Vicinagearth

REVIEW **Open Access**

Advances in 3D pre-training and downstream tasks: a survey

Yuenan Hou[1*], Xiaoshui Huang[1], Shixiang Tang[1], Tong He[1] and Wanli Ouyang[1]

Abstract

Recent years have witnessed a signifcant breakthrough in the 3D domain. To track the most recent advances in the 3D field, in this paper, we provide a comprehensive survey of recent advances in the 3D feld, which encompasses a wide collection of topics, including diverse pre-training strategies, backbone designs and downstream tasks. Compared to the previous literature review on point cloud, our survey is more comprehensive. Our survey consists of the 3D pre-training methods, various downstream tasks, popular benchmarks, evaluation metrics as well as several promising future directions. We hope the survey can serve as the cornerstone for both academia and industry.

Keywords 3D, Advances, Pre-train, Downstream tasks

1 Introduction

Recent years have witnessed a significant breakthrough in the 3D domain [1]. Compared with the 2D image, the 3D data can more precisely describe the real world as it contains accurate 3D measurement. Several primary 3D representations [2], including point clouds, voxels, depth/normal maps, neural fields and meshes, have been proposed and employed in various competitive algorithms [3–5]. Till now, a survey that summarizes the most recent advances in the 3D field is not available.

To track the most recent trends in the 3D field, in this paper, we provide a comprehensive survey of recent advances in the 3D field, which encompasses a wide collection of topics, including diverse pre-training strategies, backbone designs and downstream tasks. Compared to the previous literature review on point cloud [1], our survey is more comprehensive. Our survey consists of the 3D pre-training methods, various downstream tasks, popular benchmarks, evaluation metrics as well as several promising future directions. Specifically, [1] contains point cloud classification, detection, tracking and segmentation while ours also includes matching and registration. Our survey provides a thorough review of prevalent 3D pre-training paradigms which can benefit various downstream tasks. We also point out several promising future directions worth exploring in the 3D field. We hope the survey can serve as the cornerstone for both academia and industry.

The remainder of the survey is organized as below: in Section 2, we will first have a brief overview of the basic terminology and frequently used techniques of the 3D domain. Then, we will present the pre-training methods, downstream tasks, benchmarks and evaluation criterion in Sections 3, 4, 5, and 6 in order. Eventually, future work and conclusion will be provided in Sections 7 and 8, respectively.

The contributions of our survey are summarized as follows:

- To our knowledge, we provide the most comprehensive survey of recent advances in the 3D field.
- Our survey consists of various 3D pre-training algorithms, diverse downstream tasks, mainstream benchmarks and primary evaluation metrics.

*Correspondence:
Yuenan Hou
houyuenan@pjlab.org.cn
[1] Shanghai AI Laboratory, 701 Yunjin Road in Xuhui District, Shanghai 200000, China

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Relevance to Vicinagearth. The earth we are living is in the 3D space. Development in the 3D field will inevitably affect the lives of many humans as it can greatly facilitate the deployment of the advanced computer vision algorithms in the real-world applications. Therefore, making a comprehensive summary of recent advances in the 3D field is vital to both industry and academia of the Vicinagearth community.

2 Preliminary

In this section, we will introduce some common terminology, including 3D representations, sparse convolution, transformer, coordinate system, etc.

2.1 3D representations

Point clouds, voxels, depth/normal maps, neural fields and meshes are primary 3D representations used by contemporary perception and generation algorithms. Each representation has its own pros and cons. We summarize the comparison between these 3D representations in terms of computation efficiency, storage efficiency and representation capability in Table 1. Selecting a proper representation is important for the task at hand. In the following sentences, we will briefly introduce these representations.

Point Clouds. Point clouds are sets of points in the 3D space that represent the surface or structure of an object or scene. Other properties, such as color and intensity, can also be provided in addition to the 3D positions. Point clouds are usually obtained by depth sensors and they are widely used in diverse 3D tasks. However, these point clouds are irregular and thus are hard to be directly processed by conventional neural networks that are designed for regular and structured data such as images. To efficiently process these irregular points, sparse convolution is proposed and universally employed in modern 3D networks.

Voxels. Voxel grids represent the 3D space as a collection of regular grids of volumetric elements called voxels that are akin to pixels in the 2D space. They can store various attributes, such as occupancy or color. Owing to the regularity of voxel grids, they can be directly processed by standard convolution networks. Voxels are widely adopted in scenarios that require volumetric representations, e.g., medical imaging, 3D printing and simulations.

Depth Maps. They are also known as depth images. The depth map encodes the depth or distance information of a scene or object with respect to a reference point. Each pixel in the depth map stores a value that represents the distance from the camera or the reference point to the corresponding point in the scene. They provide valuable information for various real-world applications, such as 3D reconstruction, 3D tracking and depth-based rendering.

Normal Maps. Normal maps encode surface normals of the object's surface. Surface normals represent the direction perpendicular to the surface at a particular point and are vital to realistic shading and lighting calculations in computer graphics.

Meshes. Polygonal meshes are one of the most common representations in 3D graphics and computer vision. They are composed of a collection of polygons, such as triangles or quadrilaterals, that describe the surface of an object. Each polygon is defined by its vertices and their connectivity, forming a mesh structure. The explicit connectivity information provided by meshes is beneficial to the relationship modeling among points. Polygonal meshes can represent both the shape and fine-grained details of an object.

