Overview and Trends in Self-Supervised Learning on Point Clouds

Anshu Garg[1] and Mahdi Chamseddine[2]

[1] a_garg19@cs.uni-kl.de
[2] mahdi.chamseddine@dfki.de

Abstract. Self-Supervised Learning recently became popular among researchers to address the problem of a vast amount of labelled data. It is used effectively to learn the semantic local and global features of input data. Deep Neural Networks are efficient on a large amount of labelled Point Cloud input data, which has challenges such as manual labelling, costly and need lots of time. This paper provides a literature review to show that SSL can be used to address the above problem and enhance the performance of SOTA models. Four SSL techniques are discussed in this paper along with the experimental results of respective papers to justify the utilisation of SSL methods. The first technique shows rearranging the puzzled voxelised point cloud learns semantic features of input data. The second technique captures local and global features by representing the point cloud in a hierarchical structure. Third technique reconstructs the deformed point cloud and provides meaningful feature vector of the input point cloud. The last approach solves the pretext task of predicting the rotation angle of point cloud with respect to canonical orientation, which helps the model learn features useful in the downstream task. SSL has encouraging results discussed along with each technique.

Keywords: Self-Supervised Learning, Point Cloud, Object Classification, Semantic Segmentation, Downstream task, Pretext task.

1 Introduction

Nowadays, every technology is inspired by Machine Learning. Computer Vision and Natural Language Processing (NLP) domains have shown groundbreaking advancement in the last few years. Now, the error rate by the human in CV is 3% [8]. In the research of making machines smart like Humans, there are a variety of learning techniques introduced so far like Supervised Learning (SL), Unsupervised Learning, Reinforcement Learning, Deep Learning and more. Best results are provided by SL as most supervision is provided. Some successful Deep Learning models are AlexNet, ResNet, GoogleNet, PointCNN, PointNet++, DGCNN and more. But the biggest challenge is the availability of the enormous amount of labelled data for training these massive millions of parameters models. One of the recent technique introduced to address this problem is Self-Supervised Learning (SSL).

SSL aims to provide some kind of supervision from unlabelled data itself to models to solve some particular task [7]. This technique is inspired by the learning process of the human being. Since childhood children learn basic concepts one by one, they learn to solve complex problems later own [1]. On a similar principle, in SSL, the idea is to teach models to understand data from simple tasks called Pretext or Auxiliary task, before making it work on the actual problem called Downstream task. Different types of Pretext tasks are important for different kinds of data. For image-based data various SSL tasks are Distortion in the image then reconstruct the original image, find the position of a patch in the image relative to other patches, predict the RGB colour of Grayscale image and more. For video-based data, the learning motion of the camera for tracking objects determines if the sequence of frames from a video has the correct order, map the colours from input frame to Grayscale target frame and more [17].

Some of the most visible advantages of SSL are as follows. As per the trend, researchers are working on improving the current model's (Supervised and Unsupervised) performance using SSL. Models are trained to consume less amount of labelled data, thus saving lots of manual efforts. By using SSL techniques, lots of unused data can be utilised in training. Multiple rounds of pre-training with SSL methods provide a greater understanding of the semantics of data. Some of the most significant successful application examples of SSL are Transformers in NLP, BERT, RoBERTa, GPT2 and Meena Chatbot [3]. Application areas of SSL are many, and in this paper, our focus is the use of SSL on 3D Point Cloud. We will discuss some of the latest work on Point Cloud in Section 2. Section 3 will discuss some latest SSL pre-training tasks used by different research papers and their experiments. Finally, in Section 4, we will provide the Conclusion and Future Scope.

