Data Efficient 3D Learner via Knowledge Transferred from 2D Model

Ping-Chung Yu, Cheng Sun, and Min Sun

National Tsing Hua University
{pingchungyu, chengsun}@gapp.nthu.edu.tw, sunmin@ee.nthu.edu.tw

Abstract. Collecting and labeling the registered 3D point cloud is costly. As a result, 3D resources for training are typically limited in quantity compared to the 2D images counterpart. In this work, we deal with the data scarcity challenge of 3D tasks by transferring knowledge from strong 2D models via RGB-D images. Specifically, we utilize a strong and well-trained semantic segmentation model for 2D images to augment RGB-D images with pseudo-label. The augmented dataset can then be used to pre-train 3D models. Finally, by simply fine-tuning on a few labeled 3D instances, our method already outperforms existing state-of-the-art that is tailored for 3D label efficiency. We also show that the results of mean-teacher and entropy minimization can be improved by our pre-training, suggesting that the transferred knowledge is helpful in semi-supervised setting. We verify the effectiveness of our approach on two popular 3D models and three different tasks. On ScanNet official evaluation, we establish new state-of-the-art semantic segmentation results on the data-efficient track. Code: https://github.com/bryanyu1997/Data-Efficient-3D-Learner.

Keywords: knowledge transfer, 3D semantic segmentation, point cloud recognition, 3D pre-training, label efficiency

1 Introduction

Nowadays, 3D sensors are in demand by applications like AR/VR, 3D reconstruction, and autonomous driving. To have a high-level scene understanding (*e.g.*, recognition, semantic segmentation) on the captured 3D data, deep-learning-based models are typically employed for their outstanding performance. As 3D sensors become easier accessible, the architecture of deep 3D models [8,14,29,30,36,39,42] also progresses steadily for better result qualities. In this work, we investigate an orthogonal direction to model architecture design—we present a novel 3D models pre-training strategy to improve performance in a model agnostic manner.

In 2D vision tasks, model pre-training on ImageNet [10] has become a commonly applied strategy for achieving better performances across different downstream tasks. However, there is no standard large-scale dataset like ImageNet to pre-train 3D models due to the considerable effort to acquire and label a diverse set of point cloud data compared to the 2D counterpart. As a result, 3D models are typically trained from scratch, which hinders the performance, especially under the data scarcity scenario of the registered point cloud.

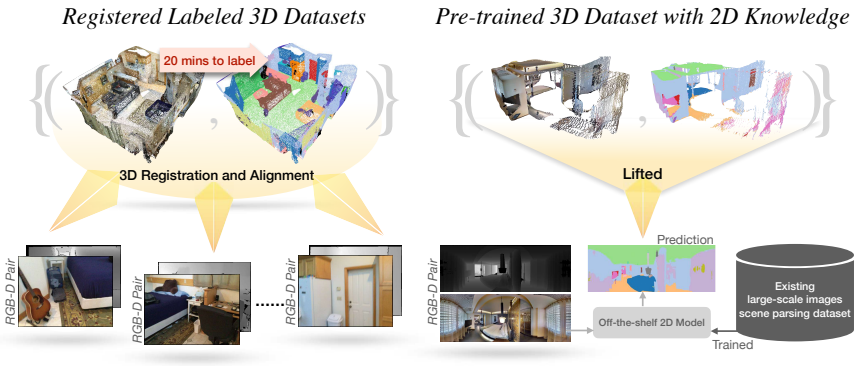

Fig. 1. Left-panel: collecting labeled 3D data is challenging due to the need for (a) robust 3D registration and alignment processes, and (b) a time-consuming human labeling process. Right-panel: a large amount of RGB-D data serve as the bridge between 2D and 3D knowledge. First, pseudo semantic labels are generated by applying an off-the-shelf 2D model on RBG images. Next, the pseudo-labeled 2D data are lifted to 3D using the associated depth map. Finally, we pre-train a 3D model with the large amount of pseudo-labeled 3D data.

To avoid the burden of data labeling, self-supervised learning has emerged as an alternative for pre-training 2D models without labels [3, 6, 7, 16, 18, 19]. To reproduce this trend in 3D data, PointContrast [41] uses contrastive loss to learn the correspondence between two point clouds with visual overlap, improving results on the downstream 3D tasks. However, the diversity and scale of 3D datasets, even releasing the dependency on labeling, are still not comparable to the 2D datasets. For instance, ScanNet [9], which is used in PointContrast's pre-training, has only about a thousand indoor scenes, while ImageNet has more than a million images covering a thousand different classes. As a result, the accuracy improvement by self-supervised learning on point clouds is still limited for 3D tasks.

To address the issue of limited resources of point clouds, we present a novel 3D model pre-training approach via transferring the learned knowledge of a 2D model via RGB-D datasets (see Fig. 1). The single view depth sensor is much cheaper than ever and could be widely popular as a built-in function of phones to capture various scenes. Using RGB-D data as the bridge to transfer knowledge from strong 2D models to 3D models is thus a valuable direction to explore. Specifically, we employ a 2D semantic segmentation model, which is trained on a large and diverse scene parsing dataset, to augment the RGB-D images with pseudo-labels. We then train a 3D model to take the 3D point cloud lifted from RGB-D as input and reproduce the pseudo-labels. By doing so, the 3D models can learn from the strong 2D teacher model and also see a large variety of scenes captured by the RGB-D data.