Neural Fields. It is a continuous neural implicit representation. The neural network is used to map the features (e.g., 3D position) to attributes (e.g., color). As to the storage cost, it is much cheaper since only the network parameters are required to be stored. Surface rendering and volume rendering are two primary techniques that render an image from a neural field.

2.2 Sparse convolution

As opposed to 2D images that are dense, compact, and have regular spatial resolutions, 3D point cloud is sparse and unordered. Since convolution can only process signals that have regular shapes, voxelization [6] is devised which divides the 3D space into many small cubes (*i.e.*, voxels) and the information of points within the same voxel is aggregated by max pooling or average pooling. After the voxelization operation, it is natural to lift vanilla 2D convolution to the 3D convolution and then process

Table 1 Comparison between different 3D representations in terms of computation efficiency, storage efficiency and representation capability

3D Representation	Point Clouds	Voxels	Depth/Normal Maps	Neural Fields	Mesh
Computation Efficiency	***	*	*****	*	****
Storage Efficiency	**	**	****	*****	***
Representation Capability	***	**	*	*****	****

More * means better performance

the point cloud with 3D convolution. However, such practice will cause enormous computation cost when the quantity of point cloud is huge, which is intolerable in the real-world application. To efficiently extract features from the large amount of point clouds, sparse convolution is proposed (a.k.a spconv) and applied to perform 3D detection from point cloud [7]. The core idea of sparse convolution is to perform convolution merely on non-empty regions. Under this circumstance, the computation cost on large and empty areas is remarkably reduced.

Albeit the efficient computation of sparse convolution, it suffers from the dilation problem that will break the sparsity of the original input signal, as observed in [8]. To resolve the above-mentioned dilation issue, sub-manifold sparse convolution is put forward that performs convolution when the center of the convolution kernel lies in the non-empty region. Although spconv will lead to dilation problem, the dilation property of spconv is also beneficial to endow the deep model with the contextual information around the non-empty grids. Therefore, sparse convolution and sub-manifold sparse convolution are both employed in many modern 3D sparse convolution networks.

2.3 Other terminology

Transformer. The advent of transformer [9] has revolutionized many natural language and computer vision fields. Self attention and cross attention are two primary operators in the transformer. Self attention can enhance the features itself whilst cross attention is usually applied to achieve information enhancement between different features. Query, Key and Value are three widely used items in the self/cross attention operations.

Indoor v.s. outdoor. The point cloud of the indoor task is usually dense, has uniform density and small scale. Take ScanNet as example. The data is captured from the RGB-D camera and the room scan is obtained through reconstruction. However, in the outdoor scenarios, the point cloud is sparse, large scale and has varying density, which poses great challenges for both discriminative and generative models.

2.4 Coordinate system

Coordinate system plays a pivotal role in the 3D field. For indoor tasks, RGB-D data captured by RGB-D cameras at different views should be transformed into the same coordinate system. We can either take the world coordinate or the coordinate system of cameras at a particular view as the reference system. For outdoor tasks, it mainly involves the coordinate system of different sensors (such as LiDAR, cameras and Radar), world coordinate, ego-vehicle coordinate and other-vehicle coordinate (vehicle-road coordination). To transform between different coordinates, we can utilize sensor extrinsic and intrinsic parameters. For fusing multiple point cloud frames, we can first transform point clouds of other frames to the world coordinate and then from world coordinate to the reference frame coordinate. In this condition, we can directly concatenate these point cloud frames and obtain the multi-scan point clouds.

3 Pre-training

Pre-training aims to learn effective and general representation that can be smoothly transferred to various downstream tasks. For example, ImageNet pre-trained models have gained immense popularity due to their versatility and efficacy in various applications. The key to their success lies in the ImageNet dataset's vast and diverse collection of over 14 million images, spanning thousands of categories. This rich dataset provides a comprehensive base for training, enabling these models to learn a wide array of features and patterns. However, due to the lack of labeled data, most methods in 3D adopt a self-supervised manner. Like supervised learning, it uses a form of labeling, but these labels are created from the data without human intervention, akin to unsupervised learning. The key idea is to design a pretext task. According to the types of tasks, self-pretraining in 3D can be categorized into contrastive-based [10, 11], MAE-based [12, 13], and rendering-based [14, 15].

3.1 Contrastive

Contrastive Learning is a technique in self-supervised machine learning that focuses on differentiating between similar and dissimilar data pairs. It operates by mapping input data into an embedding space, where the model is trained to minimize the distance between positive pairs and maximize it between negative pairs, typically using loss functions like InfoNCE loss [16]. This approach, commonly used in fields like 2D/3D representation learning, improves the generalization and robustness of models, making them effective even with less labeled data. However, the performance for contrastive learning is often contained by sample selection, especially for negative pairs. Conventional contrastive learning for 3D modality can be roughly categorized into three classes: 3D-to-3D [10, 11], 3D-to-2D [17, 18], and 3D-to-text [19–21].

