CGL	–	–	64.2	69.6	66.0	72.1
+ LCA	✓	–	67.4	73.6	69.7	78.8
+ GGF	–	✓	66.9	73.2	68.9	77.8
Ours	✓	✓	**68.3** (+3.2)	**73.8** (+3.2)	**70.7** (+4.5)	**79.1** (+6.7)

Table 4: Ablation study of the CGL module. "w/o CGL" denotes results without using the CGL module.

Settings	S3DIS		ScanNetV2	
	mIoU	mAcc	mIoU	mAcc
Shared Encoder & Decoder	64.3	69.6	64.7	68.9
Shared Encoder Only	64.8	70.1	65.6	71.3
Shared Decoder Only	**68.3**	**73.8**	**70.7**	**79.1**

Table 5: Ablation study of the shared decoder in DDN.

outdoor autonomous driving benchmark. As shown in Table 3, DiCoSeg demonstrates remarkable weakly supervised segmentation performance on this dataset, outperforming AADNet by 9.3% mIoU at the 0.1% annotation ratio and surpassing DGNet by 8.4% at 0.01%. These results further underscore the strong generalization capability of DiCoSeg across diverse and challenging environments.

Ablation Study

Ablation on CGL. Given the CGL module's essential role in diffusion denoising, we performed a staged ablation on S3DIS and ScanNetV2 (0.1% annotations) in Table 4. Omitting CGL ("w/o CGL") in the diffusion framework resulted in a marked performance drop relative to the SparseUnet baseline, underscoring the necessity of conditional guidance. Incorporating "+ LCA" improves mIoU by 2.3% and 3.5% on S3DIS and ScanNetV2, respectively. In contrast, "+ GGF" yields smaller gains of 1.8% and 2.7% in mIoU. These results suggest that local contextual guidance is more effective than global guidance for weakly supervised semantic recovery. "Ours" achieves the best performance with the full module, indicating that the joint training of LCA and GGF enables more effective semantic fusion and conditional alignment within CGL.

Ablation on Shard Head. Table 5 shows an ablation study on parameter sharing between diffusion denoising and semantic segmentation networks under a 0.1% annotation setting. Full parameter sharing causes unstable training and poor results, with mIoU scores of 64.3% and 64.7%, mainly due to noise disrupting geometric feature learning. Sharing only the encoder improves mIoU to 64.8% and 65.6%, but noise still affects feature learning when sharing the decoder. In contrast, sharing the decoder achieves the best results, showing effective fusion of contextual semantics from the diffusion network that boosts segmentation performance.

Further Analysis

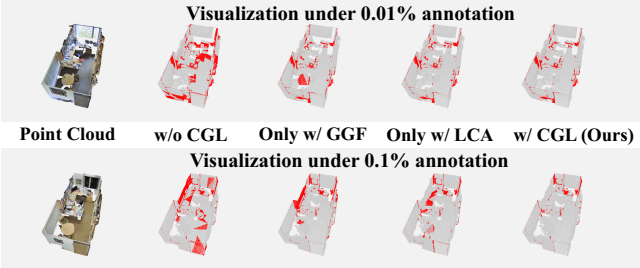


Figure 7: Qualitative comparison under 0.1% and 0.01% label rates on S3DIS, showing CGL's (LCA+GGF) impact. Red: prediction errors; gray: correct predictions.

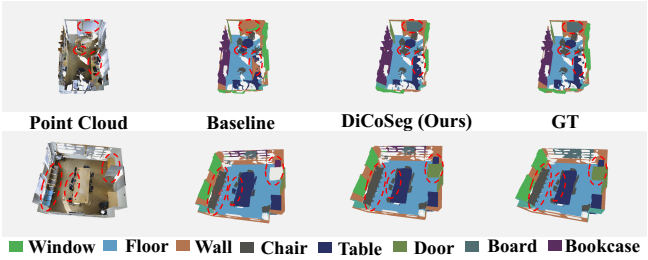


Figure 8: Qualitative comparison on S3DIS under the 0.1% annotation setting. The rightmost column: The ground truth.

Contextual Uncertainty. Figure 7 compares segmentation errors under 0.1% and 0.01% annotation rates with different fusion conditions. The results show: (1) as conditional features become richer (left → right), major error regions shrink and segmentation improves; (2) reasoning over occluded objects such as tables and bookcases is noticeably enhanced. Figure 8 compares the qualitative results of our method and the baseline. The baseline struggles with spatial reasoning (e.g., occluded chairs and tables) and unclear boundaries (e.g., boards and walls), while DiCoSeg yields more stable and accurate segmentation.

