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

Xiaoyang Wu[1] Zhuotao Tian[2] Xin Wen[1] Bohao Peng[2] Xihui Liu[1] Kaicheng Yu[3,4] Hengshuang Zhao[1*]

[1]The Univeristy of Hong Kong [2]The Chinese Univeristy of Hong Kong
[3]Westlake University [4]Alibaba Group

https://github.com/Pointcept/Pointcept

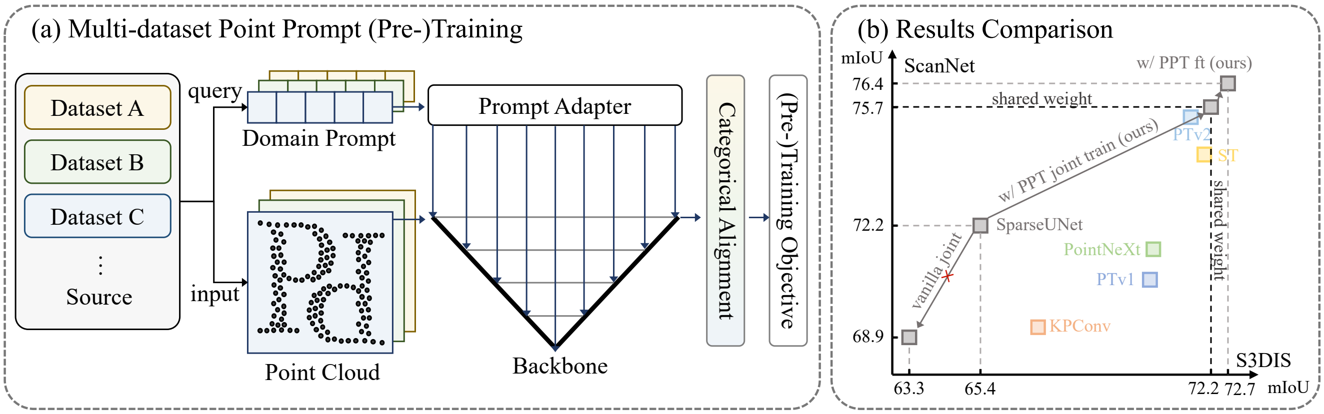

Figure 1. **Multi-dataset synergistic training with Point Prompt Training (PPT).** (a) Our *PPT Framework* is comprised of two key components: 1. The domain prompt adapter adapts the backbone to various dataset-specific contexts with a set of domain-specific prompts; 2. The categorical alignment process empowers the model to effectively undergo training within multiple category spaces concurrently in the supervised setting. (b) The *Result Comparison* plot reveals that PPT delivers state-of-the-art performance across both datasets only with one single shared-weight backbone, and fine-tuning on any single specific dataset can further enhance the results.

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.*

*Corresponding author.

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 exhibit 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 $\mathcal{D}_i = \{(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i)\}$, where $1 \leq i \leq n$, n stands for the number of datasets, and $(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i)$ represents data-label pairs that construct a dataset. Our goal is to train a model $f(\boldsymbol{x}; \theta)$ parameterized by θ, such that the cumulative loss across all datasets is minimized:

$$\operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i) \in \mathcal{D}_i} \mathcal{L}\left(f\left(\boldsymbol{x}_j^i; \theta\right), \boldsymbol{y}_j^i\right), \quad (1)$$

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-

Table 1. **Datasets summary and joint training transfer among tasks.** The entry at row i and column j indicates the semantic segmentation mIoU on dataset i using a SparseUNet [16, 18] joint-trained on datasets i and j. The *All* column represents combining all data sources. The diagonal elements represent using only the original dataset i. Note that Structured3D is originally panoramic images, and we converted it to point cloud data following Swin3D [116]. Moreover, we compute the sampling ratio based on each dataset's best performance necessary iteration number. The effects of different sampling strategies are further explored in the ablation study (Sec. 4.1) and Appendix.

	Dataset details				Baseline results w/ diff. joint training data				Ours
Target data	Source	Sparsity	Complexity	Scans	ScanNet	S3DIS	Struct.3D	All	(All)
ScanNet	Real	Sparse	Large rooms	1613	<u>72.2</u>	71.8	65.9	68.9 (-3.3)	**75.7** (+3.5)
S3DIS	Real	Dense	School office	272	64.1	<u>65.4</u>	62.8	63.3 (-2.1)	**72.2** (+6.8)
Struct.3D	Synth.	Dense	Suite	3500	73.7	74.2	<u>74.5</u>	72.9 (-1.6)	**75.8** (+1.3)

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 Arikitscene [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].

