

Towards Foundation Models for 3D Scene Understanding: Instance-Aware Self-Supervised Learning for Point Clouds

Bin Yang^{1,2}, Mohamed Abdelsamad¹, Miao Zhang¹, Alexandru Paul Condurache^{1,2}

¹Bosch Research, Robert Bosch GmbH, Stuttgart, Germany

²Institute for Neuro- and Bioinformatics, University of Lübeck, Lübeck, Germany

{Bin.Yang3, Mohamed.Abdelsamad, Miao.Zhang5, AlexandruPaul.Condurache}@de.bosch.com

Abstract

Recent advances in self-supervised learning (SSL) for point clouds have substantially improved 3D scene understanding without human annotations. Existing approaches emphasize semantic awareness by enforcing feature consistency across augmented views or by masked scene modeling. However, the resulting representations transfer poorly to instance localization, and often require full finetuning for strong performance. Instance awareness is a fundamental component of 3D perception, thus bridging this gap is crucial for progressing toward true 3D foundation models that support all downstream tasks on 3D data. In this work, we introduce *PointINS*, an instance-oriented self-supervised framework that enriches point cloud representations through geometry-aware learning. *PointINS* employs an orthogonal offset branch to jointly learn high-level semantic understanding and geometric reasoning, yielding instance awareness. We identify two consistent properties essential for robust instance localization and formulate them as complementary regularization strategies, *Offset Distribution Regularization (ODR)*, which aligns predicted offsets with empirically observed geometric priors, and *Spatial Clustering Regularization (SCR)*, which enforces local coherence by regularizing offsets with pseudo-instance masks. Through extensive experiments across five datasets, *PointINS* achieves on average +3.5% mAP improvement for indoor instance segmentation and +4.1% PQ gain for outdoor panoptic segmentation, paving the way for scalable 3D foundation models.

1. Introduction

Self-supervised learning (SSL) has achieved great success in 2D visual representation learning [3, 5, 14, 36] over the past decade. It enables models to extract powerful features from large-scale unlabeled data that generalize well across diverse downstream tasks. Extending this paradigm to 3D

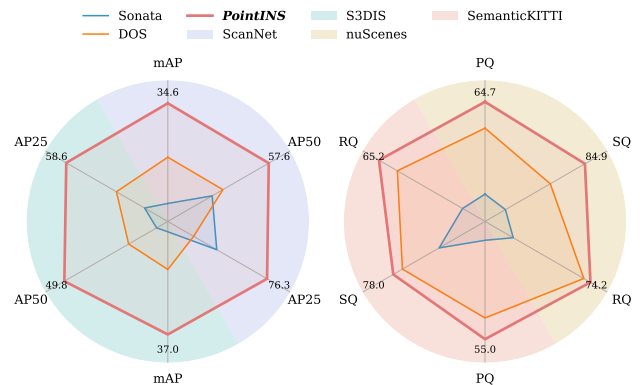


Figure 1. We achieve superior results over state-of-the-art self-supervised approaches (DOS [2] and Sonata [34]) on both indoor instance (left) and outdoor panoptic segmentation (right).

scene understanding holds significant potential, especially in domains such as autonomous driving and robotics, where large amounts of 3D data are available but manual annotation remains costly and labor-intensive [10, 17, 22, 30].

Recent works have demonstrated promising progress in 3D SSL by applying contrastive [25, 26, 35] and masked modeling [2, 15, 34]. These methods achieve impressive performance on semantic segmentation benchmarks. Their success largely stems from enforcing multi-view consistency, where models are trained to produce aligned embeddings across independently augmented views of the same point cloud [5, 15, 34]. This objective effectively captures global semantic structure (see Fig. 2). Nevertheless, such semantic-driven objectives inherently overlook instance awareness, an ability of a representation to preserve fine-grained geometric relationships. Consequently, these approaches underperform on instance-oriented tasks such as instance and panoptic segmentation. Closing this gap is essential for advancing unified 3D foundation models that can generalize across diverse tasks and domains.

A key challenge lies in balancing semantic invariance with the geometric sensitivity needed for learning instance awareness. Prior works [15, 34] have identified that point

cloud SSL can collapse to trivial cues like normals or poses of points, thus, most methods enforce strong invariance to avoid such geometric shortcuts. While we acknowledge this concern, we argue that the geometric proximity required for instance-aware learning represents a high-level relational property rather than a low-level geometric cue. We refer to this capability as *geometric reasoning*. Beyond learning consistency for semantics, SSL features should encode where a point should direct, ideally toward the centroid of the instance it belongs to. Our hypothesis aligns with supervised instance and panoptic segmentation frameworks [16, 20, 21, 42], where semantic categorization guides subsequent geometric grouping into class-agnostic instances. These architectures typically use a shared backbone with parallel semantic and offset branches: semantics restrict the candidate regions for instance boundaries, while offsets refine the spatial separation. By reinforcing each other, they jointly enhance the model’s overall understanding of 3D scenes.

Building upon this insight, we propose *PointINS*, the first self-supervised framework explicitly designed to learn both semantic consistency and geometric reasoning. Our framework augments the SSL framework with an additional instance-localization branch that predicts point-wise geometric offsets toward underlying instance centers.

However, predicting the offsets without labels remains challenging. Without ground-truth guidance, the model struggles to infer meaningful geometric relationships, and may even collapse to trivial solutions [3, 5]. At the same time, we observe that offsets inherently exhibit several stable statistical and structural properties that can be exploited to guide and regularize the learning process. To this end, we introduce two complementary regularization strategies: **Offset Distribution Regularization (ODR)** provides a global constraint by aligning predicted offsets with empirical geometric priors observed in real scenes, while **Spatial Clustering Regularization (SCR)** introduces a local constraint by enforcing consistent centroid alignment among

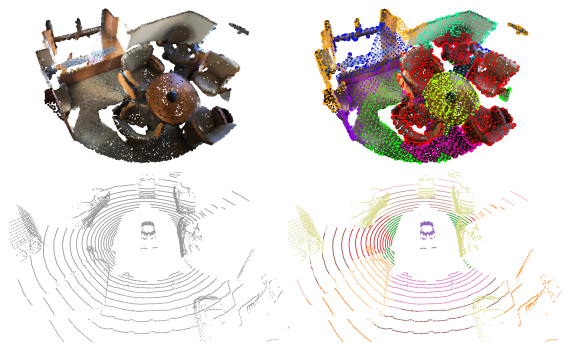


Figure 2. Example of an indoor/outdoor scene and K-means clustering over their point features extracted from a self-supervised pre-trained model [34].

points through pseudo-instance masks derived from the model’s semantic understanding. Together, these components provide both global distributional alignment and local geometric coherence to enable robust instance-aware representation learning.