We demonstrate the effectiveness of our pre-training on semantic segmentation for scene point cloud of the popular ScanNet dataset. Annotating 3D scene point clouds is a demanding task, which takes more than 20 minutes to label a single scene. Some recent approaches emerges to learn from fewer 3D labels to reduce the labeling cost. However, the lack of large-scale pre-training hinders their performances. We show that our pre-training with simple fine-tuning on the scarce label can already outperform existing results tailored for 3D data efficiency [21, 26, 41]. We also show our pre-training can boost the performance of the widely used semi-supervised techniques, which suggests that our pre-training provides an opportunity for future 3D research on both data efficiency and semi-supervised learning to build upon our results for better quality. Finally, despite pre-training on a scene level, we also evaluate our models on object-level tasks and observe improved performances on 3D object classification and shape part segmentation. This suggests that our pre-training is also well transferred to different tasks and input 3D scales.

We summarize our contributions as follows:

- We introduce a pre-training strategy to transfer knowledge from a strong 2D scene parsing model via RGB-D images to 3D models.
- We demonstrate the effectiveness of our pre-training under limited data scenario across two models (*i.e.*, O-CNN [37], SparseConv [14]) and three different tasks (*i.e.*, 3D object classification, point-cloud part segmentation, and indoor point-cloud segmentation).
- By simply finetuning our pre-trained model on a few labels, we establish new state-of-the-art results on ScanNet [9] official evaluation on data-efficient setups, verifying that our pre-training results in data-efficient 3D learners.

2 Related work

Deep 3D models for point cloud understanding. As the 3D analysis studies flourish, several deep models are proposed to extract point cloud features for a high-level understanding. Existing approaches can be classified into point-based and voxel-based methods. PointNet [29] and PointNet++ [30] are the pioneering point-based methods to apply multi-layer perceptron layers directly on point clouds. After that, several convolution-based models [1, 24, 36, 39] are proposed, achieving better quality. Recently, attention-based models [42] have become a new effective way for point cloud processing. On the other hand, the voxel-based method is attractive for its computational-friendly 3D data representations. A discretized step is typically applied on the point cloud before employing the voxel-based models [28, 40].

As the cubic memory complexity limits the resolution of dense voxel grids, 3D sparse CNNs [8, 13, 14] emerges to achieve a feasible space-time complexity for scene-level point clouds, where the CNNs are working on occupied voxels only. OctNet [32] and O-CNN [37] further use the octree data structure to achieve a higher grid resolution efficiently.

As the deep architectures for 3D point clouds progress steadily, all of them are still trained from scratch due to the lack of a large-scale point cloud dataset for pre-training. To sidestep the issue of 3D resources, this work presents a model agnostic pre-training strategy using 2D resources to improve performances.

Data-efficient 3D. Data-efficient learning restricts the amount and the variety of labeled data of the target task, which is helpful for tasks on scene-level point clouds. Registered scene point clouds are hard to acquire and time-consuming to label (*e.g.*, 20 more minutes for a ScanNet scene). Thus, data-efficient solutions are always welcome. Existing works have explored self-supervised pre-training and semi-supervised learning using unlabeled scene point clouds to improve performance. PointContrast [41] sub-samples partial scans and use contrastive learning to pre-train 3D models to identify point correspondents. CSC [21] further improves the pre-training by incorporating spatial scene contexts into the objective. As negative pairs sampling in contrastive learning could be ambiguous, ViewPointBN [27] proposed to use correlation as the objective instead. The aforementioned self-supervised pre-training achieves good label efficiency in a target task agnostic manner. In OTOC [26], pseudo-label-based semi-supervised learning is employed to simultaneously learn from both labeled and unlabeled data, achieving superior label efficiency. We use the abundant 2D resources to pre-train instead of 3D, achieving state-of-the-art label efficiency on scene point clouds.

Knowledge transferred from 2D. Most recently, DepthContrast [45] and Contrastive Pixel-to-point [25] are also proposed to pre-train 3D models using 2D RGB-D datasets. DepthContrast trains 3D models to discriminate 3D point clouds projected from different RGB-D images. Contrastive Pixel-to-point treat each 3D point as instance and apply contrastive loss to learn from pixel features extracted by a trained 2D model.

One challenge of these contrastive-based methods is the ambiguity of the selected negative pairs. Negative pairs sampling is a crucial problem for contrastive pre-training to avoid collapsing solutions [7, 11]. The sampling strategy requires careful designs [6, 19] in the contrastive pre-training, even for images. As the diversity of the employed indoor pre-training data is still limited, the model could have difficulty differentiate two room with similar setup or two 3D points belonging to the same stuff (*e.g.*, walls). Conversely, we directly train the 3D models to reproduce the pseudo labels generated by a strong and well-trained image scene parser, which is more straightforward but non-trivial to learn as we have clear and informative supervision for each 3D point.

