To Supervise or Not to Supervise: Understanding and Addressing the Key Challenges of Point Cloud Transfer Learning

Souhail Hadgi[1], Lei Li[2], and Maks Ovsjanikov[1]

[1] LIX, Ecole Polytechnique, IP Paris
[2] Technical University of Munich

Abstract. Transfer learning has long been a key factor in the advancement of many fields including 2D image analysis. Unfortunately, its applicability in 3D data processing has been relatively limited. While several approaches for point cloud transfer learning have been proposed in recent literature, with contrastive learning gaining particular prominence, most existing methods in this domain have only been studied and evaluated in limited scenarios. Most importantly, there is currently a lack of principled understanding of both *when* and *why* point cloud transfer learning methods are applicable. Remarkably, even the applicability of standard *supervised* pre-training is poorly understood. In this work, we conduct the first in-depth quantitative and qualitative investigation of supervised and contrastive pre-training strategies and their utility in downstream 3D tasks. We demonstrate that layer-wise analysis of learned features provides significant insight into the downstream utility of trained networks. Informed by this analysis, we propose a simple geometric regularization strategy, which improves the transferability of supervised pre-training. Our work thus sheds light onto both the specific challenges of point cloud transfer learning, as well as strategies to overcome them.

1 Introduction

Transfer learning is an integral part of the success of deep learning in 2D computer vision, enabling the use of powerful pre-trained architectures, which can be adapted with limited data on a broad range of downstream tasks and datasets [6, 20, 57]. The advancement of transfer learning can be attributed to several factors such as the availability of well-established architectures (*e.g.* deep CNNs [20], ViTs [11]), an abundance of source data to train on (*e.g.* ImageNet [10] among myriad others), as well as the development of different pre-training strategies [23].

Unfortunately, the extension of such approaches to the 3D domain presents several important challenges. First, and perhaps most importantly, large-scale labeled datasets are scarce in 3D due to the complexity of data acquisition and annotation as well as significant domain-specificity that creates imbalances across data (*e.g.* synthetic CAD shapes [5, 24, 48], real scenes [9], non-rigid shapes [4],

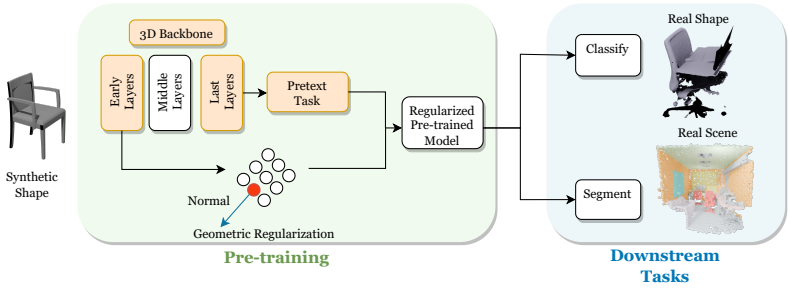


Fig. 1: Analyzing and improving point cloud transfer learning. In this work, we perform the first in-depth study of the different components of network pre-training that influence the outcome of point cloud transfer learning (components in solid orange). This includes the source data domain in relation to the downstream target data, the choice of architecture, the importance of early vs later layers, and the pre-training design choices. We also show that improvements to the pipeline can be achieved through regularization of the early layers, by promoting the prediction of *geometric* properties.

etc.). Such imbalances, coupled with limited training data, can strongly hinder the applicability of transfer learning solutions [51].

Another key challenge in learning with 3D data is the nature of the data *representation*, which most often comes in the form of unstructured point clouds, potentially with attributes (*e.g.* spatial, color). This limits the use of architectures that rely on regular grid structure, and furthermore, poses challenges due to significant variability in *sampling* properties (density, acquisition artefacts, etc.). As a result, while a large number of different architectures have been proposed for 3D deep learning [16], there is a lack of universal solutions applicable to an arbitrary scenario.

In spite of these challenges, several works have recently emerged aiming to enable transfer learning for point clouds, especially by focusing on new pre-training strategies, *e.g.*, [21, 49, 51]. Among these strategies, several approaches have shown that *self-supervised strategies* (*e.g.* contrastive learning, point cloud reconstruction) can improve performance on different downstream data (*e.g.* shapes, scenes) and on select set of tasks and datasets.

Despite the growing number of these approaches, unfortunately, most existing methods are only presented and evaluated in very specific settings, and missing comparison to baselines such as supervised pre-training. Specifically, there are several questions not addressed in existing literature, including: 1) The utility of supervised compared to contrastive pre-training, 2) The role of the architecture's inherent properties in the success of point cloud transfer learning, and 3) The impact of regularization strategies on the efficacy of point cloud pre-training.

