Global-Local Bidirectional Reasoning for Unsupervised Representation Learning of 3D Point Clouds

Yongming Rao[1,2,3], Jiwen Lu[1,2,3]*, Jie Zhou[1,2,3,4]
[1]Department of Automation, Tsinghua University, China
[2]State Key Lab of Intelligent Technologies and Systems, China
[3]Beijing National Research Center for Information Science and Technology, China
[4]Tsinghua Shenzhen International Graduate School, Tsinghua University, China
raoyongming95@gmail.com; {lujiwen, jzhou}@tsinghua.edu.cn

Abstract

Local and global patterns of an object are closely related. Although each part of an object is incomplete, the underlying attributes about the object are shared among all parts, which makes reasoning the whole object from a single part possible. We hypothesize that a powerful representation of a 3D object should model the attributes that are shared between parts and the whole object, and distinguishable from other objects. Based on this hypothesis, we propose to learn point cloud representation by bidirectional reasoning between the local structures at different abstraction hierarchies and the global shape without human supervision. Experimental results on various benchmark datasets demonstrate the unsupervisedly learned representation is even better than supervised representation in discriminative power, generalization ability, and robustness. We show that unsupervisedly trained point cloud models can outperform their supervised counterparts on downstream classification tasks. Most notably, by simply increasing the channel width of an SSG PointNet++[1], our unsupervised model surpasses the state-of-the-art supervised methods on both synthetic and real-world 3D object classification datasets. We expect our observations to offer a new perspective on learning better representation from data structures instead of human annotations for point cloud understanding.[2]

1. Introduction

Facilitating machines to understand the 3D world is crucial to many important real-world applications, such as autonomous driving, augmented reality and robotics. One core problem on 3D geometric data such as point clouds is learning powerful representations that are discriminative, generic and robust. To tackle this problem, current state-of-the-arts on point cloud analysis [2, 26, 28, 33, 38, 43, 49, 51, 54] are established with the help of extensive human-annotated supervised information. However, manually labeled data require the high cost of human labor and may limit the generalization ability of the learned models. Therefore, unsupervised learning is an attractive direction to obtain generic and robust representations for 3D object understanding.

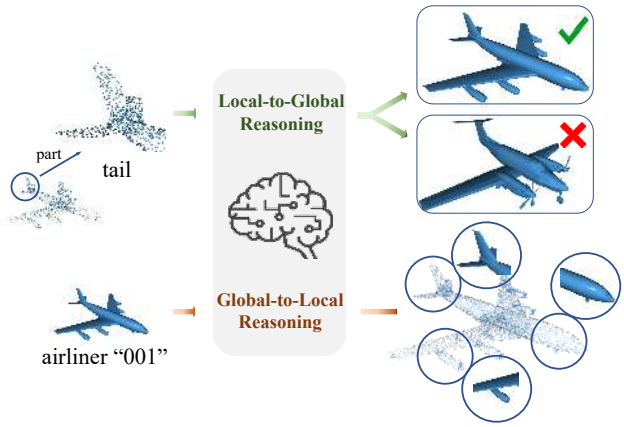

Figure 1: **Illustration of our main idea.** We propose to learn representation *unsupervisedly* from data structures by training the networks to solve two problems: reasoning the whole object from a single part and reasoning detailed structures from the global representation.

Learning useful representations from unlabeled data is a fundamental and challenging problem for point cloud analysis. While several efforts have been devoted to learn representation of a point cloud without human supervision [1, 8, 14, 18, 26, 31, 47, 55, 56] , these methods are mainly based on self-supervision signals provided by generation or reconstruction tasks, including self-reconstruction [1, 8, 14, 26, 47, 55, 56], local-to-global reconstruction [18, 31] and distribution estimation [1, 26]. These methods have proven to be effective

*Corresponding author
[1]Single-Scale Grouping PoinetNet++ [38].
[2]Code: https://github.com/raoyongming/PointGLR

in capturing structural and low-level information of point clouds, but usually fail to learn high-level semantic information from point clouds. Therefore, unsupervised models still perform far behind the state-of-the-art supervised model. The goal of this work is to explore an unsupervised learning algorithm that can learn both structural information and semantic knowledge to promote the quality of unsupervisedly learned representation.

Different from images where local patches are noisy and usually independent from the whole image (for example, given a patch of a dog, we cannot identify whether this image is about animals or the people nearby), the underlying semantic and structural information is shared in all parts of a 3D object. This distinct property of 3D objects makes reasoning the whole object from a single part possible. Based on this observation, we hypothesize that a powerful representation of a 3D object should model the underlying attributes that are shared between parts and the whole object and distinguishable from other objects. As shown in Figure 1, given a point cloud of a tail of an airplane, a good representation of the tail should reflect the type of the corresponding airplane. Simultaneously, the representation of the whole airplane should contain all the necessary details to infer the local structures of this airplane.

