Self-Supervised Learning for Point Clouds Data: A Survey

Changyu Zeng[a,b,c], Wei Wang[b], Anh Nguyen[c] and Yutao Yue[a,b,c,*]

[a]Institute of Deep Perception Technology, JITRI, Wuxi, 214000, China
[b]Department of Computing, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
[c]Department of Computer Science, University of Liverpool, Liverpool, L69 7ZX, United Kingdom

ARTICLE INFO

Keywords:
Self-Supervised Learning
Computer Vision
Point Clouds
Representation Learning
Pretext Task
Transfer Learning

ABSTRACT

3D point clouds are a crucial type of data collected by LiDAR sensors and widely used in transportation applications due to its concise descriptions and accurate localization. Deep neural networks (DNNs) have achieved remarkable success in processing large amount of disordered and sparse 3D point clouds, especially in various computer vision tasks, such as pedestrian detection and vehicle recognition. Among all the learning paradigms, Self-Supervised Learning (SSL), an unsupervised training paradigm that mines effective information from the data itself, is considered as an essential solution to solve the time-consuming and labor-intensive data labelling problems via smart pre-training task design. This paper provides a comprehensive survey of recent advances on SSL for point clouds. We first present an innovative taxonomy, categorizing the existing SSL methods into four broad categories based on the pretexts' characteristics. Under each category, we then further categorize the methods into more fine-grained groups and summarize the strength and limitations of the representative methods. We also compare the performance of the notable SSL methods in literature on multiple downstream tasks on benchmark datasets both quantitatively and qualitatively. Finally, we propose a number of future research directions based on the identified limitations of existing SSL research on point clouds.

1. Introduction

With the rapid development of 3D data processing technologies, an increasing number of relevant applications have emerged in both industrial and daily usage, such as indoor navigation (El-Sheimy and Li, 2021), autonomous driving (Li, Ma, Zhong, Liu, Cao, Li and Chapman, 2020), and object modeling (Yang, Liu, Hu, Wang and Lin, 2019). LiDAR is one of the indispensable types of sensors to capture disordered 3D point cloud data from traffic scenes, which has enabled more challenging tasks like pedestrian detection (Matti, Ekenel and Thiran, 2017) and road semantic segmentation (Wu, Zhou, Zhao, Yue and Keutzer, 2019) based on the strong inference ability of deep neural networks (DNNs).

However, several well known problems in the supervised point cloud DNNs hinder their further development and practical uses. For example, accurate environment perception via DNNs requires millions of labeled data as the input, while point cloud annotating is labor-intensive and time-consuming due to its disordered and sparse nature (Dai, Chang, Savva, Halber, Funkhouser and Nießner, 2017). Besides, manual labeling by human experts or users inevitably leads to mistakes such as mislabeling and omission. Another long-standing problem is that the supervised learning paradigm struggles to capture the underlying patterns of new data and fails to generalize the pre-training model to downstream tasks because of overfitting caused by noisy labels (Sariyildiz, Kalantidis, Alahari and Larlus, 2022).

The aforementioned issues motivate research in extracting effective feature representations from point clouds via Self-Supervised Learning (SSL) to learn implicit while better representations without manual annotations. Not only does it solve the problem of the error-prone and expensive labeling process, but also relieve the domain adaptation (DA) issues (Csurka, 2017) with improved model generalization ability. Under the SSL paradigm, basic geometric as well as advanced semantic information can be extracted

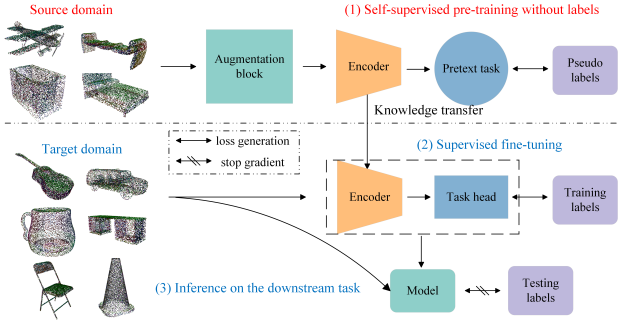

Figure 1: The general pipeline of SSL used in the point cloud data. (1) Pre-training stage: point cloud data is firstly pre-processed through the augmentation block and then fed into the point-specific encoder to learn feature representations. The features are utilized to complete well-design pretext tasks, where the output will be compared with the pseudo labels derived from the original data to generate a loss and to update encoder parameters via back-propagation; (2) Supervised fine-tuning stage: the well-trained encoder is transferred to the target domain. A task head is trained with the training labels in a supervised manner to complete the downstream tasks; (3) Inference stage: the encoder and task head are concatenated as a model to execute inference on the test set. The effectiveness of the SSL pre-training framework can be evaluated based on the performance of the model on the downstream tasks.

*Corresponding author
yueyutao@idpt.org (Y. Yue)
ORCID(s): 0000-0003-4532-0924 (Y. Yue)

as knowledge and migrated to downstream tasks under the transfer learning setup. This process approximates human learning that discovers objective principles of the world by observing phenomena and summarizing them into a system of experience and knowledge.

Fig. 1 shows a general pipeline of SSL on point cloud data. The goal of SSL is to pre-train an encoder on an unlabeled, large-scale point cloud dataset (source domain), and to transfer the well-trained network to other datasets (target domain) in various downstream tasks. A complete SSL framework usually contains the following important modules.

- **Data augmentation**: The raw input is augmented via some easy-to-implement pre-processing operations such as translation, rotation, flip, and adding noise (Zhang, Lin, Li, Jia and Zhang, 2022c). The objective is to expand the size and diversity of the raw data and to provide subjects for subsequent pretext tasks. The details will be discussed in Section 2.4.

- **Encoder**: The encoder is a point-specific deep network that captures the hierarchical representation of the input point cloud data. We will introduce some commonly used point cloud encoders that learn either from downsampling layer-by-layer (Qi, Su, Mo and Guibas, 2017a; Qi, Yi, Su and Guibas, 2017b) or local areas to capture the association between different blocks (Zhou and Tuzel, 2018; Wang, Sun, Liu, Sarma, Bronstein and Solomon, 2019). The details will be discussed in Section 2.5.

- **Pretext task**: At the core of the framework is the design of a pretext task that mines the hidden self-supervision signal via the interactions between the encoder and data. This part is also the focus of the survey and will be discussed in detail in Section 3.

- **Knowledge transfer**: The well-trained encoder will be transferred to another dataset with the knowledge gained in the source domain after completing the pretext task. A task head is constructed and trained by a small amount of labelled data in the target domain as the supervision signals to fine-tune the whole architecture. The details will be discussed in Section 4.

- **Downstream task**: To evaluate the effectiveness of the SSL framework, the pre-trained encoder will be transferred and evaluated on another dataset for performance evaluation, e.g. object classification, part segmentation, and object detection. The details will be discussed in Section 4.

Thriving progress has been made on point cloud SSL recently, and new models, algorithms and benchmark datasets are emerging quickly and continuously. A systematic review on this exciting topic, especially the research published in the past three years, is urgently needed. In our study, we find that the survey in (Xiao, Huang, Guan and Lu, 2022) employed a similar methodology but focused on unsupervised representation learning. However, it lacked a review on the state-of-the-art SSL models, and in particular, a detailed demonstration of most recent published works. Therefore, we are motivated to provide a comprehensive review on the recently published, representative research on point cloud SSL. Our contributions can be summarized as follows:

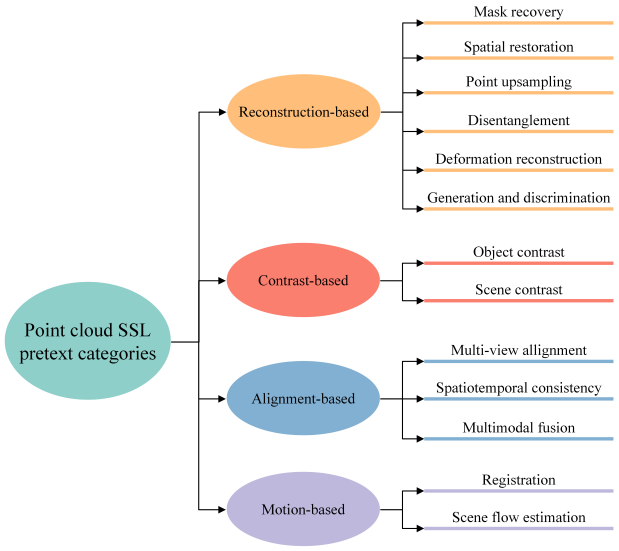

Figure 2: Taxonomy of SSL for point cloud data based on pretext tasks.

- **Systematic and novel taxonomy:** We propose a novel and systematic taxonomy for categorizing the diverse kinds of point cloud SSL methods to provide a clear and holistic view on the state of the art. Taking into consideration the characteristics of popular pretext tasks, the taxonomy groups current methods into four broad categories. Each broad category is further subdivided into more fine-grained sub-categories according to the methods in feature utilization as shown in Fig. 2.

- **Comprehensive and detailed summary:** We conduct a comprehensive review of the state of the art, including the background of SSL and point clouds, commonly used datasets and models, pretext tasks, and downstream tasks with performance comparison.

- **Exhaustive dataset summary and evaluation comparison:** We summarize the unique characteristics of 18 most frequently utilized datasets in the point cloud research. More importantly, we compare the performance of different SSL methods on these datasets according to various downstream tasks.

- **Future directions:** Based on our investigation, we summarize and discuss the major limitations and challenges in the current research and propose potential future directions which would hopefully motivate more

theoretical and practical research towards more intelligent and effective SSL approaches for point cloud data processing.

The rest of the paper is organized as follows: Section 2 introduces the preliminaries for this survey to equip the readers with the necessary background knowledge on SSL and point cloud data. Section 3 represents the main body of the study and provides an exhaustive and detailed analysis on the state of the arts methods according to the structure of the proposed taxonomy. Section 4 illustrates an evaluation and comparison study on the performance of different SSL methods on the frequently utilized downstream tasks and benchmark datasets. Section 5 discusses the limitations and challenges of current research and proposes potential future directions, and Section 6 concludes the paper.

2. Background

2.1. Self-supervised learning in the language and image domain

We firstly describe the development history of SSL in the language and image domains. The purpose is to provide readers a general understanding on SSL. Although data types vary from domains to domains, the core idea of SSL remains the same: to leverage data characteristics for transformation processing and to make the transformed data consistent with the original input in terms of feature representation by contrasting or reconstruction.

The idea of SSL was firstly introduced in Natural Language Processing (NLP) research. After converting words into vectors, e.g. Word2Vec (Mikolov, Chen, Corrado and Dean, 2013), and utilizing the relationships between the representations and context, models could learn semantic representations from neighboring words or sentences through pretext task formulations such as next sentence prediction (Devlin, Chang, Lee and Toutanova, 2018), auto-regressive language modeling (Floridi and Chiriatti, 2020), or sentence permutation (Lewis, Liu, Goyal, Ghazvininejad, Mohamed, Levy, Stoyanov and Zettlemoyer, 2019). Landmark models such as GPT (Floridi and Chiriatti, 2020) and BERT (Devlin et al., 2018), and many variants celebrate great achievements in not only NLP but also other fields later.

In the field of image processing and computer vision, different SSL methods impose simple variations on image data and extract features by recovering it to the original input, for example, from simple tasks like relative position prediction (Doersch, Gupta and Efros, 2015; Noroozi and Favaro, 2016) and rotation angle prediction (Gidaris, Singh and Komodakis, 2018), to reconstructing blocks masked by surrounding visible pictures (Pathak, Krähenbühl, Donahue, Darrell and Efros, 2016; He, Chen, Xie, Li, Dollár and Girshick, 2021). Free semantic label-based (Faktor and Irani, 2014; Stretcu and Leordeanu, 2015; Croitoru, Bogolin and Leordeanu, 2017; Jiang, Larsson, Shakhnarovich and Learned-Miller, 2018) and cross-modal-based methods (Arandjelovic and Zisserman, 2017; Agrawal, Carreira and Malik, 2015; Jayaraman and Grauman, 2015) have been proposed, which learn representations via automatically generated semantic labels and extra information from other modalities. Recently, the research community shows a great interest on contrastive learning (Chen, Kornblith, Norouzi and Hinton, 2020; He, Fan, Wu, Xie and Girshick, 2020; Caron, Misra, Mairal, Goyal, Bojanowski and Joulin, 2020), which aims to differentiate positive and negative samples by comparison using data augmentation techniques. These research works inspired the study of SSL on point clouds, with similar ideas transferred from 2D to 3D by adapting for data peculiarities.

2.2. Properties of the point cloud data

Data properties are distinct between naturals languages, images, and point clouds. Languages are usually complex and abstract in nature, and contain ambiguous information due to its versatility and richness. It is expressed in a sequence of words, which is discrete and unstructured in the representation space (He et al., 2020). In contrast, images contain rich visual information, such as color, texture, and shape information of an object in high-dimensional space (Jing and Tian, 2020) for human perception. They are usually represented as 2D data by using a matrix of pixel values.

Simply speaking, point cloud data is similar to image data in terms of visual format and can be regarded as 3D stereo images with depth information. However, the attributes of point cloud data are completely different in geometric representation. Specifically, a point cloud is a collection of discrete, disordered, and topology-free 3D points. The most basic information contained in the points is the position coordinates (x_i, y_i, z_i) in the Euclidean space, where i is the number of points in the object. There are also other optional attributes such as color, intensity, reflectivity, etc., specifying physical properties of the points in more detail. The input order is trivial for point cloud data and does not impact the semantic meaning while it is crucial for images and language where various words or pixel sequences lead to completely divergent connotations. Additionally, point cloud data is invariant to rigid transformation, which means that it remains unchanged after rotation and translation. Some of such exclusive properties can be summarized as follows:

- **Sparsity:** The point cloud data is discretely distributed on the surface of the scanned object or scene.
- **Non-uniformity:** The distance between points is not fixed and is determined by various factors such as the instruments' sampling strategy, relative position, and scanning range.
- **Imcomplete data:** Some parts of real-scanned surfaces are incomplete due to self or external occlusion.
- **Noise:** It is inevitable that noise from environmental factors or inaccuracies in instruments will be present.
- **Permutation invariance:** The order of points does not affect the overall semantic representation of point cloud objects, so identical point cloud objects can be expressed by various matrices.