2 Related Work

Point Clouds have huge application area such as automation driving, robotics, marine survey and more. Due to the development of 3D objects capturing technology such as LiDARs, 3D Scanners, RGB-D cameras and more, 3D data is immensely available. But again the problem in utilising this data is manual efforts in labelling it. Secondly, Deep Learning works well with ordered data like images and videos, where data can be defined in Euclidean space, have grid-like structure (fixed-length, height, depth), have statistical properties. On the other hand, Point Clouds are an irregular and unordered set of vectors. They can have non-euclidean nature of data, varying density, non-familiar common coordinate system and more. Thus, working with both data is quite different. Some of the recent advancement by processing point cloud data are Object classification, segmentation and tracking. Few latest techniques for handling point cloud are discussed below.

2.1 Deep Learning on Point Cloud

After the huge success of deep neural networks on structured data, many architectures have been made to process 3D point cloud as well. Following the Volumetric approach, converting point cloud to voxels, mapping geometric data to grid structure and multi-view based methods, the most powerful models are DGCNN, PointNet, PointCNN, Pointwise CNN, SPLATNET [16], [10], [6], [5], [13]. However, the performance has been achieved on the cost of heavy computations and large memory footprints as the resolution increases. Further, on the concept of hierarchical and graph structures, Octree based CNN models were proposed. But they have a complex implementation[4].

2.2 Unsupervised Learning on Point Cloud

The goal of unsupervised learning is to learn the data representation from the point cloud. The most common approach is Variational AutoEncoders (VAE), Generative Adversarial Networks (GAN) and AutoRegressive models. Some standard techniques are to reconstruct the point cloud or generate new point clouds to learn feature representation in AE and GAN, respectively[9]. These techniques rely on similarity metrics like Chamfer Distance and Earth Mover's distance. This is the general technique in Unsupervised learning. But these approaches can suffer from problems like selecting correct similarity metrics and sampling the unordered point cloud. Many methods are not directly applicable to the raw point cloud.

2.3 Self-Supervised Learning on Point Cloud

As per recent research, SSL proved to be successful in learning the relevant characteristics of point clouds. Some of the common pre-training tasks on point cloud used by SSL techniques are discussed further. In [15], local features are tried to capture by predicting the next point in point sequence using MortonNet model. In [14], the point cloud's geometric properties are learned using the multi-task geometric learning network by predicting the point-normal vector and curvature. In [11] point cloud is broken into multiple voxels and by rearranging these voxels to original representation after shuffling, makes the model learns import features. In [9], finding the rotation of point cloud reveals important information regarding the points cloud features. Many such methods proved to be useless by researchers out of which few this paper will discuss in detail in the next section.

3 Reviewed Approaches

3.1 Reconstructing Space of Point Cloud

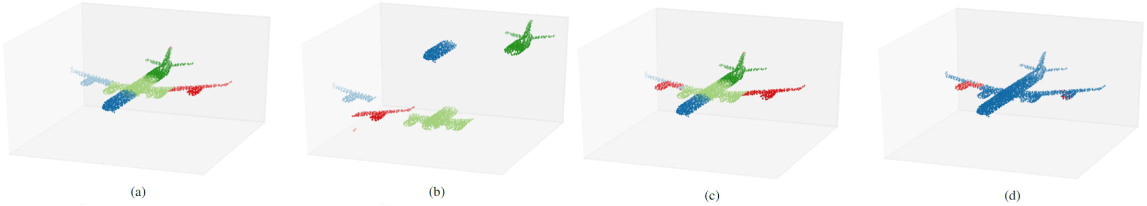

Fig. 1: (a) Point cloud split into k^3 voxels, (b)Shuffled Voxels, (c) Predicted voxel labels, (d)Correctly (blue) and incorrectly (red) predicted labels

The authors of [11], suggests a novel SSL pretext task in which neural network (NN) is trained on the raw point cloud. It learns the semantic information of the point cloud. The method is not architecture-specific, thus provides the flexibility to pre-train any Deep Learning model and improve its performance. It applies to all the domains data in which height-axis is distinctly defined. The main idea is to shuffle the point cloud voxels and rearrange those voxels in the original figure. As per this algorithm, the task is considered as point segmentation task, in which each point is labelled by its voxel Id. Firstly, the point cloud is scaled to the unit cube. Then each axis is divided into k equal parts, which forms k^3 voxels. Now, each voxel is assigned an ID which is allocated to each point in that voxel. Further, all the voxels are randomly shuffled, and the model is trained to predict the actual label for each point in the point cloud. For a generalisation of the method, data augmentation by randomly shifting each point per voxel by a small amount can be done. It is described in Fig 1.