In this paper, we propose a new scheme for unsupervised point cloud representation learning by bidirectional reasoning between local representations at different abstraction hierarchies in a network and global representation of a 3D object. Our method is simple yet effective, which can be applied to a wide range of deep learning methods for point cloud understanding. While most existing unsupervised learning methods focus on exploiting structure information by learning various autoencoders, our method aims to capture the underlying semantic knowledge shared between local structures and global shape in 3D point clouds. Specifically, the proposed Global-Local Reasoning (GLR) consists of two sub-tasks: 1) local-to-global reasoning: we formulate the problem of capturing shared attributes between local parts and global shape as a self-supervised metric learning problem, where local features are encouraged to be closer to the global feature of the same object than features of other objects, such that the distinct semantic information of each object can be extracted by local representations; 2) global-to-local reasoning: we further use the self-supervised tasks including self-reconstruction and normal estimation to learn global features that contain necessary structural information of 3D objects.

Our experimental results on several benchmark datasets demonstrate that the unsupervisedly learned point cloud representation is even more discriminative, generalizable and robust than supervised representation in downstream object classification tasks. Our unsupervisedly trained models can consistently outperform their supervised counterparts. With our unsupervised learning method, we show a simple and light-weight SSG PointNet++ [38] model can achieve very competitive results with supervised methods (92.2% classification accuracy on ModelNet40 [52]). By simply increasing the channel width, we further obtain 93.0% and 87.2% single view accuracy on ModelNet40 and ScanObjectNN [46] benchmarks respectively, surpassing the state-of-the-art unsupervised and supervised methods, while the supervised version of this model suffers from overfitting.

2. Related Work

Deep Learning on 3D Point Clouds: Recent years have witnessed rapid development on 3D point cloud analysis thanks to the deep learning techniques that are designed to consume 3D point clouds directly [28, 33, 37, 38, 49]. PointNet [26] pioneers this line of works and designed a deep network that can handle unordered and unstructured 3D points by independently learning on each point and fusing point features with max pooling. Though efficient, PointNet fails to capture local structures, which have proven to be crucial to the success of CNNs. PointNet++ [38] is proposed to mitigate this issue by developing a hierarchical grouping architecture to extract local features progressively at different abstraction levels. The subsequent works such as PointCNN [28], PointConv [51] and Relation-Shape CNN [33] also focus on local structures of point cloud and further improve the quality of captured features. Since only the relation between local and global features is needed, our method is suited for all these PointNet++ variants. While recent works push state-of-the-art of point cloud deep learning by promoting the capacity of networks, this work offers a new route to learn powerful representation in an *unsupervised* fashion, without any human annotations.

Unsupervised Representation Learning: Unsupervised learning has been an important group of methods in computer vision since the earliest day [13], which aims to learn transformations of the data that make the subsequent downstream problem solving easier [5]. Classical deep methods for unsupervised learning such as autoencoders [21], generative adversarial networks [16] and autoregressive models [35] learn representation by faithfully reconstructing the input data, which focus on low-level variations in data and is not very useful for downstream tasks like classification. Recent works on self-supervised learning present a powerful family of models that can learn discriminative representations with rich semantic knowledge. This group of methods design various problem generators such that models need to learn useful information from data in order to solve these generated problems [3, 10, 11, 19, 44]. In this work, we also follow this line and propose to learn point cloud representation by solving the global-local bidirectional reasoning problem.

There are several prior attempts on learning representa-

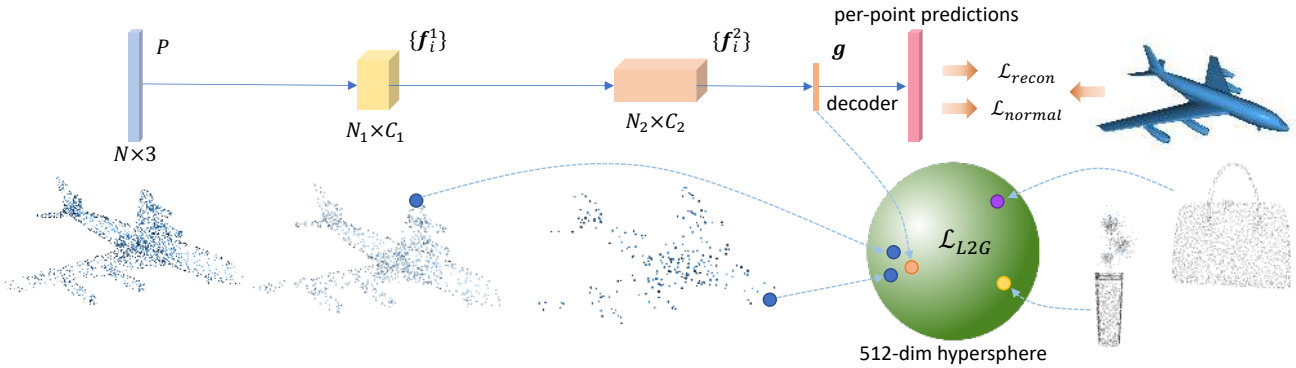

Figure 2: **The overall framework of our unsupervised feature learning approach.** The representation is learned by connecting local structures and global shape. We map the local representations at different levels and global representations to shared feature space and use a self-supervised metric learning objective to mine semantic knowledge from data. By further incorporating self-reconstruction and normal estimation tasks, a powerful representation that contains rich semantic and structural information can be learned.

tion of a point cloud without human supervision [1, 8, 14, 18, 26, 31, 47, 55, 56]. These methods discover useful information in the 3D point cloud by performing data reconstruction, which has proven to be effective in learning structural information. However, because of lacking effective semantic supervision, previous methods limit the networks' ability in downstream tasks. Our method resolves this issue by incorporating semantic supervision with structural supervision. With the exploration of high-level semantic knowledge, our method is able to learn discriminative representation like supervised method while maintaining the robustness and generalization of unsupervised representation.