We evaluate *PointINS* extensively on both indoor and outdoor scene datasets, covering semantic, instance, and panoptic segmentation under linear probing, decoder probing, and finetuning protocols. Our method achieves state-of-the-art performance among self-supervised approaches, yielding 2.5–4.6% mAP gains on indoor instance segmentation and 3.4–4.8% PQ gains on outdoor panoptic segmentation. To summarize, this work contains the following key contributions:

- We introduce a novel self-supervised training framework for point clouds that jointly learns semantic consistency and geometric reasoning.
- We identify two consistent properties essential for robust instance awareness and introduce complementary regularization strategies that prevent model collapse while enabling effective representation learning.
- Our method surpasses the state-of-the-art methods by a clear margin on indoor instance and outdoor panoptic segmentation, representing a step toward unified 3D foundation models for holistic scene understanding.

2. Related Works

2.1. 3D Instance & Panoptic Segmentation

3D instance segmentation methods typically rely on dense supervision to learn point-wise offsets or embeddings that facilitate object grouping. Early approaches [6, 16, 31] predict offset vectors pointing toward object centers and cluster points using mean-shift algorithms or connected components. Panoptic segmentation extends this paradigm by jointly predicting semantic labels and instance identities for every point in the scene [12, 23]. Many of these methods adopt a dual-branch design, where semantic and instance segmentation are learned simultaneously with a shared backbone. While highly effective, they depend on a large amount of instance and semantic annotations for fully supervised training. In this work, we seek to learn both semantic and instance-aware representations directly from unlabeled point clouds for enhancing performance on diverse segmentation tasks.

2.2. Point Cloud Self-supervised Learning

Early self-supervised learning (SSL) approaches [32, 35, 39] adopt primarily 3D sparse convolutions with U-Net-like architectures [8, 9]. These works fall broadly into two categories: contrastive learning [7, 25, 26, 32, 35] and occupancy reconstruction [1, 18, 24, 29, 39]. In contrastive frameworks, points or patches from the same scene are

treated as positive pairs and encouraged to produce similar embeddings. However, these objectives lack spatial cues to distinguish nearby objects belonging to the same semantic class. Occupancy reconstruction approaches learn to reconstruct missing geometry from visible patches. Although this enhances geometric understanding, the optimization primarily targets local completeness rather than instance awareness. Moreover, both paradigms typically require full model finetuning for downstream tasks, which is computationally expensive for large-scale 3D data. Recently, transformer-based backbones have demonstrated strong generalization and scalability across different 3D perceptual tasks [13, 33, 37, 38, 40]. This advancement motivates newer SSL methods to adopt them [34]. These methods show strong performance in semantic segmentation under linear probing setting by enforcing feature consistency across independently augmented views. However, this strong focus promotes semantic-compact but geometry-entangled representations, which suppresses the intra-class variation needed for instance-level perception.

3. Method

In this section, we detail the technical components of our approach. We first introduce the overall teacher–student architecture, which employs a prototype-based self-distillation mechanism as the semantic branch for learning category-level understanding. We then describe the offset branch, where the model learns to reason about geometric relationships among points. Finally, we present two complementary regularization strategies that stabilize training and strengthen the geometric reasoning capability essential for offset-based instance-aware learning.

3.1. Preliminary

Our framework builds upon a widely adopted teacher–student self-distillation paradigm in recent 3D SSL methods with hierarchical decoder-free architectures [2, 34]. As shown in Fig. 3, a point cloud $\mathcal{P} = (x_i, f_i)_{i=1}^N$, where x_i denotes the coordinates and f_i the feature of point i , is randomly augmented into two distinct views, $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$. Each view undergoes independent spatial and photometric augmentations. Then, we randomly mask a subset of points to create a visible subset \mathcal{P}_v as inputs for the student network, while the teacher receives the full point cloud \mathcal{P} . Both networks share identical architectures, and the teacher’s parameters are updated by the exponential moving average (EMA) of the student’s.

For the semantic branch, we adopt a prototype-based clustering mechanism [5]. Both teacher and student encode augmented views into point-wise embeddings, which are projected onto a set of learnable category prototypes $\mathcal{Q} = q_{k=1}^K$ by computing the similarity between embeddings and prototypes. Then, the similarity is transformed

into soft assignments via a temperature-scaled softmax. The student is trained to align its distributions with those of the teacher via a Kullback–Leibler (KL) divergence loss. As the student process only visible subset of points, we select the corresponding tokens in teacher for computing the loss. To enforce the semantic consistency across views, the loss is computed second time for cross-view distillation ($\mathcal{P}^{(2)} \rightarrow \mathcal{P}_v^{(1)}$ and $\mathcal{P}^{(1)} \rightarrow \mathcal{P}_v^{(2)}$).

3.2. Learning Offset without Labels

To inject instance awareness into self-supervised learning, we introduce an offset branch that predicts a 3D vector for each point, which directs toward the geometric center of its underlying instance. Unlike the semantic branch, which learns view-consistent embeddings, this branch captures view-dependent geometric relationships and is therefore sensitive to spatial transformations such as rotation, flipping, and scaling. To maintain geometric consistency, we track the transformations applied during data augmentation and invert them to map the predicted offsets back to the original coordinate.

Offset Distribution Regularization (ODR) Regressing offsets without supervision can lead to collapsed predictions [5]. Therefore, directly introducing offset learning into the self-supervised stage risks unstable optimization. To mitigate this issue, we introduce Offset Distribution Regularization (ODR), which constrains the predicted offsets to match statistically grounded distributions observed in real scenes. Each offset vector $\mathcal{O} \in \mathbb{R}^3$ can be decomposed into two components: (1) the magnitude, which measures the distance to the instance centroid, and (2) the direction, a unit vector pointing toward that centroid. From our analysis of existing scene datasets, we observe two consistent trends: (1) offset magnitudes follow a stable long-tailed distribution reflecting scene layout and object scale, and (2) offset directions are approximately uniformly distributed over the unit sphere (see Fig. 4). These observations motivate us to adopt them as global statistical priors for regularizing the predicted offsets.