By performing the first in-depth investigation of the most prominent approaches, we observe that in the *linear probing setting*, supervised pre-training leads to superior performance compared to contrastive learning across almost all tested architectures, even under relatively significant domain shift. This sug-

gests that within 3D deep learning, contrastive learning does not always lead to more "universally useful" features. At the same time, we also observe that in the full *fine-tuning scenario*, contrastive pre-training can outperform supervised learning. By analyzing gradients of pre-trained networks, we show that the *early layers* learned with supervised pre-training are not easily adaptable across new datasets, compared to those learned with contrastive learning. This sheds new and specific light on the differences between these pre-training strategies in the context of point cloud analysis. Finally, we demonstrate that a simple regularization scheme applied to the early layers during supervised pre-training can overcome this limitation and leads to networks that can be fine-tuned efficiently.

To summarize, our main contributions are as follows:

1. We perform the first comparison, within a single consistent evaluation framework, of supervised and contrastive pre-trained models using different point cloud backbones, assessing their transfer learning performance on common downstream data and tasks. We separate between the linear probing and fine-tuning settings and analyze the difference in performance.
2. We observe a clear quantifiable difference in architecture's layer properties and find that, remarkably, for 3D architectures, *even the first layers* have the ability to discriminate between downstream shape classes. Moreover we show that supervised pre-training can lead to early layers that do not easily adapt to new data.
3. To counteract this effect, we introduce a simple *layer-wise* geometric regularization procedure on supervised pre-training, which leads to performance improvement and outperforms contrastive learning in several settings. We finally strongly correlate the improvement to the key signals identified in layer gradient norm.

Throughout all our analysis, we take special care to make all models and evaluation settings consistent and comparable.

2 Related Work

Transfer learning has facilitated myriad applications in 2D image analysis, such as segmentation [19, 28], detection [37], style transfer [14], and medical image analysis [30, 43], among others. This can be largely attributed to the success of deep neural networks [20, 25] on the large-scale visual recognition dataset ImageNet [10]. In contrast, research on generic 3D representation learning, driven by the development of 3D deep learning techniques [8, 33, 47] in recent years, is still emerging.

Contrastive learning has garnered growing attention recently as a promising unsupervised representation learning paradigm, which can be conceptualized as a dictionary look-up process [18]. In this process, a query and a set of keys are encoded by some network, and a contrastive loss seeks to maximize the similarity between the query and its single positive key and minimize that between the query and its all negative keys.

For contrastive pre-training in 3D, several studies have contributed to the design of effective learning setups. In particular, Xie *et al.* introduced PointContrast [51], which utilizes two rigidly transformed views of point clouds and performs point-level feature contrasting in the two views. They demonstrated that the pre-trained network brings noticeable improvement in downstream 3D learning tasks, including object classification [5], part segmentation [52], semantic segmentation [3,9], and object detection [9,40]. Follow-up works investigated this direction further by constraining point contrasting in local regions [21], adopting patch-level [12], object-level [36], object and point-level, or even scene-level [22,56] contrasting, as well as incorporating image data into the 3D feature contrasting via point-pixel pairs [26, 27]. However, most of the aforementioned works focus on demonstrating effective representation learning with specific network architectures and data modalities, lacking a systematic study of both when and why contrastive pre-training enables 3D transfer learning. In response to this gap, our work delves into the contrastive pre-training strategies across various backbones and downstream tasks, aiming to reveal the essential aspects for a successful 3D transfer learning.

To achieve self-supervised pre-training, several studies have examined an alternative reconstruction-based learning strategy, such as recovering missing points [46, 54], solving 3D jigsaw tasks [2, 7, 13, 38, 39], or augmenting auto-encoding with clustering and classification tasks [17]. We refer the interested reader to [41] for a more comprehensive study of this pre-training paradigm.

Finally, we note that several works have aimed to analyze the factors behind successful transfer learning approaches in 2D image tasks [30,35,57]. These works point to the special importance of early and mid-level features and to layer-wise analysis of learned representations. Such analysis has not yet been performed for 3D data, and our main objective is to fill this gap by performing a first comprehensive investigation of this topic. As we mention below, our work highlights the special nature of learning on 3D data, thus shedding light on the challenges of point cloud transfer learning and possible ways to address them.

3 Datasets and Architectures

3.1 Datasets

Pre-training Dataset. Prior 3D transfer learning approaches [1,21,46,51] have typically focused on pre-training datasets with limited domain shifts to downstream tasks, such as within the synthetic shape setting [5, 48] or the real scene scenario [3,9]. Interestingly, among them, PointContrast [51] argued against pre-training on synthetic data and stated that supervised pre-training is an upper bound to information gain from pre-training (see Sec. 3.1 therein). Our experimental findings (Sec. 5), however, suggest that this is not always the case and help to refine this statement. We thus select ShapeNet [5] to evaluate the utility of pre-training with synthetic data, and focus on its transfer utility in downstream tasks with varying domain shifts, as elaborated below.

Downstream Datasets and Tasks. To comprehensively investigate different 3D transfer learning scenarios, we consider downstream tasks encompassing both 1) *synthetic* vs. *real* data and 2) *object* vs. *scene*-level 3D data, which exhibit a wide range of similarity to ShapeNet (51,300 annotated 3D synthetic shapes in 55 categories), used for pre-training:

- **ModelNet40** [48] is a synthetic object-level dataset with 3D shapes resembling those in ShapeNet. It consists of 12,311 shapes and 40 object classes for the shape classification task, and 20 of these classes can be found in ShapeNet.
- **ScanObjectNN** [45] is a real object-level dataset with scanned 3D shapes featuring noise and background, vastly different from the clean shapes in ShapeNet. There are 15,000 shapes and 15 object classes for the shape classification task, and 9 of these classes can be found in ShapeNet.
- **S3DIS** [3] is a real scene-level dataset consisting of 3D scans of 6 large-scale indoor areas in office buildings and 13 semantic categories for the semantic segmentation task.