3. Approach

The core of 3D point cloud understanding is to learning discriminative, generic and robust representations that can capture the underlying shape. To achieve this goal in an unsupervised manner, we propose to point cloud representation by solving a bidirectional reasoning problem between the local structures and the global shape. The overall framework of our method is presented in Figure 2.

3.1. Hierarchical Point Cloud Feature Learning

We begin by reviewing the hierarchical point cloud feature learning framework firstly proposed in PointNet++ [38], on which our method is built.

Consider a set of 3D points $P \subset \mathbb{R}^3$ with N elements, in which each point p_i is represented by a 3D coordinate. To learn features based on these 3D coordinates, PointNet [37] proposes to use a symmetric function f that is invariant to point permutations to transfer point set into feature space:

$$f(P) = \mathcal{A}(h(p_1), h(p_2), ..., h(p_N)), \quad (1)$$

where h is a multi-layer perceptron network that processes each point independently and shares parameters for all points and $\mathcal{A}$ is a symmetric aggregation function like max pooling to summarize features from each point. Since each point is processed independently by h, the structural information among points is captured only by the aggregation function $\mathcal{A}$. Therefore, PointNet lacks the ability to capture local context. To address this issue, PointNet++ and its variants [28, 33, 51] use a hierarchical structure to learn point cloud feature progressively at different abstraction levels. Specifically, at the ℓ-th level, point set is abstracted by using iterative furthest point sampling [38] to produce a new set $P^\ell \subset P^{\ell-1}$ with fewer points and we can extract the local geometrical feature $\mathbf{f}_i^\ell$ by applying a small PointNet on the local point subset around the centroid for each point $p_i^\ell \in P^\ell$. The global representation of the point cloud $\mathbf{g}$ is then obtained by applying another small PointNet model on the points and features at the highest abstraction level.

Almost all previous works [2, 26, 28, 33, 38, 43, 49, 51, 54] on supervised point cloud learning employ an end-to-end training paradigm, where the representation is learned directly from the annotated labels. Although achieved promising performance, these methods neglect the intrinsic semantic and structure information contained in the point clouds themselves. In this work, we focus on exploring this property of point cloud and provide a very competitive alternative for point cloud representation learning.

To discover the structure and semantic information from data without human annotations, we propose two problems for the networks to solve: *local-to-global reasoning* and *global-to-local reasoning*, which aim to *unsupervisedly* learn semantic and structural knowledge respectively.

3.2. Local-to-Global Reasoning

Humans are able to recognize many objects even when only a small part of the object is presented. This fact inspires us to exploit the relation between local parts and global shape

as a free and plentiful supervisory signal for training a rich representation for point cloud understanding. Therefore, the goal of local-to-global reasoning is to mine the shared semantic knowledge among different abstraction hierarchies of point clouds. Since global representation usually can better capture the semantic information of 3D objects than local representations, local-to-goal reasoning operates by predicting global representation from local ones. To evaluate the predictions, we formulate the prediction as a self-supervised metric learning problem and use a multi-class N-pair loss [40] to supervise the prediction task. Inspired the idea of instance discrimination [53], to learn the distinct semantic information for each object, we treat the global representation of the current object as the *positive* sample and use the global representation of other objects as the *negative* samples. In the following, we describe the details of the local-to-global reasoning.

Prediction Networks: Since the local features $\{\mathbf{f}_i^\ell, \forall i, \ell\}$ and global feature $\mathbf{g}$ have different numbers of channels, we cannot directly measure the similarity of them. Thus, we first use prediction networks $\{\phi^\ell, \forall \ell\}$ and φ to embed them into a shared feature space, respectively. The prediction networks can be implemented as multi-layer perceptron (MLP) networks and the prediction networks are shared at each abstraction level.

Self-Supervised Metric Learning: A straightforward method to optimize the predictions is to minimize the absolute overall differences between $\phi^\ell(\mathbf{f}_i^\ell)$ and $\varphi(\mathbf{g})$, i.e. minimize $\sum_{i,\ell} ||\phi^\ell(\mathbf{f}_i^\ell) - \varphi(\mathbf{g})||$. However, this objective may lead to degenerate representations that map all inputs to a constant value. Therefore, we choose to supervise the *relative* quality of the predictions with an unsupervised metric learning task. Specifically, for each embedded local representation $\mathbf{f}_i^\ell$, we enforce its embedding to be closer to the embedded global representation of the same object than any other object. The local-to-global reasoning objective can be written as:

$$\mathcal{L}_{\text{G2L}}^{i,\ell} = \log(1+\sum_{\mathbf{g}_k \neq \mathbf{g}} \exp(s\phi^\ell(\mathbf{f}_i^\ell)^\top \varphi(\mathbf{g}_k) - s\phi^\ell(\mathbf{f}_i^\ell)^\top \varphi(\mathbf{g})) \tag{2}$$