3.1.1 3D-to-3D

Xie et al. [10] proposed PointContrast, the pipeline of which is shown in Fig. 1. It encourages the network to learn effective representations that are invariant to different view transformations and noise, which are critical in 3D perception tasks. It enhances the performance of various downstream tasks like object detection and

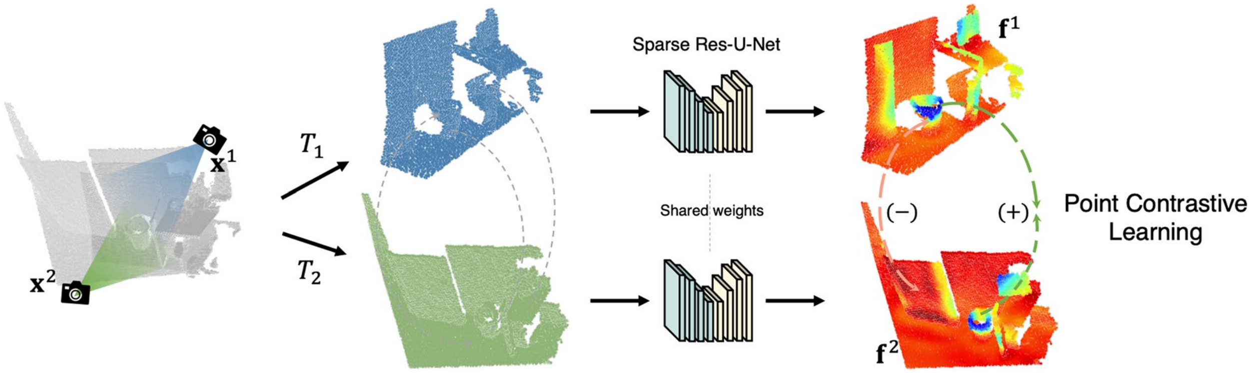

Fig. 1 PointContrast allows the network to learn invariant features under various 3D transformations

segmentation in 3D scenes. Hou et al. [11] improved the representation learning by designing a novel method to select negative pairs, which are distributed with different context surroundings. The method also proved its effectiveness with both limited annotations and scene numbers. SegContrast [22] learns powerful representations by contrasting features at the level of cluster, which is obtained by using DBSCAN.

3.1.2 3D-to-2D

CrossPoint [17] facilitates a 3D-to-2D correspondence by bridging the gap between the 3D point cloud and their rendered images. In addition, it also introduces an intramodal correspondence by imposing invariance to point cloud augmentations. SLidR [18] proposed to encode the knowledge from a fixed 2D backbone to a 3D backbone. The 2D network is pretrained with a large amount of image data. It also utilized superpixel to group pixels with close visual clues.

3.1.3 3D-to-text

Inspired by CLIP [19], many methods have been proposed to bridge the gap between language understanding and 3D visual perception. One of the key strengths of CLIP is its ability to generalize from the training data to a wide variety of visual tasks without task-specific training data. PointCLIP [20] proposed to obtain multiview images by projecting and encode rich 2D knowledge by utilizing a pre-trained 2D CLIP model. A global representation is extracted and forced to align with text embeddings. In doing so, zero-shot perception tasks can be achieved by retrieving the text representation. PointCLIPV2 [21] greatly boosts the accuracy by applying a more realistic render to produce depth images.

3.2 Masked modeling based pretraining methods

Inspired by the great success of masked image modeling in large visual pretraining on images, pretraining by masked autoencoder has emerged in the field of 3D perception for indoor and outdoor scenarios, including both multiview images and point clouds. The core idea of this paradigm is to pretrain perception models by reconstructing the input data from the extracted representations by the masked, corrupted, or noisy inputs. The schematic illustration is shown in Fig. 3. However, due to the sparse and unstructured nature of the point cloud, it is still challenging to apply MAE for 3D pretraining.

VoxelMAE transfers unstructured point sets to structured voxels. The objective is to predict the occupancy of the masked voxel. Inspired by the BERT model, PointMAE [23] first tokenized point cloud by sampling a set of key points via farthest point sampling(FPS). The feature of each token is obtained by grouping neighboring points collected by KNN. An L2 loss for predicting the features of masked tokens is used for pretraining. PointM2AE further boosts the features by adapting both the encoder and decoder into pyramid architectures. This modification allows for the progressive modeling of spatial geometries, enabling the capture of intricate details as well as high-level semantic information of 3D shapes. With the pretrained backbone fixed, the linear classification achieves SOTA on the ModelNet40 dataset, demonstrating the effectiveness of the pretraining method. GD-MAE [12] extended the idea to a convolutional backbone, which is still widely used in self-driving scenarios. The method devised a novel masking strategy that effectively prevents knowledge leakage during the downsampling stage.

3.3 Rendering for pretrain

Differentiable rendering refers to the process of rendering images from 3D models in such a way that the rendering process itself is differentiable, meaning the gradients of certain output image parameters can be computed with respect to the input 3D model parameters, enabling gradient-based optimization techniques to be used. Differentiable rendering is instrumental in tasks like 3D reconstruction from 2D images, photorealistic image synthesis, and enhancing the realism in augmented reality applications. It's increasingly vital due to its ability to bridge the gap between 3D models and 2D image analysis, driving advancements in both graphics and machine learning domains. Ponder [14] describes the 3D surface in a volume and optimizes the 3D using the 2D image projection, with the help of the camera parameters. The whole framework uses a NeRF-like structure and the process is differentiable, as shown in Fig. 2. PonderV2 [15] extends the idea with more data and has been successfully applied to outdoor cases. Not only can it be applied to pretrain a 3D backbone, but can be used for a 2D backbone. In doing so, it shows superior performance compared with traditional contrastive- and MAE-based methods (Fig. 3).

