Towards Large-scale 3D Representation Learning
with Multi-dataset Point Prompt Training

Xiaoyang Wu1 Zhuotao Tian2 Xin Wen1 Bohao Peng2 Xihui Liu1 Kaicheng Yu3,4 Hengshuang Zhao1*
1The University of Hong Kong 2The Chinese University of Hong Kong
3Westlake University 4Alibaba Group

https://github.com/Pointcept/Pointcept

Abstract
The rapid advancement of deep learning models is often attributed to their ability to leverage massive training data. In contrast, such privilege has not yet fully benefited 3D deep learning, mainly due to the limited availability of large-scale 3D datasets. Merging multiple available data sources and letting them collaboratively train a single model is a potential solution. However, due to the large domain gap between 3D point cloud datasets, such mixed supervision could adversely affect the model’s performance and lead to degenerated performance (i.e., negative transfer) compared to single-dataset training. In view of this challenge, we introduce Point Prompt Training (PPT), a novel framework for multi-dataset synergistic learning in the context of 3D representation learning that supports multiple pre-training paradigms. Based on this framework, we propose Prompt-driven Normalization, which adapts the model to different datasets with domain-specific prompts and Language-guided Categorical Alignment that decently unifies the multiple-dataset label spaces by leveraging the relationship between label text. Extensive experiments verify that PPT can overcome the negative transfer associated with synergistic learning and produce generalizable representations. Notably, it achieves state-of-the-art performance on each dataset using a single weight-shared model with supervised multi-dataset training. Moreover, when served as a pre-training framework, it outperforms other pre-training approaches regarding representation quality and attains remarkable state-of-the-art performance across over ten diverse downstream tasks spanning both indoor and outdoor 3D scenarios.

1. Introduction
The rapid advancement of deep learning models in various domains, e.g., 2D vision [27, 48, 92, 102] and natural language processing [1, 46, 66, 93], are often attributed to the availability of massive training data, which enable them to learn rich and discriminative representations and generalize well to a wide spectrum of downstream applications. Such privilege, in contrast, has not yet fully benefited 3D vision, primarily due to two challenges: previous representation learning frameworks exhibit constraints in processing larger-scale point cloud data efficiently (i.e., they build on raw frames rather than the scene-level point cloud [35, 110]), and current 3D datasets are often limited in scale (e.g., the commonly used ScanNet [21] only contains 1.6K scans, while image datasets are often at million
scale [23, 80]). As a complement to one recent work [108] which explores the first problem, we tackle the second challenge: scaling up 3D representation learning with limited data in separated domains.

A potential approach to circumvent the data scarcity issue is to merge multiple available data sources and train on them collaboratively (termed multi-dataset synergistic training) to supervise a single model, which is expected to leverage the information from all sources and learn more generalizable representations. However, large domain gaps exist between 3D datasets, and directly combining multiple data sources can lead to negative transfer, a phenomenon where differences in data distribution among the sources adversely affect the model’s performance. As shown in Tab. 1, naively joint training with merged data (ScanNet [21], S3DIS [2], and Structured 3D [124]) leads to degenerated performance on the target dataset. In other words, leveraging additional training data from other datasets could be harmful. Though similar problems have been studied in 2D scene understanding [47, 95, 99, 117, 127], the large domain gap between 3D datasets, and their sparse and heavily long-tailed nature makes it a much harder task that requires non-trivial solutions.

To tackle the challenge, we present a novel framework, termed **Point Prompt Training** (PPT), specifically designed for multi-dataset synergistic training within the 3D representation learning context (see Fig. 1a). Unlike the 2D counterparts that adopt prompt learning to adapt pre-trained models to specific downstream tasks [42, 45, 118, 126], our framework tackles pre-training directly. Moreover, the proposed framework is universal, supporting both supervised and unsupervised pre-training, and evaluation on the target dataset could be done either directly (if the target dataset is included in supervised pre-training) or via transfer learning.