3.2 Architectures

To provide an extensive study that tackles diverse 3D backbones, we follow a broadly adopted categorization of 3D deep neural networks [50] and select a diverse set of five representative architectures as follows:

- **Point-based**. PointNet [33] is a pioneering 3D backbone that is characterized by its *global* nature in feature learning, which has been shown to be beneficial in robust point cloud analysis [42], but, as we show below, can overfit to pre-training data. In addition, we consider PointMLP [29], which is a recent extension of PointNet and its hierarchical version of PointNet++ [34], and which effectively incorporates local geometry and residual MLPs.
- **Graph-based**. DGCNN [47] leverages graph convolution operations with graphs constructed in the feature space, allowing it to capture both local and global properties. This architecture has been shown to generalize well between real and synthetic data [45] and has also demonstrated strong transfer learning performance across various pre-training strategies [1, 38, 46].
- **Sparse convolutions**. MinkowskiNet [8] is built upon generalized sparse convolutions, which operate on voxels and benefit from the sparsity of point clouds. This architecture is the backbone of PointContrast [51] for 3D transfer learning. Below we compare it to other baselines and examine how its reliance on density impacts its transfer learning capabilities.
- **Transformer**. Following the success of transformers in NLP and 2D vision tasks, 3D transformer backbones have emerged for point cloud processing. For the first time, we study the transfer learning capabilities of Point Cloud Transformer (PCT) [15] in downstream tasks.

Our primary objectives in analyzing these architectures are three-fold: firstly, to carry out a first comprehensive analysis of supervised pre-training performance across diverse architectures. Secondly, we explore the relation between

architectural properties and their performance in 3D transfer learning, considering both linear probing and fine-tuning scenarios. Thirdly, we aim to identify shared properties that either hinder or contribute to successful transfer learning.

We adapt the layer configurations of these architectures to ensure consistency across all explored pre-training strategies (Sec. 4). Further details on the architecture configurations are provided in the supplementary materials.

4 Pre-training Approaches

As mentioned above, in our analysis, we focus on comparing supervised and contrastive learning approaches. Although our primary focus is on these pre-training strategies, we include additional results to alternative methods (including shape-level contrastive learning and a reconstruction-based approach PointDAE [55]) in the supplementary materials.

Supervised pre-training is a standard approach used in 2D computer vision tasks to obtain generalizable feature embeddings. Interestingly, its utility in the context of 3D transfer learning has not been extensively investigated in previous studies with a few exceptions such as PointContrast [51] as mentioned above. In training, we use the standard cross-entropy loss as a pretext objective:

$$L_{ce} = -\sum_{i}(y_i \cdot \log(p_i)), \tag{1}$$

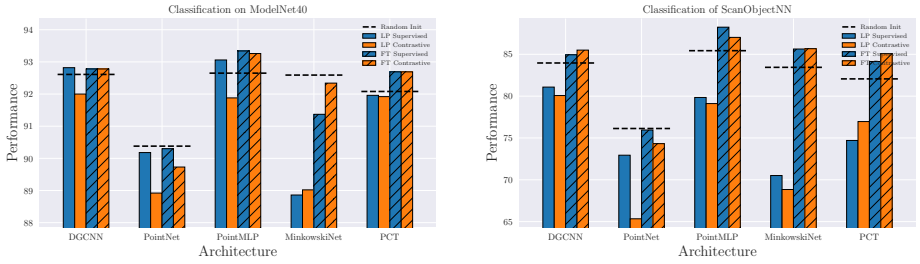
where y_i is the ground truth label, and p_i is the predicted probability of each class i for an input shape. We also use data augmentation (translation, scale and rotation) to avoid the orientation bias introduced by the oriented shapes of ShapeNet, which are also present in the ModelNet40 downstream data.

Contrastive pre-training is employed for comparisons with supervised pre-training. We use *point-level* contrastive learning, based on PointContrast [51]. Specifically, for contrastive pre-training, we first introduce a set of invariances by applying data augmentation. We start by normalizing a point cloud and perform random geometric transformations, including translation with magnitude of 0.5, scaling between 80% and 125%, and rotation of magnitude 45°. We also simulate partial data by cropping 50% the point cloud.

More details about transformations can be found in the supplementary materials. By combining these transformations, we start from an input point cloud and generate two augmented views P_i and P_j.