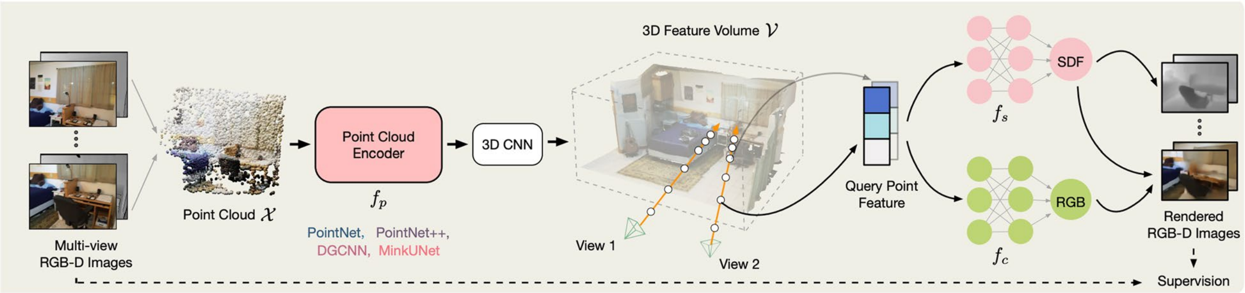

Fig. 2 The framework of Ponder, which is a classic rendering-based pipeline

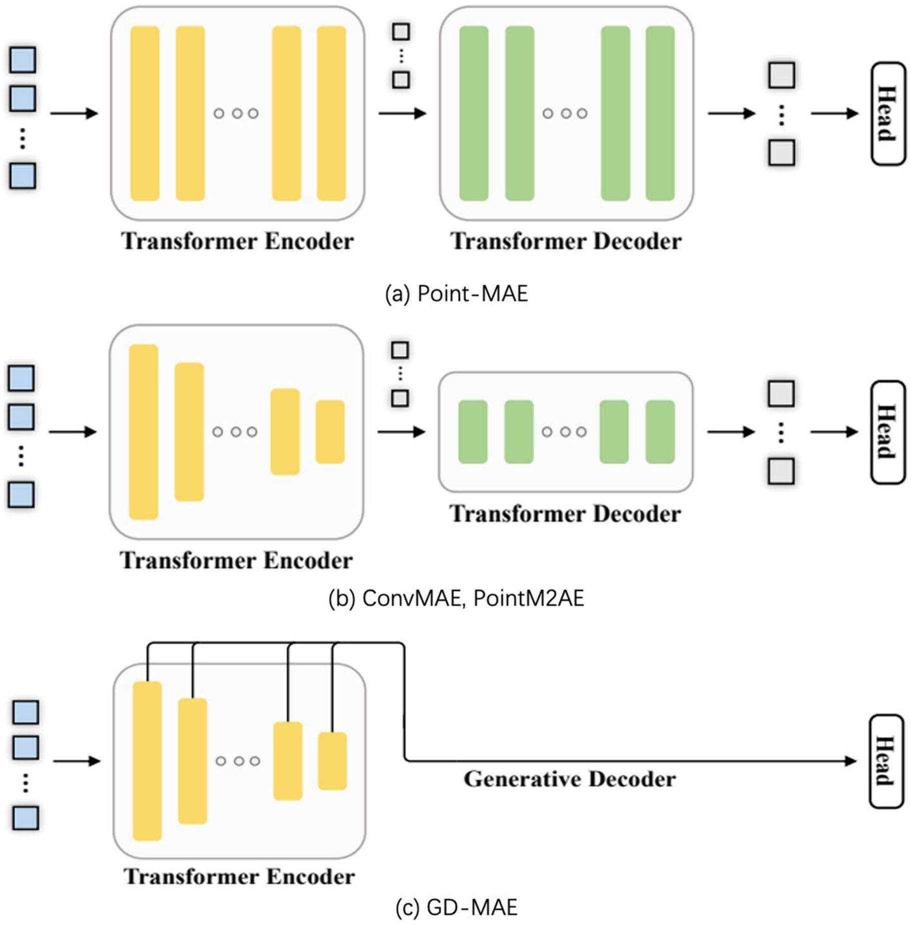

Fig. 3 Summary of the MAE pipelines

3.4 A brief summary

The selection of the 3D pre-training algorithms heavily depends on the type of the downstream tasks. Specifically, many factors, such as the sample quantity, scene diversity, dataset scale, range and point density, have a non-negligible effect on the selection of the pre-training paradigm. Since the quantity and diversity of the 3D data is relatively limited, leveraging the rich information from other modalities is also beneficial in the pre-training stage.

4 Downstream tasks

In this section, we provide the advances in different downstream tasks, including classification, segmentation, detection, tracking, matching and registration.