We achieve this using the **probabilistic integral transform (PIT)**, a non-parametric method that maps scalar samples to a target distribution while maintaining their relative order. Given the predicted offset magnitudes $\{\mathcal{M}_i\}_{i=1}^N \in \mathbb{R}$, we convert them to PIT-normalized values as follows. Let $\pi(i)$ denote the rank of \mathcal{M}_i in ascending order:

$$\pi(i) = \text{rank}(\mathcal{M}_i), \quad 1 \leq \pi(i) \leq N,$$

which we convert into probability levels via

$$u_i = \frac{\pi(i) - 0.5}{N}, \quad u_i \in (0, 1).$$

We then transform these into target-aligned magnitudes by applying the inverse cumulative distribution function (CDF)

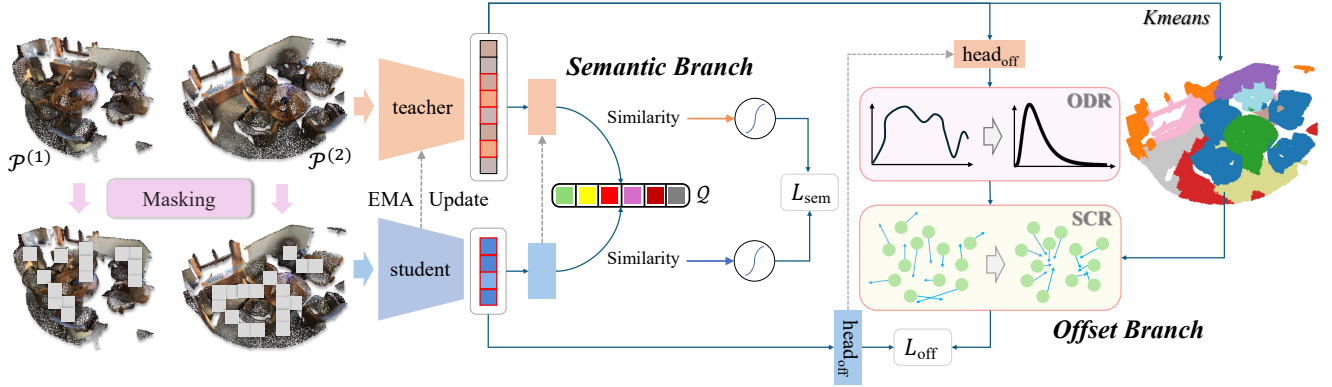


Figure 3. Overview of *PointINS*: A point cloud is augmented to two independent views and they are randomly masked. The teacher processes the full input and the student receives only visible points. Both networks share the same architecture. In the semantic branch, features are computed similarity with prototypes \mathcal{Q} . A KL-divergence loss L_{sem} is then applied for distillation. In the offset branch, an offset head maps features into 3D offset vectors. Teacher offsets are first regularized by ODR to align with empirically observed geometric priors. Next, segments obtained from K-means-clustering are used to extract pseudo-instance masks. Those masks help to enhance instance awareness by regularizing local coherence of points. Finally, an offset loss L_{off} is computed as the second distillation.

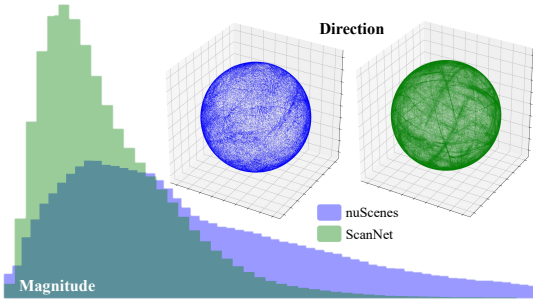


Figure 4. Offset distributions of ScanNet [11] and nuScenes [12]

F^{-1} of the empirical distribution observed in real scenes:

$$\tilde{\mathcal{M}}_i = F^{-1}(u_i).$$

This procedure strictly preserves rank order (i.e., $\mathcal{M}_i < \mathcal{M}_j \iff \tilde{\mathcal{M}}_i < \tilde{\mathcal{M}}_j$) while aligning the predicted magnitudes to the desired long-tailed distribution. Since PIT operates on scalar values, we apply it independently to each coordinate of direction $\{\mathcal{D}_i\}_{i=1}^N \in \mathbb{R}^3$ to match the uniform distribution on each axis. The complete offset after PIT-normalization is then reconstructed as

$$\tilde{\mathcal{O}}_i = \tilde{\mathcal{M}}_i \cdot \tilde{\mathcal{D}}_i.$$

This regularization encourages geometrically plausible offset predictions while preventing model collapse, thereby stabilizing training without labels.

Spatial Clustering Regularization (SCR) While effective, ODR alone cannot reason about local coherence for instance localization, i.e. points in the local neighborhood should predict offsets converging toward a common centroid when they inherently belong to the same instance. Without additional constraint, predicted offsets may remain scattered irregularly. To address the limitation, we introduce another regularization technique, so-called Spatial

Clustering Regularization (SCR). In our empirical study, we found that features learned by recent SSL frameworks naturally exhibit strong semantic awareness even in the early training stage (see Fig. 5). Inspired by this observation, we use these features to generate pseudo-instance masks that guide the model toward instance-aware learning at the pre-training stage.

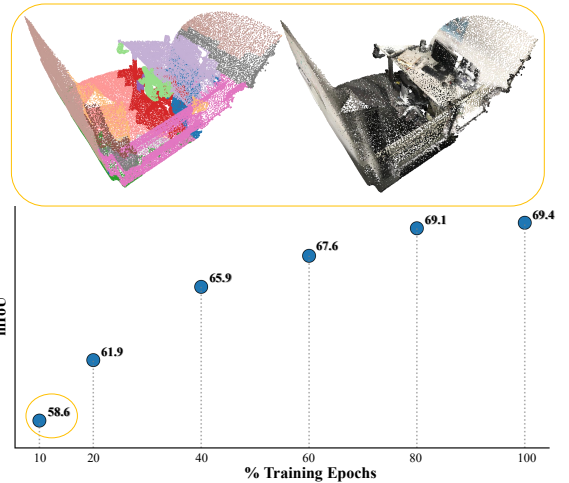


Figure 5. We monitor the linear probing (LP) performance of a model pre-trained with a recent SSL method [34]. Remarkably, the model reaches 85% of its final LP performance within just 10% of the total training.