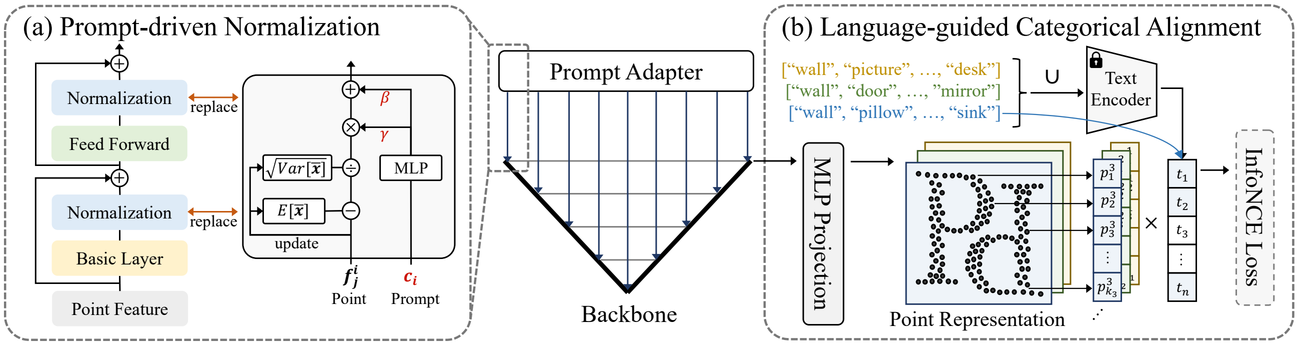

Figure 2. **Prompt adapter and categorical alignment.** (a) As a prompt adapter, *Prompt-driven Normalization* adaptly encodes domain-specific prompts into the scale and shift vectors in normalization layers. This adaptation helps adapt the model to the specific dataset domain. (b) *Language-guided Categorical Alignment* aligns point representations to a unified category-language embedding, shared across all datasets and extracted by a pre-trained text encoder.

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

$$\underset{\theta, \mathcal{C}}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x}_j^i, \boldsymbol{y}_j^i) \in \mathcal{D}_i} \mathcal{L}\big(f\big(\boldsymbol{x}_j^i, \boldsymbol{c}_i; \theta\big), \boldsymbol{y}_j^i\big). \quad (2)$$

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 $*$. 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 γ and β values in normalization:

$$\mathrm{PDNorm}(\boldsymbol{x}, \boldsymbol{c}) = \frac{\boldsymbol{x} - \mathrm{E}[\bar{\boldsymbol{x}}]}{\sqrt{\mathrm{Var}[\bar{\boldsymbol{x}}] + \epsilon}} \cdot \gamma(\boldsymbol{c}) + \beta(\boldsymbol{c}), \quad (3)$$

where $\gamma(\boldsymbol{c})$ and $\beta(\boldsymbol{c})$ are linear projections, $\bar{\boldsymbol{x}}$ for computing $\mathrm{E}[\bar{\boldsymbol{x}}]$ and $\mathrm{Var}[\bar{\boldsymbol{x}}]$ is contingent on the specific normalization employed by the backbone. It's important to note that $\mathrm{E}[\bar{\boldsymbol{x}}]$ and $\mathrm{Var}[\bar{\boldsymbol{x}}]$ 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 $\gamma(\boldsymbol{c})$ and $\beta(\boldsymbol{c})$ 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

case	none	add	c.a.	p.n.
joint	<u>68.9</u>	70.9	73.5	**75.7**
f.t.	<u>73.6</u>	73.8	75.4	**76.4**

(a) **Prompt adapter.** Here *c.a.* denotes *cross attention*, and *p.n.* denotes *prompt-driven normalization*. *p.n.* is a superior domain prompt adapter.

zero-init	w/o	w/
joint	75.2	**75.7**
f.t.	75.6	**76.4**

lr scaler	1	0.1	0.01
joint	75.4	**75.7**	75.2
f.t.	76.0	**76.4**	75.8

(b) **Zero-initialization & Learning rate scaling.** Zero-init. benefits our training scheme, and 0.1 is a suitable lr scaler for domain prompting.

location	initial	encoder	decoder	all
joint	68.7	74.2	73.2	**75.7**
f.t.	73.9	74.9	74.7	**76.4**