Any Deep Learning network like PointCNN, PointNet, DGCNN or more which are designed to solve downstream point segmentation task can be pre-trained by this method. Since the pretext task is point segmentation; thus, it provides some benefits. Firstly, no need to sample point cloud, secondly, the model doesn't have to rely on possibly incorrect similarity metrics, no need for the 2D rendering of the point cloud. Hence, it applies to any raw point cloud. Thus, solves the problem of Unsupervised techniques. An added advantage is the flexibility of density of point cloud, which provides a model to learn per-point embedding. Some of the interesting results of Part Segmentation experiments are shown in Fig 2.

3.2 Hierarchical Representation of Point Cloud

Idea/Method In paper [12], two novel SSL methods have been proposed along with network architecture. They claim to improve the performance when less amount of labelled data is available called Few-shot learning. They represented the point cloud in the hierarchy of different radius balls

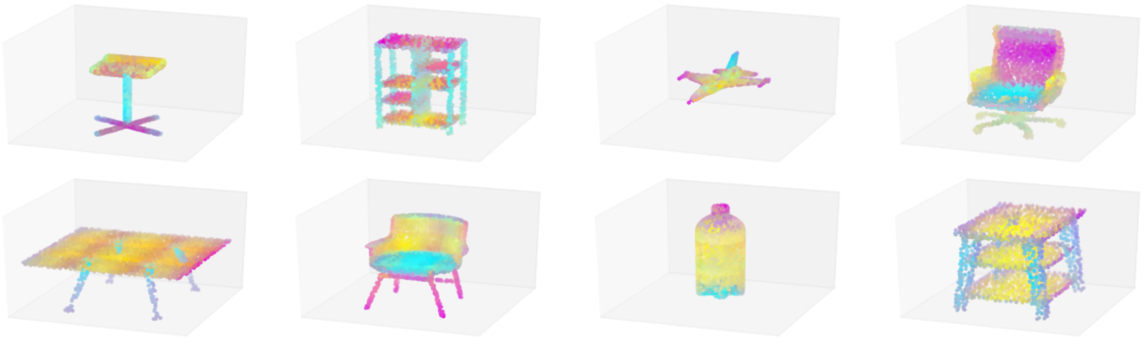

Fig. 2: Visualisation of Part Segmentation Results by DGCNN + pre-training by this SSL method

for each level. Starting from top-level with a single ball to lower levels with smaller radius balls covering the complete point cloud, as shown in Fig 3. This hierarchical structure is called cover tree T. It increases the input data complexity as learning is done at multiple levels and multiple scales, thus learning is more detailed. The idea is to generate proxy labels from this hierarchy.

The first SSL method is Regression Task (R), where the proxy label is the distance between the pair of the balls which belongs to the same level. It is designed to learn the overall inter ball spatial features. The second SSL method is Classification Task (C). It is designed to learn intra ball features that is between succeeding parent and child balls. The integer label is the quadrant of the centre of child ball with respect to the quadrant of the centre of the parent ball. These methods do not get affected by variation in density and sparsity of point cloud.

According to the paper [12], the idea of few-shot learning is n shot, and k way leaning, where n is the number of labelled sub-point clouds and k is the number of classes of labels. Y is the set of classes which labels can belong to. Upon sampling k classes from Y, further sample n sub-point clouds of each class. Thus, nk total number of the labelled point cloud is generated. Other parameters which define T are expansion constant (e) and ϵ called base of e, minimize e such that e balls of radius $1/\epsilon$ cover all the balls of the point cloud. Each level in the hierarchy is assigned an integer label i. ϵ^i, is the radius of each ball which covers all points in the point cloud, called covering. Their centres are set as individual nodes for that level in T. All balls centre (C) at level i, denoted as C_i are nodes and among those c_i, j denotes jth centre ball.