and

$$\begin{aligned} \mathcal{L}_{\text{G2L}} &= \frac{1}{M}\sum_{i,\ell} \mathcal{L}_{\text{G2L}}^{i,\ell} \\ &= -\frac{1}{M}\sum_{i,\ell} \log \frac{\exp(s\phi^\ell(\mathbf{f}_i^\ell)^\top \varphi(\mathbf{g}))}{\sum_k \exp(s\phi^\ell(\mathbf{f}_i^\ell)^\top \varphi(\mathbf{g}_k))}, \end{aligned} \tag{3}$$

where $\{\mathbf{g}_k, k = 1, 2, ..., m\}$ are the global representations of different point sets in the mini-batch with batch size m and M is the number of local features. Inspired by the studies on metric learning for face recognition [9, 30, 48] that perform metric learning on features on a hypersphere, we normalize the outputs of prediction networks before computing similarities and use a constant value $s = 64$ [9] to re-scale the features. Empirically, our experiments show that forcing features to be distributed on a hypersphere with a radius of s will significantly stabilize the training process and improve the discriminative ability of the learned features.

Discussions: The proposed local-to-global reasoning is connected to mutual information maximization methods [3, 19, 22, 44] for unsupervised image representation learning. The multi-class N-pair loss can be viewed as a variant of InfoNCE [36]. Therefore, minimizing the $\mathcal{L}_{\text{G2L}}$ maximizes the lower bound of the mutual information between local representations and global representation. From this perspective, our method captures the underlying semantic knowledge of a 3D object by maximizing the mutual information of features at different hierarchies. Unlike previous works that performs adversarial learning between the mutual information estimator and the feature encoder [22] or maximizes the mutual information of seen patches and unseen patches [19], different views of images [3] or different modalities of images [44] , our work explores the distinct property of point clouds by connecting local and global structures of a 3D object. Furthermore, our local-to-global loss offers a metric learning view of InfoNCE, which is different from previous works that are based on Noise-Contrastive Estimation [34]. Benefiting our modifications inspired by metric learning and face recognition methods, we observe that our loss is more effective and stable than previous methods on point cloud understanding tasks in our experiments.

3.3. Global-to-Local Reasoning

Since discovering knowledge that is helpful for downstream tasks from unlabeled data is usually quite intractable, local-to-global reasoning may not necessarily lead to useful representations. This fact is also pointed out by studies on mutual information maximization methods [44, 45], where evidence shows that larger mutual information may not guarantee a better performance for downstream tasks [45]. Intuitively, since the local-to-global reasoning only supervises the local representation to be close to the global one, the quality of global representation is critical. This is, if the global representation is well initiated, decent supervision to local representation will be offered, thus creating a *virtuous circle* for the learning of local and global features. On the contrary, the learning process may obtain unpredictable results for the bad initial state of global representation. To avoid this issue, we propose an auxiliary global-to-local reasoning task to supervise the networks for learning useful representation corporately. Specifically, we employ two low-level generation tasks, including self-reconstruction and normal estimation as two self-supervision signals, such that global representation needs to capture the basic structural information of point clouds.

Self-Reconstruction: Self-reconstruction, or point autoencoding, is a widely used technique for unsupervised point cloud representation learning [1,8,14,26,47,55,56]. To perform self-reconstruction, we adopt the folding-based [55] decoder D to deform the canonical 2D grid onto 3D coordinates of a point cloud conditioned on the global representation $\mathbf{g}$. The reconstruction error is defined as Chamfer Distance [12]:

$$\mathcal{L}_{\text{recon}} = \sum_{p \in P} \min_{x \in D(\mathbf{g})} ||x-p||_2 + \sum_{x \in D(\mathbf{g})} \min_{p \in P} ||x-p||_2. \quad (4)$$

Normal Estimation: Normal estimation is a more challenging task that requires a higher level understanding of the underlying surface information of a 3D shape. Different from previous works [33] that pursue the estimation precision, we use this task as a supervisory signal to improve global representation. Thus, we simply concatenate the 3D coordinates with the global representation and employ a shared light-weight MLP σ to produce the estimated normals. The cosine loss is used to measure the estimation error:

$$\mathcal{L}_{\text{normal}} = 1 - \frac{1}{N} \sum_{i} \cos(\sigma([p_i, \varphi(\mathbf{g})]), p_i^{\text{normal}}). \quad (5)$$

Combining the local-to-global reasoning and the global-to-local reasoning, we arrive at the global-local bidirectional reasoning objective:

$$\mathcal{L}_{\text{GLR}} = \mathcal{L}_{\text{L2G}} + \mathcal{L}_{\text{G2L}} = \mathcal{L}_{\text{L2G}} + \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{normal}}. \quad (6)$$

3.4. Point Cloud Analysis with GLR

Unsupervised Learning with GLR: Point cloud representation can be *unsupervisedly* learned by enforcing networks to solve the proposed global-local reasoning (GLR) problems, where the representation can be used in various downstream point cloud analysis applications like object classification. The quality of unsupervisedly learned representation is usually evaluated by linear separability of classification task, where a supervised linear SVM [6] model or single-layer neural network is trained on unsupervised representations to measure the test accuracy. For PointNet++ [38] model and its variants, we use the aggregated representation for classification task, which is obtained by summarizing embedded global and local representations:

$$\mathbf{f} = [\mathcal{A}(\{\phi^1(\mathbf{f}_i^1)\}), ..., \mathcal{A}(\{\phi^L(\mathbf{f}_i^L)\}), \varphi(\mathbf{g})], \quad (7)$$

where we use a max pooling operation $\mathcal{A}$ to aggregate local features of each abstraction level from 1 to L and concatenate these features with the global feature.