Specifically, we begin by applying K-means clustering to the point-wise features $\mathbf{F} = \{f_i\}_{i=1}^N$ extracted by the teacher network and obtain K class-wise segments:

$$\mathcal{S} = \{S_1, S_2, \dots, S_K\} \quad (1)$$

For each point, we compute its predicted centroid by adding the PIT-normalized offset to its 3D coordinate:

$$\hat{c}_i = x_i + \tilde{\mathcal{O}}_i, \quad (2)$$

Within each segment S_k , we further refine the structure by building a local k -nearest-neighbor graph over these predicted centroids and applying Breadth-First Search (BFS) to break the cluster into multiple connected components:

$$\mathcal{I}_k = \{I_{k,1}, I_{k,2}, \dots\}, \quad I_{k,j} \subseteq S_k. \quad (3)$$

These spatially coherent components serve as pseudo-instances. For each pseudo-instance $I_{k,j}$, we update the centroid as:

$$\bar{c}_{k,j} = \frac{1}{|I_{k,j}|} \sum_{i \in I_{k,j}} x_i, \quad (4)$$

The updated centroid $\bar{c}_{k,j}$ then becomes the new supervision target for each point, and the corresponding target offset is recalculated as:

$$\mathcal{O}_i^* = \bar{c}_{k,j} - x_i, \quad (5)$$

This regularization reinforces local geometric consistency and further stabilize the instance-aware learning in the self-supervised setting.

Complementarity of Two Regularization Strategies SCR enforces local coherence by grouping spatially adjacent points into pseudo-instances, thereby complementing the global structural regularization of ODR. Conversely, ODR enhances the stability of SCR. Since SCR relies on spatial proximity for instance cutting, scattered or unstable offset predictions can lead to irregular grouping. ODR addresses this by constraining the offsets via geometric priors, which provides a stable geometric anchor that guides SCR toward consistent clustering. This complementarity between these two regularization strategies is demonstrated qualitatively in Fig. 6 and quantitatively in Sec. 4.

Offset Self-distillation After deriving refined centroids through ODR and SCR, we use them to supervise the student network via self-distillation. Following the practice in instance segmentation methods [16, 42], the offset loss consists of an ℓ_1 loss that penalizes deviations in offset magnitude, and a cosine similarity loss that aligns the direction of predicted offsets with the target. Formally, the offset loss is:

$$\mathcal{L}_{\text{offset}} = \frac{1}{N} \sum_{i=1}^N (\|o_i - \mathcal{O}_i^*\|_1 + (1 - \cos(o_i, \mathcal{O}_i^*))), \quad (6)$$

where o_i is the predicted offset of student network and \mathcal{O}_i^* is the updated target offset. Same as the semantic branch, a cross-view loss is applied to enforce offset consistency between two augmented views. For stable optimization, offset loss is delayed to join the training after few warm-up epochs and a loss weight λ_{off} is integrated for balancing. The full procedure is summarized in Algo. 1.

Design Rationale Our key insight is to formulate instance-aware learning as a regularized self-distillation problem,

Algorithm 1 Pre-training procedure

Require: Full Point cloud \mathcal{P} and Visible Subset \mathcal{P}_v , Teacher T , Student S , Prototypes \mathcal{Q} , Projection Head \mathcal{H} , Offset Head \mathcal{H}_{off}

1: **Student/Teacher features:**

$$\mathbf{F}_S^{(1)} \leftarrow S(\mathcal{P}_v^{(1)}), \quad \mathbf{F}_S^{(2)} \leftarrow T(\mathcal{P}_v^{(2)})$$

$$\mathbf{F}_T^{(1)} \leftarrow T(\mathcal{P}^{(1)}), \quad \mathbf{F}_T^{(2)} \leftarrow T(\mathcal{P}^{(2)})$$

2: **Semantic self-distillation:**

$$L_{\text{sem}} \leftarrow \text{KLDiv}(\sigma(\mathcal{H}(\mathbf{F}_S^{(1)}), \mathcal{Q}), \sigma(\mathcal{H}(\mathbf{F}_T^{(1)}), \mathcal{Q})),$$

where $\sigma(\cdot)$ is softmax operator.

3: **Offset distribution regularization (ODR):**

$$\hat{\mathcal{O}}_T \leftarrow \mathcal{H}_{\text{off}}(\mathbf{F}_T^{(1)}); \quad \tilde{\mathcal{O}}_T \leftarrow \text{ODR}(\hat{\mathcal{O}}_T)$$

$$\hat{c}_i \leftarrow x_i + \tilde{\mathcal{O}}_{T,i}$$

4: **Spatial clustering regularization (SCR):**

$$\{S_k\} \leftarrow \text{KMeans}(\mathbf{F}_T^{(1)})$$

for each segment S_k do: $\{I_{k,j}\} \leftarrow \text{BFS}(\{\hat{c}_i\}_{i \in S_k})$

5: **Update the centroids**

$$\bar{c}_{k,j} \leftarrow \frac{1}{|I_{k,j}|} \sum_{i \in I_{k,j}} x_i$$

6: **Get the new teacher offsets**

$$\mathcal{O}_i^* \leftarrow \bar{c}_{k,j} - x_i \quad \text{for all } i \in I_{k,j}$$

7: **Offset self-distillation:**

$$o_S \leftarrow \mathcal{H}_{\text{off}}(\mathbf{F}_S^{(1)})$$

$$L_{\text{off}} = \mathcal{L}_{\text{offset}}(o_S, \mathcal{O}^*)$$

8: **Optimization:**

$$L \leftarrow L_{\text{sem}} + \lambda_{\text{off}} L_{\text{off}}$$

Update S by backprop and T by EMA

where ODR and SCR are applied on the teacher side to progressively refine offset predictions toward geometrically valid solutions, providing a stable supervisory signal without interfering with the student’s representation learning. We choose point-wise offsets as the new target for their simplicity, enabling seamless integration into existing SSL frameworks with minimal additional model capacity.

4. Experiment

Implementation Details We build our method upon the DOS [2] pre-training framework. The backbone follows a lightweight, decoder-free variant of Point Transformer V3 (PTv3) [33]. Multi-scale features from the encoder are up-sampled to a common resolution and concatenated. The offset branch consists of two MLP layers that map the concatenated features to 3D offset vectors. For downstream evaluation, we follow PointGroup [16] for training and test under three standard protocols: linear probing, where the backbone is frozen and only a head is trained; decoder probing, where a standard decoder is trained; and full finetuning. In SCR, we use $K = 20$ clusters for the K-means step. We set the offset loss weight $\lambda_{\text{off}} = 0.25$ and apply a warm-up schedule for the first 10% of training epochs. Additional implementation details are provided in the Appendix.