4.1 Classification

Point cloud classification is the most fundamental task for point cloud understanding and requires the deep model to estimate the object category from the given point cloud frame / room scan. PointNet [3] makes the first attempt to apply deep learning techniques to achieve point cloud classification (Fig. 4). PointNet++ [4] builds the hierarchical architecture upon PointNet to enhance the performance. Many subsequent works [24–28] have been built upon the main idea of PointNet and achieve impressive classification performance. For instance, DGCNN [28] designs the EdgeConv that explicitly constructs a local graph and learns the embeddings for the edges. The detailed computation of the EdgeConv is shown in Fig. 5.

4.2 Segmentation

The objective of LiDAR segmentation is to assign a category to each point of the input point cloud sequence. Point, voxel and range images are three typical representations for LiDAR segmentation. PointNet [3] is the seminal work that applies a shared Multi-Layer Perceptron (MLP) to perform point cloud understanding directly from raw point cloud. MinkowskiUNet [29] employs the voxel representation and designs the U-Net [30] architecture for the semantic scene understanding task, as shown in Fig. 6. SPVCNN [31] introduces an additional point branch for the voxel-based MinkowskiUNet to compensate for the missing information of the original points. Cylinder3D [32] replaces the cubic partition with

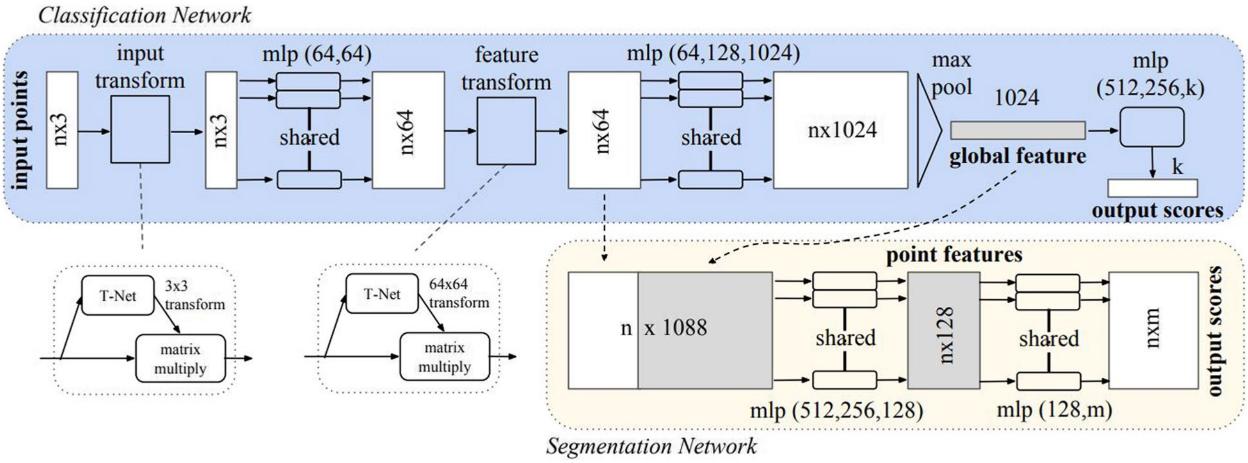

Fig. 4 Framework overview of PointNet, which is the pioneering work in point cloud understanding

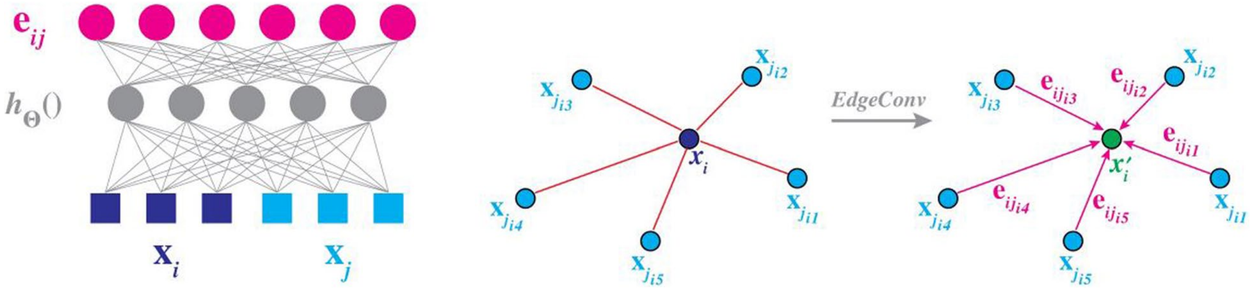

Fig. 5 Illustration of EdgeConv proposed by DGCNN that is the pioneering work in graph-based point cloud understanding