Based on this framework, we explore multi-dataset synergistic training for 3D representation learning from two perspectives: learning a domain prompt adapter that allows the network to model the intrinsic variance within different data sources while maintaining optimal generalizable representations and forming a unified label space that avoids inconsistency in categorical supervision and allows aligned guidance between datasets. Multiple design options are investigated, and we adopt the **Prompt-driven Normalization** and **Language-guided Categorical Alignment** as our final strategies.

The effectiveness of PPT is demonstrated through extensive experiments, which show that our proposed method can overcome the negative transfer associated with synergistic learning and produce generalizable representations. Notably, PPT attains state-of-the-art performance across various benchmarks, including ScanNet [21] and S3DIS [2], using a shared-weight model trained on multiple indoor datasets. Additionally, it achieves comparable state-of-the-art results on SemanticKITTI [6], nuScenes [8], and Waymo [86] using a shared-weight model trained on diverse outdoor datasets. Furthermore, as a pre-training strategy, PPT outperforms other techniques in terms of representation quality, demonstrating superior performance across an array of tasks encompassing both indoor and outdoor scenarios (partially in Fig. 1b).

In conclusion, as an effort toward large-scale 3D representation learning, this work introduces the multi-dataset synergistic training setting, points out the negative transfer issue in naive baselines, and presents a unified point prompt training framework that addresses this problem with Prompt-driven Normalization and Language-guided Categorical Alignment.

### 2. Multi-dataset Synergistic Training

In this section, we briefly demonstrate the setting (Sec. 2.1) in multi-dataset synergistic training for 3D representation learning and uncover the challenges in this setup through a pilot study (Sec. 2.2).

#### 2.1. Problem Setup

**Training objective.** In the context of supervised multi-dataset synergistic learning, the objective is to learn a single model capable of effectively performing downstream tasks on multiple datasets. Specifically, denote each dataset as \(D_i = \{(x_i^j, y_i^j)\} \), where \(1 \leq i \leq n\), \(n\) stands for the number of datasets, and \((x_i^j, y_i^j)\) represents data-label pairs that construct a dataset. Our goal is to train a model \(f(x; \theta)\) parameterized by \(\theta\), such that the cumulative loss across all datasets is minimized:

\[
\min_{\theta} \frac{1}{|D_i|} \sum_{(x_i^j, y_i^j) \in D_i} \mathcal{L}(f(x_i^j; \theta), y_i^j),
\]

where \(\mathcal{L}\) denotes the sample-wise loss function. Besides, substituting the supervised loss function with an unsupervised objective allows for reformulation in the context of unsupervised learning.

**Task.** The nature of 3D scene understanding has a higher level of complexity and richer contextual information [35, 110], which requests a challenging and versatile task for developing and evaluating advanced learning techniques. Specifically, we mainly target scene-level semantic segmentation for supervised training, which requires dense labeling on individual points or voxels in 3D scenes, thus intricate contextual perception is required to accomplish this element-wise recognition task. This characteristic makes semantic segmentation a promising foundation for further exploring scene-wise and object-wise recognition tasks, i.e., classification and detection.

**Dataset.** In our initial investigation into multi-dataset collaborative learning for 3D perception, we consider Scan-
Net [21], S3DIS [2], and Structured3D [124] as the datasets of interest, all of which include segmentation annotations. ScanNet and S3DIS represent the most commonly used real-world datasets in the realm of 3D perception, while Structured3D is a larger-scale synthetic RGB-D dataset that we specifically incorporated to establish an experimental context for addressing the domain gap between synthetic and real data, ultimately aiming to achieve mutual gains across datasets. As illustrated in the left side of Tab. 1, although all three datasets represent indoor point cloud scenes, they exhibit distinct characteristics in terms of data scale, scene variety, and point cloud density. Our objective is to examine methods for overcoming the domain gap among these diverse datasets, facilitating collaborative learning across multiple sources and thereby taking an essential step towards large-scale representation learning for 3D perception.