Fig. 3: Cover Tree Data Structure for two different levels to store Point Cloud in hierarchical manner

Label for R method, for all pairs of balls (except self pair) denoted by setting S^i. l_2 norm calculated between each pair is assigned as the label for that pair. Label for C method, for each parent-child pair (between C_i and C_{i-1}) denoted by set $S^{i,i-1}$ a label out of 1,2,3,4 is assigned.

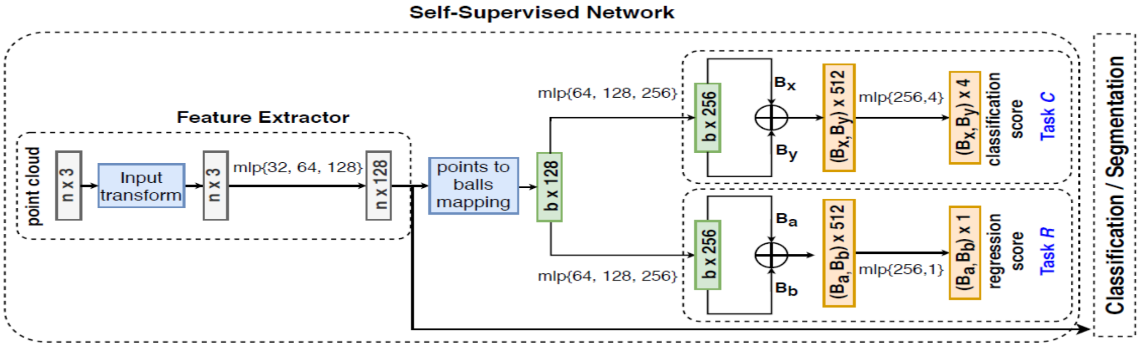

Fig. 4: Self - Supervised Network Architecture - trained by Classification and Regression task together. Feature Extractor provides Embedding which can be used for downstream task.

The Deep Neural Network model architecture (Fig 4) has SSL network, and output is provided for the downstream task of object classification or object segmentation. Any model capable of these tasks can be considered such as PointNet, DGCNN and more, which is initialised with point-embedding obtained from feature extractor. SSL Network's feature extractor is trained by combined loss from both tasks R and C. Firstly input is normalised then a feature vector is extracted by MLP layers. Ball in feature space is formed corresponding to each ball in the point cloud and passed to both R and C network. Each ball of 256 in a pair is joined together to form 512 vector and further passed for regressing the l_2 norm distance and for Classification. Some very interesting Part Segmentation experiments results are shown in Fig. 5.

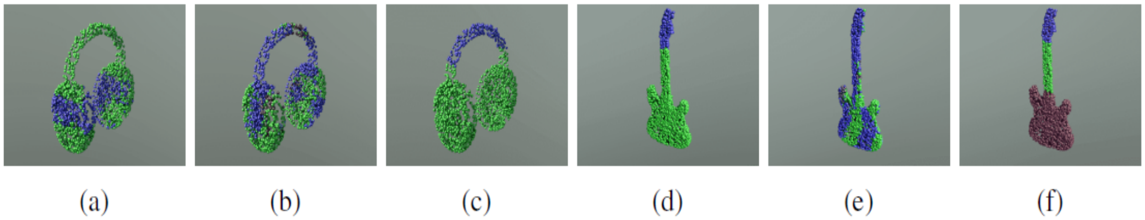

Fig. 5: Part Segmentation (a),(d) Random initialisation DGCNN result, (b),(e) VoxelSSL DGCNN result, (c),(f) this SSL method+DGCNN result