(c) **Prompt location.** An in-depth prompt adapter that runs through the entire backbone is necessary.

length	128	256	512	1024
joint	75.2	**75.7**	**75.7**	75.5
f.t.	75.9	**76.4**	76.1	76.2

(d) **Prompt length.** Domain prompt with 256 dimensions achieves a good balance.

case	decoupled	unionized	l.g.	l.g. w/ tpl.
joint	74.4	75.3	75.7	**75.8**
f.t.	74.7	75.8	**76.4**	76.0

(e) **Categorical alignment.** *l.g.* denotes *languages-guided categorical alignment*, which achieves best f.t. results, and *tpl.* stands for *template*.

criteria	L2	Disc. [22]	TC [76]	InfoNCE
joint	12.8	65.4	70.2	**75.7**
f.t.	72.1	73.5	73.2	**76.4**

(f) **Language-guidance criteria.** *TC* represents *Text-supervised Contrastive* loss. While *InfoNCE* loss makes *l.g.* the best C.A..

ratio	4:2:1	2:2:1	2:1:1	1:1:1
ScanNet	75.7	**75.8**	73.3	74.6
S3DIS	**72.2**	71.9	71.2	71.9
Struct.3D	**75.8**	73.5	72.7	74.7

(g) **Sampling ratio.** The ratios below indicate the sampling rate of Struct.3D, ScanNet, and S3DIS, and 4:2:1 achieved the best result.

data	ScanNet	S3DIS	Struct.3D	all
ScanNet	<u>72.2</u>	73.9	**76.0**	75.7
S3DIS	69.6	<u>65.4</u>	70.2	**72.2**
Struct.3D	74.8	75.0	<u>74.5</u>	**75.8**

(h) **Joint training data.** Joint training results with different joint training data similar to Tab. 1. Datasets benefit each other in our PPT framework.

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 (see attributes of datasets in Tab. 1). Note that though they differ in data source, sparsity, complexity, and scale, our final framework allows consistent benefit from different data sources regardless of large domain gaps.

Indoor Sem. Seg.		ScanNet [21]		ScanNet200 [76]		S3DIS Area5 [2]	
Methods	Params.	Val mIoU	Test mIoU	Val mIoU	Test mIoU	mIoU	mAcc
StratifiedFormer [50]	18.8M	74.3	73.7	-	-	72.0	78.1
PointNeXt [73]	41.6M	71.5	71.2	-	-	70.5	77.2
PTv1 [122]	11.4M	70.6	-	27.8	-	70.4	76.5
PTv2 [107]	12.8M	75.4	75.2	30.2	-	71.6	77.9
SparseUNet [16]	39.2M	72.2	73.6	25.0	25.3	65.4	71.7
+ PC [110]	39.2M	74.1 (+1.9)	-	26.2 (+1.2)	-	70.3 (+4.9)	76.9 (+5.2)
+ CSC [35]	39.2M	73.8 (+1.6)	-	26.4 (+1.4)	24.9 (-0.4)	72.2 (+6.8)	-
+ MSC [108]	39.2M	75.5 (+3.3)	-	28.8 (+3.8)	-	70.1 (+4.7)	77.2 (+5.5)
+ PPT Unsup. (f.t.)	41.0M	75.8 (+3.6)	-	30.4 (+5.4)	-	71.9 (+6.5)	**78.3** (+6.6)
+ PPT Sup. (joint)	41.0M	75.7 (+3.5)	**76.6** (+3.0)	-	-	72.2 (+6.8)	78.0 (+6.3)
+ PPT Sup. (f.t.)	41.0M	**76.4** (+4.2)	-	**31.9** (+6.9)	**33.2** (+7.9)	**72.7** (+7.3)	78.2 (+6.5)
PTv3 [17]	46.2M	77.5	77.9	35.2	37.8	73.4	77.7
+ PPT Sup. (f.t.)	46.3M	**78.6** (+1.1)	**78.3** (+0.4)	**36.0** (+0.8)	**39.3** (+1.5)	**74.7** (+1.3)	**79.6** (+1.9)

Table 3. **Indoor semantic segmentation results.** Our method builds on SparseUNet [16] and 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.