Conclusion

In this paper, we propose DiCoSeg, a diffusion-based contextual reconstruction framework for point cloud semantic segmentation with limited annotations. DiCoSeg integrates a diffusion-based denoising network (DDN) and a context-aware guidance learning (CGL) module to enable semantic reconstruction from sparse supervision. The DDN employs conditional denoising to recover semantic structures from noise, while the CGL enhances contextual modeling by fusing local and global geometric cues. A shared prediction head further enables efficient single-step inference. Extensive experiments on three benchmarks demonstrate that DiCoSeg achieves state-of-the-art performance under weak supervision, validating the effectiveness and generalizability of diffusion models for 3D scene understanding.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. U24A20330, 62361166670, and 62306238, and by the Fundamental Research Funds for the Central Universities.

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE international conference on computer vision*, 9297–9307.

Cheng, M.; Hui, L.; Xie, J.; and Yang, J. 2021. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1140–1147.

Choy, C.; Gwak, J.; and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3075–3084.

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

Gupta, A.; Savarese, S.; Ganguli, S.; and Fei-Fei, L. 2021. Embodied intelligence via learning and evolution. *Nature communications*, 12(1): 5721.

Hane, C.; Zach, C.; Cohen, A.; Angst, R.; and Pollefeys, M. 2013. Joint 3D scene reconstruction and class segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 97–104.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hou, J.; Graham, B.; Nießner, M.; and Xie, S. 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 15587–15597.

Hu, Q.; Yang, B.; Fang, G.; Guo, Y.; Leonardis, A.; Trigoni, N.; and Markham, A. 2022. Sqn: Weakly-supervised semantic segmentation of large-scale 3d point clouds. In *European Conference on Computer Vision*, 600–619. Springer.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2020. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11108–11117.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 17853–17862.

Hui, L.; Yang, H.; Cheng, M.; Xie, J.; and Yang, J. 2021. Pyramid point cloud transformer for large-scale place recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 6098–6107.

Jiang, H.; Dang, Z.; Wei, Z.; Xie, J.; Yang, J.; and Salzmann, M. 2023a. Robust outlier rejection for 3d registration with variational bayes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1148–1157.

Jiang, H.; Salzmann, M.; Dang, Z.; Xie, J.; and Yang, J. 2023b. Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems*, 36: 21285–21297.

Jiang, H.; Shen, Y.; Xie, J.; Li, J.; Qian, J.; and Yang, J. 2021. Sampling network guided cross-entropy method for unsupervised point cloud registration. In *Proceedings of the IEEE international conference on computer vision*, 6128–6137.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35: 23593–23606.

Le, D.-T.; Shi, H.; Cai, J.; and Rezatofighi, H. 2024. Diffusion model for robust multi-sensor fusion in 3d object detection and bev segmentation. In *European Conference on Computer Vision*, 232–249. Springer.

Li, M.; Xie, Y.; Shen, Y.; Ke, B.; Qiao, R.; Ren, B.; Lin, S.; and Ma, L. 2022. Hybridcr: Weakly-supervised 3d point cloud semantic segmentation via hybrid contrastive regularization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 14930–14939.

Lian, J.; Du, X.; Liu, J.; Hui, L.; and Yang, J. 2025. Cross-modal driven object restoration for 3D point cloud backdoor defense. *IEEE Transactions on Information Forensics and Security*, 20: 11006–11018.

Liu, L.; Zhuang, Z.; Huang, S.; Xiao, X.; Xiang, T.; Chen, C.; Wang, J.; and Tan, M. 2023. Cpcm: Contextual point cloud modeling for weakly-supervised point cloud semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 18413–18422.

Liu, Z.; Qi, X.; and Fu, C.-W. 2021. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1726–1736.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2837–2845.

Pan, Z.; Gao, W.; Liu, S.; and Li, G. 2024. Distribution guidance network for weakly supervised point cloud semantic segmentation. *Advances in neural information processing systems*, 37: 32400–32420.

Pan, Z.; Zhang, N.; Gao, W.; Liu, S.; and Li, G. 2025. Point cloud semantic segmentation with sparse and inhomogeneous annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6354–6362.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qian, G.; Li, Y.; Peng, H.; Mai, J.; Hammoud, H.; Elhoseiny, M.; and Ghanem, B. 2022. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 35: 23192–23204.