For point-level contrastive learning, we contrast between *points* rather than entire objects. This models point-level information and also naturally leads to more data for constructing positive and negative pairs. Positive pairs are obtained by matching points from different views. We use the PointInfoNCE Loss, introduced in [51] as the pretext objective:

$$\mathcal{L}_{Contrastive} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp\left(\mathbf{h}_i \cdot \mathbf{h}_j/\tau\right)}{\sum_{(\cdot,k)\in\mathcal{P}} \exp\left(\mathbf{h}_i \cdot \mathbf{h}_k/\tau\right)}, \tag{2}$$



(a) ModelNet40 **(b)** ScanObjectNN

Fig. 2: Evaluation on (a) ModelNet40 and (b) ScanObjectNN classification tasks of different pre-trained models, using linear probing (LP – solid bars) and fine-tuning (FT – dashed bars) settings. Random Init is a randomly initialized model. Transfer learning performance depends on pre-training scheme, architecture and evaluation protocol.

where $\mathbf{h}_i$ are point features obtained by attaching a decoder to the encoder f_θ, τ is a temperature parameter, and $\mathcal{P}$ is the set of matched points.

More pre-training details are provided in the supplementary materials, where we also review the same contrastive scheme applied at shape-level, rather than the point-level.

5 Transferring to Downstream Tasks

5.1 Approach

In order to evaluate the effectiveness of supervised pre-training relative to self-supervised methods, such as contrastive learning, we employ both *linear probing* and *fine-tuning* strategies of pre-trained models across various architectures, datasets, and tasks. Linear probing assesses the quality of the representations learned during pre-training by training a linear classifier on top of the frozen backbone weights, while fine-tuning involves adjusting all weights of pre-trained models to new tasks or datasets. Our first goal is to establish a comprehensive comparison for supervised vs. contrastive pre-training, which has been largely omitted in existing literature [1, 38, 46], and to identify the scenarios in which supervised pre-training can be beneficial.

In the full fine-tuning setting, we adhere to standard learning practices for each architecture. The precise details of the fine-tuning parameters, including learning rates, batch sizes, and optimization algorithms, are documented in the supplementary materials.

5.2 Results

Our first set of results are summarized in Figures 2a, 2b and Table 1. Below, we discuss several key observations.

Table 1: Evaluation on S3DIS semantic segmentation fine-tuning of pre-trained models. IoU metric used. *Random init* is a randomly initialized model.

Architecture	Random Init	Supervised	Contrastive
DGCNN	49.82	49.07	**49.99**
PointNet	**46.21**	37.46	43.48
PointMLP	56.59	55.7	**58.00**
MinkowskiNet	**66.89**	64.07	60.97
PCT	50.8	50.7	**52.34**

Transfer learning via linear probing. We first note that supervised pre-training consistently demonstrates superior performance across the majority of architectures in the linear probing setting (solid bars in Figures 2a, 2b). This suggests that supervised pre-training *can* produce general-purpose, discriminative features that are useful even under fairly significant domain shift (e.g., synthetic shapes in ShapeNet vs. real scans in ScanObjectNN). Interestingly, linear probing with supervised pre-training can even surpass results obtained with full fine-tuning when the domain shift is small, as seen in ModelNet40 (Figure 2a). We find the utility of supervised pre-training in this scenario noteworthy as it can refine the common belief that contrastive learning leads to more "universally useful" features.

Unique architectural properties. While supervised pre-training generally leads to useful final layer features, the results significantly depend on the choice of architectures. Specifically, we observe that the hierarchical nature of DGCNN and PointMLP results in more general features compared to global architectures such as PointNet. In addition, PCT's transformer architecture is adept at capturing global dependencies which overfits less with unsupervised (contrastive) pre-training.

Transfer learning with fine-tuning. We observe a notable shift of behavior when performing full fine-tuning, compared to linear probing, as shown in Figures 2a, 2b (dashed bars). We remark that for several architectures, such as DGCNN, MinkowskiNet and PCT, contrastive pre-training leads to better results compared to supervised learning under full fine-tuning (with a $\sim$0.1% accuracy advantage in the on-par scenarios). This suggests that the utility of pre-trained models highly depends on *how* those models are used in the downstream tasks. Specifically, we note that the utility of final layer features (such as with linear probing) is not always a good predictor for the performance under fine-tuning, as shown, e.g., in the case of DGCNN. In Section 6 we analyze this behavior in depth and observe the critical role of *early layers* and their ability to adapt to downstream data. Our analysis also highlights that the types of solutions obtained with supervised and contrastive pre-training are different, making a distinction between generalizable (features ready to be used in multi-task settings) and adaptable features.

