Context Prediction for Unsupervised Deep Learning on Point Clouds

Jonathan Sauder ¹  Bjarne Sievers ¹

Abstract

Point clouds provide a flexible and natural representation usable in countless applications such as robotics or self-driving cars. Recently, deep neural networks operating on raw point cloud data have shown promising results on supervised learning tasks such as object classification and semantic segmentation. While massive point cloud datasets can be captured using modern scanning technology, manually labelling such large 3D point clouds for supervised learning tasks is a cumbersome process. This necessitates effective unsupervised learning methods that can produce representations such that downstream tasks require significantly fewer annotated samples. We propose a novel method for unsupervised learning on raw point cloud data in which a neural network is trained to predict the spatial relationship between two point cloud segments. While solving this task, representations that capture semantic properties of the point cloud are learned.

Our method outperforms previous unsupervised learning approaches in downstream object classification and segmentation tasks and performs on par with fully supervised methods.

1. Introduction

Point clouds provide a natural and flexible representation of objects in metric spaces. They can also be easily captured by modern scanning devices and techniques. Algorithms that can extract semantic information from point clouds are crucial to countless applications such as robotics and self-driving cars. Traditionally, systems for such tasks have relied on the approximate computation of geometric features such as faces, edges or corners (Van Kaick et al., 2011; Guo et al., 2014) and hand-crafted features encoding statistical properties (Aubry et al., 2011; Rusu et al., 2008). However, these approaches are often tailored to specific tasks, thus not providing the necessary flexibility for modern applications. Recently, Convolutional Neural Networks (CNNs) which are domain-independent have shown promising performance on point clouds in supervised learning tasks such as object classification and semantic segmentation (Qi et al., 2017a;b; Wang et al., 2018; Li et al., 2018), outperforming conventional approaches.

The advent of scalable 3D point cloud scanning technologies such as LiDAR scanners and stereo cameras gives rise to massive point cloud datasets, possibly spanning large entities such as entire cities or regions. However, manually annotating such massive amounts of data for supervised learning tasks such as semantic segmentation poses problems due to typical real-world point clouds reaching billions of points and petabytes of data, opposing the innate limitations of user-interfaces for 3D data labelling (e.g. drawing bounding boxes) on 2D screens. Therefore, it is of large interest to develop unsupervised methods which can reduce the number of annotated samples needed for strong performance on downstream tasks.

Deep neural networks have been shown to learn powerful hierarchical representations in unsupervised settings across multiple domains (Lee et al., 2009; Mikolov et al., 2013). Such methods can generally be classified into methods that learn a mapping of samples to latent space, such as Autoencoders (Hinton & Salakhutdinov, 2006), or Generative Adversarial Nets (GANs) (Goodfellow et al., 2014) and methods in which representations are learnt from a common...
context between objects. Methods that learn from context have been famously applied in natural language processing in the form of word embeddings (Mikolov et al., 2013) and have also been introduced to images (Doersch et al., 2015). On point clouds, previous unsupervised deep learning has focused on GANs and Autoencoders.

In this work we present an unsupervised learning method based on context prediction operating on raw point cloud data. We train a neural network to predict the spatial relationship between two point cloud segments. Solving this task requires learning representations which encode the semantics of the point cloud segments. An example is given in Figure 1. Our method circumvents the need for computationally expensive similarity measures on point clouds and provides flexibility in specific network architecture design. Various neural network models that operate on raw point clouds and have been successfully applied to supervised tasks can be used as a drop-in component in our setup. As such networks learn global (scene-level) and local (per-point) representations, our method is also able to learn per-point features in an unsupervised manner.

In a series of experiments, we show that our method learns powerful representations of point clouds. Our method outperforms previous unsupervised methods in a downstream object classification task in a transfer learning setting. We also explore per-point features and show that a simple Multi-Layer Perceptron (MLP) trained on the frozen embeddings obtained from unsupervised training reaches competitive performance with fully supervised approaches. We highlight our main contributions as:

- **We present a flexible unsupervised learning method operating on raw point clouds in which a neural network is trained to predict the spatial context between two point cloud segments. Our method avoids computationally expensive and possibly flawed reconstruction losses or similarity metrics on point clouds.**

- **We demonstrate the ability of the obtained embeddings. Our method outperforms state-of-the-art unsupervised methods in a downstream object classification task.**