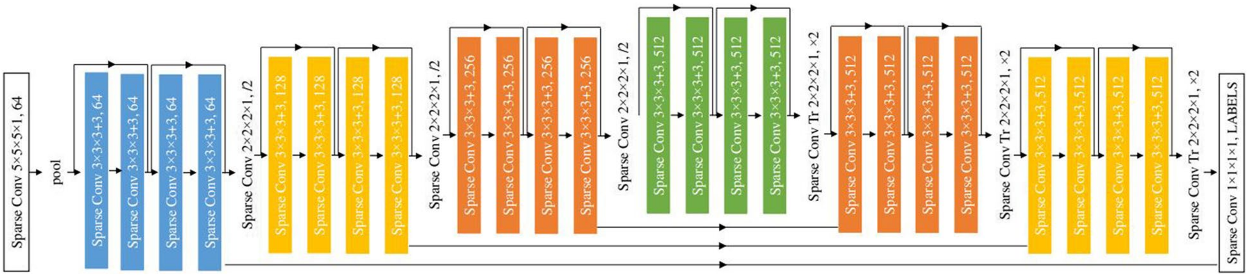

Fig. 6 Architecture of MinkowskiUNet-32, whose variants serve as the backbone for many competitive LiDAR segmentation models

the cylindrical partition to better utilize the property of varying density of the point clouds, as shown in Fig. 7. SphereFormer [33] devises the transformer-based network to leverage its global receptive field. Another line of methods concentrate on the range view representation as it is dense, compact and can resort to the contemporary and top-performing 2D segmentors owing to its 2D form. RangeViT [34] and RangeFormer [35] introduces the powerful transformer architecture for range-view-based segmentation. RPVNet [36] fully utilizes the information of range, point and voxel representations to yield the segmentation results.

Albeit that point cloud can provide accurate spatial positions of objects in the 3D, it lacks color and texture that are critical to the categorical recognition of objects. To compensate for the shortcoming, the attention of both academic and industrial communities has been shifted to the combination of the point cloud and RGB images, forming the multi-modal fusion methods [37, 38]. Since multi-modal fusion utilizes the strength of both worlds, it can benefit more robust and comprehensive perception of the surrounding environment. For instance, UniSeg [37] takes three point cloud representations and the images as input, and designs the learnable cross-modal fusion and cross-view fusion modules to fully leverage the valuable information in these input signals.

Recent years, there emerges a trend that builds the foundation model to tackle diverse tasks with single architecture. Point Transformer V3 [39] is the pioneering work that achieves impressive indoor and outdoor, segmentation and detection tasks simultaneously. It is built upon the transformer architecture and overcomes the heavy computation cost introduced by the attention calculation using the efficient serialized neighbor searching mechanism.

4.3 Detection

3D detection aims to estimate the 3D spatial locations and categories of objects in the 3D space. According to the type of the input signal, contemporary 3D detection

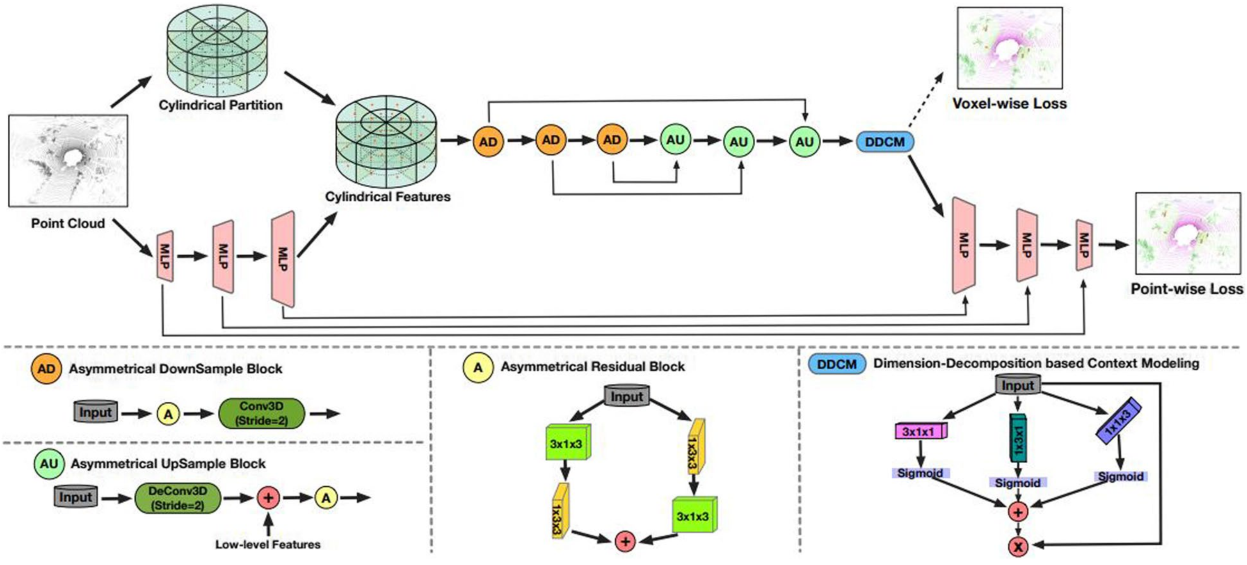

Fig. 7 Overview of Cylinder3D

algorithms can be categorized into the following groups, *i.e.*, LiDAR-based, image-based and multi-modal fusion methods. As to LiDAR-based detectors, they usually adopt the point [5], voxel [6], point-voxel [40] and range image representation [41]. Point-based detectors, such as PointRCNN [5], directly estimate the 3D spatial positions and categories of objects from the point cloud, as shown in Fig. 8. Because handling tens of thousands of points is tedious and time-consuming, the voxel representation is presented in VoxelNet, where unordered points are rasterized into a fixed number of regular cubes (a.k.a voxels). To reduce the information loss caused by the rasterization process, point-voxel-based detectors, e.g., PVRCNN [40], are designed to harness the strength of both representations.