Hybrid Learning with GLR: Since supervisedly learned global representation can be viewed as a good initialization for the proposed GLR framework, our method is also compatible with supervised learning methods, where GLR serves as an auxiliary loss to further improve the robustness of representations.

Implementation: All of our models is trained on a single GTX 1080ti GPU with deep learning library Pytorch [42]. To show our method can be used for various point cloud networks, we consider two baseline models: PointNet++ [38] and Relation-Shape CNN (RSCNN) [33]. Note that for both baseline models, we use the Single-Scale Grouping (SSG) [38] as the point grouping module, which is more than $3\times$ smaller than Multi-Scale Grouping (MSG) [38] module used in original PointNet++ model. Besides, we divide the MLP used in each set abstraction layer into two fully connected layers and use them before and after aggregation operation, respectively. Our experiments show this modification can reduce computation and improve performance while keeping the number of parameters unchanged. For unsupervised learning setting, we train a linear SVM [6] on unsupervised representations of the training data and report the classification accuracy on the test set. For supervised learning and hybrid learning settings, we use the aforementioned aggregated representation for fair comparison and employ a two-layer classifier where dropout technique [41] with a ratio of 50% is used for each layer. Our models are trained using Adam [24] optimizer with a base learning rate of 0.001, and we decay the learning rate by 0.7 every 20 epochs. The models are trained for 200 epochs, where the momentum for Batch Normalization [23] layers starts with 0.9 and decays with a rate of 0.5 every 20 epochs, following the practice of [33, 38]. Detailed model configurations can be found in Supplementary Material.

4. Experiments

We extensively evaluate our method on several widely used point cloud classification benchmark datasets including ModelNet10/40 [52], ScanObjectNN [46] and ScanNet [7]. We start by evaluating our method on the discriminative power, generalization ability and robustness across datasets and comparing with the state-of-the-art unsupervised and supervised methods. We then provide detailed experiments to analyze our method on model design and complexity. Finally, we visualize the learned representations to have an intuitive understanding of our method. The following describes the details of the experiments, results and analysis.

4.1. Unsupervised Point Cloud Recognition

Setups: We tested our method on ModelNet40/10 [52] and ScanObjectNN [46] benchmarks to compare with the state-of-the-arts. ModelNet40 and ModelNet10 comprise 9832/3991 training objects and 2468/908 test objects in 40 and 10 classes respectively, where the points are sampled from CAD models. ScanObjectNN [46] is a real-world data, where 2902 3D objects are extracted from scans. To conduct cross dataset evaluation, we used the "object-only" split in all our experiments. ScanNet [7] was also used in our cross data evaluation experiments, where we followed the

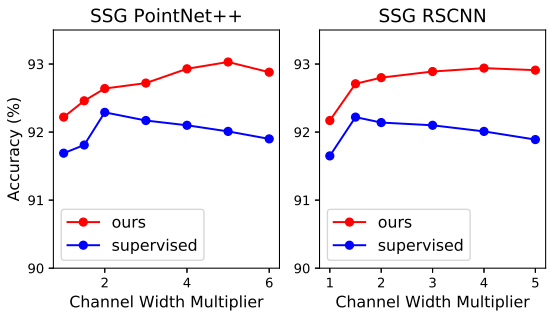

Figure 3: ModelNet40 classification accuracy (%) of our unsupervised models and their supervised counterparts.

Table 1: Classification accuracy (%) of three different training strategies on ModelNet40.

Backbone	Unsupervised	Supervised	Hybrid
PN++ (Small)	92.22	91.69	**92.42**
PN++ (Large)	**93.02**	92.01	92.76
RSCNN (Small)	92.17	91.65	**92.26**
RSCNN (Large)	**92.94**	92.14	92.78

practice in [28] to obtain point cloud from indoor scenes according to the instance segmentation labels. In all our experiments, we sample 1024 points for each point cloud for training and evaluation and all our results are measured using a single view without using the multi-view voting trick to show the neat performance of different models. Surface normal information was used to provide unsupervised signals for our models trained on ModelNet and we did not use it as input. For the models trained on ScanObjectNN and ScanNet, we only used the self-reconstruction loss for global-to-local reasoning.

Comparisons with the supervised counterparts: We first compared our method with the supervised baselines as presented in Figure 3, where we report the classification accuracy on ModelNet40 using the basic models (1×) and wider models (1.5 to 6× channel width). Note that we used the same network architecture and training settings for our models and their counterparts for a fair comparison. Clearly, our unsupervised models with different channel widths consistently outperform the supervised counterparts. As increasing the model capacity, our models can achieve better performance and reach the highest accuracy using 5× PointNet++ and 4× RSCNN backbones. In the following experiments, we denote the basic 1× models and the best models as "Small" and "Large" models respectively. Besides, we further compared three different training strategies: unsupervised learning, supervised learning and hybrid learning, which are presented in Table 1. We see hybrid learning can outperform both supervised and unsupervised models when the networks are small, but the unsupervised method achieves the best performance when large networks are used.