**Evaluation.** As a proof of concept, we consider joint training by default, in which the model is jointly trained on all datasets under the supervised setting, and directly evaluated on all datasets without fine-tuning. In the final experiments, we will also consider two standard transfer learning settings: 1) supervised pre-training, where the model supervised pre-trained during joint training is further fine-tuned on the target dataset; and 2) unsupervised pre-training, where the model is unsupervised pre-trained on all datasets, and fine-tuned on each target dataset for evaluation.

### 2.2. Pilot Study: Uncovering the Negative Transfer

As a pioneering effort, MSC [108] involved unsupervised pre-training using a combination of two indoor datasets, ScanNet [21] and Arikitiscene [5]. However, even with the addition of three times more data, the performance improvement over the single-dataset pre-training baseline on ScanNet was relatively limited. To investigate the underlying causes of this limited performance gain, we take a step back and reassess this phenomenon by studying a straightforward supervised multi-dataset learning setup, i.e., the joint training setting aforementioned in Sec. 2.1.

**Negative transfer** [10] refers to the phenomenon where learning from one dataset may negatively impact the performance on another dataset due to differences in data distribution. Despite restricting our focus to indoor scene point clouds, a significant negative transfer occurs during direct multi-dataset mixed segmentation training. As illustrated in Tab. 1 (right side), we conduct training by pairwise merging the three datasets as well as a combination of all, and evaluate the model’s performance on each related individual dataset. The experimental results reveal that direct merging training data gives rise to negative transfer between datasets, underscoring the challenges associated with attaining effective collaborative learning across multiple datasets in the 3D domain.

### 3. Point Prompt Training

Due to the risk of negative transfer discussed in Sec. 2.2, adapting a single model to diverse domains with distinct contexts still remains a significant challenge. Nevertheless, recent advances suggest that prompt tuning may be a viable approach for effectively adapting pre-trained models with large-scale datasets to downstream tasks. Inspired by this, we propose a different paradigm named Point Prompt Training (PPT) to mitigate negative transfer and enable multi-dataset training.

As shown in Fig. 2, PPT has two essential components: (1) a prompt adapter, which adapts a single model to varying contexts of different datasets using a set of learnable domain-specific prompts, and (2) a categorical alignment process, which enables the model to be decently trained within multiple category spaces simultaneously with supervised learning. Details of them are presented as follows.

### 3.1. Learning with Domain Prompting

**Issues with prompt tuning.** In the prompt tuning paradigm [59], a model pre-trained by a large-scale dataset is fine-tuned for specific tasks or datasets by incorporating additional information or context through prompts. These prompts facilitate the model’s adaptation to new tasks with minimal parameter changes, often outperforming that with full fine-tuning [42, 125, 126] and laying the ground for a unified foundation model [7].
However, in 3D perception, the lack of a large-scale pre-trained model hinders the applications of prompt tuning. Furthermore, prompt tuning aims to address the domain gap between pre-training and fine-tuning datasets rather than improving the model’s ability to fit multiple datasets simultaneously during either pre-training or fine-tuning. To tackle this issue, we introduce a novel method termed domain prompting. Instead of merely fine-tuning prompts on pre-trained models, we incorporate learnable prompt tokens as conditions for varying dataset contexts and (pre-)train the domain prompt with backbone cooperatively.

**Domain prompting.** Specifically, for each interested dataset $D_i$, we generate a learnable $d$-dimensional vector as the domain-specific prompt. The collection of $n$ contexts is denoted as $\mathcal{C} = \{c^i \in \mathbb{R}^d | i \in \mathbb{N}, 1 \leq i \leq n\}$. Then the multi-dataset training objective in Eq. 1 becomes:

$$\arg\min_{\theta, \mathcal{C}} \sum_{i=1}^{n} \frac{1}{|D_i|} \sum_{(x^i_j, y^i_j) \in D_i} \mathcal{L}(f(x^i_j, c^i; \theta), y^i_j).$$ (2)

These learnable domain prompts facilitate the discovery of distribution differences among datasets, enabling the backbone to surmount domain gaps encountered in multi-dataset training. As a result, the model focuses more on learning the representations that can be decently shared across datasets. This method fosters mutual benefits among distinct datasets and promotes a collaborative synergy between the backbone model and the prompts. Similar to VPT [42], we also observe that the shared prompt within each domain can achieve comparable or even better performance than the independent ones for different backbone blocks, and we put the discussion in the Appendix. We believe this approach can benefit both supervised and unsupervised pre-training, as well as fine-tuning, by addressing the negative transfer that may exist within multiple datasets.