Since the collection and annotation of LiDAR data is expensive, recent trends favour the image-based detector [42]. Compared to the point cloud, images are cheaper and contain rich color and texture information. As a representative, BEVFormer [43], which follows the detection pipeline of Tesla, achieves impressive performance for multi-view detection and serves as the cornerstone of many subsequent works (Fig. 9). BEVFormer follows the top-down pipeline that takes the BEV features as query and employs the cross-attention mechanism to extract useful information from the front-view features.

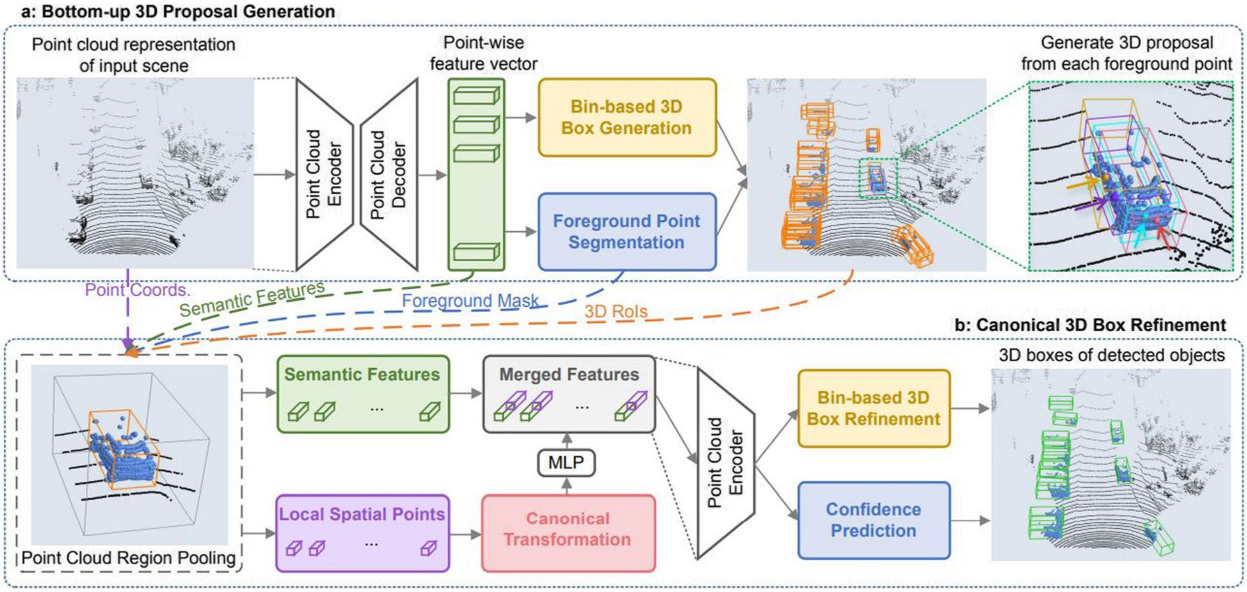
a: Bottom-up 3D Proposal Generation
Point cloud representation of input scene
Point Cloud Encoder
Point Cloud Decoder
Point-wise feature vector
Bin-based 3D Box Generation
Foreground Point Segmentation
Generate 3D proposal from each foreground point
Point Coords.
Semantic Features
Foreground Mask
3D RoIs
b: Canonical 3D Box Refinement
Point Cloud Region Pooling
Semantic Features
Merged Features
MLP
Local Spatial Points
Canonical Transformation
Point Cloud Encoder
Bin-based 3D Box Refinement
Confidence Prediction
3D boxes of detected objects