Table 2: Comparisons of the classification accuracy (%) of our method against the state-of-the-art **unsupervised** 3D representation learning methods on ModelNet40 and ModelNet10. † indicates that the model is trained on ShapeNet.

Method	Input	Accuracy	
		MN40	MN10
TL Network [15]	voxel	74.40	-
VConv-DAE [39]	voxel	75.50	80.50
3DGAN [50]	voxel	83.30	91.00
VSL [29]	voxel	84.50	91.00
VIPGAN [17]	views	91.98	94.05
LGAN† [1]	points	85.70	95.30
LGAN [1]	points	87.27	92.18
FoldingNet† [55]	points	88.40	94.40
FoldingNet [55]	points	84.36	91.85
MRTNet† [14]	points	86.40	-
3D-PointCapsNet [56]	points	88.90	-
MAP-VAE [18]	points	90.15	94.82
Ours w/ PN++ (Small)	points	92.22	94.82
Ours w/ PN++ (Large)	points	**93.02**	**95.53**
Ours w/ RSCNN (Small)	points	92.17	94.60
Ours w/ RSCNN (Large)	points	92.94	95.50

We conjecture that the supervised models are prone to overfitting more severely to the training set. All these results reveal that our unsupervised representation is more discriminative and generalizable than its supervised counterpart.

Comparisons with the unsupervised state-of-the-arts: To show the effectiveness of the proposed global-local reasoning method, we compared several variants of our models with the state-of-the-art unsupervised representation learning methods in Table 2. Except for point-based methods, we also compare with some advanced voxel and view based methods. Note that we only use ModelNet40 as the training data, while some methods are trained on larger ShapeNet [52] dataset. Nevertheless, our models outperform all other methods by a large margin. As can be observed, our small PointNet++ model surpasses state-of-the-art methods and our large model significantly advances the best point cloud model (MAP-VAE) by 2.87% on ModelNet40.

Comparisons with the supervised state-of-the-arts: More notably, our method can even achieve very competitive results compared to state-of-the-art supervised methods *in an unsupervised manner*. We compared our method with the supervised methods on both the widely used synthetic dataset ModelNet and the recently proposed real-world dataset ScanObjectNN. Our unsupervised representation was trained on ModelNet40 and a linear SVM is then trained on the target dataset to produce predictions. The results are summarized in Table 3 and Table 4. Surprisingly, our unsupervised learned representation can outperform all other

Table 3: Comparisons of the *single-view* classification accuracy (%) of our method aganist the state-of-the-art **supervised** point cloud models on **ModelNet40**. We also list results that use more points, normal information ("nor") or/and multi-view voting trick ("vote") in gray as references. Besides, we show the supervised baselines of our models.

Method	#Points	Supervised	Acc.
PointNet [37]	1k	✓	89.2
PointNet++ [38]	1k	✓	90.5
PointNet++ [38] (vote)	1k	✓	90.7
SO-Net [27]	1k	✓	92.5
PointCNN [28]	1k	✓	92.5
DGCNN [49]	1k	✓	92.9
DensePoint [32]	1k	✓	92.8
DensePoint [32] (vote)	1k	✓	93.2
RSCNN [33]	1k	✓	92.9
RSCNN [33] (vote)	1k	✓	93.6
DGCNN [49]	2k	✓	93.5
PointNet++ [38] (vote, nor)	5k	✓	91.9
SO-Net [27] (nor)	5k	✓	93.4
KPConv [43]	$\sim$ 6.8k	✓	92.9
PN++ (Large)	1k	✓	92.1
Ours w/ PN++ (Large)	1k	✗	**93.0**
RSCNN (Large)	1k	✓	92.0
Ours w/ RSCNN (Large)	1k	✗	92.9

Table 4: Comparisons of the *single-view* classification accuracy (%) of our method aganist the state-of-the-art **supervised** point cloud models on **ScanObjectNN**.

Method	Supervised	Accuracy
3DmFV [4]	✓	73.8
PointNet [37]	✓	79.2
SpiderCNN [54]	✓	79.5
PointNet++ [38]	✓	84.3
DGCNN [49]	✓	86.2
PointCNN [28]	✓	85.5
Ours w/ PN++ (Large)	✗	**87.2**
Ours w/ RSCNN (Large)	✗	86.9

state-of-the-arts methods in the single-view setting[3] on both datasets. Since only a linear classifier is applied, these results demonstrate that our representation is much more discriminative than the supervised representation on the test set. Moreover, we observe that our representation can achieve very strong results on ScanObjectNN without finetuning. As the categories in ModelNet and ScanObjectNN are different, this evidence indicates that our method can discover semantic knowledge shared in different kinds of objects.

[3]Here we borrow the concept of "view" from image recognition literatures, where the number of views represents the number of augmented inputs (*e.g.* rotated or scaled point clouds) used during testing.