2D3DNet [12] trains 3D models with only 2D supervision, which requires the 2D teacher to be a specialist for the downstream task and the 3D model has to co-work with the 2D teacher to achieve good results. Our pre-trained 3D model does not work directly in the downstream 3D task, but it is generally benefit to different downstream task and our 3D model works independently after pre-training.

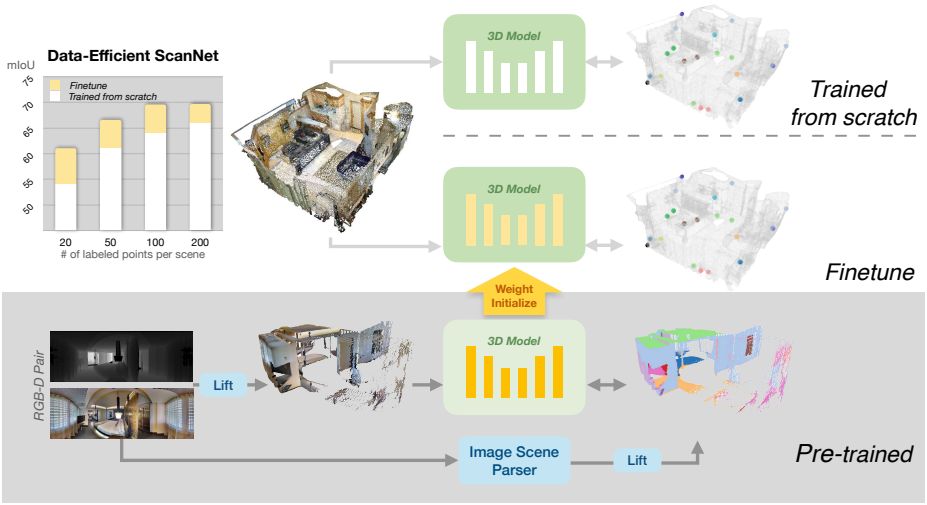

Fig. 2. Overview of our approach. **(Below)** We use a strong and well-trained image scene parser to augment single view RGB-D datasets with pseudo-labels, which is used to pre-train a 3D model in an architecture agnostic manner. **(Top)** Our pre-training improves the results of the limited annotation training.

3 Approach

Our goal is to pre-train 3D models in an architecture agnostic manner. An overview of the proposed approach is illustrated in Fig. 2. Below, we first give a general introduction to 3D models in Section 3.1. In Section 3.2, we detail the proposed pre-training approach. We also apply our pre-trained model on some semi-supervised techniques in Section 3.3.

3.1 3D model

Our pre-training approach is model agnostic and does not require specific 3D architecture designs. Here, we describe a general 3D encoder-decoder to extract deep features from point clouds. A 3D point cloud $\boldsymbol{x}^{(\text{pts})}$ represents a 3D scene or object by a set of 3D coordinates (with additional color or normal features). For voxel-based models, a pre-process input layer before the first model layer is needed to discretize the coordinates into a regular 3D grid. The encoder E, consisting of a sequence of 3D convolution layers, batch normalizations, and nonlinear activation (*e.g.*, ReLU), is a bottom-up way to map the point clouds into a down-sampled latent features $\boldsymbol{z}^{(\text{e})} = E(\boldsymbol{x}^{(\text{pts})})$, which has the high-level understanding of the input scene or object. The decoder D then upsamples and incorporates the low-level feature into $\boldsymbol{z}^{(\text{e})}$ to have a holistic point-wise deep feature $\boldsymbol{z}^{(\text{d})} = D(\boldsymbol{z}^{(\text{e})})$.

Classification. To classify a whole input point cloud into a pre-defined set of classes, we discard the decoder D as we do not need point-level features. The high-level features $\boldsymbol{z}^{(\mathrm{e})}$ are first aggregated into a single latent vector via global max pooling, which is then followed by a classification head ClsHead to map the latent vector into a categorical distribution: $\boldsymbol{y}^{(\mathrm{cls})} = \mathrm{ClsHead}(\mathrm{GlobalMaxPool}(\boldsymbol{z}^{(\mathrm{e})}))$, where the ClsHead consists of a linear layer and a Softmax layer.

Point-level semantic segmentation. The semantic segmentation head simply maps the point features into per-point class distribution by a linear and a Softmax layer: $\boldsymbol{y}^{(\mathrm{ss})} = \mathrm{SegHead}(\boldsymbol{z}^{(\mathrm{d})})$.

3.2 Knowledge transferred from 2D

Our approach aims to pre-train a 3D model by transferring the knowledge of a strong 2D model learned from a large-scale 2D dataset. However, the 2D model $F^{(\mathrm{2D})}$ takes images $\boldsymbol{x}^{(\mathrm{img})}$ as input while the 3D model takes point clouds $\boldsymbol{x}^{(\mathrm{pts})}$. To learn from 2D models, we use RGB-D dataset $\left\{\left(\boldsymbol{x}^{(\mathrm{img})}, \boldsymbol{d}^{(\mathrm{img})}\right)\right\}$ as the bridge, where $\boldsymbol{d}^{(\mathrm{img})}$ is the depth map of a image. Our idea is to use a well-trained 2D model to generate pseudo-label for the images in the RGB-D dataset

$$\boldsymbol{t} = F^{(\mathrm{2D})}\left(\boldsymbol{x}^{(\mathrm{img})}\right) \ , \tag{1}$$

and then we lift the image to point cloud using their depth maps

$$\boldsymbol{x}^{(\mathrm{img2pts})} = \mathrm{Lift}\left(\boldsymbol{x}^{(\mathrm{img})}, \boldsymbol{d}^{(\mathrm{img})}\right) \tag{2}$$

so that we can train the 3D model using the augmented RGB-D dataset $\left\{\left(\boldsymbol{x}^{(\mathrm{img2pts})}, \boldsymbol{t}\right)\right\}$. Below, we introduce the explored two different sources of RGB-D data and detail our pre-training strategy.