Table 1
Summary of commonly used point cloud datasets. Abbreviations for suitable tasks: Cls (Classification); Seg (Semantic Segmentation); Det (Object Detection); Com (Semantic Scene Completion); Rec (Surface Reconstruction); CM (Cross-Modal tasks); Pos (Pose estimation); Tra (Object Tracking)

Year	Name	#Samples	#Categories	Types	Suitable tasks	Highlights
2012	KITTI (Geiger, Lenz and Urtasun, 2012)	Over 200K objects	8	RGB	Cls/Det/CM	Comprehensive outdoor driving dataset
2015	ModelNet (Wu, Song, Khosla, Yu, Zhang, Tang and Xiao, 2015)	12,311 models	40	CAD	Cls/Seg/Rec	Frequently used in classification and few-shot
2015	ShapeNet (Chang, Funkhouser, Guibas, Hanrahan, Huang, Li, Savarese, Savva, Song, Su et al., 2015)	57,448 models	55	CAD	Cls/Seg	Commonly employed as the pre-training dataset
2015	SUN RGBD (Song, Lichtenberg and Xiao, 2015)	10,335 images	37	RGB-D	Seg/Det/Pos	A RGB-D scene understanding benchmark suite
2016	SceneNN (Hua, Pham, Nguyen, Tran, Yu and Yeung, 2016)	100 scenes	-	RGB-D	Det/Rec/Pos	Using unique triangle meshes shape contour
2016	ObjectNet3D (Xiang, Kim, Chen, Ji, Choy, Su, Mottaghi, Guibas and Savarese, 2016)	44,147 shapes	100	CAD	Det/CM/Pos	Well-aligned 2D-3D dataset
2016	S3DIS (Armeni, Sener, Zamir, Jiang, Brilakis, Fischer and Savarese, 2016)	272 scans	13	Point	Cls/Seg/CM	Large-scale indoor space scanning dataset
2017	ScanNet (Dai et al., 2017)	1513 scenes	20	RGB-D	Cls/Seg/Com	Rich labels for scene understanding tasks
2017	Semantic3D.net (Hackel, Savinov, Ladicky, Wegner, Schindler and Pollefeys, 2017)	Over 4B points	8	Point	Cls/Seg/Det	High quality resolution and scope outdoor dataset
2018	Pix3d (Sun, Wu, Zhang, Zhang, Zhang, Xue, Tenenbaum and Freeman, 2018)	10,069 3D-2D pairs	9	CAD	Rec/CM/Pos	Pixel-level image-shape pairs dataset
2019	ABC (Koch, Matveev, Jiang, Williams, Artemov, Burnaev, Alexa, Zorin and Panozzo, 2019)	1M objects	-	CAD	Seg/Rec	Providing a benchmark for surface normal estimation
2019	ScanObjectNN (Uy, Pham, Hua, Nguyen and Yeung, 2019)	2,902 objects	15	Point	Cls/Seg	Challenging real-world scenario with noise
2019	PartNet (Mo, Zhu, Chang, Yi, Tripathi, Guibas and Su, 2019)	26,671 models	24	Points	Seg/Rec	Producing fine-grained multi-level 3D part objects
2020	RobustPointSet (Taghanaki, Luo, Zhang, Wang, Jayaraman and Jatavallabhula, 2020)	73,843	40	Mesh	Cls	Benchmark to evaluate the robustness of classifiers
2020	Waymo (Sun, Kretzschmar, Dotiwalla, Chouard, Patnaik, Tsui, Guo, Zhou, Chai, Caine et al., 2020)	12M objects	-	Point	Det/CM/Tra	Suitable for cross-modal and transfer learning
2020	NuScenes (Caesar, Bankiti, Lang, Vora, Liong, Xu, Krishnan, Pan, Baldan and Beijbom, 2020)	1K scenes	23	Point	Det/CM/Tra	Containing additional annotations and scenes
2021	SensatUrban (Hu, Yang, Khalid, Xiao, Trigoni and Markham, 2021)	4B points	13	Point	Seg	Data collected by UAV over UK landscape
2022	STPLS3D (Chen, Hu, Hugues, Feng, Hou, McCullough and Soibelman, 2022)	62 scenes	18	Point	Seg/Det	Covering both real and synthetic aerial point clouds

- **Transformation immutability:** Point clouds remain immutable through rigid transformations such as rotation and translation.

- **Points interaction:** There are correlations, either strong or weak, between points in global and local regions.

2.3. Point Cloud Dataset

Quality benchmark datasets (e.g. complete, well-varied, and densely-labeled) play essential roles in SSL research. This section lists the most commonly used point cloud datasets and summarize them in Table 1 in terms of sample number, object categories, suitable tasks, and highlights. These datasets may contain synthetic and real scanned data, in single frames and time series and from individual objects and complex scenes. There are also a few datasets for complex traffic scenarios (e.g. automatic driving) containing extra data in different modalities, such as from images or radars.

- **KITTI** (Geiger et al., 2012) is a benchmark suite for autonomous driving vision tasks. The dataset was collected using several pieces of equipment, including four video cameras, a laser scanner, and a localization system. It includes not only point clouds but also stereo and optical flow data. There are more than 200,000 annotated point cloud scenarios consisting of cars and pedestrians, providing a novel and challenging benchmark for 3D object detection and orientation estimation.

- **ModelNet** (Wu et al., 2015) is the most widely used 3D point cloud CAD dataset for object classification and few-shot learning. It contains 12,311 single objects from 40 categories, with each point composed of six dimensions of information, including XYZ spatial coordinates and RGB values.

- **ShapeNet** (Chang et al., 2015) is a relatively large-scale repository of 3D CAD objects frequently employed as a pre-training dataset. It contains more than 3 million samples categorized into 55 classes under the WordNet synsets (Miller, 1995) criteria. The annotations in the dataset are versatile, including rigid alignments, parts, physical sizes, and key points.

- **SUN RGBD** (Song et al., 2015) is an RGB-D scene understanding benchmark suite containing 10,335 samples at a comparable scale to PASCAL VOC (Everingham, Van Gool, Williams, Winn and Zisserman). It has 146,617 2D polygons and 64,595 3D bounding boxes densely annotated to indicate object orientation, room layout, as well as scene category for overall scene awareness.

- **S3DIS** (Armeni et al., 2016) is a 3D indoor venue dataset that consists of scanning of 272 rooms in 6 areas overlaying a 6,000 m^2 area. It has 13 semantic categories labeled by fine-grained point-wise annotations carrying full 9D information, including XYZ, RGBs, and normalized location coordinates.

- **ScanNet** (Dai et al., 2017) is a 3D RGB-D dataset that comprises 2.5M views in 1,513 scenes acquired in 707 indoor environments. Various tests containing semantic voxel labeling and CAD model retrieval proved that ScanNet provides quality data for 3D scene understanding.

- **ScanObjectNN** (Uy et al., 2019) was proposed as a collection of real-world indoor point cloud scenes to break the performance saturation of 3D object classification on synthetic data. This dataset introduces new challenges for 3D object classification due to the presence of background noise and occlusions that require networks' ability on context-based reconstructions and partial observations.

- **Waymo** (Sun et al., 2020) is a large autonomous driving dataset produced by Waymo in collaboration with Google Inc. The dataset consists of 1,150 urban and suburban geography scenes spanning 20 seconds, which are collected via well-synchronized and calibrated LiDARs and cameras.

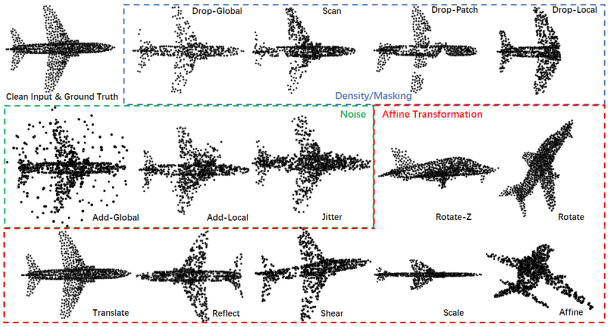

Figure 3: Illustration of the commonly used data augmentation methods for point cloud data. There are a total of 14 sub-categories of data augmentation methods that could be classified as three general corruption families. The figure is adapted from (Zhang et al., 2022c).

- **NuScenes** (Caesar et al., 2020) is another remarkable multimodal dataset provided by the full sensor suite including cameras, radars, and LiDARs. Compared to other autonomous driving datasets, it contains additional annotations like pedestrian pose, vehicle state, and also scenes from nighttime and rainy weather.

2.4. Point cloud data augmentation

Data augmentation is a crucial technique for enhancing DNNs performance by increasing the amount and diversity of training samples. For SSL tasks, it not only prevents the model from overfitting but also facilitates capturing robust and invariant representations of point clouds under multiple transformations. In this section, we will introduce the commonly used data augmentation methods and compare the effectiveness of each methods via a metric called task relatedness.

Essentially, data augmentation is a process of generating new data by adding interventions or corruptions without destroying the original semantic expressions. For point clouds, augmentation methods are based on the properties mentioned in Section 2.2 and can be classified into three general groups: density/masking, noise, and affine transformation (Zhang et al., 2022c). These three corruption families could be further divided into 14 sub-categories as shown in Fig. 3.

Density/masking is the most frequent data augmentation method adopted in mask autoencoder (MAE) type SSL research (He et al., 2021; Pang, Wang, Tay, Liu, Tian and Yuan, 2022; Yu, Tang, Rao, Huang, Zhou and Lu, 2021). Based on the principle that point cloud data is sparse with uneven density, randomly removing a certain percentage of points while preserving part of the semantic expression presents a challenging learning objective for such MAE-based tasks. On the contrary, the noise based methods impose interventions on the original clean input to increase the difficulty of feature extraction. Affine transformation leverages point cloud invariance characteristics to shift the spatial coordinates of each points. This has significant impact on the input since the basic position information completely changes.

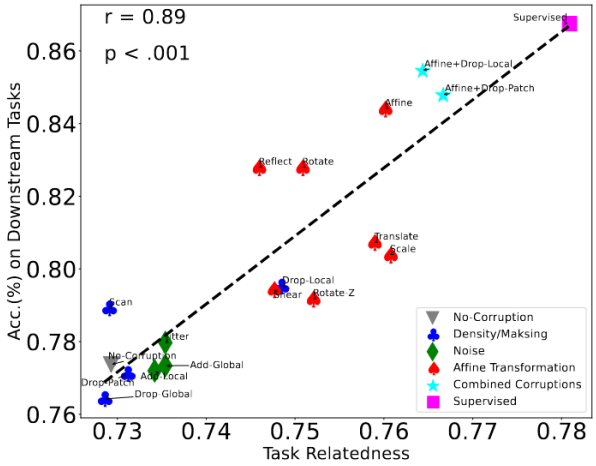

Figure 4: Illustration of the relationship between task relatedness and classification accuracy on downstream tasks. r and p are the coefficients to measure the linear relationship and statistical significance for the Pearson correlation, respectively. The figure is adapted from (Zhang et al., 2022c).

The work in (Chen et al., 2020; Zhang et al., 2022c) investigated the effectiveness of the aforementioned augmentation methods as pretext data preprocessing on downstream classification tasks. Task relatedness is employed as the evaluation metric to statistically measure the performance of SSL models on downstream tasks, which provides valuable advice for proxy data augmentation selection. Following (Zamir, Sax, Shen, Guibas, Malik and Savarese, 2018), for each pretext task c, its task relatedness to downstream task t is defined as:

$$A_{c\to t} := \mathbb{E}_{x\in X}\mathcal{I}_t(R_c(E_c(x)), f_t(x)) \tag{1}$$

Where x is a sample in a point cloud dataset X; E_c is the model's encoder pre-trained on task c; R_c is a readout function, which indicates the classification head composed of several fully connected (FC) layers; f_t is the labeling function; $\mathcal{I}_t$ is accuracy measurement estimating whether the downstream output $R_c(E_c(x))$ conforms to the ground truth $f_t(x)$.

To further explore the relationship between task relatedness and classification accuracy on downstream tasks, Pearson correlation coefficient r and p-value are utilized to estimate the linear relationship as well as statistical significance (Fraser, 1976), respectively, where $|r| > 0.5$ refers to a strong correlation and $p < 0.05$ is considered statistically significant. Fig. 4 demonstrates the statistically significant linear relationship between task relatedness and classification accuracy on downstream tasks when $r = 0.89$ and $p < 0.001$. The results reveal a counter-intuitive fact that frequently used density/mask and noise-based data augmentation methods are ineffective for downstream tasks either in accuracy and task relatedness. Conversely, the seemingly simple affine transformation enhances task relatedness to point cloud classification, resulting in higher accuracy.

Table 2
Commonly used deep networks for extracting point cloud features.

Model	Year	Architecture	Contributions
PointNet (Qi et al., 2017a)	2017	CNN	Pioneer in direct processing of raw point clouds with lightweight architecture
PointNet++ (Qi et al., 2017b)	2017	CNN	Aggregating local neighborhood by multi-scale and multi-resolution sampling and groping
VoxelNet (Zhou and Tuzel, 2018)	2018	3D CNN	Partitioning disordered point clouds into regular voxels for local feature learning
DGCNN (Wang et al., 2019)	2019	Graph CNN	Constructing a dynamic local graph to capture edge features around a neighbor
PCT (Guo, Cai, Liu, Mu, Martin and Hu, 2021)	2021	Transformer	Successfully capturing the long-range dependencies between point patches
GANs (Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio, 2014)	2014	GAN	Generating synthetic data through adversarial training

Furthermore, combining corruptions of affine transformation and mask can approach the performance of supervised benchmarks. Hence, using affine transformation-based methods for data augmentation is preferable for in SSL pretraining.

2.5. Popular deep models for point clouds

SSL techniques designed for languages and images need to be revised and extended for point clouds. For instance, traditional CNN networks cannot handle irregular and discrete point cloud data well since there is no guarantee that a corresponding point exists at the same relative position of the convolution. In this section, we briefly introduce five point cloud networks that are frequently used as feature extraction encoders in the literature and summarize their respective characteristics in Table 2.

2.5.1. PointNet

To reduce data size and computation complexity, Qi et al. proposed PointNet (Qi et al., 2017a), which is the pioneering work to extract features directly on raw point clouds. It is widely deployed as the feature extractor (Wang, Liu, Yue, Lasenby and Kusner, 2021b; Poursaeed, Jiang, Qiao, Xu and Kim, 2020; Sauder and Sievers, 2019b) due to its simple and lightweight network structure. Taking advantage of the point permutation invariance, PointNet aligns the input points to a canonical space and aggregates global features by symmetric functions such as max pooling.

However, it fails to capture local structures induced by the metric space in which the points reside, thereby limiting its ability to recognize fine-grained patterns and generalize to complex scenes. The updated version PointNet++ (Qi et al., 2017b) was then put forward several months later. It adopts multi-scale, multi-resolution sampling, and groping strategies to propagate features from one level to another, which improves the feature learning ability further. Furthermore, the point patch generation strategy combining Farthest Point Sampling (FPS) and K-Nearest Neighbor (KNN) provides a template for point cloud cropping preprocessing for subsequent studies (Yu et al., 2021; Pang et al., 2022; Zhang, Lin, He, Chen, Jia and Zhang, 2022b).