Table 5: **Cross dataset evaluation.** We evaluate generalization ability of unsupervised and supervised representations to unseen datasets. We report the classification accuracy (%) measured using a linear SVM trained on the target dataset. (Sup.: supervised)

Task	Sup.	Ours	Δ
ModelNet10 → ModelNet30	85.45	92.34	+6.89
ModelNet30 → ModelNet10	91.32	95.47	+4.15
ModelNet40 → ScanObjectNN	65.92	87.22	+21.30
ScanObjectNN → ModelNet40	78.76	90.80	+12.04
ModelNet40 → ScanNet	77.31	89.23	+11.92
ScanNet → ModelNet40	80.38	91.32	+10.94
ScanObjectNN → ScanNet	84.31	87.96	+3.63
ScanNet → ScanObjectNN	82.44	85.43	+2.99

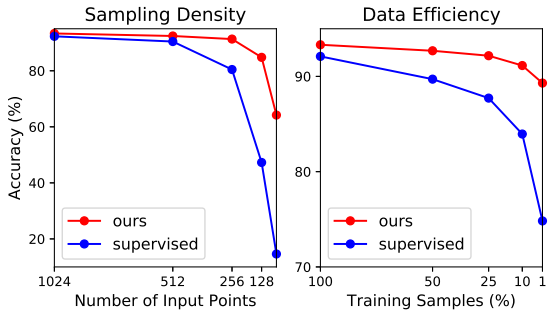

Figure 4: The **robustness** of our method on sampling density and the number of training samples compared to the supervised baseline.

Cross Dataset Evaluation: To further explore the generalization ability of the learned representation, we conducted extensive cross data evaluation experiments on ModelNet, ScanObjectNN and ScanNet, which are varying in categories and sources. Our experiments were conducted based on the unsupervised representations of the PointNet++ large model and we compared the results with the supervised version of this model. Specifically, we trained the features using supervised or unsupervised learning methods on the source dataset and used a linear SVM trained on the target dataset to perform classification. The results are presented in Table 5, where we used the rest 30 categories in ModelNet40 apart from 10 categories in ModelNet10 to form the ModelNet30 dataset. We see the unsupervisedly learned representation has much stronger transferability than the supervised counterparts and our models generalize well to various unseen data since we learn from *data structures instead of labels*. Our method can maintain strong performance even in cross data evaluation, which reflects the unsupervised representation can be a *generic* representation of 3D objects cross datasets.

Robustness Analysis: The robustness of our method on sampling density and the number of training samples is shown in Figure 4. For the former, we tested the model

Table 6: **Ablation study** of our method. We report the classification accuracy (%) on ModelNet40. ($\mathcal{L}_{\text{L2G}}$: local-to-global reasoning, $\mathcal{L}_{\text{recon}}$: self-reconstruction, Agg: multi-level feature aggregation in Eq. (7), $\mathcal{L}_{\text{normal}}$: normal estimation, SN: training on ShapeNet.)

Model	$\mathcal{L}_{\text{L2G}}$	$\mathcal{L}_{\text{recon}}$	Agg.	$\mathcal{L}_{\text{normal}}$	SN	Acc.
A		✓				86.77
B	✓					90.02
C	✓	✓				90.96
D	✓	✓	✓			91.69
E	✓	✓	✓	✓		92.22
F	✓	✓	✓	✓	✓	**92.30**

Table 7: **Complexity analysis.** We report the FLOPs and GPU inference throughput with batch size 16. Measured on NVIDIA GTX 1080Ti GPU. (pc/s: point cloud(s) per second)

Model	FLOPs	Throughput	Acc.
MSG PN++ [38]	1.68G	113pc/s	90.5
SSG RSCNN [33]	**0.30G**	634pc/s	**92.2**
Our PN++ (Small)	0.31G	**731pc/s**	**92.2**
MSG PN++ [38] (12 votes)	14.15G	9pc/s	90.7
SSG RSCNN [33] (10 votes)	**2.95G**	63pc/s	92.7
Our PN++ (Large)	5.65G	**194pc/s**	**93.0**

trained with 1024 points with sparser points of 1024, 512, 256, 128 and 64. Note that different from previous works [33, 38], we did not perform random input dropout during training. For the latter, we trained the representation with randomly sampled 100%, 50%, 25%, 10% and 1% ModelNet40 training set and trained the linear classifier on the whole set. We used the PointNet++ large model in this experiment. Generally, we see our models are much more robust than their supervised versions. Notably, our method can maintain decent performance even when using only 10% (983 samples) and 1% (98 samples) training samples and achieve 91.4% and 89.3% accuracy on ModelNet40 respectively.

Visualization: To have an intuitive understanding of our method, we visualized the unsupervised learn features. The results are presented in the supplementary material.

4.2. Method Design Analysis

Ablation Study: To examine the effectiveness of our designs, we conducted a detailed ablation study based on the small PointNet++ network. The results are summarized in Table 6. The baseline model A can be viewed as a variant of FoldingNet [55], which was trained by self-reconstruction loss only and gets a low classification accuracy of 86.77%. We see the model trained by the proposed local-to-global reasoning task (model B) can significantly improve the baseline model by 3.25%. This convincingly verifies its effectiveness. Then, when incorporating these two losses, the accuracy can be further improved to 90.96%. We also observe a 0.73% improvement by aggregating local and global representations (model D). Our full model can be obtained by adding normal estimation supervision (model E), which achieves a notable 92.22% accuracy on ModelNet40 with a very lightweight network. In addition, we also investigated the training set size by adding more training data (model F) from ShapeNet [52], but obtaining a slight improvement on accuracy (0.08%). We conjecture that ModelNet is large enough for learning a good representation. Thus we conducted most of the experiments on ModelNet.