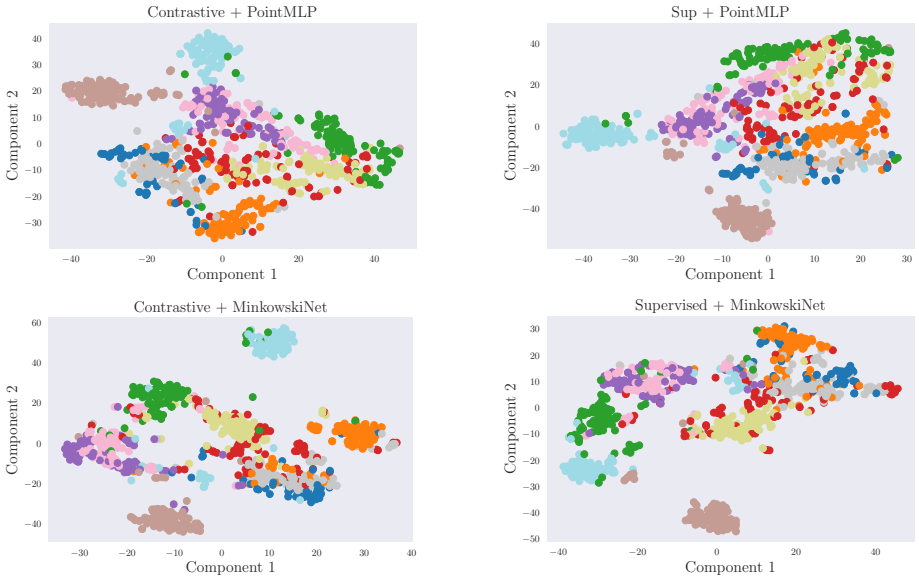


Fig. 3: t-SNE plots of the first-layer feature activation for different architectures and pre-training schemes. We use the *ModelNet10* evaluation set, which is a subset of ModelNet40 containing 10 classes, each represented by a different color. *Clusters* are formed even in the feature space of first layers, which implies their discriminative capability. Visualization on additional architectures can be found in the supplementary.

In addition, we note that for most architectures, including MinkowskiNet, supervised pre-training is *not* an "upper bound" to information gain from pre-training, differently from the claim made in [51], and that other pre-training strategies can lead to better results.

Unique architectural properties. PointNet, with its global feature aggregation strategy, tends to overfit to the source training dataset, and is thus not able to extract information that is useful for fine-tuning. MinkowskiNet's performance is influenced by variations in point cloud density and scale. Its ability to process large-scale scenes explains its high semantic segmentation performance (Table 1), but highlights its dependence on the sampling density for effective transfer learning.

6 Feature Analysis

The inconsistency in the effectiveness of pre-training approaches across transfer learning scenarios leads us to explore the intrinsic properties that contribute to the transferability of model features. Our focus is particularly on the behavior of early layers, given their established importance in 2D transfer learning [53].

Discriminative capability of early layers. To illustrate the behavior of the early layers, we plot a t-SNE visualization of the first-layer features of pre-trained

Table 2: Linear SVM classification of features extracted from the first and last layers for different architectures and pre-training strategies. Downstream task is classification on ModelNet40. The 1st layer's evaluation displays remarkably high *accuracy*, not that far from the last layer's evaluation accuracy.

Architecture	Supervised		Contrastive	
	1st Layer	Last Layer	1st Layer	Last Layer
DGCNN	81.80	90.92	81.15	89.87
PointNet	79.37	88.24	79.29	86.46
PointMLP	82.69	91.32	74.59	89.7
MinkowskiNet	77.51	88.4	70.74	87.72
PCT	78.9	90.55	77.9	89.87

models in Figure 3 on the ModelNet10 dataset. We observe the appearance of discernible clusters even in these early layers and without any fine-tuning. This clustering in the feature space suggests that *early layers* possess a class-discriminative capacity. This points to the potential of early layers to contribute significantly to the downstream task, unlike in the 2D domain where early layers are typically "universal" and do not need to be adapted [31].

To analyze this behavior quantitatively, we used a linear SVM to evaluate the utility of early layer features against those from output layers. The results, presented in Table 2 highlight the remarkable classification performance of the early layers even with a simple linear probing on a new downstream dataset. This behavior is observed for all studied architectures. As a point of comparison, we evaluated the *first layer* of the EfficientNetB0 [44] backbone pre-trained with ImageNet [10] on the Oxford-IIIT Pet Dataset [32], resulting in 10% classification accuracy (against 92.28% obtained using the output layers), which showcases the low discriminative capability of early layers in the 2D domain.

As observed in Section 5, the utility of final layer features is not always a good predictor for the performance under fine-tuning for 3D transfer learning. Coupled with the discriminative power of early layers revealed above, this leads us to investigate how the network weights change and what factors contribute to successful fine-tuning.

Gradient norm of layer weights. To understand how the models behave in the fine-tuning setting, we analyze the norm of the gradients of pre-trained networks with respect to downstream data. Figure 4 shows the layer-wise gradient norm, computed using a cross-entropy loss on the downstream training data using frozen pre-trained weights. We observed that for contrastive pre-training, early layers exhibit a higher gradient norm compared to the last layer. This finding further substantiates the need, not only for *last layers*, but also for *early layers* to adapt to new data/tasks when dealing with 3D data for successful transfer learning.