- **To the best of our knowledge, we are the first to explore per-point unsupervised embeddings in part segmentation and semantic segmentation tasks. A simple classifier trained on frozen obtained unsupervised point embeddings performs on par with fully supervised approaches.**

### 2. Related Work

#### 2.1. Deep Learning on Point Clouds

Deep neural networks have shown impressive performance on regularly structured data representations such as images and time series. However, point clouds are unordered sets of vectors, therefore exemplifying a class of problems posing challenges for deep learning for which the term *geometric deep learning* (Bronstein et al., 2017) has been coined. Although deep learning methods for unordered sets (Vinyals et al., 2015; Zaheer et al., 2017) have been proposed and also applied to point clouds (Ravanbakhsh et al., 2016), these approaches do not leverage spatial structure.

To address this problem, popular point cloud representations suitable for deep learning include volumetric approaches (Maturana & Scherer, 2015), in which the containing space is voxelized to be suitable for 3D CNNs and multi-view approaches (Savva et al., 2017), in which 3D point clouds are rendered into 2D images fed into 2D CNNs. However, voxelized representations can be difficult to use when the point cloud density varies, and as such are constrained by the resolution and limited by the computational cost of 3D convolutions. Despite multi-view approaches having shown strong performance in classification of standalone objects, it is unclear how to extend them to work reliably in larger scenes (e.g. with covered objects) and on per-point tasks such as part segmentation (Qi et al., 2017a).

A more recent approach, pioneered by PointNet (Qi et al., 2017a), is feeding raw point cloud data into neural networks. As point clouds are unordered sets, such networks have to be permutation invariant - PointNet achieves this by using the max-pooling operation to form a single feature vector representing the global context from a variable amount of points. PointNet++ (Qi et al., 2017b) proposes an extension that introduces local context by stacking multiple PointNet layers. Inspired by advancements in deep learning on graphs (Masci et al., 2015; Yi et al., 2017), further improvements were made by introducing Dynamic Graph CNNs (DGCNNs) (Wang et al., 2018), in which a graph convolution is executed on edges of the k-nearest neighbor graph of the point clouds, which is dynamically recomputed in feature space after each layer. Even better performance was achieved by PointCNN (Li et al., 2018), which uses a hierarchical convolution that is trained to learn permutation invariance. All neural networks operating on raw point cloud data naturally provide per-point embeddings, making them particularly useful for point segmentation tasks. Our proposed method can leverage these methods as it is flexible with regards to the use of specific neural network architecture.
2.2. Unsupervised Deep Learning

Deep learning algorithms have demonstrated the ability to learn powerful internal hierarchical embeddings through unsupervised learning tasks (Lee et al., 2009). These representations can be directly used in downstream tasks or as strong initializers for supervised tasks (Mikolov et al., 2013; Erhan et al., 2010). In cases where large amounts of data are available but annotated samples are scarce, unsupervised learning can significantly reduce the number of annotated training samples required for strong performance in various tasks (Yang et al., 2018), making effective unsupervised methods for point clouds particularly desirable. Unsupervised methods for deep learning can be loosely categorized into methods that learn a mapping between samples and a latent space such as Generative Adversarial Nets (Goodfellow et al., 2014) and Autoencoders (Hinton & Salakhutdinov, 2006), and methods that learn through the common context between a set of objects such as unsupervised word embeddings (Mikolov et al., 2013).

GANs have shown impressive results in the image domain (Brock et al., 2018), and have also been applied to point clouds (Wu et al., 2016). However, such approaches work on volumetric point cloud representations as GANs rely on the ability to generate artificial samples which is non-trivial for unordered sets such as raw point cloud data. A volumetric Autoencoder for point clouds titled VConv-DAE (Sharma et al., 2016) has also been proposed, as well as another approach trains a GAN on the latent space mapping obtained through training a separate Autoencoder (Achlioptas et al., 2017). The quality of the embeddings produced by these approaches is conventionally evaluated in object classification as a downstream task, where a linear SVM is trained on the embeddings obtained from unsupervised training on a larger, unlabelled dataset. The best performance thus far on such a task is achieved by FoldingNet (Yang et al., 2018), an Autoencoder that learns to fold a regular 2D grid around a point cloud.