2.5.2. VoxelNet

VoxelNet (Zhou and Tuzel, 2018) is a generic point-specific network that uses voxels (i.e. finite unit cubes), to divide and access a local representation of point clouds for 3D detection tasks (Li, Yu, Meng, Caine, Ngiam, Peng, Shen, Lu, Zhou, Le et al., 2022; Min, Zhao, Xiao, Nie and Dai, 2022; Hess, Jaxing, Svensson, Hagerman, Petersson and Svensson, 2022). This network partitions disordered point clouds and performs feature learning in quantified and fixed-size 3D structures. One innovation is the stacking Voxel Feature Encoding (VFE) layers which encode interaction between points within a voxel and grasp descriptive appearance information. The output of each VFE layer is the concatenation of point-wise features and locally aggregated features so that local features are better captured. However, the expensive computation of voxel construction and quantization artifacts constrain the model from capturing high-resolution or fine-grained representations.

2.5.3. DGCNN

A point with its neighbors can reflect the geometry property of a local point cloud. Such a local relationship could be expressed by a graph network. Therefore, Wang et al. proposed a dynamic graph-based CNN network (DGCNN) (Wang et al., 2019) that encodes the edge features between vertices. Instead of learning point representations directly, DGCNN represents the interactions between points and their edges in both Euclidean and semantic space, and learns the graph structure dynamically. This graph network-based architecture has served as a backbone in many subsequent point cloud SSL models with notable results (Poursaeed et al., 2020; Sauder and Sievers, 2019b; Afham, Dissanayake, Dissanayake, Dharmasiri, Thilakarathna and Rodrigo, 2022).

2.5.4. GAN

Generative Adversarial Network (GAN) (Goodfellow et al., 2014) is a widely used framework in reconstruction-based pretext tasks for point cloud knowledge mining. It consists of two components: the generator, which generates point clouds similar to the training data, and the discriminator, which distinguishes between generated and real points. These two modules are trained under an adversarial paradigm without supervision. The framework can be formulated as a two-player minimax game:

$$\min_G \max_D E_{x\in X}[log(D(x))] + E_{z\in Z}[log(1-D(G(z)))] \quad (2)$$

where D and G denotes the discriminator and the generator, and X and Z represent the data and noise distribution, respectively.

2.5.5. Transformers

Transformers have become one of the most prevalent architectures in many fields. They benefit from the multi-head self-attention mechanism, which allows them to capture long-range dependencies between point patches and

discover implicit regional correlations. The state-of-the-art performance on SSL point cloud classification and part segmentation has been achieved by transformer-based models such as the one proposed by Zhang et al. (Zhang et al., 2022b). Furthermore, point cloud transformer (PCT) (Guo et al., 2021), a variant adapted specifically for point clouds, enhances local feature extraction with the support of farther point sampling and nearest neighbor search, and further improves performance on various downstream tasks.

2.6. Pseudo labels

Pseudo labels are introduced in point cloud SSL due to the absence of ground truth labels. It facilitates the calculation of loss with the output of the pretext tasks, which is then used for updating encoders via backpropagation. Information contained in pseudo labels is often considered as a more reliable and informative source for pretext tasks to learn point cloud representation than tags. For instance, the label 'airplane' only indicates the shape of objects without descriptions like colors, poses, and differences from other samples in same category. In contrary, these attributes are implicitly contained in point clouds and could be captured as pseudo label in SSL tasks.

Different methods define pseudo labels in different ways. In most reconstruction-based pretext tasks, pseudo labels are point cloud itself which provides a rebuilding objective for pretext task. In contrast-based methods, pseudo labels are a multidimensional matrix carrying collection information and are typically generated using clustering methods such as memory bank (Wu, Xiong, Yu and Lin, 2018), online dictionary (He et al., 2020), and prototype approaches (Caron et al., 2020), representing mean and variance of all or part of the features of point cloud dataset. For some alignment-based prediction or motion-based tasks pre-trained on temporal point cloud datasets, pseudo labels are geometric information like position, pose, and orientation in a number of frames before and after.

2.7. Loss functions

Appropriate and easily-differentiable loss functions are critical to facilitate backpropagation and optimization for encoders. In reconstruction-based pretext tasks, the symmetric function, Chamfer distance (CD), is commonly employed to assess the distance between each point in one set and its corresponding nearest point in the other. More formally, for two non-empty subsets X and Y, Chamfer distance $d_{CD}(X, Y)$ is defined as:

$$d_{CD}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y}||x-y||^2 + \frac{1}{|Y|}\sum_{y\in Y}\min_{x\in X}||x-y||^2 \tag{3}$$

Here, x and y represent the points in the reconstruction point set X and the original input point set Y, respectively; $||\cdot||$ denotes the L2 distance between two points and $|\cdot|$ refers to the number of points. The smaller the CD value, the more similar the two point sets are, and the better the SSL algorithm performs.

For contrast-based pretext tasks, the objective is to discriminate the similarities and differences between each point cloud samples on the overall semantic level. A cross-entropy like loss function to encourage the positive samples to be close to each other (and negative ones to be far from each other) is needed. InfoNCE (NCE stands for Noise-Contrastive Estimation) is a contrastive loss function that estimates the mutual information between a pair of samples, and can be formulated as:

$$L_q = -\log\frac{\exp(q\cdot k_+/\tau)}{\sum_{i=0}^{K}\exp(q\cdot k_i/\tau)} \tag{4}$$

where q indicates the encoded query (feature); k indicates a set of $K+1$ encoded samples $\{k_0, k_1, k_2, \ldots, k_K\}$, which could be regarded as the prototypes of historical samples; τ is the temperature parameter controlling the sharpness of the distribution. Assuming there is only one positive sample k_+ in the set k matching the query q, the others K samples are all negative. InfoNCE aims to assign the query q into the positive sample k_+ in the $K+1$ classification problem (He et al., 2020). In other words, the loss function tries to maximize the logits of $q\cdot k_+$ and minimize the value of the denominator.

3. Self-Supervised Learning pretext tasks for point cloud

We classify the current point cloud SSL research into four general categories based on the nature of the pretext tasks: reconstruction-based, contrast-based, alignment-based, and motion-based methods, as shown in Fig. 2. These categories can be further divided into more fine-grained sub-categories according to the different ways in which the features are extracted and used. The following sections summarize the principles and peculiarities of various proxy tasks in details. It should be noted that some methods may reside in multiple sub-categories.

3.1. Reconstruction-based methods

Reconstruction-based methods learn point cloud representations by reconstructing the corrupted point clouds and recovering the original ones as much as possible. Global features as well as the mappings between local and global areas are learned during the reconstruction process. According to different types of corruption and reconstruction objects, we further divide them into six sub-categories: mask recovery, spatial restoration, point sampling, disentanglement, deformation reconstruction, and generation and discrimination. Summary about the methods under these six sub-categories is shown in Table 3.

3.1.1. Mask recovery

The core idea of reconstruction is to mask a portion(s) of the point cloud and recover such missing part via an encoder-decoder architecture. Similar to the image inpainting task

Table 3
Summary of reconstruction-based point cloud SSL methods. D & G stand for Generation and Discrimination.

Year	Method	Sub-categories	Contributions
2021	Point-BERT (Yu et al., 2021)	Mask recovery	Reconstructing missing point tokens with BERT-style transformer
2022	Point-MAE (Pang et al., 2022)	Mask recovery	Shifting masked tokens to decoder to avoid early leakage
2022	MaskSurf (Zhang et al., 2022b)	Mask recovery	Estimating surfel position and per-surfel orientation simultaneously
2022	Voxel-MAE (Min et al., 2022)	Mask recovery	Performing additional binary voxel classification for complicated semantics awareness
2019	3D jigsaw (Sauder and Sievers, 2019b)	Spatial restoration	Rearranging randomly disorganized point clouds
2019	CloudContext (Sauder and Sievers, 2019a)	Spatial restoration	Predicting relative structural position between two given patches
2020	Orientation estimation (Poursaeed et al., 2020)	Spatial restoration	Predicting and recovering rotation angle around an axis
2019	PU-GAN (Li, Li, Fu, Cohen-Or and Heng, 2019)	Point upsampling / D & G	Utilizing self-attention unit for feature aggregation and quality enhancement
2021	SSPU-Net (Zhao, Hui and Xie, 2021)	Point upsampling	Leveraging shape coherence between sparse input and generated dense point cloud
2022	UAE (Zhang, Shi, Deng and Wu, 2022a)	Point upsampling	Gaining both advanced semantic information and basic geometric structure
2022	SPU-Net (Liu, Liu, Liu and Han, 2022d)	Point upsampling	Integrating self-attention with graph convolution network for context feature extraction
2022	PUFA-GAN (Liu, Yuan, Hou, Hamzaoui and Gao, 2022c)	Point upsampling / D & G	Employing graph filter to extract high frequency points of sharp edges and corners
2022	SSAS (Zhao, Liu, Zhong, Jiang, Gao, Li and Ji, 2022a)	Point upsampling	Achieving magnification-flexible point clouds upsampling
2022	Pose Disentanglement (Tsai, Chiang, Tsai and Chiu, 2022)	Disentanglement	Uncoupling content and pose attributes in partial point clouds
2022	CP-Net (Xu, Zhou, Xu, Wang and Qiao, 2022)	Disentanglement	Disentangling point clouds into contour and content ingredients
2022	MD (Sun, Zheng, Wang, Xu and Yang, 2022)	Disentanglement	Separating mixing point cloud into two independent objects
2018	FoldingNet (Yang, Feng, Shen and Tian, 2018)	Deformation reconstruction	Stretching 2D grid lattice to reproduce 3D surface structure
2021	Self-correction (Chen, Liu, Ni, Wang, Yang, Liu, Li and Tian, 2021)	Deformation reconstruction	Recovering shape-disorganized point regions
2021	DefRec (Achituve, Maron and Chechik, 2021)	Deformation reconstruction	Performing deformation on 2D grids to fit arbitrary 3D object surface
2018	PC-GAN (Li, Zaheer, Zhang, Poczos and Salakhutdinov, 2018)	D & G	Employing hierarchical and interpretable sampling strategy
2019	RL-GAN (Sarmad, Lee and Kim, 2019)	D & G	Introducing reinforcement learning agent to control GAN
2019	TreeGCN (Shu, Park and Kwon, 2019)	D & G	Leveraging ancestor information to boost point representation
2022	MaskPoint (Liu, Cai and Lee, 2022b)	D & G	Performing simple binary classification as proxy task

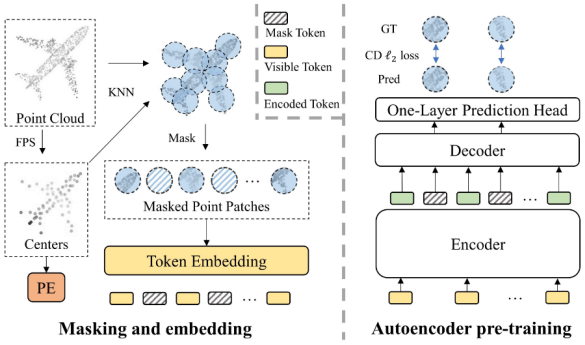

Figure 5: The general pipeline of Point-MAE. (1) The process of masking and embedding is demonstrated on the left. The point cloud patches are generated by FPS and KNN and masked randomly. Both visible and mask patches are mapped to the corresponding tokens through PointNet-based embedding layers. Also, the Position Embedding (PE) is obtained by mapping the center coordinates to the embedding dimension. (2) The autoencoder pre-training is shown on the right. The encoder only processes the visible tokens while the mask tokens are shifted and added to the input sequence of the decoder to reconstruct the masked patches. This figure is adapted from (Pang et al., 2022).

(Sarmad et al., 2019) and Mask AutoEncoder (MAE) (Hess et al., 2022), the encoder is required to capture the local geometric structure and the regional relations during the restoration process. Generally speaking, the better the reconstruction, the more effective the learned features.

Point-BERT (Yu et al., 2021), built based on BERT (Devlin et al., 2018), designs a point-specific tokenizer on discrete Variational AutoEncoder (dVAE) to map patches to discrete tokens to capture meaningful local geometric patterns. A portion of the input is randomly masked out, and a BERT-style transformer is trained to reconstruct the missing token under the supervision of point tokens obtained by the tokenizer. However, the tokenizer should be pre-trained in advance, and Point-BERT over-relies on auxiliary contrastive learning as well as data augmentation.

To address this issue, Pang et al. proposed Point-MAE (Pang et al., 2022) as a neat and efficient scheme of mask autoencoder as shown in Fig. 5. Concretely, Point-MAE employs the standard transformer as the backbone with an asymmetric encoder-decoder architecture to process random masking points with a high ratio (60%-80%). The mask tokens are shifted from the input of the encoder to the lightweight decoder, which saves considerable computation, and more significantly, avoids early leakage of location information. To further capture local geometric information, Zhang et al. introduced Mask Surfel Prediction (MaskSurf) (Zhang et al., 2022b), which estimates the surfel position (i.e., points) and per-surfel orientation (i.e., normals) simultaneously. Such a two-head pre-training paradigm has been justified to capture more effective representations than a reconstruction-only pretext. Likewise, Voxel-MAE (Min et al., 2022) transforms point clouds into volumetric representations and applies the range-aware random masking strategy on the voxel grid. Besides reconstructing the occupancy value of masked voxels, a supplementary binary voxel classification task distinguishing whether the voxel contains point clouds boosts the model to learn more complicated semantics.

3.1.2. Spatial restoration

Point clouds are the coordinate sets containing abundant spatial information that describes the structural distribution of objects and the environment in the Euclidean space. It is natural to exploit such rich spatial knowledge as the supervision signal in pretext tasks.

Sauder et al. (Sauder and Sievers, 2019b) proposed a 3D version of the jigsaw pretext to rearrange point clouds whose parts have been randomly disrupted and displayed by voxels along the axes. The goal of this pretext is to restore the original position of each patch (labeled by voxel ID) from the state of chaotic and disorderly distribution. They

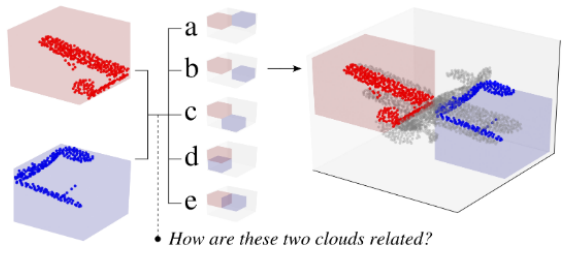

Figure 6: Illustration of the CloudContext pretext task. The pre-training model is enforced to estimate the spatial relevance between two given point cloud segments from six categories. In this case, the exact relation of these two components is 'the red part is diagonally above the blue part'. This figure is adapted from (Sauder and Sievers, 2019a).

later developed CloudContext (Sauder and Sievers, 2019a) to forecast the spatial relevance between two point cloud segments. As shown in Fig. 6, the model is trained to predict the relative structural position between two given patches from the same object, which utilizes the innate attributes of point clouds as they are not restrained by a discrete grid. By doing so, powerful per-point features can be accessed in an easy-to-implement unsupervised manner without expensive computation.