Lifting perspective images. With the provided depth maps and the camera intrinsic K, we can lift a 2D image coordinate $\left[u, v\right]$ to the 3D camera coordinate $\left[x, y, z\right]$ by:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \ , \tag{3}$$

where d is the depth value of a pixel and $\left[u, v, 1\right]^{\intercal}$ is the homogeneous coordinate.

Lifting panoramic images. Panoramic images cover more information in one shot attributed to the omnidirectional field of view compared to perspective images. Unlike perspective depth maps where z-values are recorded, panoramic depth maps directly record the distance between the observed 3D points to camera. We can lift a panoramic image to 3D by:

$$\begin{cases} x = d \cdot \cos(v) \cdot \cos(u) \ ; \\ y = d \cdot \cos(v) \cdot \sin(u) \ ; \\ z = -d \cdot \sin(v) \ , \end{cases} \tag{4}$$

where d is the recorded distance of a pixel in the panoramic depth maps, and $u \in [-\pi, \pi], v \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ are the panoramic image coordinate in UV space.

Learning from 2D scene parser via soft pseudo-label. Image scene parsing aims to classify each pixel of images, which provides a thorough and detailed understanding of the captured scenes. Besides, existing training corpora for image scene parsing [2, 48] is abundant, with more than 10k images covering a large variety of scenes and classes. Scene parsing models [5, 22, 31, 44, 46, 49] have also progressed steadily and achieved strong performances. In this work, we adopt DPT [31] for its outstanding performance, and we find it generalized well on both perspective and panoramic images.

We first attach a point semantic segmentation head $\text{SegHead}^{(\text{pre})}$ on the 3D encoder decoder to classify C classes of the trained DPT

$$\boldsymbol{y}^{(\text{img2pts})} = \text{SegHead}^{(\text{pre})}\left(D\left(E\left(\boldsymbol{x}^{(\text{img2pts})}\right)\right)\right) \ . \tag{5}$$

Instead of the one-hot decision, soft labels are learned as the target during pre-training. Soft-label is first introduced in a knowledge distillation method [20], where each pixel or point is labeled by the categorical distribution over classes. We use cross-entropy loss as the training objective for the 3D model to learn from the soft-label:

$$\mathcal{L}^{(\text{pretrain})} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(-\boldsymbol{t}_i[c]\log\left(\boldsymbol{y}_i^{(\text{img2pts})}[c]\right)\right) \ , \tag{6}$$

where i is the index to the N 3D points, and t_i is the soft-label probability of the pixel corresponding to the i-th 3D point.

Downstream tasks. After pre-training, we discard the pre-training head $\text{SegHead}^{(\text{pre})}$ and directly used the trained encoder E and decoder D in the downstream tasks.

3.3 Semi-Supervised

Semi-supervised learning is helpful to learn from scarce labels, where the improvement by our pre-training can be additive to it to achieve an even better result. We use two simple and commonly used semi-supervised techniques in this work. First, entropy minimization [15] encourages concentration of the predicted probability distribution on the unlabeled data:

$$\mathcal{L}^{(\text{mini-entropy})} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(-\boldsymbol{y}_i[c]\log\left(\boldsymbol{y}_i[c]\right)\right) \ . \tag{7}$$

Second, we use mean-teacher [35] to guide our model, where the weights of teacher model $\theta_t{}'$ is obtained by exponential moving averaging (EMA) of the weight of student model θ_t across training step t:

$$\theta_t{}' = \alpha\theta_{t-1}{}' + (1-\alpha)\theta_t \ , \tag{8}$$

where α is a smoothing hyperparameter. We use mean squared error to encourage consistency between our predictions $\boldsymbol{y}$ and teacher predictions $\boldsymbol{y}'$

$$\mathcal{L}^{(\text{consistency})} = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{y}_i - \boldsymbol{y}'_i\|^2 \ . \tag{9}$$

Different from the 2d teacher model we mention in Section 1, the mean-teacher here is a typical semi-supervised technique.

4 Experiments

In this section, we evaluate our pre-training strategy under the data scarcity scenario. The implementation details are provided in Section 4.1. To demonstrate the benefits of our approach, we conduct a series of experiments on scene understanding task in Section 4.2 and the shape analysis tasks in Section 4.4. We show the results of our pre-training on different 3D models in Section 4.3 and present the ablation experiments in Section 4.5.