All mentioned unsupervised methods for point clouds rely on similarity metrics such as the Earth Mover’s Distance (EMD) (Rubner et al., 2000) for which the computational cost is infeasible, or the closely related differentiable Chamfer (pseudo) distance. The Chamfer distance for two point clouds \( A \subset \mathbb{R}^d \), \( B \subset \mathbb{R}^d \) is the distance between each point to the closest point in the other set:

\[
d(A, B) = \sum_{a \in A} \min_{b \in B} \|a - b\|_2^2 + \sum_{b \in B} \min_{a \in A} \|a - b\|_2^2
\]

The Chamfer distance can also be inefficient to compute in large point clouds, but more importantly, the authors (Achlioptas et al., 2017) observe that specific pathological cases are handled incorrectly (Achlioptas et al., 2017). This motivates unsupervised methods such as ours which avoid such potentially problematic similarity functions.

While approaches such as GANs and Autoencoders are extremely popular in the image domain (possibly due to their generative ability), a completely different approach to unsupervised learning for images is taken by (Doersch et al., 2015). They split an image into patches and train a neural network to predict how two patches from the same image correspond, demonstrating that the learned features can be a powerful initialization for supervised training and that the learned embeddings successfully capture semantics. We build on this approach and adapt the idea of spatial context prediction to point clouds, which have certain characteristics that make them particularly well suited for such a task. The authors (Doersch et al., 2015) argue that such a context prediction task addresses the problem that predicting...
pixels is more difficult than predicting words, as a large variety of pixels can arise from the same semantic object. This holds even more true when moving from images, i.e. regular grids in 2D space, to unordered sets in 3D space. Furthermore, context prediction phrases the unsupervised learning problem as a classification task, therefore avoiding possibly problematic point cloud similarity metrics.

3. Method

3.1. Learning Representations from Context

In this paper we propose a method that learns high-level representations from context in point clouds in an unsupervised fashion. Our method works by training a neural network to predict the spatial relation between two point cloud segments. While learning to solve this unsupervised task, representations that capture the semantics of the point cloud are learned. We extend the concept of context prediction proposed by (Doersch et al., 2015) from images to point clouds, as point clouds have particularly desirable properties for such a task.

In general, a context gives rise to the semantics of a single entity. For example, a sentence puts a word in context and an image puts a pixel in context. Similarly, the context of a point in space is determined by the surrounding point cloud. From a context it is possible to extract relations, i.e. semantic relationships between two entities. In point clouds, an example for such a relation between two of its segments could be “A is part of B”. Other useful relations can be data-dependent e.g. “A is B in summer” could relate a 3D scanned leafless tree in wintertime to its green summer pendant through a temporal dimension. Spatial relations are innate to point clouds in euclidean space, e.g. “A is above B” or “A is rotated” by using a random rotation, and can be used independently of domain and data. We propose a deep neural net architecture which is flexible with regards to the set of possible relations to learn from.

An overview of our architecture is shown in Figure 2. The input is given by two point cloud segments standing in a relation to each other. The learning objective is a classification task in which each class represents a possible relation. This can be done by minimizing the multi-label cross entropy, without needing to compute any point cloud similarity metrics. The weights of the network taking the variable-sized point cloud as inputs and producing a fixed size feature vector are shared between the two point cloud inputs. The exact choice of this network is flexible: while models operating on raw point clouds, such as PointNet or DGCNN, are particularly interesting as they produce per-point embeddings and are robust to varying densities, volumetric or multi-view networks could also be used. In fact, future work proposing better-performing networks for fully supervised tasks should translate directly into better performance within a context prediction task.

3.2. Spatial Context in Point Clouds

In our experiments we focus exclusively on spatial relations, making our method universally applicable to any point cloud. An exhaustive analysis of possible relations which can lead to better results in specific tasks is out of scope for this paper. Instead, we empirically show that even with only simple spatial relations, our proposed method can achieve good results. While in images, spatial relations are limited to selecting patches from different locations in the image (Doersch et al., 2015), spatial relations in point clouds naturally provide more flexibility as they are not constrained by a discrete grid, e.g. allowing for rotation and scaling.

In all completed experiments we use the same set of ten relations. Specifically, we include relations over the height axis, i.e. “A is above B” or “A is below B”. To avoid the triviality of simply comparing the height-values, we scale each input segment to unit ball, losing global height values but preserving height values within each segment. We treat relations on the non-height axes as equal, i.e. there is no distinction between “behind” or “beside” as objects are not always facing in the same direction. We do however distinguish between “A is next to B” and “A is diagonally next to B”. Combining these, e.g. “A is diagonally below B” leads to eight spatial relations. Furthermore, we include the relations “A is B rotated” by using a random rotation, and “A is not related to B” realized by choosing two random point cloud segment within the entire training dataset.