Orientation estimation (Poursaeed et al., 2020) is another simple but effective proxy task to capture the spatial information of point clouds. With the canonical orientation provided in most datasets, the orientation estimation pretext task aims to predict and recover the rotation angle around an axis via matrix multiplication. Such a pretext requires a high-level holistic understanding of shapes and obviates the need for manual annotations.

3.1.3. Point upsampling

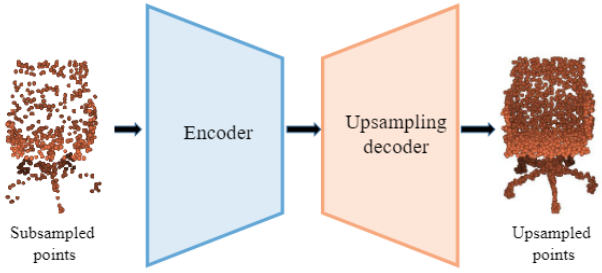

Figure 7: Overview of Upsampling AutoEncoder. The input point cloud is subsampled by a random sampling strategy and then fed into the encoder to extract point-wise features. The decoder is adopted to reconstruct the original point cloud with offset attention based on the learned representation. This figure is adapted from (Zhang et al., 2022a).

Point upsampling is the operation to upsample sparse, noisy, and non-uniform point clouds to generate a dense, complete, and high-resolution point cloud, which is challenging but also beneficial for the model to capture implicit geometric representations of the underlying surface.

PU-GAN (Li et al., 2019) is a pioneer SSL upsampling paradigm formulated based on the generative adversarial network (GAN) (Goodfellow et al., 2014) to generate a diverse range of point distributions from the latent space and upsample points over patches. An up-down-up unit is embedded in the generator to expand point features as well as a self-attention unit for quality enhancement on feature aggregation. The discriminator is inspired to gain inherent patterns and improve the uniformity of output generation according to a compound loss including adversarial, uniform, and reconstruction terms. Motivated by PU-GAN, Zhang et al. proposed the Upsampling AutoEncoder (UAE) (Zhang et al., 2022a) to gain both advanced semantic information and basic geometric structure from subsampled point clouds. As shown in Fig. 7, the encoder is devised to perform point-wise feature extraction on the subsampled point cloud, and the upsampling decoder is designed to reconstruct the original dense point cloud with offset attention (Guo et al., 2021) to refine global shape structure.

Liu et al. (Liu et al., 2022d) proposed a coarse-to-fine reconstruction framework, dubbed SPU-Net, integrating self-attention with graph convolution network (GCN) for contextual feature extraction and generating fine point sets with hierarchically learnable 2D grids. Zhao et al. (Zhao et al., 2021) introduced SSPU-Net by leveraging the shape coherence between input sparse and generated dense point clouds. In addition, it has an image-consistent loss among multi-view rendered images to capture the latent patterns of underlying point structures.

PUFA-GAN (Liu et al., 2022c), a frequency-aware framework, utilizes a graph filter to extract high frequency (HF) points of sharp edges and corners so that the discriminator could focus on the HF geometric properties and enforce the generator producing neat and more uniform upsampled point clouds. To get rid of the fixed upsampling factor restriction, Zhao et al. (Zhao et al., 2022a) presented a self-supervised arbitrary-scale (SSAS) framework with a magnification-flexible upsampling strategy. Instead of direct mapping from sparse to dense point clouds, the proposed scheme seeks the nearest projection points on the implicit surface for seed points via two functions, which are exploited to estimate the projection direction and distance, respectively.

3.1.4. Disentanglement

Models pre-trained under the SSL paradigm usually tend to learn well the low-level geometric features of point clouds, such as pose, contour, and shape information, but overlook the high-level semantic content understanding, which often leads to unsatisfactory performance in downstream tasks such as object classification that requires global discriminative capability. To tackle this issue, disentanglement-based SSL pretexts are proposed to separate the low-level geometric features from the high-level semantic embedding.

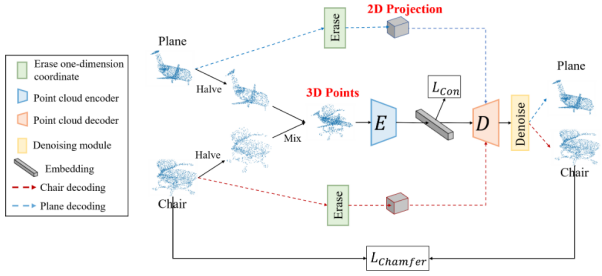

Figure 8: The schematic of Mixing and Disentangling (MD) pretext. The two input point clouds are separately halved and mixed into a hybrid object feeding to the encoder E to mine the geometry-aware embedding. The 'Erase' operation is applied to obtain the 2D projection from both original input point clouds simultaneously. The instance-adaptive decoder D receives the embedding together with the two partial projections as input to disentangle the blended shape into the original two point clouds. The chamfer distance is used to measure the reconstruction error between generated point clouds and the original ones. This figure is adapted from (Sun et al., 2022).

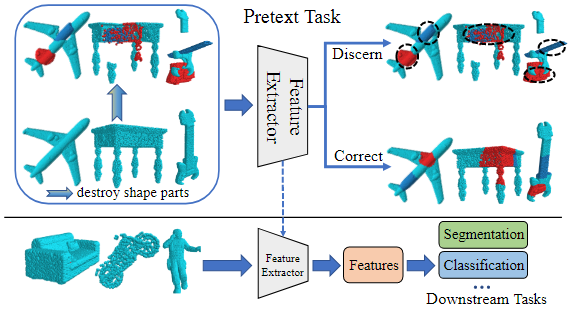

Figure 9: Demonstration of shape self-correction pretext. The input point cloud is firstly preprocessed by a shape-disorganizing module to generate a deformed point cloud and then fed to the encoder to learn the geometry-aware representation. Two separate task heads are constructed to distinguish and segment points belonging to distorted parts, and subsequently reconstruct the partial-deformed objects. The well-trained feature extractor is transferred to downstream tasks to estimate the feature capturing capability. This figure is adapted from (Chen et al., 2021).

Feature extraction is performed based on various contents using distinct modules to obtain hierarchical representations.

Tsai et al. (Tsai et al., 2022) proposed a disentanglement framework that uncouples content and pose attributes in partial point clouds to enhance both geometric and semantic feature abstraction. Two encoders are employed to learn the content and multi-view poses separately, where the gained pose representation should predict the viewing angle and navigate the partial point cloud reconstruction cooperated with the content from another specific view. Likewise, Xu et al. (Xu et al., 2022) presented a universal Contour-Perturbed Reconstruction Network (CP-Net) that disentangles a point cloud into contour and content ingredients. A concise contour-perturbed augmentation unit is exploited on the contour component and retains the content part of the point cloud. Therefore, the self-supervisor is able to concatenate the content component for advanced semantic comprehension.

Different from the above two pretexts, Mixing and Disentangling (MD) (Sun et al., 2022) blends two disparate point shapes into a hybrid object and attains geometry-aware embedding from the encoder. An instance-adaptive decoder is then leveraged to restore the original geometries based on the obtained embedding by disentangling the mixed shape. As shown in Fig. 8, except for the main encoder-decoder structure, the proposed scheme also encompasses a coordinate extracting operation 'Erase', which randomly drops one-dimension coordinate of each point to provide an extra 2D partial projection to better reconstruct the original point cloud shapes.

3.1.5. Deformation reconstruction

Point cloud deformation is a common phenomenon in real-world data scanning, which is usually caused by object distortion, sensor noise, or external occlusion. It has been discovered that SSL by reconstructing the original point cloud from the artificially deformed one (e.g. adding Gaussian noise or local translation) enables the learned model to obtain geometric perception as well as context awareness.

Chen et al. (Chen et al., 2021) proposed a shape self-correction pretext to mine implicit geometric embeddings of point clouds. The pretext assumes that a robust shape representation could identify and correct distorted regions of a shape. As shown in Fig. 9, the proposed scheme imposes destruction over certain regions by a shape-disorganizing module and sends the deformed point cloud to the feature extractor for embedding learning. Two task heads are built separately to discern the distorted components and further restore them to their original normal shapes for fine-grained geometric and contextual feature exploration.

Achituve et al. (Achituve et al., 2021) conducted the first study of SSL for domain adaptation (DA) on point cloud via Deformation Reconstruction (DefRec). By mapping the dislocating points to their original location, the model is able to obtain the latent statistical structure of the input point cloud. Moreover, the distribution gap between source and target domains is bridged by the learned representation since they are invariant to distribution shift.

FoldingNet (Yang et al., 2018) presents a novel folding-based decoder to perform deformation on the canonical 2D grid to fit an arbitrary 3D object surface. Instead of deforming the point cloud, the folding operation exerts a virtual force induced by the embedding captured from input to stretch a 2D grid lattice to reproduce the 3D surface structure. This approach tackles issues caused by point cloud's irregular attributes by applying implicit 2D grid constraints.

3.1.6. Generation and discrimination

The generation and discrimination pretext is a unique paradigm that designs a discriminator module to distinguish

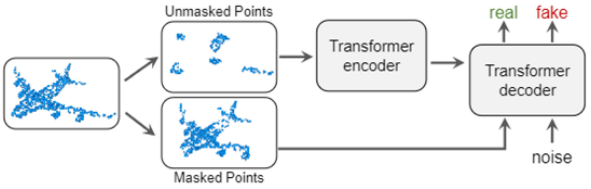

Figure 10: The general pipeline of the Mask-Point model. The reconstruction challenge is formulated as a discriminative pretext to determine whether the source of the extracted sample is a masked point cloud or a random noise. The figure is adapted from (Liu et al., 2022b).

whether the fed point cloud is reconstructed from noise distribution or truly sampled. During the adversarial training process, the generator (encoder) and discriminator (decoder) compete with each other and are updated alternatively so that both components can be transferred for downstream tasks.

PC-GAN (Li et al., 2018) is specifically designed for point clouds and employs a hierarchical and interpretable sampling strategy inspired by Bayesian and implicit generative models to tackle the issue of missing constraints on the discriminator. Sarmad et al. (Sarmad et al., 2019) introduced a reinforcement learning (RL) agent to control the GAN to extract implicit representations from noisy and partial input to generate high-fidelity and entire point clouds. Meanwhile, applying an RL agent to seek the best-fit input of GAN to produce low-dimensional latent embedding relieves the challenge of unstable GAN training. Shu et al. (Shu et al., 2019) introduced a tree-structured graph convolutional network (TreeGCN) as the generator, leveraging ancestor information to boost the representation of the point. It is more efficient in computation than using neighborhood features as adopted in regular GCNs. PU-GAN (Li et al., 2019) and PUFA-GAN (Liu et al., 2022c), both employed GANs-based models to generate dense and uniform point clouds with innovative modules for feature aggregation enhancement and high-frequency point filtering.

Liu et al. (Liu et al., 2022b) proposed a discriminative mask pretraining transformer framework, MaskPoint, which combines mask and discrimination techniques to perform simple binary classification between masked object points and sampled noise. As shown in Fig. 10, the original complete point cloud is divided into 90% masking portion and a 10% visible potion. Two kinds of query, where the real is sampled from masked point clouds while the fake is derived from random noise, are fed to the decoder for classification. During the discrimination process, the model is required to deduce the full geometry from small visible portions.

3.2. Contrast-based methods

Contrastive learning is a popular mode of SSL that encourages augmentation of the same input to have more comparable representations. The general approach is to expand the views of input point clouds (anchors) by various data augmentation techniques. In particular, it tries to enforce positive samples augmented from the same anchor more similar than negative samples from different anchors in the feature space. In this section, we will introduce contrast-based methods with representative examples and discuss their contributions and limitations. A brief summary of these methods is shown in Table 4.

3.2.1. Object contrast

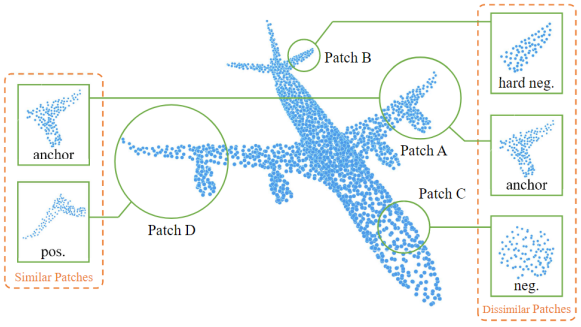

Figure 11: Illustration of a self-contrastive paradigm. Patch A is selected as the anchor and the symmetrical part Patch D is the positive sample. Patch B and C are the negative samples, where Patch B is hard to distinguish due to its comparative similarity to the anchor. The figure is adapted from (Du et al., 2021).

Traditional contrastive learning research usually focuses on instance-wise objects. The priority is on overall semantic learning through discriminative pretext tasks that capture context similarity and difference of point clouds. Such an object-contrast paradigm performs data augmentation on relatively large patches or whole single point objects to capture global geometric awareness.

Sanghi (Sanghi, 2020) proposed Info3D, which takes inspiration from Contrastive Predictive Coding (Oord, Li and Vinyals, 2018) and Deep InfoMax (Velickovic, Fedus, Hamilton, Liò, Bengio and Hjelm, 2019), to obtain rotation-insensitive representation by maximizing mutual information between 3D objects and their local chunks as well as geometrically transformed versions. Lu et al. (Lu et al., 2022) proposed the Augmentation Fusion Self-Supervised Representation Learning (AFSRL) framework, which imposes data-level augmentation and feature enhancement simultaneously to construct a stable and invariant point cloud embedding. The correspondence between augmented pairs is acquired, and the invariant semantic is maintained under perturbations during augmentation.

Zhang et al. (Zhang and Zhu, 2019) introduced a simple two-phase unsupervised GCN framework (contrasting and clustering), to capture superior point embedding by solving part contrast and object cluster tasks consecutively. Du (Du et al., 2021) presented a self-contrastive paradigm leveraging self-similar point cloud patches within a single point cloud to facilitate local shape and global context primitives capturing. As shown in Fig. 11, according to the nonlocal self-similar property of the point cloud, where regional geometry remains invariant after affine transformation, the self-similar

Table 4
Summary of contrast-based point cloud SSL methods.