**Domain prompt adapter.** With the domain prompts that possess unique characteristics specific to individual datasets, enabling the model to effectively engage with domain-specific prompts becomes another challenge. Previous research on visual prompt tuning has demonstrated that the adapters utilizing shared prompts to exert block-wise control over models are more effective than those that inject prompts at the input level [42]. Building on this insight, we investigate various designs for prompt adapters as outlined below and mark our main proposal with $\ast$. More specific illustrations and details regarding the alternative designs are available in our Appendix.

- **Direct Injection.** The domain-specific contextual cues of various datasets are encoded within their respective prompts. The incorporation of domain priors can be achieved by simply adding channel-aligned prompts to the intermediate feature maps with a linear projection.
- **Cross Attention.** Drawing inspiration from DETR [9], we leverage a cross-attention-based domain prompt adapter as another alternative design for multi-dataset training. This scheme introduces a cross-attention block with a skip connection at the beginning of each encoder-decoder stage, injecting domain-specific information into the intermediate feature maps. Our design allows broad applicability to versatile 3D backbones without structural constraints while still preserving the advantages of the VPT technique.
- **Prompt-driven Normalization*.** The objective of domain prompt adapter is to learn a shared representation that is robust and generalizable across various datasets, akin to how the style transfer methods [24, 94] retain the content essence while only transferring the contextual styles across images. Also, adapting the normalization layer to varying individual contexts is found beneficial for achieving better style transfer performance [40, 68]. With the analogy to style transfer, we introduce the context adapter of Prompt-driven Normalization (PDNorm), a novel approach to tackle the transfer challenges associated with multi-dataset train-
ing illustrated in Fig. 2a. Formally, with a given domain prompt $c$, PDNorm adaptively learns the $\gamma$ and $\beta$ values in normalization:

$$\text{PDNorm}(x, c) = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x]} + \epsilon} \cdot \gamma(c) + \beta(c),$$

(3)

where $\gamma(c)$ and $\beta(c)$ are linear projections, $\bar{x}$ for computing $\mathbb{E}[x]$ and $\text{Var}[x]$ is contingent on the specific normalization employed by the backbone. It’s important to note that $\mathbb{E}[x]$ and $\text{Var}[x]$ are statisticized independently for each dataset involved. We substitute the original backbone’s normalization layers with PdNorm layers. This approach promotes a more efficient yet effective alignment of feature distributions across datasets in the scenario of multi-dataset training.

Zero-initialization and learning rate scaling. Unlike prevalent prompt tuning methods that only adjust inserted prompts while retaining the pre-trained models, our proposed domain prompts are joint-trained with the backbone. Nevertheless, in our paradigm, the introduction of randomly initialized prompts may disrupt the representation learning of the rest of the model, resulting in unstable training with large loss values at early training stages. We conjecture that, during the initial stages of training, the model is acquiring general knowledge that can be applied across diverse domains. However, as training proceeds, the model gradually begins to generate domain-specific representations based on general representations. To address this issue, we employ zero-initialization [41] and learning rate scaling [33], ensuring stability during early training stages and yielding superior results. Specifically, we zero-initialize the $\gamma(c)$ and $\beta(c)$ parameters of PDNorm, and we start with a smaller base learning rate of prompt-related parameters to prioritize the backbone during the initial training stage. We also perform a similar design to our alternative prompt adapters for a fair comparison, and details are shown in the Appendix.

3.2. Categorical Alignment

In PPT, an additional critical issue that needs to be addressed is the inconsistency of the label space across different datasets with supervised learning. To tackle this problem, we have investigated various approaches to unify the categories for multi-dataset training as follows. Also, more details and discussions can be found in the Appendix.