Outdoor Sem. Seg.		SemanticKITTI [6]		nuScenes [8]		Waymo [86]	
Methods	Params.	Val mIoU	Test mIoU	Val mIoU	Test mIoU	Val mIoU	Val mAcc
SPVNAS [87]	10.8M	64.7	66.4	-	77.4	-	-
Cylinder3D [128]	26.1M	64.3	67.8	76.1	77.2	-	-
SphereFormer [51]	32.3M	67.8	74.8	78.4	81.9	69.9	-
SparseUNet [16]	39.2M	63.8	-	73.3	-	65.9	76.6
+ PPT Sup. (joint)	41.0M	70.9 (+7.1)	-	78.5 (+5.2)	-	70.0 (+4.1)	**79.1** (+2.5)
+ PPT Sup. (f.t.)	41.0M	**71.4** (+7.6)	-	**78.6** (+5.3)	-	**70.4** (+4.5)	78.9 (+2.3)
PTv3 [17]	46.2M	70.8	74.2	80.4	82.7	71.3	80.5
+ PPT Sup. (f.t.)	46.3M	**72.3** (+1.5)	**75.5** (+1.3)	**81.2** (+0.8)	**83.0** (+0.3)	**72.1** (+0.8)	**81.3** (+0.8)

Table 4. **Outdoor semantic segmentation results.** We also examine the efficacy of PPT in an outdoor context using SparseUNet [16] and PTv3 [17]. Our evaluation encompasses SemanticKITTI, nuScenes, and Waymo semantic segmentation benchmarks. We report on two main settings: supervised joint training and supervised pre-training. We conduct comprehensive comparisons against previous SOTAs [51, 87, 107, 128], and our method shows multiple superior results across benchmarks.

4.2. Results Comparision

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

Indoor Ins. Seg.		ScanNet Val [21]			ScanNet200 Val [76]		
PointGroup [44]	Params.	mAP@25	mAP@50	mAP	mAP@25	mAP@50	mAP
SparseUNet [16]	39.2M	72.8	56.9	36.0	32.2	24.5	15.8
+ PC [110]	39.2M	-	58.0 (+1.1)	-	-	24.9 (+0.4)	-
+ CSC [35]	39.2M	-	59.4 (+2.5)	-	-	25.2 (+0.7)	-
+ LGround [76]	39.2M	-	-	-	-	26.1 (+1.6)	-
+ MSC [108]	39.2M	74.7 (+1.9)	59.6 (+2.7)	39.3 (+3.3)	34.3 (+2.1)	26.8 (+2.3)	17.3 (+1.5)
+ PPT (f.t.)	41.0M	**76.9** (+4.1)	**62.0** (+3.1)	**40.7** (+4.7)	**36.8** (+4.6)	**29.4** (+4.9)	**19.4** (+3.6)
PTv3 [17]	46.2M	77.5	61.7	40.9	40.1	33.2	23.1
+ PPT (f.t.)	46.3M	**78.9** (+1.4)	**63.5** (+1.8)	**42.1** (+1.2)	**40.8** (+0.7)	**34.1** (+0.9)	**24.0** (+0.9)

Table 5. **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

Pct.	SC	CSC [35]	MSC [108]	PPT
1%	26.0	28.9 (+2.9)	29.2 (+3.2)	**31.3** (+5.3)
5%	47.8	49.8 (+2.0)	50.7 (+2.9)	**52.2** (+4.4)
10%	56.7	59.4 (+2.7)	61.0 (+4.3)	**62.8** (+6.1)
20%	62.9	64.6 (+1.7)	64.9 (+2.0)	**66.4** (+3.5)
100%	72.2	73.8 (+1.6)	75.3 (+3.1)	**75.8** (+3.6)

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

Pts.	SC	CSC [35]	MSC [108]	PPT
20	41.9	55.5 (+13.6)	60.1 (+18.2)	**60.6** (+18.7)
50	53.9	60.5 (+6.6)	66.8 (+12.9)	**67.5** (+13.6)
100	62.2	65.9 (+3.7)	69.7 (+7.5)	**70.8** (+8.6)
200	65.5	68.2 (+2.7)	70.7 (+5.2)	**72.2** (+6.7)
Full	72.2	73.8 (+1.6)	75.3 (+3.1)	**75.8** (+3.6)

(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 *SC* denotes train from scratch.

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. In Tab. 5, we conduct fine-tuning experiments on instance segmentation using SparseUNet [16] and PTv3 [17] as the backbone, powered by PointGroup [44]. The fine-tuning outcomes are reported on both the ScanNet [21] and ScanNet200 [76] instance segmentation benchmarks. Our findings consistently reveal the superior performance of our approach compared to the prior state-of-the-art method, MSC [108]. To be specific, PPT outperforms MSC by 2.4 points in terms of mAP@50 on the ScanNet validation split, and by 2.6 points on the ScanNet200 validation split. This underscores the effectiveness of the point representation learned by PPT in enhancing instance segmentation performance.