3.3 Deformation Reconstruction on Point Cloud

Idea/Method Achituve et al. [2], uses SSL technique for Domain Adaptation (DA). A new SSL method DefRec (Deformation Reconstruction) is devised by this paper's authors, getting inspired by the fact, scanning any object in real-world causes distortions and occlusions. They also suggest a new training technique called PCM (Point Cloud Mixup). They claim when both methods are combined together, the best results for DA have been achieved. DefRec is SSL task acting on Target (T) (unlabelled dataset) and Source (S) (labelled dataset) while PCM is Supervised task only applied on Source data. They introduced a new multi-head model architecture, where shared Encoder is used, and decoder has 2 heads one for Supervised task h_{sup} and other for SSL task h_{SSL} as shown in Fig 6 (left).

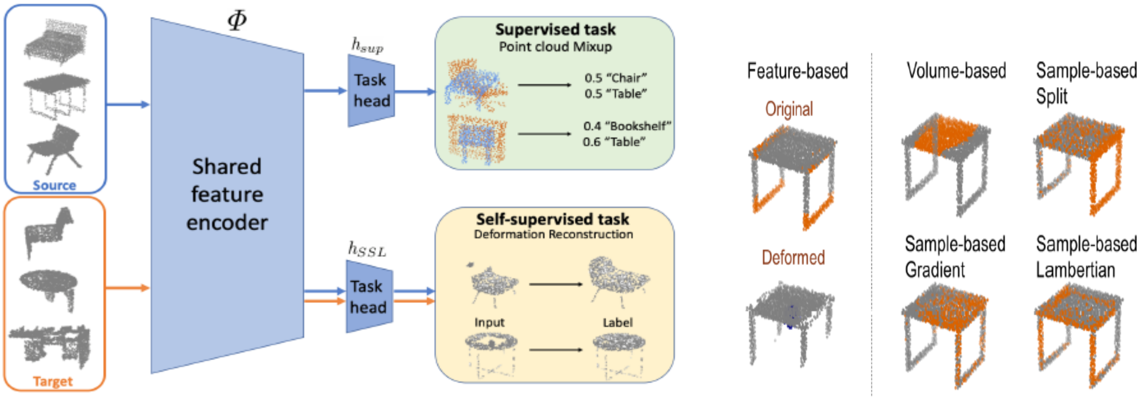

Fig. 6: Network Architecture - trained on Source (labelled dataset) and Target (unlabelled dataset) by using shared Encoder and two different heads for Supervised and Self Supervised task (left). Different deformation types (right).

In SSL task DefRec, the point cloud is deformed by removing some points from a region and dislocating them in some other part of the point cloud. The model is trained to reconstruct the original point cloud. Different types of deformation techniques are discussed in [2]. *Volume-based deformation* split input space on the basis of closeness. Either input space is broken into k^3 voxels and points of one voxel randomly chosen, or deformation is done on a region in the shape of a sphere of radius r choosing any random point as the centre. Those points are replaced by new sample points from Gaussian distribution and kept around the centre of that voxel/sphere. *Feature-based deformation* is done on the basis of semantics. Chose any random point (p) from the feature space vector extracted from some deep layer of the network, and replace (k) neighbours around (p) by putting new points at the centre of the point cloud. *Sample-based deformation* is done by sampling random points on the basis of three techniques *Split, Gradient and Lambertian.* Three deformations are shown in Fig 6 (right).

For training method PCM (head h_{sup}), two point clouds (x,x'), with labels (y,y'), are mixed up on the basis of mixup coefficient γ sampled from Beta distribution. $\gamma.n$ points are selected from x and $(1-\gamma).n$ points from x'. The new mixed up point cloud is formed $\bar{x}$ with convex label $\bar{y} = \gamma y + (1-\gamma)y'$.