4.1 Implement details

Pre-training images dataset. For perspective pre-training, we utilize SUNRGB-D [34], which consists of more than 10,000 RGB-D indoor images and the corresponding pixel-wise semantic labels from 37 categories. In addition to the depth maps, SUNRGB-D also provides the intrinsic which allow us to lift the 3D scenes. We split the dataset into 5,285 training sets and 5,050 validation sets in the pre-training procedure. We only used the provided ground-truth semantic maps in ablation experiments and used the generated pseudo-label by our approach in all other experiments. For panoramic pre-training, Matterport3D [4] provides various indoor scenes captured by panoramic images, each of which covers a omnidirectional field of view. Matterport3D contains 10,800 panoramic views with corresponding depth maps from 90 building-scale scenes, including 61 scenes for training, 11 for validation, and 18 for testing following official data splits.

Image scene parser. We employ DPT [31] as a teacher model in our pre-training. The DPT is first trained on the ADE20K [48] dataset, which has 20k images with ground-truth semantic segmentation maps covering 150 different classes. In pre-training, the DPT is fixed and used to generate soft pseudo-labels to transfer its knowledge to the 3D models.

3D models. SparseConv [14] is trained by Adam optimizer [23] for scene semantic segmentation. The models are trained on NVIDIA GTX 1080Ti GPUs for 200 epochs with batch size 4 and learning rate 0.001. O-CNN [37] is implemented for shape analysis and indoor scene semantic segmentation. During lifted scenes pre-training, the projected-points are formed as 512^3 resolution of leaf octants as the same as indoor scene semantic segmentation. For part segmentation, we train

networks for 600 epochs with batch size 32, leaning rate 0.025 and weight decay 0.0001. For scene semantic segmentation, we train models for 500 epochs with batch size 4 and leaning rate 0.05. Both tasks adopt the SGD optimizer with a momentum 0.9 and use polynomial schedulers powered by 0.9. The models are trained on several NVIDIA Tesla V100-SXM2 GPUs. As the setup for part segmentation, the classification model is trained for 300 epochs with batch size 32 and learning rate 0.05. The O-CNNs are adopted with 32^3, 64^3 and 512^3 resolution of leaf octants for object classification, shape part segmentation and scene semantic segmentation, respectively. The point-wise predictions are interpolated by linear for part segmentation and nearest for scene semantic segmentation.

Data augmentation. We apply random rotation, random scaling, random elastic distortion, random color contrast, and random color jittering to the input point cloud as data augmentation.

4.2 Data efficient scene semantic segmentation

We validate the effectiveness of our pre-training on ScanNet Data Efficient Benchmark. ScanNetV2 [9] consists of various indoor scenes formed as 3D point clouds and corresponding semantic annotations. The data is taken from 707 distinct spaces and covers 20 semantic classes. We follow the official data scarcity scenarios: *i)* Limited Annotations (LA) considers only a few labeled points in each scene, *ii)* and Limited Reconstructions(LR) considers the a few number of labeled scenes. We use O-CNN pre-trained by our approach on Matterport3D dataset for all the results submitted to official ScanNet.

Limited annotations. Following the official configuration in the *3D Semantic label with Limited Annotations benchmark*, there are four different scales, including $\{20, 50, 100, 200\}$ labeled points per training scene. We show the quantitative comparison on the testing split in Table 1. We achieve state-of-the-art results on purely supervised fine-tuning setup and semi-supervised setup. It is worth noting that our results without using unlabeled data already outperforms the semi-supervised OTOC [26] on 50, 100, and 200 labeled points per scene, which further highlights the effectiveness of our pre-training approach.

Limited reconstruction. Limited Reconstruction (LR) is constructed by limiting the number of scenes. The official subset is randomly sampled from 1201 scenes, where $\{1\%, 5\%, 10\%, 20\%\}$ of training data is subsampled (corresponding to 12, 60, 120, and 240 scenes). We compare our results on *3D Semantic label with Limited Reconstructions benchmark* in Table 2. When only 1% of the training data is available, all methods achieve similar performance. When $5\%, 10\%, 20\%$ of the training data is given, we achieve a superior mIoU compared to previous methods. Additionally, we adopt semi-supervised learning and examine the capability under the limited reconstruction scenario. We first trained our model on the given labeled data, and then we randomly sample a subset of unlabeled scenes to

Table 1. Quantitative comparisons on official ScanNet Limited Annotations(LA) track. For a fair comparisons, the results are separated based on whether the unlabeled data points are used in training.

Method	semi	# of labeled points per scene			
		20	50	100	200
PointContrast [41]	-	55.0	61.4	63.6	65.3
CSC [21]	-	53.1	61.2	64.4	66.5
ViewPointBN [27]	-	54.8	62.3	65.0	66.9
Ours	-	57.9	65.8	**71.4**	**71.1**
OTOC [26]	✓	59.4	64.2	67.0	69.4
Ours	✓	**63.9**	**69.5**	70.4	70.9

Table 2. Quantitative comparisons on offical ScanNet Limited Reconstructions(LR) track.

Method	semi	percentage of labeled scene			
		1%	5%	10%	20%
PointContrast [41]	-	25.3	43.8	55.5	60.3
CSC [21]	-	**27.0**	46.0	57.5	61.2
ViewPointBN [27]	-	25.6	45.2	56.6	62.5
Ours	-	26.6	46.7	**61.2**	64.0
Ours	✓	26.3	**50.8**	60.8	**66.3**

produce pseudo labels. The pseudo labels is generated by selecting the top-most 20% confident predictions of each class. We can then combined the labeled, pseudo-labeled, and unlabeled data points to train our models.