- **Decoupled.** One straightforward approach is to employ separate linear projection heads for each dataset. While this method is effective in handling inconsistencies, it introduces redundant parameters for decoding the same categories shared by different datasets. Besides, it overlooks the commonalities among the datasets and fails to account for their potential correlations.
- **Unionized.** Another intuitive approach is to construct a shared linear segmentation head that projects the representation space into a unified label space encompassing all datasets while the loss computation remains separate and constrained to the distinct label spaces for each dataset. This method effectively resolves the inconsistency in point representations pertaining to the shared label space across datasets.

- **Language-guided.** The aforementioned options treat each category independently and assume that they are uncorrelated. However, it is a natural fact that labels with close meanings should have similar representations [76]. Leveraging such prior information can further benefit the discovery of robust representations in our scenario. To this end, we propose language-guided categorical alignment, which aligns projected point representations with the category-language embeddings extracted by a pre-trained text encoder, such as CLIP [74]. To achieve this goal, we employ the InfoNCE [65] as alignment criteria and restrict negative samples to the specific dataset category space as shown in Fig. 2b.

4. Experiments

In this section, we conduct extensive experiments to substantiate the efficacy of our proposed framework across multiple data sources with different evaluation settings. Specifically, in Sec. 4.1, assess the effectiveness of different design choices via detailed ablation studies. After that, in Sec. 4.2, we conduct system-level comparisons with existing methods. All experiments are conducted on compute nodes equipped with 8 NVIDIA A100 GPUs.

4.1. Ablation Study

In this part, we ablate different design choices of PPT from the perspective of module design and data engineering. We employ supervised joint training with SparseUNet, train it on ScanNet, S3DIS, and Structured3D, and evaluate it on ScanNet 20-category semantic segmentation. For evaluation, we consider both direct evaluation (joint training) and fine-tuning (see details in Sec. 2.1). More details of the setting are available in the Appendix.

**Prompt adapter.** In Tab. 2a, we show results with different designs of the domain prompt adapter. Compared with the vanilla baseline (none) without a prompt adapter, all designs show effectiveness in learning good representations from multiple datasets. Moreover, compared with simpler designs like direct injection (add) and cross attention (c.a.), our novel design prompt-driven normalization (p.n.) achieves significantly stronger results, verifying its effectiveness.

**Zero-initialization and learning rate scaling.** In Tab. 2b, we verify the effect of zero initialization and learning rate scaling. Overall, it shows that zero initialization, a technique often adopted for adapting pre-trained models, could also benefit training from scratch. Besides, scaling the
Table 2. Module ablation. We adopt SparseUNet and supervised multi-dataset joint training to ablate our designs. We report both joint training and fine-tuning mIoU (%) results on ScanNet 20-category semantic segmentation. All of our designs are enabled by default, and default settings are marked in gray. The detailed setting for joint training and fine-tuning is reported in Appendix.

Learning rate for domain prompting to a relatively smaller value (0.1) than the backbone also helps training.

Prompt location. In Tab. 2c, we study the influence of injecting the prompt adapter to different stages of the backbone. Empirically, the benefit of the prompt adapter becomes higher if it is added to relatively deeper stages. Our intuition is that features in earlier stages are more related to low-level attributes, which could be easier shared across datasets. And, deeper features are more related to high-level semantics, where negative effect of the domain gap occurs and a domain adapter is needed.

Prompt length. In Tab. 2d, we ablate the feature-level length (dimension) of the prompt adapter. A larger dimension of the adapter often allows space for higher information capability, but our experiments show that the adapter is quite memory-efficient. The results with different feature dimensions do not differ much, and a small dimension of 256 is already sufficient.

Categorical alignment. In Tab. 2e, we show results with different methods for aligning the label space of different training datasets. Compared with learning separate segmentation heads for each dataset, obtaining a unionized head allows better alignment of the supervision from different datasets. Further, language guidance takes the relationship between class names into account, resolving possible conflicts, and results in a further performance boost. Besides that, we also tried a simple prompt engineering technique that augments class names to a sentence (e.g., “A point of [class],”), which does not show effectiveness in this case.