Method	ScanNet val				ScanNet200 val				S3DIS Area5			
	SemSeg.	InsSeg.			SemSeg.	InsSeg.			SemSeg.	InsSeg.		
		mIoU	mAP	AP ₅₀		AP ₂₅	mIoU	mAP		AP ₅₀	AP ₂₅	mIoU
PTv3 (sup.)	77.6	40.9	61.7	77.5	35.3	24.0	34.1	40.8	73.4	40.2	52.1	61.2
SegContrast [26](lin.)	38.4	6.4	13.7	30.6	12.4	1.8	4.7	10.2	-	-	-	-
PSA [25] (lin.)	42.9	9.7	20.4	41.9	13.9	2.4	5.3	11.4	-	-	-	-
NOMAE [1] (lin.)	47.5	9.5	20.0	42.0	14.7	2.9	6.7	12.7	-	-	-	-
Sonata [34] (lin.)	67.4	25.0	46.1	64.6	26.9	8.7	17.5	25.2	69.3	24.2	33.8	47.5
DOS [2] (lin.)	72.8	<u>28.7</u>	<u>49.8</u>	<u>68.7</u>	<u>29.1</u>	<u>10.9</u>	<u>20.6</u>	<u>27.9</u>	<u>70.6</u>	<u>28.6</u>	<u>41.0</u>	<u>52.3</u>
<i>PointINS</i> (lin.)	<u>72.4</u>	32.1	55.2	73.6	29.6	13.4	24.9	35.8	71.0	33.2	45.3	59.4
SegContrast [26](dec.)	67.1	30.5	50.3	66.9	24.8	10.5	18.0	25.5	-	-	-	-
PSA [25] (dec.)	67.1	28.9	47.3	64.5	24.3	10.9	18.0	25.5	-	-	-	-
NOMAE [1] (dec.)	68.0	28.5	49.3	68.0	30.5	15.8	24.8	31.8	-	-	-	-
Sonata [34] (dec.)	75.5	37.1	57.8	74.2	31.6	17.9	27.8	34.8	71.8	36.8	48.3	60.6
DOS [2] (dec.)	76.9	<u>38.9</u>	<u>60.2</u>	<u>75.5</u>	<u>33.7</u>	<u>18.8</u>	<u>28.5</u>	<u>36.2</u>	73.0	<u>37.2</u>	47.4	60.1
<i>PointINS</i> (dec.)	<u>76.7</u>	40.2	62.5	77.2	33.9	21.3	32.2	38.3	<u>72.9</u>	39.1	51.7	61.7
SegContrast [26](fin.)	75.5	39.8	60.5	75.2	33.8	19.6	29.2	36.5	-	-	-	-
PSA [25] (fin.)	76.2	38.8	59.7	74.3	33.5	19.4	28.0	35.6	-	-	-	-
NOMAE [1] (fin.)	75.3	37.0	59.5	76.4	33.0	19.3	28.4	36.0	-	-	-	-
Sonata [34] (fin.)	77.2	39.5	61.1	76.7	34.4	20.0	28.5	35.8	73.4	41.3	53.3	62.2
DOS [2] (fin.)	<u>78.7</u>	<u>40.5</u>	<u>62.0</u>	<u>77.3</u>	36.7	<u>22.5</u>	<u>33.5</u>	<u>38.6</u>	74.2	40.4	52.0	60.3
<i>PointINS</i> (fin.)	79.0	41.5	63.7	78.4	<u>36.6</u>	25.1	34.8	41.5	<u>73.6</u>	42.9	57.2	64.1

Table 1. Comparison of self-supervised methods on various indoor scene datasets: ScanNet [11], ScanNet200 [28] and S3DIS [19]. For fair comparison, all models deploy the same backbone of Point Transformer v3 (PTv3) [33] and no additional data is used during pre-training. The **best** and second best results in each setting are highlighted in **bold** and underline, respectively.

Method	nuScenes val				SemanticKITTI val			
	SemSeg.	PanSeg.			SemSeg.	PanSeg.		
		mIoU	PQ	SQ		RQ	mIoU	PQ
PTv3 (sup.)	80.3	69.9	86.3	80.5	69.1	58.2	78.7	67.7
SegContrast [26](lin.)	37.8	25.4	69.8	33.4	-	-	-	-
PSA [25] (lin.)	44.5	30.1	73.9	38.6	-	-	-	-
NOMAE [1] (lin.)	64.7	45.5	77.0	56.4	29.8	17.1	54.7	24.9
Sonata [34] (lin.)	59.2	50.7	79.8	61.6	46.5	34.5	65.5	44.2
DOS [2] (lin.)	74.1	57.4	82.8	68.5	67.5	49.6	71.8	60.9
<i>PointINS</i> (lin.)	74.4	62.2	84.5	72.8	<u>66.9</u>	52.8	73.7	63.2
SegContrast [26](dec.)	73.1	60.7	83.9	71.7	-	-	-	-
PSA [25] (dec.)	74.6	62.2	84.1	73.3	-	-	-	-
NOMAE [1] (dec.)	80.1	69.0	<u>85.6</u>	<u>79.3</u>	64.3	52.4	73.6	62.3
Sonata [34] (dec.)	76.8	66.0	85.2	76.8	64.5	55.1	77.5	64.8
DOS [2] (dec.)	79.2	69.1	85.3	79.0	68.6	<u>56.7</u>	<u>76.3</u>	<u>65.2</u>
<i>PointINS</i> (dec.)	80.0	70.8	86.6	80.6	<u>68.2</u>	59.2	78.2	68.7
SegContrast [26](fin.)	78.0	69.3	85.9	80.3	-	-	-	-
PSA [25] (fin.)	78.7	69.2	86.0	80.0	-	-	-	-
NOMAE [1] (fin.)	81.8	<u>71.0</u>	<u>86.5</u>	80.8	71.6	<u>59.5</u>	76.4	<u>68.9</u>
Sonata [34] (fin.)	79.6	70.0	86.4	80.5	69.6	58.2	<u>78.5</u>	67.7
DOS [2] (fin.)	81.5	70.5	86.0	<u>81.1</u>	73.1	59.2	76.3	68.6
<i>PointINS</i> (fin.)	81.1	72.3	87.4	82.3	<u>72.7</u>	60.5	79.6	69.2

Table 2. Comparison of self-supervised methods on the outdoor scene datasets: nuScenes [12] and SemanticKITTI [4].