From the comparisons in Table 1 and Table 2, we find our pre-trained models with simple supervised fine-tuning already outperform existing methods tailored for 3D data-efficiency. Combining our pre-training with semi-supervised sometimes can further improve our results. In cases that semi-supervised does not improve, we still achieve similar performance. As we only use the simplest semi-supervised techniques, a more tailored one could be more helpful. In sum, our pre-training provides a good starting point for future work to develop 3D data-efficient approach.

4.3 Pre-training on different 3D models

To verify the effectiveness of our pre-training strategy on different 3D models, we implement O-CNN [37] and SparseConv [14] for 3D scene semantic segmentation. SparseConv uses a sparse voxelized input representation and keeps the same level of sparsity throughout the model. O-CNN builds up the octree structure of 3D points representation. As the octree depth increases, the octree-based CNN layers extract the information with higher resolution. Both models work on

Table 3. mIoU of different 3D segmentation model and different pre-training strategy. The results are evaluated on ScanNet validation set. The oracle pre-training supervision is the ground-truth semantic segmentation maps provided in the pre-training dataset.

Based model	Pre-training strategy	Oracle pre-training supervision	LA (points) 20	200
SparseConv	trained from scratch	-	51.6	65.4
	perspective pre-training	✓	55.4	66.0
	perspective pre-training	-	56.2	65.7
	panoramic pre-training	-	58.5	67.3
O-CNN	trained from scratch	-	52.9	65.0
	perspective pre-training	✓	56.9	66.2
	perspective pre-training	-	56.2	65.3
	panoramic pre-training	-	55.6	67.8

occupied point sets only to ensure computational efficiency, and the networks are constructed by U-Net [14, 33] architecture to produce point-wise predictions. The overall results are summarized in Table 3. On both SparseConv and O-CNN, all different variations of our pre-training approach improved mIoU significantly over training from scratch.

4.4 Shape analysis under limited data scenario

Our approach discussed above presents a positive outcome on scene-level semantic segmentation under the limited data scenario. In this section, we apply our pre-trained model on object-level 3D tasks such as object classification and shape part segmentation. With the data scarcity scenario for shape analysis, the implementation details and results are presented in the following paragraph. For all the experiments in shape analysis, we use O-CNN pre-trained by our approach on Matterport3D, which can also show the generalizibility from scene-level to object-level of our pre-training.

Object classification Object classification is the fundamental task of shape analysis. Given the 3D shape represented as point clouds, the classification model intends to assign the category of the input objects. With our approaches, the 3D models are required to extract features by limited annotations.

Dataset. ModelNet40 [40] dataset contains 12,311 shapes from 40 object categories, and is split into 9,843 objects for training and 2,468 objects for testing. For the limited data scenario, we randomly sample the subset by ratio $\{1\%, 5\%, 20\%\}$ in each category from the training set, and evaluate on the original testing set. For semi-supervised learning, the unlabeled batches is sampled randomly from the remained data of the training set.

Results. We finetune the 3D models with panoramic images pre-trained for object classification. The results are shown in Table 4, which is divided by whether the semi-supervised training is used. Considering the model trained without

Table 4. Accuracy of 3D object classification in ModelNet40 [40] dataset under limited training data. The training data is random sampled by 1%, 5%, 20% and the results are evaluated on the same testing set.

Pre-training strategy	Semi-supervised	1%	5%	20%
Trained from scratch	-	60.2	79.6	86.5
	✓	61.1	80.4	87.6
Our pre-training	-	65.5	80.6	87.5
	✓	**69.0**	**82.9**	**89.4**

Table 5. mIoU of shape part segmentation in ShapeNet [40] dataset under sampled training data. We use category mIoU across all categories and instance mIoU across all shape instance as the evaluation metrics. The results are evaluated on ShapeNet testing set which split by [43].

Pre-training strategy	Category mIoU			Instance mIoU		
	1%	5%	20%	1%	5%	20%
Trained from scratch	63.3	67.5	73.9	67.9	72.2	76.7
Our pretraining	**64.1**	**74.1**	**76.3**	**68.5**	**76.8**	**78.9**

unlabeled data, our pre-training strategy can directly improve the accuracy compared with the model trained from scratch. In the cases that the models are trained from scratch, the improvements of using the unlabeled data are limited. Conversely, our pre-training with only supervised fine-tuning already outperform the results of semi-supervised with random weight initialization. Besides, the improvements by semi-supervised learning is much significant than the improvement from random initialization.

Shape part segmentation Shape part segmentation is more complicated than classification for shape analysis. Part segmentation is expected to generate the dense prediction and assign the part category to each point in objects.

Dataset. Yi *et al.* [43] annotates a subset of ShapeNet [40] 3D models with semantic part labels. The annotated subset of ShapeNet contains 2 to 6 parts per category, and 50 distinct parts in total among 16 shape categories. For training on the limited scenario, we build the limited subset by random sampling as ratio $\{1\%, 5\%, 20\%\}$ from each shape category, and then evaluate on the original validation set. According to the setup of pre-processing in [37], the input points are condensed by triangle faces and built in octree structure.