Complexity Analysis: Table 7 shows the model complexity in theoretical computation cost (in FLOPs) and actual inference throughput on GPU of our models and several state-of-the-art methods. We see our large model requires considerable computation cost but maintains an acceptable actual cost on GPU due to the simplicity of the SSG model. These results reveal that increasing channel width can achieve a better trade-off on speed and accuracy compared to voting. For computational cost-sensitive applications, we think our learned model can provide strong supervision to train lighter models for real-time applications by model distillation [20] or generating pseudo labels [25], which is an interesting direction for future research.

5. Conclusion

We have proposed a new scheme for unsupervised representation learning of 3D point clouds by bidirectional global-local reasoning. Comprehensive experimental studies have demonstrated our unsupervisedly learned representation can surpass its supervised counterpart and achieve state-of-the-art performance on several widely used benchmarks. We expect our method to open a new door for learning better point cloud representation from data structures instead of human annotation. Transferring the learned knowledge to more efficient models and extending our method to more point cloud analysis scenarios like segmentation and detection are interesting directions in future work.

Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by Beijing Academy of Artificial Intelligence (BAAI), in part by a grant from the Institute for Guo Qiang, Tsinghua University, in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564, and in part by Tsinghua University Initiative Scientific Research Program.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *ICML*, 2018. 1, 3, 5, 6

[2] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *arXiv preprint arXiv:1803.10091*, 2018. 1, 3

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910*, 2019. 2, 4

[4] Yizhak Ben-Shabat, Michael Lindenbaum, and Anath Fischer. 3dmfv: Three-dimensional point cloud classification in real-time using convolutional neural networks. *IEEE Robotics and Automation Letters*, 3(4):3145–3152, 2018. 7

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *T-PAMI*, 35(8):1798–1828, 2013. 2

[6] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 5

[8] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *ECCV*, pages 602–618, 2018. 1, 3, 5

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 4

[10] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pages 1422–1430, 2015. 2

[11] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In *ICCV*, pages 2051–2060, 2017. 2

[12] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613, 2017. 5

[13] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*, pages 267–285. Springer, 1982. 2

[14] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *ECCV*, pages 103–118, 2018. 1, 3, 5, 6

[15] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, pages 484–499. Springer, 2016. 6

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[17] Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. In *AAAI*, volume 33, pages 8376–8384, 2019. 6

[18] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae: unsupervised feature learning for 3d point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. *ICCV*, 2019. 1, 3, 6

[19] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019. 2, 4

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 8

[21] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. 2

[22] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019. 4

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[25] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICMLW*, volume 3, page 2, 2013. 8

[26] Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018. 1, 2, 3, 5

[27] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *CVPR*, pages 9397–9406, 2018. 7

[28] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, pages 828–838, 2018. 1, 2, 3, 6, 7

[29] Shikun Liu, Lee Giles, and Alexander Ororbia. Learning a hierarchical latent-variable model of 3d shapes. In *3DV*, pages 542–551. IEEE, 2018. 6

[30] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, pages 212–220, 2017. 4

[31] Xinhai Liu, Zhizhong Han, Xin Wen, Yu-Shen Liu, and Matthias Zwicker. L2g auto-encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In *ACM MM*, pages 989–997. ACM, 2019. 1, 3

[32] Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, and Chunhong Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. In *ICCV*, pages 5239–5248, 2019. 7

[33] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *CVPR*, pages 8895–8904, 2019. 1, 2, 3, 5, 7, 8

[34] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. In *NeurIPS*, pages 2265–2273, 2013. 4

[35] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016. 2

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, 1(2):4, 2017. 2, 3, 7

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, pages 5099–5108, 2017. 1, 2, 3, 5, 7, 8

[39] Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In *ECCV*, pages 236–250. Springer, 2016. 6

[40] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, pages 1857–1865, 2016. 4

[41] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014. 5

[42] Benoit Steiner, Zachary DeVito, Soumith Chintala, Sam Gross, Adam Paszke, Francisco Massa, Adam Lerer, Gregory Chanan, Zeming Lin, Edward Yang, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 5

[43] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. *ICCV*, 2019. 1, 3, 7

[44] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 2, 4

[45] Michael Tschannen, Josip Djolonga, Paul K. Rubensteiny, Sylvain Gelly, and Lucic Mario. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 4

[46] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, pages 1588–1597, 2019. 2, 5

[47] Diego Valsesia, Giulia Fracastoro, and Enrico Magli. Learning localized generative models for 3d point clouds via graph convolution. In *ICLR*, 2019. 1, 3, 5

[48] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 4

[49] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 2019. 1, 2, 3, 7

[50] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, pages 82–90, 2016. 6

[51] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, pages 9621–9630, 2019. 1, 2, 3

[52] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, pages 1912–1920, 2015. 2, 5, 6, 8

[53] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, pages 3733–3742, 2018. 4

[54] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. *ECCV*, 2018. 1, 3, 7

[55] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, pages 206–215, 2018. 1, 3, 5, 6, 8

[56] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *CVPR*, pages 1009–1018, 2019. 1, 3, 5, 6