Data efficient benchmark. In Tab. 6, we report results for the ScanNet Data Efficient benchmark [35], where scene reconstruction or annotation percentages are limited. Our method, integrating MSC [108], is compared with prior pre-training methods and consistently outperforms them under data efficient settings.

5. Conclusion and Discussion

This paper introduces PPT, an effort toward large-scale 3D representation learning with a novel 3D multi-dataset synergistic training setting. We identify the negative transfer issue and present a unified framework that addresses this problem with the proposed Prompt-driven Normalization and Language-guided Categorical Alignment, delivering consistent and significant performance gains.

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

[1] Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. Ext5: Towards extreme multi-task scaling for transfer learning. In *ICLR*, 2022. 1

[2] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2, 3, 7, 13, 15, 17, 18

[3] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. PASS: An imagenet replacement for self-supervised pretraining without humans. In *NeurIPS*, 2021. 13

[4] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv:2203.17274*, 2022. 13

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS Workshops*, 2021. 3, 13

[6] 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 *ICCV*, 2019. 2, 7, 18

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 3, 13

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 7, 18

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 4

[10] Rich Caruana. Multitask learning. *Machine learning*, 1997. 3

[11] Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *NeurIPS*, 2021. 13

[12] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *CVPR*, 2017. 13

[13] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Largekernel3d: Scaling up kernels in 3d sparse cnns. In *CVPR*, 2023. 18

[14] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. Scaledet: A scalable multi-dataset object detector. In *CVPR*, 2023. 13

[15] Hung-Yueh Chiang, Yen-Liang Lin, Yueh-Cheng Liu, and Winston H Hsu. A unified point-based framework for 3d segmentation. In *3DV*, 2019. 18

[16] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4D spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 3, 7, 8, 13, 16, 17, 18, 19

[17] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023. 7, 8, 16, 18, 19

[18] Spconv Contributors. Spconv: Spatially sparse convolution library. https://github.com/traveller59/spconv, 2022. 3, 16

[19] Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. In *ACL*, 2022. 13

[20] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *ECCV*, 2018. 18

[21] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1, 2, 3, 7, 8, 13, 15, 17, 18

[22] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. In *CVPR*, 2017. 6

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 13

[24] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *ICLR*, 2017. 4

[25] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, 2021. 13

[26] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv:2202.06687*, 2022. 13

[27] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv:2103.01988*, 2021. 1, 13

[28] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 13, 18

[29] Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. Ppt: Pre-trained prompt tuning for few-shot learning. In *ACL*, 2022. 13

[30] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 2021. 13

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

[32] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *ICCV*, 2019. 13

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

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

[35] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1, 2, 7, 8, 13, 18

[36] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. Point-to-voxel knowledge distillation for lidar semantic segmentation. In *CVPR*, 2022. 19

[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

[38] Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *ACL*, 2022. 13

[39] Zeyu Hu, Mingmin Zhen, Xuyang Bai, Hongbo Fu, and Chiew-lan Tai. Jsenet: Joint semantic segmentation and edge detection network for 3d point clouds. In *ECCV*, 2020. 18

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

[41] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[42] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 2, 3, 4, 13, 17

[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

[45] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2022. 2, 13

[46] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv:2001.08361*, 2020. 1, 13

[47] Dongwan Kim, Yi-Hsuan Tsai, Yumin Suh, Masoud Faraki, Sparsh Garg, Manmohan Chandraker, and Bohyung Han. Learning semantic segmentation from multiple datasets with label shifts. In *ECCV*, 2022. 2, 13

[48] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023. 1

[49] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *ICCV*, 2023. 19

[50] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *CVPR*, 2022. 7, 18

[51] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *CVPR*, 2023. 7, 8, 19

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

[53] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019. 13

[54] Huan Lei, Naveed Akhtar, and Ajmal Mian. Seggcn: Efficient 3d point cloud segmentation with fuzzy spherical kernel. In *CVPR*, 2020. 18

[55] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 13

[56] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network. In *RSS*, 2016. 13

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

[58] Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. Meta architecture for point cloud analysis. In *CVPR*, 2023. 18