As shown in Figure 4 supervised pre-trained models demonstrate a low gradient norm, with the disparity between supervised and contrastive pre-training

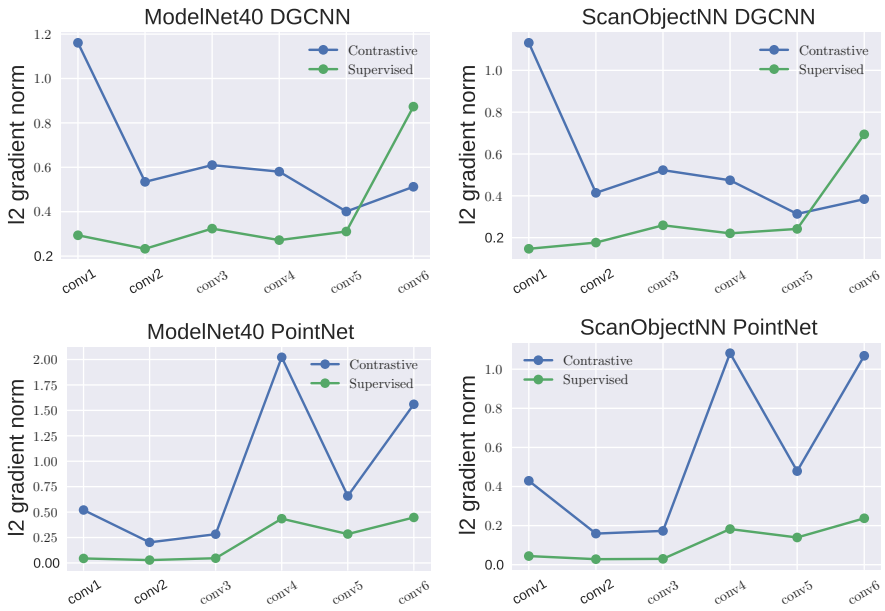


Fig. 4: Layer-wise gradient norms of pre-trained models on downstream ModelNet40 and ScanObjectNN datasets using DGCNN and PointNet. Supervised pre-training shows low gradient norms, especially in early layers. Further analysis across architectures is provided in the supplementary.

being more prominent especially in the early layers. The low gradient norm in supervised pre-training suggests that the learned features are confined to a local region of the optimization landscape. This implies that these features are *less prone to adaptation when exposed to new data or tasks* during fine-tuning, which aligns with our previous observations regarding the limited *adaptability* of supervised features under significant data shifts (Section 5.2). In contrast, features developed through contrastive pre-training, particularly in the early layers, show greater potential for adapting to new inputs, thereby enhancing their success in fine-tuning.

We note that this adaptability is not the same as feature generality, since the latter typically applies only to the final layers, which, as shown in Section 5 can be informative (in linear probing) despite not being adaptable (in fine-tuning), e.g., for supervised pre-training.

7 Layer-wise Geometric Regularization

7.1 Normal Prediction

The qualitative insights of our analysis have revealed various challenges in pre-training that can hinder effective transfer learning. Specifically, we have observed

Table 3: Evaluation of geometric regularization of supervised and contrastive pre-training. Evaluations on Shape classification on ModelNet40 and ScanObjectNN, and 3D scene segmentation on S3DIS with DGCNN architecture. Regularization improves downstream performance of supervised pre-training in the fine-tuning setting.

Pre-training	ModelNet40	ScanObjectNN	S3DIS
Supervised	92.78	84.94	49.07
Supervised + regularization	**93.34**	**85.91**	**49.88**
Contrastive	92.78	85.05	49.99
Contrastive + regularization	**93.18**	**85.39**	**50.44**

a correlation between low gradient norms in early layer weights and diminished fine-tuning performance, indicating a decreased adaptability of the learned features when moving from source to downstream tasks.

Given these insights, our objective is to improve the standard pre-training method while adhering to the pre-text task. We achieve this by implementing a straightforward regularization strategy designed to enhance the geometric and, consequently, more universally applicable characteristics of the learned features.

As highlighted in our analysis, effective regularization to pre-training should promote low-level features, particularly in early layers, that can adapt better to unseen data. Besides, the regularization needs to leverage properties intrinsic to input point clouds. We find that predicting geometric properties as an auxiliary task aligns with these criteria, as it promotes *local,* category-agnostic attributes of input point clouds. A practical method to achieve this is to predict point-wise normal vectors, a task that can be achieved without extra labels by using estimation methods such as PCA, or by extracting them from raw meshes when available.

We apply this regularization only to a set of early layers, whereas the output features of the entire architecture continue to serve the primary pre-text task. This strategy infuses purely geometric information into early layers with the goal of making them universally applicable and reducing the dependency to the pre-text task. Let $H_{\text{normal}}(\mathbf{f}_l)$ denote the normal prediction head, $\mathbf{f}_l$ the point-wise features of layer l, and $\mathbf{n}$ the ground truth normals. The combined training objective comprises the pre-training loss $L_{\text{pre-train}}$ and the regularization loss L_{regul}. We use absolute cosine similarity in L_{regul} since estimated normals are not oriented. Our total pre-training loss can be formulated as:

$$L_{\text{total}} = L_{\text{pretrain}} + L_{\text{regul}}\left(H_{\text{normal}}(\mathbf{f}_l), \mathbf{n}\right). \tag{3}$$

7.2 Evaluation

The experimental setup for pre-training and fine-tuning remains consistent with our previous evaluation. We use normals computed through PCA on local neighborhoods comprising 30 points within point clouds of size 2048. We explore