Qu, W.; Shao, Y.; Meng, L.; Huang, X.; and Xiao, L. 2024. A conditional denoising diffusion probabilistic model for point cloud upsampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20786–20795.

Qu, W.; Wang, J.; Gong, Y.; Huang, X.; and Xiao, L. 2025. An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27325–27335.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. pmlr.

Sun, Y.; Lian, J.; Yang, J.; and Luo, L. 2025a. Controllable-LPMoE: Adapting to Challenging Object Segmentation via Dynamic Local Priors from Mixture-of-Experts. *International Conference on Computer Vision*.

Sun, Y.; Xu, C.; Yang, J.; Xuan, H.; and Luo, L. 2024. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, 343–360.

Sun, Y.; Yan, J.; Qian, J.; Xu, C.; Yang, J.; and Luo, L. 2025b. Dual-Perspective United Transformer for Object Segmentation in Optical Remote Sensing Images. *International Joint Conference on Artificial Intelligence*.

Tang, L.; Chen, Z.; Zhao, S.; Wang, C.; and Tao, D. 2023. All points matter: Entropy-regularized distribution alignment for weakly-supervised 3d segmentation. *Advances in Neural Information Processing Systems*, 36: 78657–78673.

Tran, A.-T.; Le, H.-S.; Lee, S.-H.; and Kwon, K.-R. 2024. Pointct: Point central transformer network for weakly-supervised point cloud semantic segmentation. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 3556–3565.

Wang, C.; Fang, X.; and Tiwari, P. 2025. DyPolySeg: Taylor Series-Inspired Dynamic Polynomial Fitting Network for Few-shot Point Cloud Semantic Segmentation. In *Forty-second International Conference on Machine Learning*.

Wang, C.; He, S.; Fang, X.; Han, J.; Liu, Z.; Ning, X.; Li, W.; and Tiwari, P. 2025a. Point Clouds Meets Physics: Dynamic Acoustic Field Fitting Network for Point Cloud Understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22182–22192.

Wang, C.; He, S.; Fang, X.; Wu, M.; Lam, S.-K.; and Tiwari, P. 2025b. Taylor series-inspired local structure fitting network for few-shot point cloud semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7527–7535.

Wang, C.; Wu, M.; Lam, S.-K.; Ning, X.; Yu, S.; Wang, R.; Li, W.; and Srikanthan, T. 2024a. Gpsformer: A global perception and local structure fitting-based transformer for point cloud understanding. In *European conference on computer vision*, 75–92. Springer.

Wang, T.; Mao, X.; Zhu, C.; Xu, R.; Lyu, R.; Li, P.; Chen, X.; Zhang, W.; Chen, K.; Xue, T.; et al. 2024b. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 19757–19767.

Wang, Z.; Yan, Z.; Pan, J.; Gao, G.; Zhang, K.; and Yang, J. 2025c. DORNet: A Degradation oriented and regularized network for blind depth super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15813–15822.

Wang, Z.; Yan, Z.; and Yang, J. 2024. Sgnet: Structure guided network via gradient-frequency awareness for depth map super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5823–5831.

Wu, X.; Jiang, L.; Wang, P.-S.; Liu, Z.; Liu, X.; Qiao, Y.; Ouyang, W.; He, T.; and Zhao, H. 2024. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4840–4851.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, 574–591. Springer.

Xu, X.; and Lee, G. H. 2020. Weakly supervised semantic point cloud segmentation: Towards 10x fewer labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 13706–13715.

Yang, C.-K.; Wu, J.-J.; Chen, K.-S.; Chuang, Y.-Y.; and Lin, Y.-Y. 2022. An mil-derived transformer for weakly supervised point cloud segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11830–11839.

Zhang, Y.; Li, Z.; Xie, Y.; Qu, Y.; Li, C.; and Mei, T. 2021. Weakly supervised semantic segmentation for large-scale point cloud. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3421–3429.

Zheng, X.; Huang, X.; Mei, G.; Hou, Y.; Lyu, Z.; Dai, B.; Ouyang, W.; and Gong, Y. 2024. Point cloud pre-training with diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 22935–22945.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and Du, Y. 2023. Curricular contrastive regularization for physics-aware single image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5785–5794.