3.3. Model

In all experiments we use the same model architecture in the unsupervised learning task 1. We refer to our model as CloudContext. It contains a simplified version of DGCNN (Wang et al., 2018) as it provides a reasonable trade-off between performance and computational complexity. The simplified version uses no spatial transform network and consists of four EdgeConv layers on the 20-nearest neighbor graph with filter sizes 64, 64, 64, 128 and a last convolution that concatenating previous layers through skip connections with filter size 256. This leads to a per-point and per-segment embedding of size 256. Two dense layers of size 1024 and 512 follow before the final linear layer to classify the correct number of relation class labels. We use ReLU activations and batch normalization on all layers, and 50 % and 20 % dropout on the two dense layers. We use Adam as optimizer with a fixed learning rate of 0.0001. To improve generalization, we augment the input point clouds. The entire point cloud is scaled to the unit ball before being rotated.

Footnote 1: Full code is provided in supplementary material, will be made publicly available upon publication.
Table 1. Comparison of unsupervised methods in downstream object classification on the ModelNet40 and ModelNet10 dataset in terms of accuracy. All methods use a linear SVM trained on the representations learned in an unsupervised fashion on the ShapeNet dataset.

| Model                    | MN40   | MN10   |
|--------------------------|--------|--------|
| VCONV-DAE                | 75.5%  | 80.5%  |
| 3D-GAN                   | 83.3%  | 91.0%  |
| LATENT-GAN               | 85.7%  | 95.3%  |
| FOLDINGNET               | 88.4%  | 94.4%  |
| CLOUDCONTEXT (Ours)      | 89.3%  | 94.5%  |

around the height-axis by a random multiple of 90 degrees. Then each spatial dimension is independently scaled by a random variable drawn from $N(1, 0.1)$, the point cloud is entirely shifted by a random vector drawn from $N(0, 0.05)$, and each point is jittered on by adding a vector drawn from $N(0, 0.05)$.

4. Experimental Results

4.1. Object Classification

In this section, we show that the embeddings learned with our method, referred to as CloudContext, perform strongly in object classification as a downstream task and show that we outperform state-of-the-art unsupervised methods before comparing our method to fully supervised approaches.

In line with previous approaches, we evaluate our performance on the object classification problem using the ModelNet dataset (Wu et al., 2014), which contains CAD models from different categories of man-made objects. For this we use the standard train/test split, with the same uniform point sample as defined in (Qi et al., 2017a) with ModelNet40 on 40 classes containing 9843 train and 2468 test models and ModelNet10 on ten classes containing 3991 and 909 models respectively.

To obtain object-level embeddings from our method, we combine all segment-level embeddings that make up a single object into a single embedding through concatenation. For the first experiment, we train a SVM as in (Achlioptas et al., 2017; Wu et al., 2016; Yang et al., 2018) with representations obtained using a dataset different from ModelNet: The ShapeNet dataset (Chang et al., 2015) consists of 57448 models from 55 categories. From every model we use only a random sample of 2048 points as provided by (Yang et al., 2018). In the second experiment, comparing to end-to-end supervised methods, our method still proceeds unsupervised as in the first experiment, but with embeddings learned using ModelNet instead of ShapeNet in order to have only exactly the same information available as the supervised methods.

In the unsupervised setting our method outperforms all previous approaches on ModelNet40, and all except Latent-GAN on ModelNet10. However, as noted by (Yang et al., 2018), the point cloud format and sampling procedure from Latent-GAN is not publicly available, making a comparison on this accuracy inconclusive. The obtained embeddings for the ModelNet10 test data are visualized using t-SNE (Maaten & Hinton, 2008) in Figure 3. One can see that clear clusters are formed for each class which are separable except for the classes dresser (violet) vs nightstand (pink), which are visually similar when scaled to the unit cube, as done in the ShapeNet dataset. Figure 4 highlights the effectiveness of context prediction: a decrease in unsupervised training loss gives a better downstream classification accuracy, which suggests that our method is able to capture more relevant semantics with more training, enabling better linear separability.