Results We finetune the segmentation models with panoramic images pre-training in each category separately, and evaluate the performance of part segmentation by mean IoU across all categories and mIoU across all object instances. The results are shown in Table 5. As the data split in [17,38,47] are not provided, we only compare the results on our own split. With the knowledge distilled from off-the-shelf image scene parser, the pre-trained model performs better category mIoU and instance mIoU across all limited scales (1%, 5% and 20%).

4.5 Ablation study

Image modality To observe the effect of the distillation via different image modalities, we pre-train O-CNN [37] on perspective images and panoramic images, and the supervised fine-tune the O-CNN on ScanNet with limited annotations. The results are demonstrated in Table 3. For all the variations of our pre-training, we achieve significant better results than training from scratch. Note that the panoramic images pre-training rises the performance by 2.7% and 2.8% mIoU compared with baseline on 20, 200 labeled points of scene(55.6% vs 52.9% and 67.8% vs 65.0%). The results imply that models can learn non-trivial and informative representation from the strong and well-trained image scene parser via RGB-D images. The learned representation boosts the capability of scene understanding under the limited annotations.

This work can not conclude whether perspective pre-training or panoramic pre-training is better, as the number of their training data is different. The panoramic data used in this work leads us to better results, so we use panoramic pre-training in all the other experiments.

Pre-training by 2D annotations and pseudo labels In our pre-training strategy, we supervise 3D models by the pseudo-labels introduced in Section 3.2. To examine the difference between the information distilled from pseudo labels and ground-truth labels, we conduct our pre-training strategy on perspective images with ground-truth semantic labels and generated pseudo-labels. The ground-truth semantic labels are provided by SUNRGB-D [34], while the pseudo labels are predicted by a strong and well-trained image scene parser. With the 3D model trained on limited annotations scenario, the results are presented in Table 3. The perspective images attain higher performance than the model trained from scratch regardless of whether the ground-truth labels are provided. The results imply that the pseudo labels can also improve the knowledge distillation during pre-training. Additionally, with provided annotations of images, our pre-training strategy achieves better mIoU than the baseline by 4% mIoU on 20 points and 1% mIoU on 200 points. Overall speaking, the less constrained pseudo-labels pre-training can achieve comparable results to the ground-truth labels pre-training, which suggests that ground-truth semantic maps are not necessary in our pre-training approach.

Combining pre-training with semi-supervised learning We study the effectiveness of our pre-training when combining with semi-supervised learning to simultaneously learn from labeled and unlabeled data. We run semi-supervised learning with random initialized model weights and our pre-trained model weights, and then evaluate the performance on ScanNetv2 validation sets. The approach is examined on both limited annotation(LA) and limit reconstruction(LR).

For LA, the results are summarized in Table 6. In "trained from scratch" column, the semi-supervised learning slightly improves mIoU when unlabeled data is used during training. In contrast, our pre-training strategy increases the performance significantly when combining with the semi-supervised learning,

Table 6. Ablation studies of our pre-training with and without semi-supervised learning. We follow the limited training data from official. The results are presented as mIoU and compared on ScanNet validation set.

Method	semi	# of labeled points per scene			
		20	50	100	200
Trained from scratch	-	54.2	61.1	64.1	65.9
Our pre-training	-	**55.9**	**63.4**	**70.2**	**69.4**
Trained from scratch	✓	56.6	62.5	65.3	66.0
Our pre-training	✓	**61.1**	**66.6**	**69.6**	**69.7**

Table 7. mIoU of limited scene variety on ScanNet validation set.

Model	Pre-training strategy	LR(%)			
		1	5	10	20
O-CNN	Trained from scratch	18.7	39.7	52.4	59.6
	Our pre-training	26.3	44.9	56.5	62.7
	Our pre-training + Semi-supervised	**27.5**	**49.4**	**59.2**	**64.8**

particularly in the 20, 50 of labeled points per scene. Using unlabeled data with our pre-training strategy outperforms the models trained from scratch by 4.5% and 4.1%.(61.1% vs 56.6% and 66.6% vs 62.5%).

For LR, we follow Section 4.2 for semi-supervised training. As a result, Table 7 shows that our pre-training strategy achieves greater mIoU by 27.5%, 49.4%, 59.2% and 64.8% on $\{1\%, 5\%, 10\%, 20\%\}$. Consequently, unlabeled data enhances the ability of scene understanding on both limited annotations and limited reconstruction.

5 Conclusion

This work presents a new 3D deep models pre-training strategy. We use RGB-D images as the bridge to transfer the knowledge from a strong and well-trained 2D scene parsing network to 3D models. Our pre-training strategy is model agnostic, and we show its effectiveness on two popular 3D models architectures. On the official ScanNet data-efficient track, we establish new state-of-the-art results. Besides, we also show that the improvement by our pre-training is additive to other label-efficient techniques. We hope our pre-trained weights can serve as a stepping stone for future 3D approaches and encourage more exploration on how to make good use of 2D resources in 3D tasks.