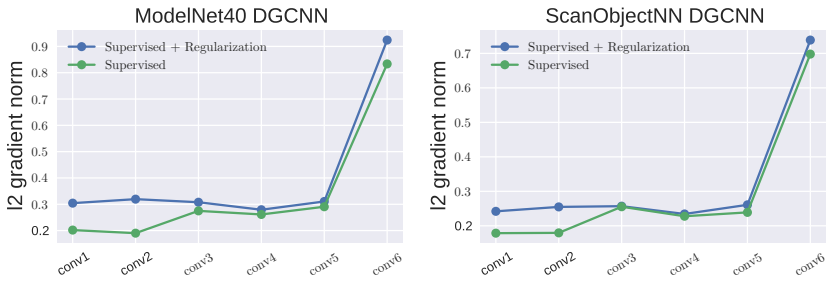


Fig. 5: Layer-wise gradient norms of supervised pre-trained models with and without regularization on downstream ModelNet40 and ScanObjectNN datasets using DGCNN. Regularization increases the gradient norm values, especially in early layers.

the optimal number of layers required for the most effective regularization, and find that regularizing the features of the first two layers yields the best results. Through empirical testing to determine the optimal layers for regularization, we discovered that targeting early layers for regularization yields better outcomes than applying it to the encoder's output layer alone. Ablation of layer choice and other architectures is included in the supplementary.

When comparing between baseline and regularized supervised pre-training, Table 3 shows notable improvement (1% gain) in all evaluated downstream data and tasks. This improvement is also correlated to an increase in gradient norm values as shown in Figure 5. These results highlight the potential of layer-wise regularization and encourages its use for future supervised pre-training.

Despite the primary goal of addressing the limitations of early layers in supervised pre-training through this regularization technique, we also explore its application in other pre-text tasks, such as contrastive learning. According to Table 3, incorporating our regularization into contrastively pre-trained models can benefit fine-tuning especially under strong domain shift. This suggests that layer-wise regularization can be a useful general approach for diverse transfer learning strategies.

8 Summary and Key Takeaway Remarks

Our findings emphasize several critical aspects of pre-training strategies and their impact on transfer learning performance across different architectures.

The choice of evaluation protocol is critical. Supervised pre-training excels in linear probing often surpassing contrastive pre-training. This highlights its potential for universal task ready features. However, this approach falls behind contrastive learning in fine-tuning, emphasizing the importance of feature *adaptability*.

The architecture also influences pre-training outcomes. We find that simpler models that capture global features such as PointNet fail at transferring learned knowledge when fine-tuned. More local point-based methods such as PointMLP

benefit greatly from pre-training, especially with supervised pre-training. Transformers benefit more from contrastive learning. Complex architectures that process large scale data such as MinkowskiNet, which is based on sparse convolutions can boost the transfer learning performance under small domain shift (*e.g.* scenes to scenes, shapes to shapes) but can be sensitive to changes in data sampling. Graph-based such as DGCNN with local and global feature extraction, can be pre-trained and fine-tuned efficiently, although they do not always exhibit the best overall performance.

Second, we observe that unlike the 2D case, features learned *even in the earliest layers* of 3D architectures tend to be class-specific, which can hinder their utility in transfer learning. This highlights the importance of adaptability of early layers for point cloud transfer learning.

Finally, regularizing exclusively early layers using geometric signals can improve pre-training in several settings, particularly for supervised pre-training. This approach addresses the lack of general-purpose low-level 3D features. This insight also opens the door for future research to focus on early layer-wise regularization rather than applying it uniformly across the entire model.

9 Conclusion, Limitations and Future Work

In this paper, we have conducted the first in-depth qualitative and quantitative investigation of the key factors when performing point cloud transfer learning with supervised vs contrastive pre-training and proposed a regularization approach informed by this analysis. We evaluated several architectures under pre-training approaches and found that their performance changes with the transfer learning protocol. We also examined early layer importance through their discriminative capability and fine-tuning adaptability, shedding light on their utility and importance for point cloud transfer learning. Finally, we correlated this analysis with a new regularization approach that targets early layers to improve on downstream performance. Overall, our work establishes a consistent evaluation framework, presents detailed analysis tools, and proposes an appropriate simple method for successful transfer learning on point cloud data, which can both inform and allow to compare future designs.

As our main focus was on the careful analysis of the factors that contribute to successful transfer learning for point clouds, we did not investigate *all* possible combinations of the source datasets, architectures, pre-training strategies, fine-tuning approaches, and downstream tasks. The resulting combinatorial complexity would incur very significant computational costs, and furthermore might obscure the possible analytical insights. Nevertheless, in the future, an analysis of other tasks and pre-training strategies can fit within and complement the framework that we established. Finally, as suggested by our analysis, there is currently a need for new architectures and pre-training approaches, that are robust under changes of sampling density and can lead to multi-scale features, which would generalize and adapt to new downstream data. Exploring the possible solution space is an exciting direction for future work.