Comparative Study Tab. 1 compares *PointINS* with recent self-supervised learning (SSL) methods across three indoor benchmarks: ScanNet [11], its long-tailed variant ScanNet200 [28], and S3DIS [19]. Overall, *PointINS* consistently outperforms prior state-of-the-art SSL frameworks across all datasets and evaluation settings. Remarkably, our method achieves substantial improvements on all three metrics of instance segmentation. On ScanNet and S3DIS, with only linear probing, *PointINS* attains 80–90% of the supervised performance, highlighting the strong generalization of the learned representations on instance-level tasks.

To further validate the generalization of our method, we benchmark on two large-scale outdoor scene datasets:

nuScenes [12] and SemanticKITTI [4]. We evaluate on the panoptic segmentation task, where both semantic and instance-level predictions jointly contribute to the final score. As shown in Tab. 2, *PointINS* consistently outperforms existing SSL baselines, achieving notable gains of +4.8 PQ on nuScenes and +3.2 PQ on SemanticKITTI. In addition to the improvements in instance-level segmentation, our method also preserves strong semantic segmentation performance, remaining competitive with state-of-the-art approaches. These results demonstrate that *PointINS* not only enhances indoor scene understanding but also generalizes robustly to complex outdoor environments.

Component Design We conduct an ablation study to validate the effectiveness of each component in our method, as shown in Tab. 4. Simply adding a new branch for offset learning yields only marginal gains. Although applying either regularization strategy alone leads to noticeable improvements, their individual effects remain limited. When the two strategies are combined, however, their complementary strengths result in a larger performance gain, achieving a +3.4% mAP on instance and +4.8% PQ increase on panoptic segmentation.

To further assess how these regularization strategies enhance geometric reasoning during pre-training, we visualize the predicted centroids in Fig. 6. The results show that the model trained with both ODR and SCR produces offsets that are spatially coherent and semantically meaningful, demonstrating a stronger capability for geometric reasoning. These observations qualitatively confirm that our regularization strategies effectively support instance-aware representation learning. Additional ablation studies on key

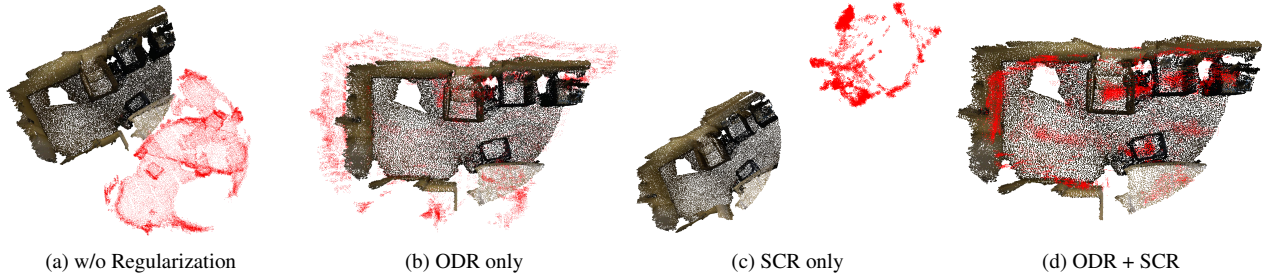


Figure 6. Visualization of predicted centroids $c_i = x_i + \mathcal{O}_i$ (no explicit regularization) from models pre-trained under four configurations. (a) Without regularization, the centroids are scattered irregularly across the scenes. (b) With ODR only, the centroids are aligned more closely with the scene structure but they are not grouped (lack of local coherence). (c) With SCR only, the centroids are partially grouped but spatially scattered. (d) With ODR and SCR, the centroids become tightly concentrated around instances.

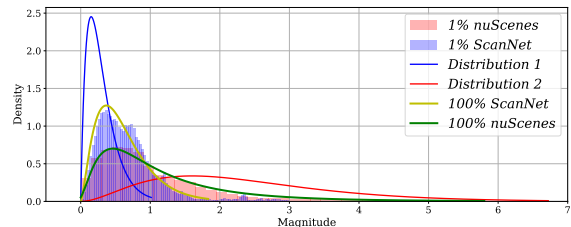
Component			InsSeg.			PanSeg.		
$+L_{\text{off}}$	ODR	SCR	mAP	AP ₅₀	AP ₂₅	PQ	SQ	RQ
			28.7	49.8	68.7	57.4	82.8	68.5
✓			28.9	49.6	69.1	58.5	82.3	69.7
✓	✓		30.2	52.1	70.8	60.4	83.6	71.5
✓		✓	30.5	51.9	71.0	60.1	83.1	71.2
✓	✓	✓	32.1	55.2	73.6	62.2	84.5	72.8

Table 3. Ablation study on *PointINS* components. Models are evaluated on ScanNet [11] for instance segmentation and nuScenes [12] for panoptic segmentation via linear probing.

hyperparameters are provided in the Appendix.

Sensitivity of ODR Offset Distribution Regularization (ODR) requires a prior distribution of instance offsets from annotated datasets, particularly their magnitudes. Since our framework focuses on self-supervised learning, we have to assume no access to labels during training. To verify the effectiveness of ODR in such scenario, we evaluate ODR’s sensitivity by testing various distribution priors, including those derived from real datasets. As shown in Tab. 5, our method exhibits strong robustness to distribution choice. Even when the prior is from the dataset with totally different scene layout (outdoor↔indoor) or it deviates significantly from the empirical one, the performance drop remains minimal while still outperforming the baseline without regularization. To further validate our self-supervised claim, we conduct an experiment in which *PointINS* is trained using offset priors estimated by clustering the 3D input points with HDBSCAN, requiring no annotations whatsoever. Performance remains consistent, confirming that an effective prior does not require fine-grained annotations, instead, a coarse estimate of typical object scales is sufficient. In practice, a long-tailed offset distribution can be reasonably assumed, as points close to instance centroids naturally dominate in 3D point clouds. We also observe that as little as 1% of annotations suffices to estimate a reliable empirical prior, suggesting that exhaustive labeling is unnecessary for achieving strong performance. To contextualize this finding, we compare against a semi-supervised baseline trained directly with 1% and 10% ground-truth labels. Notably, the semi-supervised setting only reaches comparable performance

when using approximately 10% labels, highlighting the practical advantage of our regularization-based approach.