Acknowledgements This work is supported in part by Ministry of Science and Technology of Taiwan (MOST 110-2634-F-002-051). We would like to thank National Center for High-performance Computing (NCHC) for computational and storage resource.

References

1. Boulch, A.: Convpoint: Continuous convolutions for point cloud processing. Comput. Graph. pp. 24–34 (2020) 3
2. Caesar, H., Uijlings, J.R.R., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: CVPR. pp. 1209–1218 (2018) 7
3. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: NeurIPS (2020) 2
4. Chang, A.X., Dai, A., Funkhouser, T.A., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from RGB-D data in indoor environments. In: 3DV. pp. 667–676 (2017) 8
5. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. pp. 834–848 (2018) 7
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020) 2, 4
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021) 2, 4
8. Choy, C.B., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR. pp. 3075–3084 (2019) 1, 3
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 2432–2443 (2017) 2, 3, 9
10. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 1
11. *et al.*, Z.: Barlow twins: Self-supervised learning via redundancy reduction. In: ICML (2021) 4
12. Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T.A.: Learning 3d semantic segmentation with only 2d image supervision. In: 3DV (2021) 4
13. Graham, B.: Sparse 3d convolutional neural networks. In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) BMVC. pp. 150.1–150.9 (2015) 3
14. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR. pp. 9224–9232 (2018) 1, 3, 8, 10, 11
15. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NeurIPS. pp. 281–296 (2005) 7
16. Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - A new approach to self-supervised learning. In: NeurIPS (2020) 2
17. Hassani, K., Haley, M.: Unsupervised multi-task feature learning on point clouds. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. pp. 8159–8170. IEEE (2019) 12
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 (2021) 2
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9726–9735 (2020) 2, 4

20. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 7
21. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: CVPR. pp. 15587–15597 (2021) 3, 4, 10
22. Hsiao, C., Sun, C., Chen, H., Sun, M.: Specialize and fuse: Pyramidal output representation for semantic segmentation. In: ICCV (2021) 7
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015) 8
24. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) NeurIPS. pp. 828–838 (2018) 3
25. Liu, Y.C., Huang, Y.K., Chiang, H.Y., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. arXiv preprint arXiv:2104.04687 (2021) 4
26. Liu, Z., Qi, X., Fu, C.: One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In: CVPR. pp. 1726–1736 (2021) 3, 4, 9, 10
27. Luo, L., Tian, B., Zhao, H., Zhou, G.: Pointly-supervised 3d scene parsing with viewpoint bottleneck. arXiv preprint arXiv:2109.08553 (2021) 4, 10
28. Maturana, D., Scherer, S.A.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015, Hamburg, Germany, September 28 - October 2, 2015. pp. 922–928 (2015) 3
29. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR. pp. 77–85 (2017) 1, 3
30. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: NeurIPS. pp. 5099–5108 (2017) 1, 3
31. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV. pp. 12179–12188 (2021) 7, 8
32. Riegler, G., Ulusoy, A.O., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: CVPR. pp. 6620–6629 (2017) 3
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015) 11
34. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR. pp. 567–576 (2015) 8, 13
35. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: ICLR (2017) 7
36. Thomas, H., Qi, C.R., Deschaud, J., Marcotegui, B., Goulette, F., Guibas, L.J.: Kpconv: Flexible and deformable convolution for point clouds. In: ICCV. pp. 6410–6419 (2019) 1, 3
37. Wang, P., Liu, Y., Guo, Y., Sun, C., Tong, X.: O-CNN: octree-based convolutional neural networks for 3d shape analysis. ACM Trans. Graph. pp. 72:1–72:11 (2017) 3, 8, 10, 12, 13
38. Wang, P., Yang, Y., Zou, Q., Wu, Z., Liu, Y., Tong, X.: Unsupervised 3d learning for shape analysis via multiresolution instance discrimination. In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021. pp. 2773–2781. AAAI Press (2021) 12

39. Wu, W., Qi, Z., Li, F.: Pointconv: Deep convolutional networks on 3d point clouds. In: CVPR. pp. 9621–9630 (2019) 1, 3
40. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: CVPR. pp. 1912–1920 (2015) 3, 11, 12
41. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L.J., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: ECCV. pp. 574–591 (2020) 2, 3, 4, 10
42. Yan, X., Zheng, C., Li, Z., Wang, S., Cui, S.: Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In: CVPR. pp. 5588–5597 (2020) 1, 3
43. Yi, L., Kim, V.G., Ceylan, D., Shen, I., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.J.: A scalable active framework for region annotation in 3d shape collections. ACM Trans. Graph. pp. 210:1–210:12 (2016) 12
44. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) ECCV. pp. 173–190 (2020) 7
45. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: ICCV (2021) 4
46. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. pp. 6230–6239 (2017) 7
47. Zhao, Y., Birdal, T., Deng, H., Tombari, F.: 3d point capsule networks. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 1009–1018. Computer Vision Foundation / IEEE (2019) 12
48. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR. pp. 5122–5130 (2017) 7, 8
49. Zhu, Z., Xu, M., Bai, S., Huang, T., Bai, X.: Asymmetric non-local neural networks for semantic segmentation. In: ICCV. pp. 593–602 (2019) 7