Language-guidance criteria. In Tab. 2f, we ablate the loss function for aligning with category-specific language embeddings extracted from a pre-trained text encoder. Simple L2 loss, which does not consider negative examples, could result in mode collapse. Compared with other specialized criteria, e.g., text-supervised contrastive loss proposed in [76], our method suits well with the most commonly used InfoNCE loss, highlighting its universality.

Sampling ratio. In Tab. 2g, we show the results with different sampling ratios across datasets, and experiments show that overall our method is relatively robust to this ratio. It is important to note that, in contrast to downstream tasks where the sampling ratio can significantly impact the final performance, our focus is on representation learning. Therefore, the effect of the sampling ratio may be negligible if the model is sufficiently trained on each dataset for an adequate duration [34].

Joint training data. In Tab. 2h, we show the results with different joint training data similar to Tab. 1. Datasets benefit each other in our PPT framework.
Under the unsupervised setting, our framework could be built on both convolution-based architecture SparseUNet [16] and transformer-based architecture PTv3 [17], and is evaluated on ScanNet, ScanNet200, and S3DIS benchmarks. The framework is universal, and we report on three settings: unsupervised pre-training integrated with MSC [108], supervised joint training, and supervised pre-training. Besides comparing with previous pre-training methods [35, 108, 110], we also conduct system-level comparisons against previous SOTAs [50, 73, 107, 122], and our method shows consistently better results across benchmarks even with one single share-weighted model.

### 4.2. Results Comparison

**Indoor semantic segmentation results.** In Tab. 3, we present the main results of different variants of our method on multiple standard semantic segmentation benchmarks, and compare with previous state-of-the-art methods at both system-level and module-level. Following the common practice of pre-training methods [35, 108, 110], our method is built on both convolution-based architecture SparseUNet [16] and transformer-based architecture PTv3 [17]. Under the unsupervised setting, our framework could smoothly integrate MSC [108] and enable it to benefit from joint training on multiple datasets, e.g., improving on ScanNet200 Val split by 1.6 points, and on S3DIS Area5 mIoU by 1.8 points. More importantly, the results also surpass all previous SOTAs, verifying the effectiveness and potential of large-scale unsupervised pre-training for 3D scene understanding. When further considering the supervised joint training setting, and fine-tuning upon it, our method further sees consistent performance gains across tasks and secures position as a new SOTA.

**Outdoor semantic segmentation results.** In Tab. 4, we expand our methodology to outdoor scenarios by presenting additional results of our approach on multiple outdoor semantic segmentation benchmarks. We systematically compare these results with those of previously established SOTA methods. Our method is still based on SparseUNet [16], a classic framework within the outdoor perception community, and PTv3 [17], which is the latest SOTA backbone for outdoor perception. Under the supervised joint training paradigm, our method showcases significant enhancements across all tasks when contrasted with scratch results, even with a single shared-weight model. For instance, on the SemanticKITTI Validation split, our approach elevates by 7.1 points, underscoring the potential of all-data learning in the realm of 3D understanding. Through subsequent fine-tuning on each dataset, PPT consistently sees consistent performance gains across tasks and secures position as a new SOTA.
PPT demonstrates superiority over the latest literature. For instance, it outperforms SphereFormer [51] by 5.0 points in terms of mIoU on the SemanticKITTI validation set.

**Indoor instance segmentation results.** We conduct PPT supervised pre-training on SparseUNet [16] as described in Tab. 3 and further fine-tuning on ScanNet and ScanNet200 instance segmentation driven by PointGroup [44]. We compare mAP@25, mAP@50, and mAP results with previous pre-training methods, and our method shows significant superior results across benchmarks.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor Ins. Seg.</td>
<td>39.2M</td>
<td>72.8</td>
<td>56.9</td>
<td>36.0</td>
<td>32.2</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScanNet Val [21]</td>
<td>46.2M</td>
<td>77.5</td>
<td>61.7</td>
<td>40.9</td>
<td>40.1</td>
<td>33.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ScanNet200 Val [76]</td>
<td>46.3M</td>
<td>78.9</td>
<td>63.5</td>
<td>42.1</td>
<td>40.8</td>
<td>34.1</td>
</tr>
</tbody>
</table>