Acknowledgements Parts of this work were supported by the ERC Starting Grant 758800 (EXPROTEA), ERC Consolidator Grant 101087347 (VEGA), ANR AI Chair AIGRETTE, as well as gifts from Ansys and Adobe Research.

References

1. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R.: Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9902–9912 (2022)
2. Alliegro, A., Boscaini, D., Tommasi, T.: Joint supervised and self-supervised learning for 3d real world challenges. In: International Conference on Pattern Recognition (ICPR). pp. 6718–6725 (2021)
3. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1534–1543 (2016)
4. Bogo, F., Romero, J., Pons-Moll, G., Black, M.J.: Dynamic faust: Registering human bodies in motion. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6233–6242 (2017)
5. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An information-rich 3d model repository. Tech. rep. (2015)
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
7. Chen, Y., Liu, J., Ni, B., Wang, H., Yang, J., Liu, N., Li, T., Tian, Q.: Shape self-correction for unsupervised point cloud understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8382–8391 (2021)
8. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Niessner, M.: ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition. pp. 248–255 (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Du, B., Gao, X., Hu, W., Li, X.: Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3133–3142 (2021)
13. Eckart, B., Yuan, W., Liu, C., Kautz, J.: Self-supervised learning on 3d point clouds by learning discrete generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8257 (2021)

14. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
15. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: Pct: Point cloud transformer. Computational Visual Media **7**, 187–199 (2021)
16. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE transactions on pattern analysis and machine intelligence **43**(12), 4338–4364 (2020)
17. Hassani, K., Haley, M.: Unsupervised multi-task feature learning on point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8160–8171 (2019)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
19. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
21. Hou, J., Graham, B., Nießner, M., Xie, S.: Exploring data-efficient 3d scene understanding with contrastive scene contexts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15587–15597 (2021)
22. Huang, S., Xie, Y., Zhu, S.C., Zhu, Y.: Spatio-temporal self-supervised representation learning for 3d point clouds. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6535–6545 (2021)
23. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D., Makedon, F.: A survey on contrastive self-supervised learning. Technologies **9**(1), 2 (2020)
24. Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D.: Abc: A big cad model dataset for geometric deep learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9601–9611 (2019)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. vol. 25 (2012)
26. 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)
27. Liu, Y., Yi, L., Zhang, S., Fan, Q., Funkhouser, T., Dong, H.: P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. arXiv preprint arXiv:2012.13089 (2020)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
29. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. arXiv preprint arXiv:2202.07123 (2022)
30. Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K.: What makes transfer learning work for medical images: Feature reuse & other factors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9225–9234 (June 2022)

31. Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K.: What makes transfer learning work for medical images: feature reuse & other factors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9225–9234 (2022)
32. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: The oxford-iiit pet dataset `https://www.robots.ox.ac.uk/~vgg/data/pets/`
33. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
34. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
35. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)
36. Rao, Y., Liu, B., Wei, Y., Lu, J., Hsieh, C.J., Zhou, J.: Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3283–3292 (2021)
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
38. Sauder, J., Sievers, B.: Self-supervised deep learning on point clouds by reconstructing space. Advances in Neural Information Processing Systems **32** (2019)
39. Sharma, C., Kaul, M.: Self-supervised few-shot learning on point clouds. Advances in Neural Information Processing Systems **33**, 7212–7221 (2020)
40. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
41. Sun, J., Cao, Y., Choy, C.B., Yu, Z., Anandkumar, A., Mao, Z.M., Xiao, C.: Adversarially robust 3d point cloud recognition using self-supervisions. Advances in Neural Information Processing Systems **34** (2021)
42. Taghanaki, S.A., Luo, J., Zhang, R., Wang, Y., Jayaraman, P.K., Jatavallabhula, K.M.: RobustPointSet: A dataset for benchmarking robustness of point cloud classifiers. arXiv preprint arXiv:2011.11572 (2020)
43. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging **35**(5), 1299–1312 (2016)
44. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
45. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1588–1597 (2019)
46. Wang, H., Liu, Q., Yue, X., Lasenby, J., Kusner, M.J.: Unsupervised point cloud pre-training via occlusion completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9782–9792 (2021)

47. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (TOG) **38**(5), 1–12 (2019)
48. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015)
49. Xiao, A., Huang, J., Guan, D., Lu, S.: Unsupervised representation learning for point clouds: A survey. arXiv (2022)
50. Xiao, A., Huang, J., Guan, D., Zhang, X., Lu, S., Shao, L.: Unsupervised point cloud representation learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
51. Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O.: Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In: European conference on computer vision. pp. 574–591 (2020)
52. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016)
53. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
54. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: Point-bert: Pre-training 3d point cloud transformers with masked point modeling. arXiv preprint arXiv:2111.14819 (2021)
55. Zhang, Y., Lin, J., Li, R., Jia, K., Zhang, L.: Point-dae: Denoising autoencoders for self-supervised point cloud learning. arXiv preprint arXiv:2211.06841 (2022)
56. Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10252–10263 (2021)
57. Zhao, N., Wu, Z., Lau, R.W.H., Lin, S.: What makes instance discrimination good for transfer learning? In: International Conference on Learning Representations (2021)