Year	Method	Sub-categories	Contributions
2020	Info3D (Sanghi, 2020)	Object contrast	Maximizing mutual information between objects and their transformations
2022	AFSRL (Lu, Dai, Li and Su, 2022)	Object contrast	Imposing data-level augmentation and feature enhancement simultaneously
2019	Contrasting and clustering (Zhang and Zhu, 2019)	Object contrast	Solving part contrast and object cluster tasks consecutively
2021	Hard negatives (Du, Gao, Hu and Li, 2021)	Object contrast	Leveraging self-similar point cloud patches; facilitating hierarchical context primitives capturing
2020	PointContrast (Xie, Gu, Guo, Qi, Guibas and Litany, 2020)	Scene contrast	Obtaining dense features at point-level on complex scenes by point contrast
2021	Contrastive Scene Contexts (Hou, Graham, Nießner and Xie, 2021)	Scene contrast	Introducing ShapeContext local descriptor and achieving data-efficiency
2021	CoCoNets (Lal, Prabhudesai, Mediratta, Harley and Fragkiadaki, 2021)	Scene contrast	Mapping RGB-D images to 3D points by optimizing view-contrastive prediction
2020	P4Contrast (Liu, Yi, Zhang, Fan, Funkhouser and Dong, 2020)	Scene contrast	Utilizing synergies between two modalities for better feature extraction
2021	DepthConstrast (Zhang, Girdhar, Joulin and Misra, 2021)	Scene contrast	Applying Instance Discrimination on depth maps

point cloud patches are treated as positive samples otherwise negative based on the inferred similarity score. Moreover, hard negative samples, close to positive samples in the representation space, are sampled for more discriminative and expressive representation learning.

3.2.2. Scene contrast

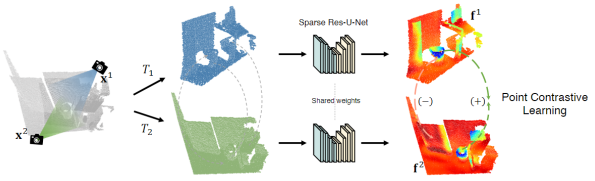

Figure 12: The illustration of PointContrast. Contrast is performed at the point-level between two transformed point clouds, where positive samples are the matched points while negative samples are the unmatched points across two views. The figure is adapted from (Xie et al., 2020).

Different from object-contrast, the scene-contrast paradigm concentrates on scenes to capture broader environmental context and neighborhood perception, which is more relevant to real-world complex scenarios.

To address the domain gap issue (i.e., it is insufficient to capture a global representation from object instances), Xie et al. (Xie et al., 2020) proposed PointContrast, a sparse residual U-Net based framework aiming to obtain dense features at the point-level on complex scenes. As shown in Fig. 12, two views x^1 and x^2 are produced from a complicated point cloud scene, where corresponding pairs are computed between these two views as the positive samples. Two rigid transformations T^1 and T^2 are utilized to increase the difficulty of the pretext which demands the network to learn the invariant embedding under random geometric shift. The contrastive loss is defined to shorten the distance between the matched points and enlarge the distance of mismatched points of the two overlapping partial scans so that the pre-training model can capture local descriptions and be universally pertinent to various advanced 3D understanding downstream tasks.

However, PointContrast only considers point correspondence matching but ignored the spatial configurations and contexts in a scene, e.g., relative pose and distance, therefore confining its transferability and scalability. To address this issue, Hou et al. (Hou et al., 2021) presented Contrastive Scene Contexts to fuse spatial information into pre-training objects by introducing ShapeContext local descriptor (Xie, Liu, Chen and Tu, 2018) partitioning and performing contrastive learning in each region. The method improves the performance and data efficiency on downstream tasks in which employing only 0.1% of point labels reaches the performance level with full supervision.

Continuous Contrastive 3D Networks (CoCoNets) (Lal et al., 2021) aims to infer latent scene representations by mapping RGB-D images to 3D point scenarios and optimizing view-contrastive prediction. P4Contrast (Liu et al., 2020), another RGB-D bi-modal SSL framework, proposes to contrast point-pixel pairs and provides additional flexibility for hard negative creation to exploit the synergies between two modalities for better feature extraction. DepthConstrast (Zhang et al., 2021) circumvents the need for point correspondences and instead applies the Instance Discrimination (Wu et al., 2018) method on depth maps combined with a momentum encoder to improve the geometric perception.

3.3. Alignment-based methods

Point cloud representation is generally invariant to transformations in terms of time flow, spatial motion, multi-view photography, etc. Based on this property, alignment-based methods have been proposed to learn the implicit embedding of point clouds by preserving the coherence of point features in spatiotemporal consistency, multi-view alignment, and multimodal fusion. A brief summary of the methods under this category is provided in Table 5.

3.3.1. Multi-view alignment

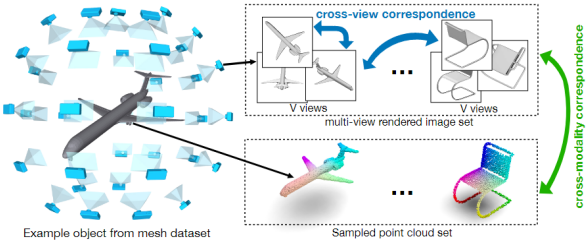

Figure 13: The schematic view of cross-modality and cross-view correspondences. The 3D point cloud objects and corresponding pairs of multi-view rendered images are sampled from the same mesh input, respectively. The relation of diverse views is captured as the supervision signal by sustaining the alignment among multi-view and cross-domain representations. The figure is adapted from (Jing et al., 2021).

Table 5
Summary of alignment-based point cloud SSL methods.

Year	Method	Sub-categories	Contributions
2020	Info3D (Sanghi, 2020)	Multi-view alignment	Maximizing mutual information between objects and their transformations
2021	OcCo (Wang et al., 2021b)	Multi-view alignment	Shielding and restoring occluded points in camera view
2021	Multi-view stereo (Yang, Alvarez and Liu, 2021)	Multi-view alignment	Generating prime depth map as self-supervision signal
2021	Cross-view (Jing, Zhang and Tian, 2021)	Multi-view alignment	Jointly learning both 3D point cloud and 2D image embedding concurrently
2022	Multi-view rendering (Tran, Hua, Tran and Hoai, 2022)	Multi-view alignment	Encouraging 2D-3D global feature distributions to be similar
2021	Order prediction (Wang, Yang, Rong, Feng and Tian, 2021c)	Spatiotemporal consistency	Sorting temporal order of sampled and disorganized point cloud clips
2021	STRL (Huang, Xie, Zhu and Zhu, 2021)	Spatiotemporal consistency	Dual-branch network to predict representation of another temporally correlated input
2022	Futrue prediction (Mersch, Chen, Behley and Stachniss, 2022)	Spatiotemporal consistency	Forecasting future point cloud scenes with lightweight model
2020	PointPainting (Vora, Lang, Helou and Beijbom, 2020)	Multimodal fusion	Projecting LiDAR points into semantic segmentation diagram for traffic scenes
2021	PointAugmenting (Wang, Ma, Zhu and Yang, 2021a)	Multimodal fusion	Replacing sub-optimal segmentation scores with high-dimension CNN features
2022	DeepFusion (Li et al., 2022)	Multimodal fusion	Exploiting cross-attention to capture long-range correlations of image-LiDAR pairs

Compared to direct processing and feature extraction on 3D point clouds, projecting point clouds into 2D images for dimension reduction and utilizing mature image networks as well as 2D SSL technologies is relatively more accessible. To ensure that the learned embeddings sufficiently represent the entire 3D point cloud objects or scenes, multi-view alignment pretexts are necessary to preserve the integrity and uniformity of the point cloud features.

Info3D (Sanghi, 2020) aims to obtain rotation-insensitive representations by maximizing mutual information between 3D objects and their local chunks for patch-level consistency. Occlusion Completion (OcCo) (Wang et al., 2021b) combines the idea of mask recovery shielding and restoring occluded points in a camera view for better spatial and semantic properties comprehension. Similarly, Yang et al. (Yang et al., 2021) introduced an SSL multi-view stereo structure generating prime depth map as pseudo-labels and refined such self-supervision from neighboring views as well as high-resolution images by multi-view depth fusion iteratively. Furthermore, the correspondence of pixel/point of the point clouds and the corresponding multi-view images are aligned for cross-modality consistency.

Jing et al. (Jing et al., 2021) proposed a novel SSL framework leveraging cross-modality and cross-view correspondences to jointly learn both 3D point cloud and 2D image embedding concurrently. As shown in Fig. 13, point cloud objects and comparable pairs of multi-view rendered images are sampled from the same mesh input. In addition to 2D-3D consistency, the contrastive notion is adopted into cross-view alignment that shortens intra-object distance while maximizing inter-object discrepancy of distinct rendered images. Similarly, Tran et al. (Tran et al., 2022) presented a dual-branch model not only agreeing upon fine-grained pixel-point local representation but also encouraging 2D-3D global feature distributions as approaching as possible by exploiting knowledge distillation.

3.3.2. Spatiotemporal consistency

Unlike previous methods, the spatiotemporal approach is more concerned with long-range spatial and temporal invariance before and after certain point cloud frames, which are 4D data (XYZ coordinate + temporal dimension), to capture intrinsic characteristics of dynamic sequences.

Motivated by the success of Xu et al.'s work (Xu, Xiao, Zhao, Shao, Xie and Zhuang, 2019) in video SSL, Wang et al. proposed the first SSL scheme to gain effective temporal

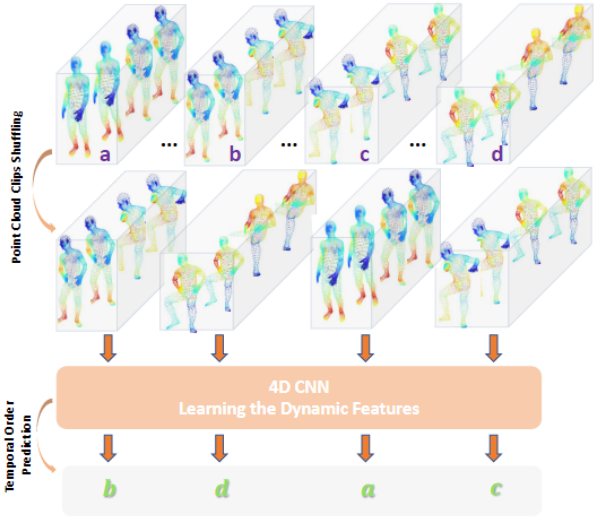

Figure 14: Demonstration of point cloud sequence order prediction. The first row is the uniformly sampled point cloud clips from the continuous point cloud sequence. Then these clips are randomly shuffled and then fed into 4D CNN in the second row to learn the dynamic features of human actions, The original temporal order is predicted in a self-supervised manner. The figure is adapted from (Wang et al., 2021c).

embeddings on dynamic point cloud data by sorting the temporal order of sampled and disorganized point cloud clips. As shown in Fig. 14, a few static point cloud frames are uniformly sampled and disordered, which are then processed by a 4D CNN to restore the disrupted fragments to the correct order on an unannotated, large-scale, sequential point cloud action recognition dataset.

Another spatiotemporal representation learning (STRL) (Huang et al., 2021) framework, inspired by BYOL (Grill, Strub, Altché, Tallec, Richemond, Buchatskaya, Doersch, Avila Pires, Guo, Gheshlaghi Azar et al., 2020), designs a dual-branch pipeline, referred to as online and target networks, to collaborate and promote each other. Specifically, the online network is enforced to predict the target network representation of another temporally correlated input, which is augmented by random spatial transformation, for spatiotemporal invariant contextual cues extraction. Taking training and inference time into account, Mersch et al. (Mersch et al., 2022) presented an innovative 3D spatiotemporal convolution encoder-decoder neural network consisting of

fewer parameters to predict future point cloud scenes. Such a lightweight model concatenates range images as input to estimate forthcoming images and per-point scores in multiple future steps, so that spatial and temporal scene information can be captured simultaneously.

3.3.3. *Multimodal fusion*

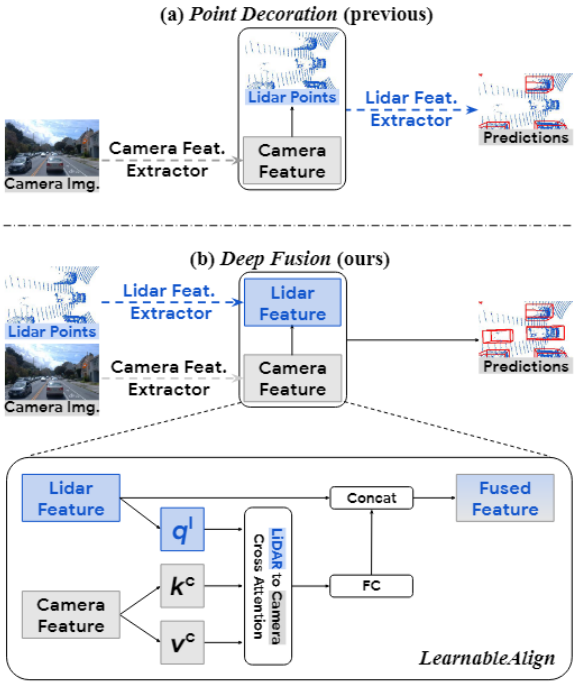

Figure 15: Demonstration of point decoration and deep fusion. (a) Previous cross-modal paradigms (Lal et al., 2021; Jing et al., 2021) decorate LiDAR points with camera feature on input-level for 3D detection. (b) DeepFusion (Li et al., 2022) fuses camera and LiDAR features extracted by respective encoders and leverages cross-attention consistency technique. The figure is adapted from (Li et al., 2022).

Rather than simply requiring coherence between 2D-3D correspondences (Lal et al., 2021; Jing et al., 2021; Tran et al., 2022), automatic driving algorithms demand sophisticated collaboration between in-vehicle sensors. For example, cameras and LiDARs provide complementary information (e.g., colorful texture visualization and distance perception) for 3D object detection. Therefore, multimodal fusion is a promising direction to exploit the potential of images and point clouds for acquiring effective traffic scene features.

Vora et al. (Vora et al., 2020), Wang et al. (Wang et al., 2021a), and Li et al. (Li et al., 2022) offered compact frameworks for tight sensor-fusion which could be implemented under the SSL paradigm without human annotations. PointPainting (Vora et al., 2020) is a sequential fusion method that projects LiDAR points onto semantic segmentation diagrams for traffic scenes with color marking. Each point is painted with a class score obtained from the image segmentation network and then can be utilized in any LiDAR detection approaches. Such a painting fusion method cleverly addresses the limitations of depth-blurring and scale ambiguity by consolidating the birds-eye and camera view.