Table 5. Data efficient results. We follow the ScanNet Data Efficient benchmark [35] and compare the validation results of the PPT unsupervised setting with previous pre-training methods. All methods are trained with SparseUNet, and 3C denotes train from scratch.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>26.0</td>
<td>28.9</td>
<td>29.2</td>
<td>31.3</td>
</tr>
<tr>
<td>5%</td>
<td>47.8</td>
<td>49.8</td>
<td>50.7</td>
<td>52.2</td>
</tr>
<tr>
<td>10%</td>
<td>56.7</td>
<td>59.4</td>
<td>61.0</td>
<td>62.8</td>
</tr>
<tr>
<td>20%</td>
<td>62.9</td>
<td>64.6</td>
<td>64.9</td>
<td>66.4</td>
</tr>
<tr>
<td>100%</td>
<td>72.2</td>
<td>73.8</td>
<td>75.3</td>
<td>75.8</td>
</tr>
</tbody>
</table>

(a) Limited reconstructions. Pct. denotes the percentage of scene reconstruction that could be used for training.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>41.9</td>
<td>55.5</td>
<td>60.1</td>
<td>60.6</td>
</tr>
<tr>
<td>50</td>
<td>53.9</td>
<td>60.5</td>
<td>66.8</td>
<td>67.5</td>
</tr>
<tr>
<td>100</td>
<td>62.2</td>
<td>65.9</td>
<td>69.7</td>
<td>70.8</td>
</tr>
<tr>
<td>200</td>
<td>65.5</td>
<td>68.2</td>
<td>70.7</td>
<td>72.2</td>
</tr>
<tr>
<td>Full</td>
<td>72.2</td>
<td>73.8</td>
<td>75.3</td>
<td>75.8</td>
</tr>
</tbody>
</table>

(b) Limited annotations. Pts. denotes the number of points per scene that are annotated for training.

Table 6. Data efficient results. We follow the ScanNet Data Efficient benchmark [35] and compare the validation results of the PPT unsupervised setting with previous pre-training methods. All methods are trained with SparseUNet, and 3C denotes train from scratch.

We discuss limitations and broader impacts as follows:

- **Module design.** As a preliminary work in 3D multi-dataset pre-training, this paper first verifies the effectiveness of this setting and opens doors for large-scale 3D representation learning. Yet current explorations are still restricted to a limited scope and the designs could be sub-optimal, thus further study on more advanced techniques is necessary. For example, one could verify the effectiveness of this framework when combined with more advanced unsupervised pre-training methods and explore more effective prompting techniques.

- **Data domain.** Our study demonstrates the potential benefit of simultaneously utilizing both synthetic and real point cloud data. It would be exciting to see this ability extended to more specific scenarios in different domains, e.g., jointly learning from both indoor and outdoor scenes.

- **Multi-task training.** Our current formulation only considers one pre-training task. Upon that, as it has shown the ability to achieve superior results across datasets with a single model, a promising direction is to enable multi-task training for 3D scene understanding with a unified framework.

**Acknowledgements**

This work is supported in part by the National Natural Science Foundation of China (No. 622014840), Alibaba Innovative Research Fund, HKU Startup Fund, and HKU Seed Fund for Basic Research.
References


[31] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. AI Open, 2022. 13


[34] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In ICCV, 2019. 6


[37] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In CVPR, 2020. 18


[40] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In ICCV, 2017. 4


[43] Li Jiang, Hengshuang Zhao, Shu Liu, Xiaoyong Shen, Chi-Wing Fu, and Jiaya Jia. Hierarchical point-edge interaction network for point cloud semantic segmentation. In ICCV, 2019. 18

[44] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In CVPR, 2020. 8


[52] Loïc Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superposition graphs. In CVPR, 2018. 18


[57] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhai Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In NeurIPS, 2018. 18


[64] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In IROS, 2019. 18