Dist. Type	ScanNet val			nuScenes val		
	mAP	AP ₅₀	AP ₂₅	PQ	SQ	RQ
w/o Regularization	28.9	49.6	69.1	57.8	82.8	68.7
Dist. 1	31.2	53.6	74.1	60.8	83.4	72.2
Dist. 2	31.7	54.2	72.8	61.2	83.9	72.1
ScanNet	32.1	55.2	73.6	62.0	84.3	72.5
nuScenes	31.3	54.9	72.7	62.2	84.5	72.8
*HDBSCAN	31.8	54.7	73.1	62.1	84.1	72.3
Semi-supervised w/o regularization						
+1% labels	29.8	50.9	70.1	59.3	83.3	70.0
+10% labels	32.3	55.2	74.1	62.0	84.2	72.1

Table 5. Sensitivity of ODR to different distribution priors applied to offset magnitudes. *ScanNet* and *nuScenes* represent empirical distributions fitted from respective ground-truth instance annotations. Histograms are from 1% of annotations. *: Distributions are fitted from unsupervised clustering of 3D point clouds.

Framework Compatibility Our method can be seamlessly integrated into other teacher–student-based frameworks. To evaluate its generalization, we apply *PointINS* to two additional SSL frameworks and report the results in Tab. 6. Notably, although these baselines like SONATA [34] already perform strongly, *PointINS* not only enhances instance segmentation but also brings consistent improvements in semantic segmentation performance. Such improvements indicate that semantic and instance objectives are inherently complementary rather than conflicting, and further highlights the broad compatibility and adaptability of our approach within the self-supervised learning paradigm for instance-aware representation learning.

Label Efficiency We evaluate label-efficient training on nuScenes [12] under extremely low annotation regimes

Method	ScanNet val				S3DIS Area5				nuScenes val				SemanticKITTI val			
	SemSeg.	InsSeg.			SemSeg.	InsSeg.			SemSeg.	PanSeg.			SemSeg.	PanSeg.		
	mIoU	mAP	AP ₅₀	AP ₂₅	mIoU	mAP	AP ₅₀	AP ₂₅	mIoU	PQ	SQ	RQ	mIoU	PQ	SQ	RQ
PTv3 (sup.)	77.6	40.9	61.7	77.5	73.4	40.2	52.1	61.2	80.3	69.9	86.3	80.5	69.1	58.2	78.7	67.7
Sonata* [34] (lin.)	72.5	30.7	53.9	72.6	72.3	26.1	36.6	45.8	66.1	54.9	81.0	66.0	62.0	50.8	76.5	61.1
DOS* [2] (lin.)	73.9	32.5	54.6	70.9	71.7	30.1	40.6	50.4	74.8	60.9	82.6	71.5	68.3	52.6	77.1	63.2
<i>PointINS</i> * (lin.)	73.5	34.6	57.6	76.3	72.6	37.0	49.8	58.6	74.6	64.7	84.9	74.2	68.0	55.0	78.0	65.2

Table 4. Comparison of self-supervised methods with additional data for pre-training.

Method	SemSeg.	InsSeg.		
	mIoU	mAP	AP ₅₀	AP ₂₅
PSA [25] + <i>PointINS</i>	47.5 (+4.6)	14.2 (+4.5)	30.8 (+10.4)	50.0 (+8.1)
Sonata [34] + <i>PointINS</i>	68.6 (+1.2)	28.4 (+3.4)	50.3 (+4.2)	67.1 (+2.5)

Table 6. Evaluation of other SSL frameworks with integration of *PointINS*. Models are evaluated on ScanNet val set under linear probing setting. (+ Δ) denotes the performance gain.

(0.1% and 1% labels), as shown in Tab. 7. *PointINS* consistently outperforms all self-supervised baselines across all metrics of panoptic segmentation. With only 0.1% labeled data, *PointINS* achieves 34.9% in PQ, surpassing supervised performance over 10%. When the label ratio increases to 1 it further improves to 42.5%. These results demonstrate that the instance-aware representations learned by *PointINS* are highly transferable, enabling strong downstream performance even with minimal supervision.

Layout of Regularization We further investigate how different regularization layouts affect performance. As shown in Tab. 8, applying SCR before ODR leads to a clear drop in instance segmentation accuracy. This suggests that global regularization (ODR) is crucial for stabilizing offset predictions before enforcing local coherence among points. We also explore applying ODR on the student side, where the regularization directly influences gradient updates in the student backbone. This configuration again results in degraded performance, likely due to conflicting gradients and unstable optimization. In contrast, regularizing the teacher outputs provides a stable and structured supervisory signal, allowing the student to adapt smoothly without disrupting representation learning [5, 27, 34].

Multi-dataset Pre-training To explore the scalability of our framework, we expand the pre-training pool by aggregating multiple datasets. For indoor scenes, we combine ScanNet [11], S3DIS [19], and Structured3D [41], yield-

Method	0.1% labels			1% labels		
	PQ	SQ	RQ	PQ	SQ	RQ
PTv3 (sup.)	24.3	69.0	30.7	34.2	71.2	40.1
NOMAE [1] (fin.)	30.4	74.9	40.9	37.4	78.5	45.2
Sonata [34] (fin.)	27.1	71.4	34.0	34.7	72.5	42.3
DOS (fin.)	33.6	74.3	41.2	41.2	75.7	49.2
<i>PointINS</i> (fin.)	34.9	75.2	42.1	42.5	82.0	49.6

Table 7. Label efficient training on nuScenes panoptic segmentation [12].

Design	SemSeg.	InsSeg.		
	mIoU	mAP	AP ₅₀	AP ₂₅
<i>PointINS</i> [◊]	72.0 (-0.4)	30.0 (-2.1)	50.7 (-4.5)	69.5 (-4.1)
<i>PointINS</i> [†]	71.9 (-0.5)	30.9 (-1.2)	53.9 (-1.3)	72.8 (-0.8)
<i>PointINS</i>	72.4	32.1	55.2	73.6

Table 8. Comparison of different regularization layouts. All models are evaluated on ScanNet via linear probing. [◊]: order of ODR and SCR is reversed. [†]: ODR is added on the student side as a regularization loss via KL divergence.

ing approximately 24k point clouds (significantly smaller than the 140k scenes used by Sonata*). For outdoor scenes, we follow prior protocols [2, 34] and mix nuScenes [12], SemanticKITTI [4], and Waymo [23] during pre-training. The results in Tab. 4 show that performance consistently improves as the scale of pre-training data increases. Remarkably, under this scale-up setting, *PointINS* approaches supervised performance under linear probing, particularly on instance and panoptic segmentation. This strong scalability suggests that *PointINS* progresses a promising step toward unified 3D foundation models.