[59] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 2023. 3, 13

[60] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv:2011.13677*, 2020. 13

[61] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv:2103.10385*, 2021. 13

[62] Youquan Liu, Lingdong Kong, Xiaoyang Wu, Runnan Chen, Xin Li, Liang Pan, Ziwei Liu, and Yuexin Ma. Multi-space alignments towards universal lidar segmentation. In *CVPR*, 2024. 19

[63] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, 2015. 13

[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

[65] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 5, 15

[66] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023. 1
[67] Chunghyun Park, Yoonwoo Jeong, Minsu Cho, and Jaesik Park. Fast point transformer. In *CVPR*, 2022. 18
[68] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv:2212.09748*, 2022. 4
[69] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation. In *CVPR*, 2024. 18, 19
[70] Gilles Puy, Alexandre Boulch, and Renaud Marlet. Using a waffle iron for automotive point cloud semantic segmentation. In *ICCV*, 2023. 19
[71] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 13, 18
[72] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 13, 18
[73] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *NeurIPS*, 2022. 7, 18
[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 15
[75] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2022. 13
[76] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022. 5, 6, 7, 8, 17
[77] Aditya Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *ECCV*, 2020. 13
[78] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *NeurIPS*, 2019. 13
[79] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, 2021. 13
[80] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 2
[81] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. 13
[82] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 15
[83] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, 2017. 13
[84] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*, 2015. 13
[85] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017. 13
[86] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 2, 7
[87] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*, 2020. 7, 19
[88] Maxim Tatarchenko, Jaesik Park, Vladlen Koltun, and Qian-Yi Zhou. Tangent convolutions for dense prediction in 3d. In *CVPR*, 2018. 18
[89] Lyne Tchapmi, Christopher Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. In *3DV*, 2017. 18
[90] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 13, 18
[91] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 13
[92] Yonglong Tian, Olivier J Henaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *CVPR*, 2021. 1, 13
[93] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023. 1
[94] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, 2017. 4
[95] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *TPAMI*, 2021. 2, 13
[96] Chengyao Wang, Li Jiang, Xiaoyang Wu, Zhuotao Tian, Bohao Peng, Hengshuang Zhao, and Jiaya Jia. Groupcontrast: Semantic-aware self-supervised representation learning for 3d understanding. In *CVPR*, 2024. 18
[97] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *TPAMI*, 2022. 13
[98] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. Graph attention convolution for point cloud semantic segmentation. In *CVPR*, 2019. 18

[99] Li Wang, Dong Li, Han Liu, Jinzhang Peng, Lu Tian, and Yi Shan. Cross-dataset collaborative learning for semantic segmentation in autonomous driving. In *AAAI*, 2022. 2, 13

[100] Peng-Shuai Wang. Octformer: Octree-based transformers for 3D point clouds. In *SIGGRAPH*, 2023. 18

[101] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. Deep parametric continuous convolutional neural networks. In *CVPR*, 2018. 18

[102] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 1

[103] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV*, 2019. 13

[104] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. In *NeurIPS*, 2022. 13

[105] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 18

[106] Wenxuan Wu, Li Fuxin, and Qi Shan. Pointconvformer: Revenge of the point-based convolution. In *CVPR*, 2023. 18

[107] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 7, 13, 14, 18, 19

[108] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *CVPR*, 2023. 2, 3, 7, 8, 13, 18

[109] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *NeurIPS*, 2021. 13

[110] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pretraining for 3d point cloud understanding. In *ECCV*, 2020. 1, 2, 7, 8, 13, 18

[111] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *CVPR*, 2021. 13

[112] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, 2021. 18

[113] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, 2020. 18

[114] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*, 2022. 19

[115] Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. Modeling point clouds with self-attention and gumbel subset sampling. In *CVPR*, 2019. 18

[116] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv:2304.06906*, 2023. 3, 15, 18

[117] Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 2023. 2, 13

[118] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv:2210.07225*, 2022. 2, 13

[119] Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *CVPR*, 2023. 13

[120] Feihu Zhang, Jin Fang, Benjamin Wah, and Philip Torr. Deep fusionnet for point cloud semantic segmentation. In *ECCV*, 2020. 18

[121] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, 2019. 13, 18

[122] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 7, 13, 18

[123] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. Object detection with a unified label space from multiple datasets. In *ECCV*, 2020. 13

[124] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, 2020. 2, 3, 15

[125] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022. 3

[126] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022. 2, 3, 13

[127] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 2, 13

[128] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, 2021. 7, 19