For loss (L_{SSL}) calculation by h_{SSL}, Chamfer Distance is used. After Encoder ϕ on deformed input point cloud $\hat{x}$, the output is $\phi(\hat{x})$, which is further fed to h_{SSL} giving final output as $h_{SSL}(\phi(\hat{x}))$. Loss is calculated between $\hat{x}$ and $h_{SSL}(\phi(\hat{x}))$. For loss (L_{sup}) calculation via h_{sup} head, $\phi(\bar{x})$ is passes to it and cross entropy loss is calculated between $\bar{y}$ and $h_{sup}(\phi(\bar{x}))$. Total loss is $L_{SSL} + \lambda L_{SSL}$ if both methods are used. For DA the aim is to bridge the gap between the Source and Target data distribution.

The authors of [2] used DGCNN [16] for feature extraction and h_{sup} with similar configuration. For feature extractor (64,64,128,256) convolution layers with one fully connected (FC) layer of 1024 size provides global features. h_{sup} was designed with 3 FC layers of size (512, 256, 10), where 10 is the total number of classes to be classified. For h_{SSL} head, (256, 256, 128, 3) 1D convolution layers were used.

3.4 Orientation Estimation of Point Cloud

Idea/Method Poursaeed et al. [9], suggests a new SSL method. Every input is rotated. The Classifier is trained to classify rotation angle, out of K total rotation angles, to bring back point

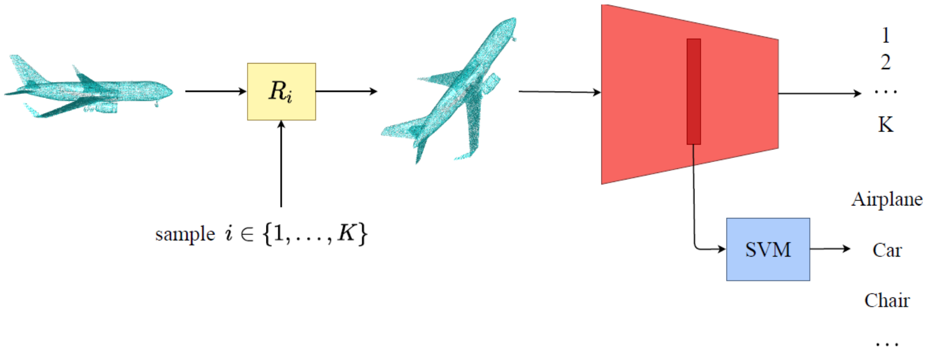

Fig. 7: Model Architecture - trained to predict rotation angle of input point cloud. Feature embedding extracted from model are used for shape classification task.

cloud to its authoritative orientation. Apart from Classification, the paper [9] suggests that pretext task can be solved as a Regression problem and train model to regress the value of rotation, with respect to original orientation, by outputting 4 values (3 rotation axis and 1 rotation angle). The learned features are important for solving the downstream task of keypoint prediction and object classification even by simple models such as linear SVM.

For the classification task, K values are uniformly chosen by mathematical concepts of a regular polyhedron with each vertex provides an angle of rotation with respect to the standard coordinate system. For K = 6, possible rotation angle values are obtained directly from both positive and negative axis of x,y,z coordinate axis of a regular hexagon. For k = 18, consider 6 above angles and 12 angle bisectors for both positive and negative axis. For k = 32, 12 vertices and rest other face centres of the regular icosahedron is considered. For k = 54, $(1+\sqrt{5}/2)$ is chosen as the golden ratio and used to find rotation angle by golden spiral distributed uniformly, shown in Fig 8.

For Regression task, 6D representation of rotation. PointNet model's MLP is modified to produce 6 dim rotation result which is further converted to a special orthogonal group of 3-dimensional rotation by a mapping function according to paper [18].

PointNet and DGCNN models are used with the last layer modified to output according to Classification (k probabilities) or Regression task (4 values). For Classification task Cross-Entropy loss was used for back-propagation while in Regression L_2 loss is calculated by equal weightage to axis and angle. The features extracted from pretext task are used to train a simple linear SVM model on keypoint prediction and shape classification downstream task. The model layout is shown in Fig 7.