PointAugmenting (Wang et al., 2021a) adopts a late cross-modal fusion mechanism based on PointPainting, replacing the sub-optimal segmentation scores with high-dimension CNN features containing rich outlook hints and larger receptive fields to emphasize the delicate details. Moreover, a simple yet effective cross-modal data augmentation pastes virtual objects into images and point clouds for alignment between the camera and LiDAR. However, both PointPainting and PointAugmenting simply decorate LiDAR points with camera embeddings as shown in Fig. 15(a). To improve the performance on downstream tasks, DeepFusion (Li et al., 2022) proposed an end-to-end cross-modal fusion on the feature level, focusing on consistency improvement. As shown in Fig. 15(b), a block named LearnableAlign is introduced to exploit cross-attention to dynamically capture long-range correlations during the image-LiDAR fusion process to enhance the model's recognition and localization capability.

3.4. Motion-based methods

Various point cloud frames contain rich geometric patterns and kinematic schemas that are concealed in the movement of objects or scenes. The motion-based SSL paradigm focuses on dynamically capturing the intrinsic motion characteristics from spatial variations by taking advantage of traditional registration and scene flow estimation as pretexts. A brief summary on the methods under this category is shown in Table 6.

3.4.1. *Registration*

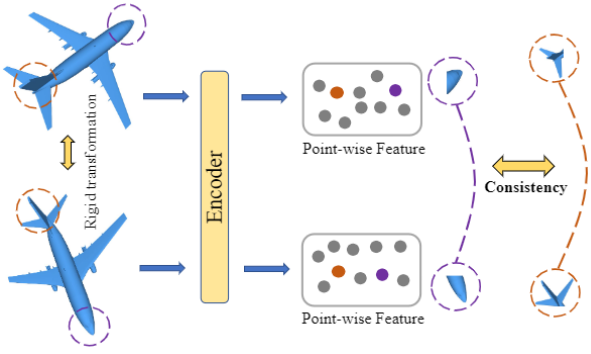

Figure 16: Demonstration of deep versatile descriptors. The input consists of two point clouds before and after rigid transformations, where the common point components are utilized to train the encoder for global and local representation learning. The figure is adapted from (Liu et al., 2022a).

Point cloud registration is the task to merge two point clouds X and Y into a globally consistent coordinate system via estimating the rigid transformation matrix, which can be formulated as:

Table 6
Summary of motion-based point cloud SSL methods.

Year	Method	Sub-categories	Contributions
2019	PRNet (Wang and Solomon, 2019)	Registration	Pioneer for partial-to-partial point cloud registration enabling coarse-to-fine refinement
2021	Part mobility (Shi, Cao and Zhou, 2021)	Registration	Converting points to trajectories to derive the rigid transformation hypotheses
2022	SuperLine3D (Zhao, Yang, Huang, Chen, Ma, Li and Liu, 2022b)	Registration	Obtaining precise line representation under arbitrary scale perturbations
2022	DVD (Liu, Chen, Xu, Qiu and Chu, 2022a)	Registration	Learning local and global point embedding jointly
2020	PointPWC-Net (Wu, Wang, Li, Liu and Fuxin, 2020)	Scene flow estimation	Discretizing cost volume onto 3D point clouds in a coarse-to-fine fashion
2020	Just go with the flow (Mittal, Okorn and Held, 2020)	Scene flow estimation	Optimizing two SSL losses based on nearest neighbors and cycle consistency
2021	Self-Point-Flow (Li, Lin and Xie, 2021)	Scene flow estimation	Converting pseudo label matching problem as optimal transport task

$$R, t = \underset{R \in SO(3), t \in \mathbb{R}^3}{\arg\min} \|\psi(RX + t) - \psi(Y)\|_2. \tag{5}$$

where $R \in SO(3)$ and $t \in \mathbb{R}^3$ indicate rotation matrix and translation vector, respectively; ψ is the feature extraction network learning the hierarchical informative features from dynamic point clouds. Unlike the classic ICP registration method (Besl and McKay, 1992) which iteratively searches correspondences and estimates rigid transformation, SSL registration can obtain informative point cloud features without high-quality ground-truth correspondences.

PRNet (Wang and Solomon, 2019) is a partial-to-partial registration method that enables coarse-to-fine refinement iteratively. Based on co-contextual information, the framework boils down the registration problem as a key point detection task, which aims to recognize the matching points from two input clouds. Shi (Wang and Solomon, 2019) presented a part mobility segmentation approach to understand the essential attributes of the dynamic object. Instead of directly processing the sequential point clouds, the raw input is converted to trajectories by point correspondence between successive frames to derive rigid transformation hypotheses. Analogously, Zhao et al. (Zhao et al., 2022b) proposed an SSL line segmentation and description for LiDAR point clouds, called SuperLine3D, providing applicable line features for global registration without any prior hints. Compared to point embedding constrained by limited resolution, this segmentation model is capable of obtaining precise line representation under arbitrary scale perturbations.

Motivated by the observation that the local distinctive geometric structures of two subsets of point clouds can improve representations, Liu et al. (Liu et al., 2022a) introduced the deep versatile descriptors (DVDs) which learn local and global point embeddings jointly. As shown in Fig. 16, the co-occurring point cloud local regions, which retain the structural knowledge under rigid transformations, are regarded as the input of DVD to extract latent geometric patterns restrained by local consistency loss. To further enhance the model's capability of transformation awareness, reconstruction and normal estimation are added as auxiliary tasks for better alignment.

3.4.2. Scene flow estimation

Scene flow estimation is a vital computer vision task. For point clouds, its objective is to estimate the motion of objects by computing dense correspondences between consecutive LiDAR scans of a scene over time. The variation of points can be represented as 3D displacement vectors to describe the motions in terms of scene flow.

Figure 17: Demonstration of just going with the flow. The nearest neighbor loss is utilized to push the predicted flow (green) close to the pseudo-ground truth (red) of the frame at $t+1$. The cycle consistency loss is the penalty term to estimate the flow between predicted points (green) in the opposite direction to the original points (blue) in frame at t for temporal alignment. The figure is adapted from (Mittal et al., 2020).

Wu et al. (Wu et al., 2020) introduced the notion of cost volume and proposed a learnable point-based network called PointPWC-Net. The cost volume is discretized as input point pairs to reduce computational complexity; additionally, an efficient upsampling strategy and wrap layers are employed. Mittal et al. (Mittal et al., 2020) proposed a novel SSL scene flow estimation network to achieve safe navigation during interactions with highly dynamic environments by optimizing two loss components based on the nearest neighbors and cycle consistency. As shown in Fig. 17, the nearest neighbor loss encourages the points predicted based on current moment t flowing toward occupied regions of the future frame at $t+1$. The cycle consistency loss ensures that the points of the future frame $t+1$ can be restored in the reverse direction back to frame t to avoid degenerate solutions by maintaining temporal consistency. Self-Point-Flow (Li et al., 2021) employs more than 3D point coordinates, surface normal, and color in one-to-one matching to generate pseudo labels and formulates the pseudo label generation issue as an optimal transport problem. It leverages a random walk module to refine annotation quality by imposing local alignment.

4. Downstream tasks

One of the primary objectives of SSL is to pre-train a backbone network and transfer it to solve the problems in downstream tasks. Therefore, performance of the model in downstream tasks could reflect the effectiveness of SSL to a certain degree. The evaluation criteria indicate whether the SSL methods can extract useful knowledge from pretext tasks with large-scale unlabeled point cloud data. In this section, we introduce four commonly used downstream tasks and provide the widely used evaluation metrics. In addition, we summarize and compare the performance of the aforementioned representative SSL methods in the corresponding downstream tasks.

4.1. Object classification

Object classification is a fundamental and prevalent downstream task that requires the model to output a most likely label for the given point cloud object to assess the overall semantic awareness of the pre-trained model. The two commonly used metrics for this task are Overall Accuracy (OA) and Mean Class Accuracy (mAcc). OA is the ratio of correctly classified objects to the total number of objects, and mAcc is the average of each class's accuracy. Object classification can be divided into three protocols based on task settings:

- **Few-shot:** Few-shot learning (FSL) is a challenging task that involves training with limited information provided by the downstream dataset. Specifically, the *n*-way, *m*-shot setting is employed, where *n* is the number of classes randomly selected from the dataset and *m* is the number of objects randomly sampled for each class. The trained model is evaluated on the test split. Few-shot protocol performance of reviewed SSL methods is shown in Table 7.
- **Fine-tuning:** The pre-trained feature extractor serves as the initial downstream backbone encoder, and the entire network is re-trained in a supervised manner with labels from the downstream datasets. Fine-tuning protocol performance of proposed SSL methods is presented in Table 8.
- **Linear classification:** The pre-trained feature extractor is frozen by stopping the backpropagation gradients. Linear classifiers are trained in a supervised manner with downstream datasets. Linear classification protocol performance of proposed SSL methods is shown in Table 9.

4.2. Part segmentation

Part segmentation is a fine-grained task that aims to distinguish and separate various components of an object, such as plane wings or desk legs. This task usually requires a model that can extract local point-level features more effectively than the overall discriminative ability required for object recognition. The popular evaluation criteria of point cloud part segmentation is the mean Intersection over Union (mIoU), which computes the ratio of the intersection of the predicted and ground truth part labels to the union of the two, across all categories (mIoU_C) or all instances (mIoU_I). Table 10 summarizes the results of part segmentation on the ShapeNetPart dataset based on SSL pre-training models and supervised fine-tuning in terms of mIoU_C (%), mIoU_I (%).

Table 7
Summary of few-shot protocol performance of representative SSL methods on ModelNet40 (Wu et al., 2015) and ScanObjectNN (Uy et al., 2019). The results are reported in terms of OA (%).

Method	Backbone	5-way		10-way	
		10-shot	20-shot	10-shot	20-shot
	ModelNet40				
Point-MAE (Pang et al., 2022)	Transformer	96.3	97.8	92.6	95.5
Point-BERT (Yu et al., 2021)	Transformer	94.6	96.3	91.0	92.7
OcCo (Wang et al., 2021b)	PointNet	89.7	92.4	89.3	89.7
OcCo (Wang et al., 2021b)	DGCNN	90.6	92.5	82.9	86.5
OcCo (Wang et al., 2021b)	Transformer	94.0	95.9	89.4	92.4
3D jigsaw (Sauder and Sievers, 2019b)	PointNet	66.5	69.2	56.9	66.5
3D jigsaw (Sauder and Sievers, 2019b)	DGCNN	34.3	42.2	26.0	29.9
MaskSurf (Zhang et al., 2022b)	Transformer	96.5	98.0	93.0	95.3
MaskPoint (Liu et al., 2022b)	Transformer	95.0	97.2	91.4	93.4
	ScanObjectNN				
Point-MAE (Pang et al., 2022)	Transformer	63.9	77.0	53.6	61.6
OcCo (Wang et al., 2021b)	PointNet	70.4	72.2	54.8	61.8
OcCo (Wang et al., 2021b)	DGCNN	72.4	77.2	57.0	61.6
3D jigsaw (Sauder and Sievers, 2019b)	PointNet	58.6	67.6	53.6	48.1
3D jigsaw (Sauder and Sievers, 2019b)	DGCNN	65.2	72.2	45.6	48.2
MaskSurf (Zhang et al., 2022b)	Transformer	65.3	77.4	53.8	63.2

4.3. Semantic segmentation

Semantic segmentation requires a model to assign a semantic label to each points in the point cloud in order to group meaningful regions. It is frequently implemented on complicated outdoor or indoor scenes with background noise. mIoU, OA, and mAcc are commonly employed as estimation indicators to judge the feature extraction capability of pre-training models on the S3DIS dataset (Armeni et al., 2016), which contains six large-scale indoor venues, with the following two protocols. Performance of representative methods on semantic segmentation under the two protocols are shown in Table 11.

- **Area 5 test**: The SSL pre-trained model is fine-tuned on all areas except the largest area 5, which is chosen as the test set.
- **Six-fold cross validation**: Areas 1-6 are selected in turn as the test set and fine-tuned in the remaining 5 areas.

4.4. Object detection

Object detection is a task that involves localizing the 6 Degrees-of-Freedom (DoF) bounding box of an object and differentiating its category in a complex scene. The evaluation metric used is the average precision (AP), which measures the precision of the 3D bounding box at various recall levels. The threshold is usually set to 0.25 and 0.5. Table 12 summarizes the object detection performance of the SSL pre-training models on the SUN RGB-D (Song et al., 2015) and ScanNet (Dai et al., 2017) datasets.

Table 8
Summary of fine-tuning protocol performance of representative SSL methods on ModelNet40 (Wu et al., 2015) and ScanObjectNN (Uy et al., 2019). ScanObjectNN has three challenges. The results are reported in terms of OA (%).

Method	Year	Pretext type	Backbone	Pre-train dataset	ModelNet40	ScanObjectNN		
						OBJ-BG	OBJ-ONLY	PB-T50-RS
Supervised	2017	-	PointNet (Qi et al., 2017a)	-	89.2	73.3	79.2	68.0
	2017		PointNet++ (Qi et al., 2017b)		90.7	82.3	84.3	77.9
	2019		DGCNN (Wang et al., 2019)		92.9	82.8	86.2	78.1
	2017		Transformer (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017)		91.4	79.86	80.55	77.24
Point-MAE (Pang et al., 2022)	2022	Reconstruction	Transformer	ShapeNet	93.8	90.02	88.29	85.18
Point-BERT (Yu et al., 2021)	2021	Reconstruction	Transformer	ShapeNet	93.2	87.43	88.12	83.07
3D jigsaw (Sauder and Sievers, 2019b)	2019	Reconstruction	DGCNN	ShapeNet	92.4	82.0	82.1	-
MaskSurf (Zhang et al., 2022b)	2022	Reconstruction	Transformer	ShapeNet	93.40	91.22	89.17	85.81
CloudContext (Sauder and Sievers, 2019a)	2019	Reconstruction	DGCNN	ShapeNet	90.8	-	-	-
UAE (Zhang et al., 2022a)	2022	Reconstruction	DGCNN	ShapeNet	93.2	-	-	-
MD (Sun et al., 2022)	2022	Reconstruction	DGCNN	ModelNet40	93.39	-	-	-
Self-correction (Chen et al., 2021)	2021	Reconstruction	PointNet	ShapeNet	90.0	-	-	-
Self-correction (Chen et al., 2021)	2021	Reconstruction	RSCNN	ShapeNet	93.0	-	-	-
MaskPoint (Liu et al., 2022b)	2022	Reconstruction	Transformer	ShapeNet	93.8	88.1	89.3	84.3
Info3D (Sanghi, 2020)	2020	Contrast	PointNet	ShapeNet	90.20	-	-	-
Info3D (Sanghi, 2020)	2020	Contrast	DGCNN	ShapeNet	93.03	-	-	-
OcCo (Wang et al., 2021b)	2021	Alignment	PointNet	ModelNet40	90.1	-	-	-
OcCo (Wang et al., 2021b)	2021	Alignment	DGCNN	ModelNet40	93.0	82.1	83.2	-
Cross-view (Jing et al., 2021)	2021	Alignment	DGCNN	ModelNet40	93.0	82.2	83.0	-
Multi-view rendering (Tran et al., 2022)	2022	Alignment	PointNet	ModelNet40	89.5	78.5	80.5	-
Multi-view rendering (Tran et al., 2022)	2022	Alignment	DGCNN	ModelNet40	93.2	84.5	84.3	-
STRL (Huang et al., 2021)	2021	Alignment	DGCNN	ShapeNet	93.1	-	-	-

Table 9
Summary of linear classification protocol performance of representative SSL methods on ModelNet10/40 (Wu et al., 2015). The results are reported in terms of OA (%).