5. Conclusion

In this work, we introduce *PointINS*, a novel self-supervised learning framework that enables point cloud encoders to jointly learn semantic invariance and instance-aware geometric reasoning without labels. By introducing two complementary regularization strategies applied to the geometric learning, our method prevents model collapse while enriching feature representations with geometric priors and enforcing local coherence. *PointINS* achieves state-of-the-art performance in instance and panoptic segmentation across five benchmarks. We believe this work marks an important step toward scalable 3D foundation models capable of holistic scene understanding.

Limitations and Future Work While *PointINS* achieves substantial improvements over existing self-supervised methods under the linear probing setting, a noticeable gap still remains compared to fully supervised performance. Scaling up pre-training with more data or jointly using indoor and outdoor datasets could advance further progress. In addition, extending the framework to incorporate spatiotemporal cues for 4D geometric reasoning represents a promising avenue for future exploration.

Acknowledgements

We are grateful to Professor Abhinav Valada of the Robot Learning Lab at the University of Freiburg for his invaluable guidance, insightful discussions, and continuous support throughout this work. As the PhD advisor of our second author, Mohamed Abdelsamad, his expertise and thoughtful feedback significantly shaped the direction and quality of this research.

References

- [1] Mohamed Abdelsamad, Michael Ulrich, Claudius Gläser, and Abhinav Valada. Multi-scale neighborhood occupancy masked autoencoder for self-supervised learning in lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22234–22243, 2025. [2](#), [6](#), [8](#)
- [2] Mohamed Abdelsamad, Michael Ulrich, Bin Yang, Miao Zhang, Yakov Miron, and Abhinav Valada. Dos: Distilling observable softmaps of zipfian prototypes for self-supervised point cloud representation learning. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2026. [1](#), [3](#), [5](#), [6](#), [8](#)
- [3] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. [1](#), [2](#)
- [4] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019. [6](#), [8](#)
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. [1](#), [2](#), [3](#), [8](#)
- [6] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *International Conference on Computer Vision (ICCV)*, pages 15467–15476, 2021. [2](#)
- [7] Prakash Chandra Chhipa, Richa Upadhyay, Rajkumar Saini, Lars Lindqvist, Richard Nordenskjold, Seiichi Uchida, and Marcus Liwicki. Depth contrast: Self-supervised pretraining on 3dpm images for mining material classification. In *European Conference on Computer Vision (ECCV)*, pages 212–227. Springer, 2022. [2](#)
- [8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3075–3084, 2019. [2](#)
- [9] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. [2](#)
- [10] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Nova: Learning to see in novel viewpoints and domains. In *2019 International Conference on 3D Vision (3DV)*, pages 116–125. IEEE, 2019. [1](#)
- [11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [4](#), [6](#), [7](#), [8](#)
- [12] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *RAL*, 7(2):3795–3802, 2022. [2](#), [4](#), [6](#), [7](#), [8](#)
- [13] Chenheng He, Ruihuang Li, Shuai Li, and Lei Zhang. Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8417–8427, 2022. [3](#)
- [14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. [1](#)
- [15] Pedro Hermosilla, Christian Stippel, and Leon Sick. Masked scene modeling: Narrowing the gap between supervised and self-supervised learning in 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14835–14844, 2025. [1](#)
- [16] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2020. [2](#), [5](#)
- [17] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21705–21715, 2023. [1](#)
- [18] Georg Kispel, David Schinagl, Christian Fruhwirth-Reisinger, Horst Possegger, and Horst Bischof. Maeli: Masked autoencoder for large-scale lidar point clouds. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 3383–3392, 2024. [2](#)
- [19] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4558–4567, 2018. [6](#), [8](#)
- [20] Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering

- pseudo heatmap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11809–11818, 2022. 2
- [21] Xiaoyan Li, Gang Zhang, Boyue Wang, Yongli Hu, and Bao-cai Yin. Center focusing network for real-time lidar panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13425–13434, 2023. 2
- [22] Christian Löwens, Thorben Funke, André Wagner, and Alexandru Paul Condurache. Unsupervised point cloud registration with self-distillation. In *British Machine Vision Conference (BMVC)*, 2024. 1
- [23] Jieru Mei, Alex Zihao Zhu, Xinchen Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *European Conference on Computer Vision (ECCV)*, pages 53–72. Springer, 2022. 2, 8
- [24] Chen Min, Liang Xiao, Dawei Zhao, Yiming Nie, and Bin Dai. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles (IV)*, 9(7): 5150–5162, 2023. 2
- [25] Barza Nisar and Steven L Waslander. Psa-ssl: Pose and size-aware self-supervised learning on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6670–6679, 2025. 1, 2, 6, 8
- [26] Lucas Nunes, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. *RAL*, 7(2):2116–2123, 2022. 1, 2, 6
- [27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 8
- [28] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision (ECCV)*, 2022. 6
- [29] Xiaoyu Tian, Haoxi Ran, Yue Wang, and Hang Zhao. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13570–13580, 2023. 2
- [30] 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 (CVPR)*, pages 2697–2707, 2022. 1
- [31] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2708–2717, 2022. 2
- [32] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9415–9424, 2023. 2
- [33] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024. 3, 5, 6
- [34] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3, 4, 6, 7, 8
- [35] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European Conference on Computer Vision (ECCV)*, pages 574–591. Springer, 2020. 1, 2
- [36] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 1
- [37] Bin Yang and Alexandru Paul Condurache. Flares: fast and accurate lidar multi-range semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (CVPR)*, pages 3451–3461, 2026. 3
- [38] Bin Yang, Patrick Pfreundschuh, Roland Siegwart, Marco Hutter, Peyman Moghadam, and Vaishakh Patil. Tulip: Transformer for upsampling of lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15354–15364, 2024. 3
- [39] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Gd-mae: generative decoder for mae pre-training on lidar point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9403–9414, 2023. 2
- [40] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 3
- [41] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision (ECCV)*, 2020. 8
- [42] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13194–13203, 2021. 2, 5