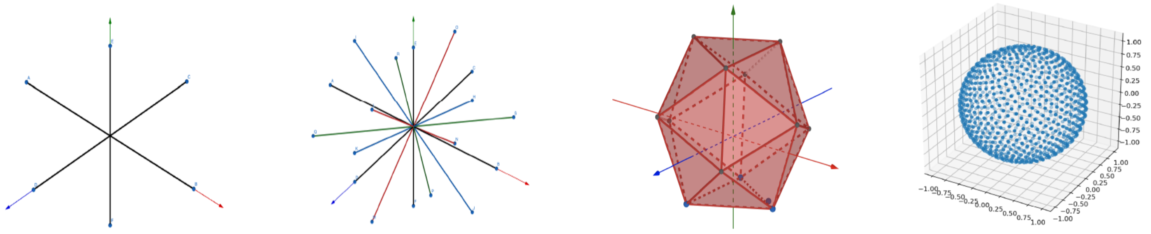

Fig. 8: (1) K = 6, (2) K = 18, (3) K = 32, (4) K = 54 Rotation Angles

4 Conclusion

After reviewing different SSL techniques, SSL provides promising results. It can be used to improve the performance of SOTA models. SSL can address the problem of the requirement of the huge amount of labelled dataset. A large amount of unlabelled dataset can be utilised for Machine Learning tasks. Different SSL methods focusing on learning different features in the dataset, there learning can be combined together and can be utilised in Transfer Learning. In general, SSL allows models to learn the generic features of the dataset. This technique can be utilised in many application areas like sim-to-real scenarios, Domain Adaptation, Transfer Learning and more. Even on the unstructured point cloud, dataset SSL provides promising results, thus on structured data SSL results are expected to be far better, which is altogether different data domain. SSL in general address the problem of Deep Learning. Next revolution in AI is unpredictable, but SSL is one of the potential fields by Yann LeCun in [1].

References

1. Aaai 20 / aaai 2020 keynotes turing award winners event / geoff hinton, yann le cunn, yoshua bengio, Feb 2020.
2. Idan Achituve, Haggai Maron, and Gal Chechik. Self-supervised learning for domain adaptation on point-clouds. *arXiv preprint arXiv:2003.12641*, 2020.
3. month=Mar Ben Dickson, year=2020. Self-supervised learning: The plan to make deep learning data-efficient.
4. Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
5. Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 984–993, 2018.
6. Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in neural information processing systems*, pages 820–830, 2018.
7. Nbro and David. What is self-supervised learning in machine learning?, Mar 1968.
8. Benedict Neo. The future of machine learning, Dec 2019.
9. Omid Poursaeed, Tianxing Jiang, Quintessa Qiao, Nayun Xu, and Vladimir G Kim. Self-supervised learning of point clouds via orientation estimation. *arXiv preprint arXiv:2008.00305*, 2020.
10. Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.
11. Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *Advances in Neural Information Processing Systems*, pages 12962–12972, 2019.
12. Charu Sharma and Manohar Kaul. Self-supervised few-shot learning on point clouds. *arXiv preprint arXiv:2009.14168*, 2020.
13. Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2530–2539, 2018.
14. Lulu Tang, Ke Chen, Chaozheng Wu, Yu Hong, Kui Jia, and Zhixin Yang. Improving semantic analysis on point clouds via auxiliary supervision of local geometric priors. *arXiv preprint arXiv:2001.04803*, 2020.
15. Ali Thabet, Humam Alwassel, and Bernard Ghanem. Mortonnet: Self-supervised learning of local features in 3d point clouds. *arXiv preprint arXiv:1904.00230*, 2019.
16. Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.
17. Lilian Weng. Self-supervised representation learning, Nov 2019.
18. Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.