Method	Pretext type	Backbone	ModelNet10/40
Point-MAE (Pang et al., 2022)	Reconstruction	Transformer	- / 91.41
Orientation estimation (Poursaeed et al., 2020)	Reconstruction	PointNet	- / 88.6
Orientation estimation (Poursaeed et al., 2020)	Reconstruction	DGCNN	- / 90.75
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	PointNet	91.61 / 87.31
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	DGCNN	94.52 / 90.64
MaskSurf (Zhang et al., 2022b)	Reconstruction	Transformer	- / 92.26
CloudContext (Sauder and Sievers, 2019a)	Reconstruction	DGCNN	94.5 / 89.3
UAE (Zhang et al., 2022a)	Reconstruction	DGCNN	95.6 / 92.9
Pose Disentanglement (Tsai et al., 2022)	Reconstruction	PointNet	- / 90.1
Pose Disentanglement (Tsai et al., 2022)	Reconstruction	DGCNN	- / 92.0
CP-Net (Xu et al., 2022)	Reconstruction	RSCNN	- / 91.9
FoldingNet (Yang et al., 2018)	Reconstruction	GNN	94.4 / 88.4
Self-correction (Chen et al., 2021)	Reconstruction	PointNet	93.3 / 89.9
Self-correction (Chen et al., 2021)	Reconstruction	RSCNN	95.0 / 92.4
PC-GAN (Li et al., 2018)	Reconstruction	GAN	- / 87.5
Info3D (Sanghi, 2020)	Contrast	PointNet	- / 89.8
Info3D (Sanghi, 2020)	Contrast	DGCNN	- / 91.6
AFSRL (Lu et al., 2022)	Contrast	GNN	- / 91.5
Contrasting and clustering (Zhang and Zhu, 2019)	Contrast	DGCNN	93.8 / 86.8
Hard negatives (Du et al., 2021)	Contrast	DGCNN	- / 89.6
OcCo (Wang et al., 2021b)	Alignment	DGCNN	- / 89.2
Cross-view (Jing et al., 2021)	Alignment	GNN	- / 89.8
Multi-view rendering (Tran et al., 2022)	Alignment	PointNet	- / 89.7
Multi-view rendering (Tran et al., 2022)	Alignment	DGCNN	- / 91.7
STRL (Huang et al., 2021)	Alignment	PointNet	- / 88.3
STRL (Huang et al., 2021)	Alignment	DGCNN	- / 90.9
PRNet (Wang and Solomon, 2019)	Motion	DGCNN	- / 85.2

Table 10
Summary of performance of representative methods on part segmentation using ShapeNetPart (Armeni et al., 2016).

Method	Type	Backbone	$mIoU_C$	$mIoU_I$
Supervised	-	PointNet (Qi et al., 2017a)	83.39	83.7
		PointNet++ (Qi et al., 2017b)	81.85	85.1
		DGCNN (Wang et al., 2019)	82.33	85.2
		Transformer (Vaswani et al., 2017)	83.42	85.1
Point-MAE (Pang et al., 2022)	Reconstruction	Transformer	84.19	86.1
Point-BERT (Yu et al., 2021)	Reconstruction	Transformer	84.11	85.6
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	PointNet	-	82.2
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	DGCNN	-	85.3
MaskSurf (Zhang et al., 2022b)	Reconstruction	Transformer	84.36	86.1
CloudContext (Sauder and Sievers, 2019a)	Reconstruction	DGCNN	-	81.5
UAE (Zhang et al., 2022a)	Reconstruction	DGCNN	-	85.6
Pose Disentanglement (Tsai et al., 2022)	Reconstruction	PointNet	- /83.8	
Pose Disentanglement (Tsai et al., 2022)	Reconstruction	DGCNN	- / 85.1	
MD (Sun et al., 2022)	Reconstruction	DGCNN	-	85.5
Self-correction (Chen et al., 2021)	Reconstruction	PointNet	-	84.1
Self-correction (Chen et al., 2021)	Reconstruction	RSCNN	-	85.2
MaskPoint (Liu et al., 2022b)	Reconstruction	Transformer	84.4	86.0
AFSRL (Lu et al., 2022)	Contrast	GNN	- / 85.7	
Hard negatives (Du et al., 2021)	Contrast	DGCNN	- / 82.3	
PointContrast (Xie et al., 2020)	Contrast	U-Net	-	85.1
OcCo (Wang et al., 2021b)	Alignment	PointNet	-	83.4
OcCo (Wang et al., 2021b)	Alignment	DGCNN	-	85.0
Cross-view (Jing et al., 2021)	Alignment	DGCNN	79.1	83.7
Multi-view rendering (Tran et al., 2022)	Alignment	PointNet	-	83.3
Multi-view rendering (Tran et al., 2022)	Alignment	DGCNN	-	84.7
PRNet (Wang and Solomon, 2019)	Motion	DGCNN	78.8 / 82.5	

Table 11
Summary of performance of representative methods on semantic segmentation using S3DIS (Armeni et al., 2016).

Method	Type	Backbone	OA	mAcc	mIoU
Supervised	-	PointNet (Qi et al., 2017a)	78.6	49.0	47.7
		DGCNN (Wang et al., 2019)	84.1	-	56.1
		Transformer (Vaswani et al., 2017)	86.8	68.6	60.0
Area 5 test					
Point-MAE (Pang et al., 2022)	Reconstruction	Transformer	87.4	69.4	61.0
OcCo (Wang et al., 2021b)	Alignment	PointNet	-	83.6	44.5
OcCo (Wang et al., 2021b)	Alignment	DGCNN	-	87.0	49.5
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	PointNet	-	82.5	43.6
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	DGCNN	-	86.8	48.2
MaskSurf (Zhang et al., 2022b)	Reconstruction	Transformer	88.3	69.9	61.6
PointContrast (Xie et al., 2020)	Contrast	SR-UNet	-	77.0	70.9
Contrastive Scene Contexts (Hou et al., 2021)	Contrast	DGCNN	-	-	73.8
DepthConstrast (Zhang et al., 2021)	Contrast	PointNet++	-	72.1	64.8
Multi-view rendering (Tran et al., 2022)	Alignment	PointNet	-	85.0	46.7
Multi-view rendering (Tran et al., 2022)	Alignment	DGCNN	-	87.0	49.9
Six-fold cross validation					
OcCo (Wang et al., 2021b)	Alignment	PointNet	82.0	-	54.9
OcCo (Wang et al., 2021b)	Alignment	DGCNN	84.6	-	58.0
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	PointNet	80.1	-	52.6
3D jigsaw (Sauder and Sievers, 2019b)	Reconstruction	DGCNN	84.1	-	55.6
CloudContext (Sauder and Sievers, 2019a)	Reconstruction	DGCNN	78.9	-	47.6
Multi-view rendering (Tran et al., 2022)	Alignment	PointNet	-	83.2	52.1
Multi-view rendering (Tran et al., 2022)	Alignment	DGCNN	-	87.5	59.0

5. Future directions

Although self-supervised learning has shown great success for point cloud processing, we have identified some of its deficiencies and limitations. We argue that SSL should not be studied in isolation but rather in conjunction with advanced techniques from other domains. In this section, we discuss a number of future research directions that have the potential to improve the SSL learning capability and performance on downstream tasks.

5.1. Few-shot and zero-shot learning

There have been a good number of publicly available, labelled datasets for SSL research. However, real scenarios often face the data shortage or quality challenges, such as damaged labels, missing information, and uneven assortment. Few-shot learning (FSL) (Garcia and Bruna, 2017) is considered as a potential solution that allows the network to train under the situations with very small amount of data. It is also possible to identify new sample types that have not been seen before in a test task without training samples. This method is often referred to as the zero-shot learning (ZSL). Both SSL and FSL (ZSL) (Romera-Paredes and Torr, 2015) can free models from the reliance on large annotated datasets

Table 12
Summary of performance of representative methods on object detection using SUN RGB-D (Song et al., 2015) and ScanNet (Dai et al., 2017). The pre-training input only contains the point cloud geometry.

Method	Type	Backbone	SUN RGB-D		ScanNet	
			AP_{25}	AP_{50}	AP_{25}	AP_{50}
Point-BERT (Yu et al., 2021)	Reconstruction	Transformer	-	-	61.0	38.3
MaskPoint (Liu et al., 2022b)	Reconstruction	Transformer	-	-	64.2	42.0
PointContrast (Xie et al., 2020)	Contrast	SR-UNet	57.5	34.8	59.2	38.0
PointContrast (Xie et al., 2020)	Contrast	VoteNet	59.2	38.0	57.5	34.8
DepthConst (Zhang et al., 2021)	Contrast	PointNet++	-	-	61.3	-
DepthConst (Zhang et al., 2021)	Contrast	VoteNet	64.0	42.9	61.6	35.5
DepthConst (Zhang et al., 2021)	Contrast	H3DNet	69.0	50.0	63.5	43.4
Multi-view rendering (Tran et al., 2022)	Alignment	DGCNN	58.1	35.1	60.3	39.2
STRL (Huang et al., 2021)	Alignment	VoteNet	58.2	-	-	-

and reduce the cost. In addition, the combination of these two could potentially improve the generalization capability the models.

5.2. Multiple modality interaction and fusion

Despite of the assorted modalities in many existing datasets, for example, for outdoor autonomous driving (Geiger et al., 2012; Sun et al., 2020; Caesar et al., 2020), researchers normally only focus on and make use of the point cloud data while ignoring the connections and alignment relationships with data of other modalities. We have seen some recent research works design models (Vora et al., 2020; Wang et al., 2021a; Li et al., 2022) for multi-model data alignment and fusion, primarily point clouds and images. We anticipate more research to focus on cross-modal SSL with more diverse modalities, e.g., natural language, radar and voice, exploiting the unique characteristics of each modality and the synergy among them to build transportation systems, e.g. autonomous driving and traffic scene analysis, with more artificial general intelligence.

5.3. Hierarchical feature extraction

To cope with sophisticated downstream tasks with somehow conflicting objectives, for example, object classification which requires overall semantic understanding and part segmentation which requires fine-grained geometrical awareness, SSL models should have the capability for both global perception and local analysis. This necessitates hierarchical feature extraction; in particular, interactions between feature representations on different levels in the hierarchy need to be considered to discover the implicit relations. Therefore, we suggest that hierarchical feature extraction should be embedded in the SSL paradigm to improve the model's capability to capture both global and local features from point clouds.

5.4. Multiple tasks pre-training

Up to now, most point cloud SSL methods have only one specific pre-training pretext while few works train diverse tasks concurrently. The main resistance is that multi-tasking has to consider the compatibility and synergy between various pretexts simultaneously, and fit each loss item for steady parameter updating. This is also one of the very reasons why a model performs well on one downstream task but not others. Indeed, distinct proxies could provide useful information from various perspectives of point clouds so that jointly training multiple tasks could facilitate the network to learn more comprehensive representations; obviously, more research on multi-task SSL is needed to push the research one step further.

5.5. Theory and interpretability

Similar to traditional deep learning, point cloud SSL lacks sufficient theoretical support and has poor interpretation. The process of model training is conducted as a black-box, making it difficult for human users to analyze the results. Most of the technical works demonstrate their contributions via ablation studies and draw conclusions empirically. Such 'tried and tested' methods do not have theoretical support and are therefore difficult to verify, generalize and replicate. We suggest that future studies should include more inquiries into explainable theory, for example, the well-established theories from mutual information (Sayed, Brattoli and Ommer, 2018) or causal inference (Wang, Lin, Feng, He, Lin and Chua, 2022), which can be applied in the design of network structures and loss functions.

6. Conclusion

Point cloud self-supervised learning fundamentally moves away from models' dependency on manual annotations. The learning paradigm focuses on the design of pre-training pretext tasks to enable the models to extract effective features and achieves performance competitive to the supervised learning paradigms in many downstream tasks. This paper extensively surveys recent representative deep neural network-based methods for self-supervised learning from point cloud data. A novel taxonomy is proposed to systematically classify the current research, especially the works publishes in the recent three years. Besides detailed analysis on the representative methods, we provide summaries on the commonly used datasets and performance comparison to make the survey more comprehensive. Future research directions are also discussed to hopefully provide an insightful view on the issues that the research community should pay attention to. We hope that our work provides a valuable reference on point cloud SSL research and could motivate researchers to further explore this promising topic.

7. Acknowledgement

This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND).

References

Achituve, I., Maron, H., Chechik, G., 2021. Self-supervised learning for domain adaptation on point clouds, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 123–133.

Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K., Rodrigo, R., 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9902–9912.

Agrawal, P., Carreira, J., Malik, J., 2015. Learning to see by moving, in: Proceedings of the IEEE international conference on computer vision, pp. 37–45.

Arandjelovic, R., Zisserman, A., 2017. Look, listen and learn, in: Proceedings of the IEEE international conference on computer vision, pp. 609–617.

Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1534–1543.

Besl, P.J., McKay, N.D., 1992. Method for registration of 3-d shapes, in: Sensor fusion IV: control paradigms and data structures, Spie. pp. 586–606.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11621–11631.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A., 2020. Unsupervised learning of visual features by contrasting cluster assignments. Advances in Neural Information Processing Systems 33, 9912–9924.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al., 2015. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 .

Chen, M., Hu, Q., Hugues, T., Feng, A., Hou, Y., McCullough, K., Soibelman, L., 2022. Stpls3d: A large-scale synthetic and real aerial photogrammetry 3d point cloud dataset. arXiv preprint arXiv:2203.09065 .

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR. pp. 1597–1607.

Chen, Y., Liu, J., Ni, B., Wang, H., Yang, J., Liu, N., Li, T., Tian, Q., 2021. Shape self-correction for unsupervised point cloud understanding, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8382–8391.