A main motivation for unsupervised training is that in a
large dataset, a very small number of labelled samples can suffice to achieve strong performance in machine learning tasks. We evaluate context prediction in such a scenario by limiting the number of training samples available for SVM training in the ModelNet object classification task. We sample according to the following procedure: first we randomly sample one object per class, and then sample the remaining objects uniformly out of the entire training set. The results are shown in Figure 5. Using only 1% of training data, equivalent to roughly three samples per class, our model is able to achieve 61.1% accuracy on the test set. When using 10% of available training samples, this accuracy rises up to 82.6%.

In the second experiment we compare our approach to fully supervised approaches. All models, including our unsupervised approach, are trained on the ModelNet40 train dataset. We train a kernel SVM on the obtained embeddings in our approach. The results are shown in Table 2. Our unsupervised approach can outperform the fully supervised PointNet++.

### 4.2. Part Segmentation

In this section we explore the per-point embeddings obtained through unsupervised training in a part segmentation task. Again, we train our model in an unsupervised fashion on the ShapeNet dataset. The task is then to correctly classify points as belonging to an object part on the ShapeNet Part dataset (Yi et al., 2016), which is a subset of the full ShapeNet containing 16881 3D objects from 16 categories, annotated with 50 parts in total. We use the official train / validation / test splits. Each point is represented as the concatenation of the frozen point embedding with the point’s corresponding segment embedding obtained through unsupervised training. Following the same procedure as in (Qi et al., 2017a;b; Wang et al., 2018), we also concatenate the class label of the object. We then train a 3-hidden layer fully connected MLP with layer sizes of 1024, ReLU activations, batch normalization and 50% dropout on each hidden layer on the embeddings. Part segmentation is evaluated on the mean Intersection-over-Union (mIoU) metric, calculated by averaging IoUs for each part in an object before averaging the obtained values for each object class. The results are shown in Table 3. Our approach shows competitive performance with fully supervised methods. CloudContext outperforms PointNet in six object classes and can outperform every evaluated methods in two classes.

### 4.3. Semantic Segmentation

The semantic segmentation task is evaluated on the Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset (Armeni et al., 2017). The dataset consists of 3D point cloud scans from 6 indoor areas totalling 272 rooms. The points are classified into 13 semantic classes such as board, chair, ceiling, beam, and clutter. For fair comparability, we use the exact same procedure as in (Qi et al., 2017a; Wang et al., 2018), in which each room is split into blocks of 1m × 1m area and each point is given as a 9D vector containing XYZ coordinates, RGB color values, and normalized coordinates indicating the position in the room. A 6-fold cross validation is performed in which each of the areas is used as test set once. We train in an unsupervised manner on each training split data, and use the same setup for classification on the obtained embeddings as in the part segmentation task. The results are summarized in Table 4. Despite the embeddings being learned in an unsupervised manner, CloudContext is able to slightly outperform PointNet in classification accuracy and performs on par with PointNet in terms of mIoU. This demonstrates that our method is applicable to point clouds that capture scenes that go beyond simple free-standing simple objects such as included in the ShapeNet or ModelNet datasets.

### 5. Discussion

While learning to predict spatial relations of point cloud segments in an unsupervised manner, a neural network learns object-level and per-point semantic representations of objects. For example, to successfully predict that a part of...
The specific neural network architecture used for context prediction could instead benefit from learning in a contiguous point cloud setting, allowing to learn from relations with varying granularity. One of the main motivations for unsupervised learning in point clouds is reducing the number of labelled samples required for strong performance in downstream tasks. Our method succeeds in this aspect.

The authors (Doersch et al., 2015) observe that when using context prediction for images, the neural network quickly finds trivial solutions in low-level information such as similar textures or chromatic aberration. However, we observe that in none of our experiments, the neural network quickly learns to correctly identify the relations without producing useful embeddings for downstream tasks. Our results suggest that trivial solutions are not detrimental to unsupervised learning in a point cloud setting.

### 6. Conclusion

In this paper we propose an unsupervised method for learning representations in point clouds from context. In this method, a neural network learns to predict the spatial relationship between two point cloud segments. While learning to solve this task, a neural network learns representations encoding the semantics of the point cloud. We propose a generalizable setup, in which a concrete neural network architecture can be flexibly used as a drop-in component. We use a simplified DGCNN within our setup and expect to see improvements with context prediction in the future as advancements in deep learning architectures for point clouds are made. We demonstrate that the embeddings obtained through unsupervised training can lead to strong performance in downstream tasks: in an object classification task, our method outperforms state-of-the-art unsupervised methods. We also explore unsupervised per-point embeddings in point segmentation tasks: training a simple classifier on the obtained unsupervised embeddings leads to competitive performance with fully supervised methods. Using our unsupervised method in downstream tasks can result in strong performance even when the number of labelled samples is extremely small. This is useful as point clouds in practice such as 3D LiDAR scans contain massive amounts of data, but are cumbersome to label. In the scope of this paper we focus exclusively on simple spatial relations. An exhaustive analysis of possible relations and experimenting with input data formats other than pre-processed chunks of fixed sizes that could improve learning is left as future work.
References

Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. J. Representation learning and adversarial generation of 3d point clouds. *CoRR*, abs/1707.02392, 2017. URL http://arxiv.org/abs/1707.02392.

Armeni, I., Sax, A., Zamir, A. R., and Savarese, S. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, February 2017.

Aubry, M., Schlickewei, U., and Cremers, D. The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1626–1633. IEEE, 2011.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096, 2018. URL http://arxiv.org/abs/1809.11096.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Chang, A. X., Funkhouser, T. A., Guibas, L. J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. URL http://arxiv.org/abs/1512.03012.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1422–1430, 2015.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Guo, Y., Bennamoun, M., Sohel, F., Lu, M., and Wan, J. 3d object recognition in cluttered scenes with local surface features: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2270–2287, 2014.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, pp. 609–616, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553453. URL http://doi.acm.org/10.1145/1553374.1553453.

Li, Y., Bu, R., Sun, M., Wu, W., Di, X., and Chen, B. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, pp. 828–838, 2018.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Masci, J., Boscaini, D., Bronstein, M., and Vandergheynst, P. Geodesic convolutional neural networks on riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 37–45, 2015.

Maturana, D. and Scherer, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pp. 922–928. IEEE, 2015.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Computer Vision and Pattern Recognition (CVPR)*, abs/1612.00593, 2017a. URL http://arxiv.org/abs/1612.00593.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pp. 5099–5108, 2017b.

Ravanbakhsh, S., Schneider, J., and Poczos, B. Deep learning with sets and point clouds. *arXiv preprint arXiv:1612.04500*, 2016.

Rubner, Y., Tomasi, C., and Guibas, L. J. The earth mover’s distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

Rusu, R. B., Blodow, N., Marton, Z. C., and Beetz, M. Aligning point cloud views using persistent feature histograms. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pp. 3384–3391. IEEE, 2008.
Savva, M., Yu, F., Su, H., Kan ezaki, A., Furuya, T., Ohbuchi, R., Zhou, Z., Yu, R., Bai, S., Bai, X., et al. Large-scale 3d shape retrieval from shapenet core55: Shrec’17 track. In Proceedings of the Workshop on 3D Object Retrieval, pp. 39–50. Eurographics Association, 2017.

Sharma, A., Grau, O., and Fritz, M. Vconv-dae: Deep volumetric shape learning without object labels. In European Conference on Computer Vision, pp. 236–250. Springer, 2016.

Van Kaick, O., Zhang, H., Hamarneh, G., and Cohen-Or, D. A survey on shape correspondence. In Computer Graphics Forum, volume 30, pp. 1681–1707. Wiley Online Library, 2011.

Vinyals, O., Bengio, S., and Kudlur, M. Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391, 2015.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. Dynamic graph CNN for learning on point clouds. CoRR, abs/1801.07829, 2018. URL http://arxiv.org/abs/1801.07829.

Wu, J., Zhang, C., Xue, T., Freeman, B., and Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In Advances in Neural Information Processing Systems, pp. 82–90, 2016.

Wu, Z., Song, S., Khosla, A., Tang, X., and Xiao, J. 3d shapenets for 2.5d object recognition and next-best-view prediction. CoRR, abs/1406.5670, 2014. URL http://arxiv.org/abs/1406.5670.

Yang, Y., Feng, C., Shen, Y., and Tian, D. Foldingnet: Point cloud auto-encoder via deep grid deformation. In Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), volume 3, 2018.

Yi, L., Kim, V. G., Ceylan, D., Shen, I., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L., et al. A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (TOG), 35(6):210, 2016.

Yi, L., Su, H., Guo, X., and Guibas, L. J. Syncspeccnn: Synchronized spectral cnn for 3d shape segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2282–2290, 2017.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., and Smola, A. J. Deep sets. In Advances in Neural Information Processing Systems, pp. 3391–3401, 2017.