Croitoru, I., Bogolin, S.V., Leordeanu, M., 2017. Unsupervised learning from video to detect foreground objects in single images, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4335–4343.

Csurka, G., 2017. Domain adaptation for visual applications: A comprehensive survey. arXiv preprint arXiv:1702.05374 .

Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5828–5839.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .

Doersch, C., Gupta, A., Efros, A.A., 2015. Unsupervised visual representation learning by context prediction, in: Proceedings of the IEEE international conference on computer vision, pp. 1422–1430.

Du, B., Gao, X., Hu, W., Li, X., 2021. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 3133–3142.

El-Sheimy, N., Li, Y., 2021. Indoor navigation: State of the art and future trends. Satellite Navigation 2, 1–23.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., . The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Faktor, A., Irani, M., 2014. Video segmentation by non-local consensus voting., in: BMVC, p. 8.

Floridi, L., Chiriatti, M., 2020. Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines 30, 681–694.

Fraser, D.A.S., 1976. Probability and statistics: Theory and applications. Technical Report.

Garcia, V., Bruna, J., 2017. Few-shot learning with graph neural networks. arXiv preprint arXiv:1711.04043 .

Geiger, A., Lenz, P., Urtasun, R., 2012. Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE. pp. 3354–3361.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations. international conference on learning representations .

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Advances in neural information processing systems 27.

Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems 33, 21271–21284.

Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M., 2021. Pct: Point cloud transformer. Computational Visual Media 7, 187–199.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M., 2017. Semantic3d. net: A new large-scale point cloud classification benchmark. arXiv preprint arXiv:1704.03847 .

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2021. Masked autoencoders are scalable vision learners. arXiv preprint arXiv:2111.06377 .

He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9729–9738.

Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., Svensson, L., 2022. Masked autoencoders for self-supervised learning on automotive point clouds. arXiv preprint arXiv:2207.00531 .

Hou, J., Graham, B., Nießner, M., Xie, S., 2021. Exploring data-efficient 3d scene understanding with contrastive scene contexts, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15587–15597.

Hu, Q., Yang, B., Khalid, S., Xiao, W., Trigoni, N., Markham, A., 2021. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4977–4987.

Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K., 2016. Scenenn: A scene meshes dataset with annotations, in: 2016 fourth international conference on 3D vision (3DV), Ieee. pp. 92–101.

Huang, S., Xie, Y., Zhu, S.C., Zhu, Y., 2021. Spatio-temporal self-supervised representation learning for 3d point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6535–6545.

Jayaraman, D., Grauman, K., 2015. Learning image representations tied to ego-motion, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1413–1421.

Jiang, H., Larsson, G., Shakhnarovich, M.M.G., Learned-Miller, E., 2018. Self-supervised relative depth learning for urban scene understanding, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 19–35.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence 43, 4037–4058.

Jing, L., Zhang, L., Tian, Y., 2021. Self-supervised feature learning by cross-modality and cross-view correspondences, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1581–1591.

Koch, S., Matveev, A., Jiang, Z., Williams, F., Artemov, A., Burnaev, E., Alexa, M., Zorin, D., Panozzo, D., 2019. Abc: A big cad model dataset for geometric deep learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9601–9611.

Lal, S., Prabhudesai, M., Mediratta, I., Harley, A.W., Fragkiadaki, K., 2021. Coconets: Continuous contrastive 3d scene representations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12487–12496.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L., 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461 .

Li, C.L., Zaheer, M., Zhang, Y., Poczos, B., Salakhutdinov, R., 2018. Point cloud gan. arXiv preprint arXiv:1810.05795 .

Li, R., Li, X., Fu, C.W., Cohen-Or, D., Heng, P.A., 2019. Pu-gan: a point cloud upsampling adversarial network, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 7203–7212.

Li, R., Lin, G., Xie, L., 2021. Self-point-flow: Self-supervised scene flow estimation from point clouds with optimal transport and random walk, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15577–15586.

Li, Y., Ma, L., Zhong, Z., Liu, F., Cao, D., Li, J., Chapman, M.A., 2020. Deep learning for lidar point clouds in autonomous driving: A review. IEEE Transactions on Neural Networks .

Li, Y., Yu, A.W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q.V., et al., 2022.

Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17182–17191.

Liu, D., Chen, C., Xu, C., Qiu, R., Chu, L., 2022a. Self-supervised point cloud registration with deep versatile descriptors. arXiv preprint arXiv:2201.10034 .

Liu, H., Cai, M., Lee, Y.J., 2022b. Masked discrimination for self-supervised learning on point clouds. arXiv preprint arXiv:2203.11183 .

Liu, H., Yuan, H., Hou, J., Hamzaoui, R., Gao, W., 2022c. Pufa-gan: A frequency-aware generative adversarial network for 3d point cloud upsampling. IEEE Transactions on Image Processing 31, 7389–7402.

Liu, X., Liu, X., Liu, Y.S., Han, Z., 2022d. Spu-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization. IEEE Transactions on Image Processing 31, 4213–4226.

Liu, Y., Yi, L., Zhang, S., Fan, Q., Funkhouser, T., Dong, H., 2020. P4Contrast: Contrastive Learning with Pairs of Point-Pixel Pairs for RGB-D Scene Understanding. arXiv e-prints , arXiv:2012.13089doi:10.48550/arXiv.2012.13089, arXiv:2012.13089.

Lu, Z., Dai, Y., Li, W., Su, Z., 2022. Joint data and feature augmentation for self-supervised representation learning on point clouds. arXiv preprint arXiv:2211.01184 .

Matti, D., Ekenel, H.K., Thiran, J.P., 2017. Combining lidar space clustering and convolutional neural networks for pedestrian detection, in: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. doi:10.1109/AVSS.2017.8078512.

Mersch, B., Chen, X., Behley, J., Stachniss, C., 2022. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks, in: Conference on Robot Learning, PMLR. pp. 1444–1454.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 .

Miller, G.A., 1995. Wordnet: a lexical database for english. Communications of the ACM 38, 39–41.

Min, C., Zhao, D., Xiao, L., Nie, Y., Dai, B., 2022. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. arXiv preprint arXiv:2206.09900 .

Mittal, H., Okorn, B., Held, D., 2020. Just go with the flow: Self-supervised scene flow estimation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11177–11185.

Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H., 2019. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 909–918.

Noroozi, M., Favaro, P., 2016. Unsupervised learning of visual representations by solving jigsaw puzzles, in: European conference on computer vision, Springer. pp. 69–84.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 .

Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L., 2022. Masked autoencoders for point cloud self-supervised learning. arXiv preprint arXiv:2203.06604 .

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. computer vision and pattern recognition .

Poursaeed, O., Jiang, T., Qiao, H., Xu, N., Kim, V.G., 2020. Self-supervised learning of point clouds via orientation estimation, in: 2020 International Conference on 3D Vision (3DV), IEEE. pp. 1018–1028.

Qi, C.R., Su, H., Mo, K., Guibas, L.J., 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 652–660.

Qi, C.R., Yi, L., Su, H., Guibas, L.J., 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems 30.

Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning, in: International conference on machine learning, PMLR. pp. 2152–2161.

Sanghi, A., 2020. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning, in: European Conference on Computer Vision, Springer. pp. 626–642.

Sariyildiz, M.B., Kalantidis, Y., Alahari, K., Larlus, D., 2022. Improving the generalization of supervised models. arXiv preprint arXiv:2206.15369 .

Sarmad, M., Lee, H.J., Kim, Y.M., 2019. Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5898–5907.

Sauder, J., Sievers, B., 2019a. Context prediction for unsupervised deep learning on point clouds. arXiv preprint arXiv:1901.08396 2, 5.

Sauder, J., Sievers, B., 2019b. Self-supervised deep learning on point clouds by reconstructing space. Advances in Neural Information Processing Systems 32.

Sayed, N., Brattoli, B., Ommer, B., 2018. Cross and learn: Cross-modal self-supervision, in: German Conference on Pattern Recognition, Springer. pp. 228–243.

Shi, Y., Cao, X., Zhou, B., 2021. Self-supervised learning of part mobility from point cloud sequence, in: Computer Graphics Forum, Wiley Online Library. pp. 104–116.

Shu, D.W., Park, S.W., Kwon, J., 2019. 3d point cloud generative adversarial network based on tree structured graph convolutions, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 3859–3868.

Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 567–576.

Stretcu, O., Leordeanu, M., 2015. Multiple frames matching for object discovery in video., in: BMVC, p. 3.

Sun, C., Zheng, Z., Wang, X., Xu, M., Yang, Y., 2022. Self-supervised point cloud representation learning via separating mixed shapes. IEEE Transactions on Multimedia .

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al., 2020. Scalability in perception for autonomous driving: Waymo open dataset, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2446–2454.

Sun, X., Wu, J., Zhang, X., Zhang, Z., Zhang, C., Xue, T., Tenenbaum, J.B., Freeman, W.T., 2018. Pix3d: Dataset and methods for single-image 3d shape modeling, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2974–2983.

Taghanaki, S.A., Luo, J., Zhang, R., Wang, Y., Jayaraman, P.K., Jatavallabhula, K.M., 2020. Robustpointset: A dataset for benchmarking robustness of point cloud classifiers. arXiv preprint arXiv:2011.11572 .

Tran, B., Hua, B.S., Tran, A.T., Hoai, M., 2022. Self-supervised learning with multi-view rendering for 3d point cloud analysis, in: Proceedings of the Asian Conference on Computer Vision, pp. 3086–3103.

Tsai, M.S., Chiang, P.Z., Tsai, Y.H., Chiu, W.C., 2022. Self-supervised feature learning from partial point clouds via pose disentanglement, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 1031–1038.

Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K., 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 1588–1597.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Velickovic, P., Fedus, W., Hamilton, W.L., Liò, P., Bengio, Y., Hjelm, R.D., 2019. Deep graph infomax. ICLR (Poster) 2, 4.

Vora, S., Lang, A.H., Helou, B., Beijbom, O., 2020. Pointpainting: Sequential fusion for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 4604–4612.

Wang, C., Ma, C., Zhu, M., Yang, X., 2021a. Pointaugmenting: Cross-modal augmentation for 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11794–11803.

Wang, H., Liu, Q., Yue, X., Lasenby, J., Kusner, M.J., 2021b. Unsupervised point cloud pre-training via occlusion completion, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9782–9792.

Wang, H., Yang, L., Rong, X., Feng, J., Tian, Y., 2021c. Self-supervised 4d spatio-temporal feature learning via order prediction of sequential point cloud clips, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3762–3771.

Wang, W., Lin, X., Feng, F., He, X., Lin, M., Chua, T.S., 2022. Causal representation learning for out-of-distribution recommendation, in: Proceedings of the ACM Web Conference 2022, pp. 3562–3571.

Wang, Y., Solomon, J.M., 2019. Prnet: Self-supervised learning for partial-to-partial registration. Advances in neural information processing systems 32.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. Acm Transactions On Graphics (tog) 38, 1–12.

Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K., 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE. pp. 4376–4382.

Wu, W., Wang, Z.Y., Li, Z., Liu, W., Fuxin, L., 2020. Pointpwc-net: Cost volume on point clouds for (self-) supervised scene flow estimation, in: European conference on computer vision, Springer. pp. 88–107.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1912–1920.

Wu, Z., Xiong, Y., Yu, S.X., Lin, D., 2018. Unsupervised feature learning via non-parametric instance discrimination, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3733–3742.

Xiang, Y., Kim, W., Chen, W., Ji, J., Choy, C., Su, H., Mottaghi, R., Guibas, L., Savarese, S., 2016. Objectnet3d: A large scale database for 3d object recognition, in: European conference on computer vision, Springer. pp. 160–176.

Xiao, A., Huang, J., Guan, D., Lu, S., 2022. Unsupervised representation learning for point clouds: A survey. arXiv preprint arXiv:2202.13589 .

Xie, S., Gu, J., Guo, D., Qi, C.R., Guibas, L., Litany, O., 2020. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding, in: European conference on computer vision, Springer. pp. 574–591.

Xie, S., Liu, S., Chen, Z., Tu, Z., 2018. Attentional shapecontextnet for point cloud recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4606–4615.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., Zhuang, Y., 2019. Self-supervised spatiotemporal learning via video clip order prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10334–10343.

Xu, M., Zhou, Z., Xu, H., Wang, Y., Qiao, Y., 2022. Cp-net: Contour-perturbed reconstruction network for self-supervised point cloud learning. arXiv preprint arXiv:2201.08215 .

Yang, J., Alvarez, J.M., Liu, M., 2021. Self-supervised learning of depth inference for multi-view stereo, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7526–7534.

Yang, Y., Feng, C., Shen, Y., Tian, D., 2018. Foldingnet: Point cloud auto-encoder via deep grid deformation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 206–215.

Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S., 2019. Reppoints: Point set representation for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 9657–9666.

Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J., 2021. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. arXiv preprint arXiv:2111.14819 .

Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S., 2018. Taskonomy: Disentangling task transfer learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3712–3722.

Zhang, C., Shi, J., Deng, X., Wu, Z., 2022a. Upsampling autoencoder for self-supervised point cloud learning. arXiv preprint arXiv:2203.10768 .

Zhang, L., Zhu, Z., 2019. Unsupervised feature learning for point cloud understanding by contrasting and clustering using graph convolutional neural networks, in: 2019 international conference on 3D vision (3DV), IEEE. pp. 395–404.

Zhang, Y., Lin, J., He, C., Chen, Y., Jia, K., Zhang, L., 2022b. Masked surfel prediction for self-supervised point cloud learning. arXiv preprint arXiv:2207.03111 .

Zhang, Y., Lin, J., Li, R., Jia, K., Zhang, L., 2022c. Pointdae: Denoising autoencoders for self-supervised point cloud learning. arXiv preprint arXiv:2211.06841 .

Zhang, Z., Girdhar, R., Joulin, A., Misra, I., 2021. Self-supervised pretraining of 3d features on any point-cloud, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10252–10263.

Zhao, W., Liu, X., Zhong, Z., Jiang, J., Gao, W., Li, G., Ji, X., 2022a. Self-supervised arbitrary-scale point clouds upsampling via implicit neural representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1999–2007.

Zhao, X., Yang, S., Huang, T., Chen, J., Ma, T., Li, M., Liu, Y., 2022b. Superline3d: Self-supervised line segmentation and description for lidar point cloud, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, Springer. pp. 263–279.

Zhao, Y., Hui, L., Xie, J., 2021. Sspu-net: Self-supervised point cloud upsampling via differentiable rendering, in: Proceedings of the 29th ACM International Conference on Multimedia, pp. 2214–2223.

Zhou, Y., Tuzel, O., 2018. Voxelnet: End-to-end learning for point cloud based 3d object detection. computer vision and pattern recognition .