Fig. 8 Overall architecture of PointRCNN

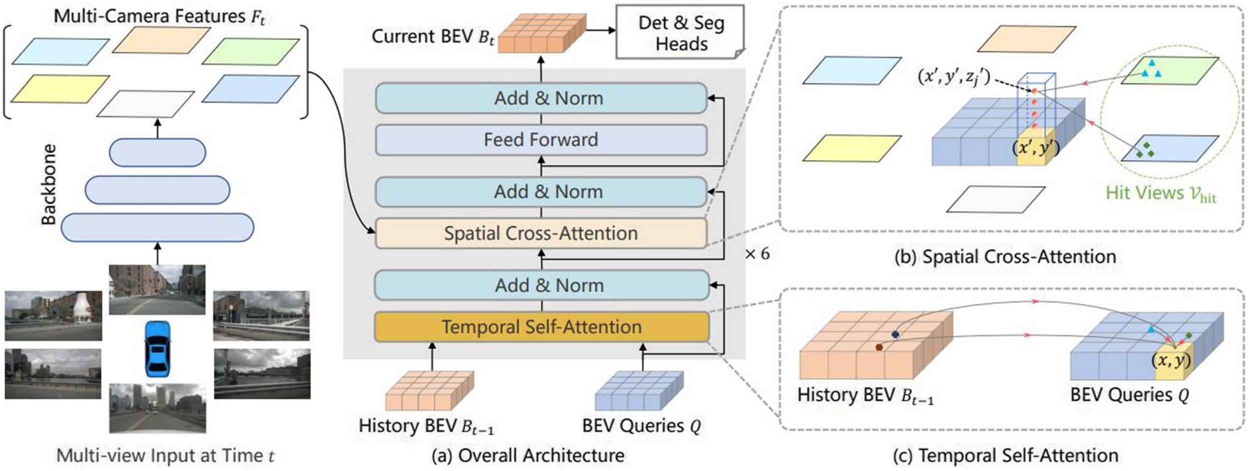

Fig. 9 Overall architecture of BEVFormer

However, considering that images lack precise spatial measurement compared with the point cloud counterpart, there emerges surging interest in the combination of both point cloud and images, forming the multi-modal fusion detectors. PointPainting [44] and PointAugmenting [45] are the classical multi-modal detectors that incorporate the rich visual information of RGB images into the LiDAR-based detectors. LoGoNet [46] devises the local-to-global fusion module and surpasses many competitive multi-modal detectors in the popular Waymo leaderboard.

The aforementioned detectors are mainly built in the onboard settings where the computation and storage resources are limited. Recently, the offboard detectors have drawn surging attention from both academic and industrial communities which can achieve impressive detection performance given unlimited resources. Take the notable 3DAL [47] as an example. It fuses all point cloud frames of one sequence and surpasses the detection accuracy of humans in the challenging Waymo benchmark. The top-performing offboard detectors [47, 48] can serve as the auto labelling function that can significantly reduce the expensive cost of the labelling process.

4.4 Tracking

3D object tracking targets at assigning the same object the same instance id across multiple frames. It relies on the single-frame detection performance and the sufficient utilization of the valuable temporal information. Recent efforts have been paid on handling the occlusion problem, reducing false predictions as well as making better use of the temporal and contextual information hidden in the input sequences [49–51]. The overview of the classical CenterPoint framework is shown in Fig. 10. The development in the single-frame 3D detection field also promotes the advances in the 3D tracking domain. The offboard 3D detection usually utilizes the tracking philosophy to improve the detection accuracy in a consecutive sequence of frames.

4.5 Matching

Matching aims to find the correspondences between 3D points of two point cloud frames. The matching task usually consists of two consecutive steps: feature extraction and data association. Before the advent of deep learning, the feature extraction is hand-crafted and the data association techniques are mostly optimization-based ones. The FPFH [52], ESF [53] are two widely used features that used in the pre deep learning era. The other feature extractors such as 3DSIFT [54], PFH [55] can be found in this survey [56]. After deep learning is prevalent, the performance of point feature is largely improved, regarding the distinguishability and recall. The typical example is 3DMatch [57]. As depicted from Fig. 11, 3DMatch uses a Siamese Style 3D ConvNets to extract point-wise features and trains the network by minimizing the distance of corresponding 3D point pairs (matches) while pulling apart the distance of noncorresponding 3D point pairs. Based on this idea, many variants are subsequently proposed. OctNet [58] proposes a Octree-based neural network to relieve the memory consumption of a 3D CNN. FCGF [59] leverages the sparse convolution operation to reduce the memory consumption and improve the efficiency of point feature extraction. IMFNet [60] leverages the multi-modal information to further improve the discrimination of point features.

4.6 Registration

The objective of 3D registration is to estimate the transformation matrix between two point cloud frames. The

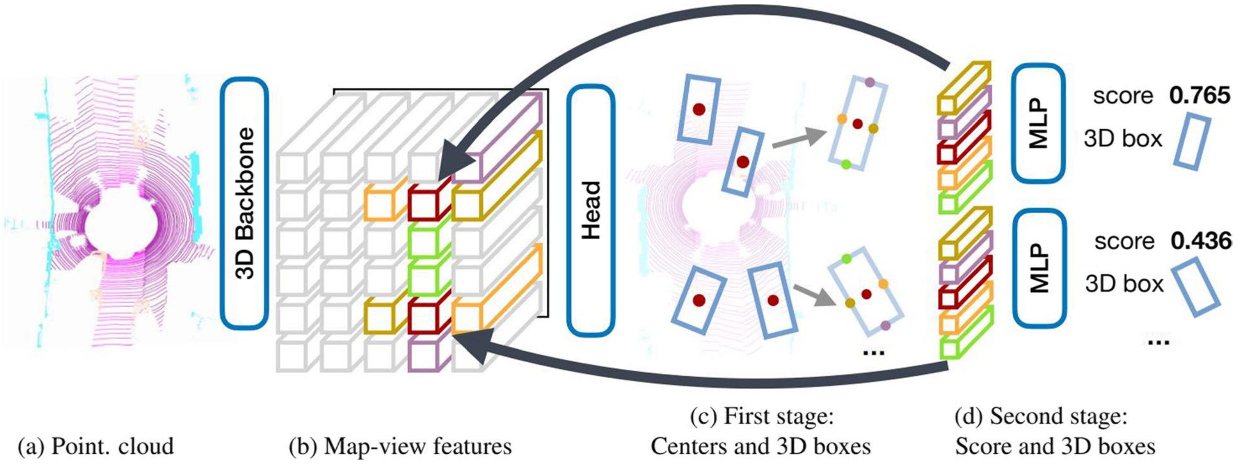

Fig. 10 Schematic overview of the CenterPoint algorithm

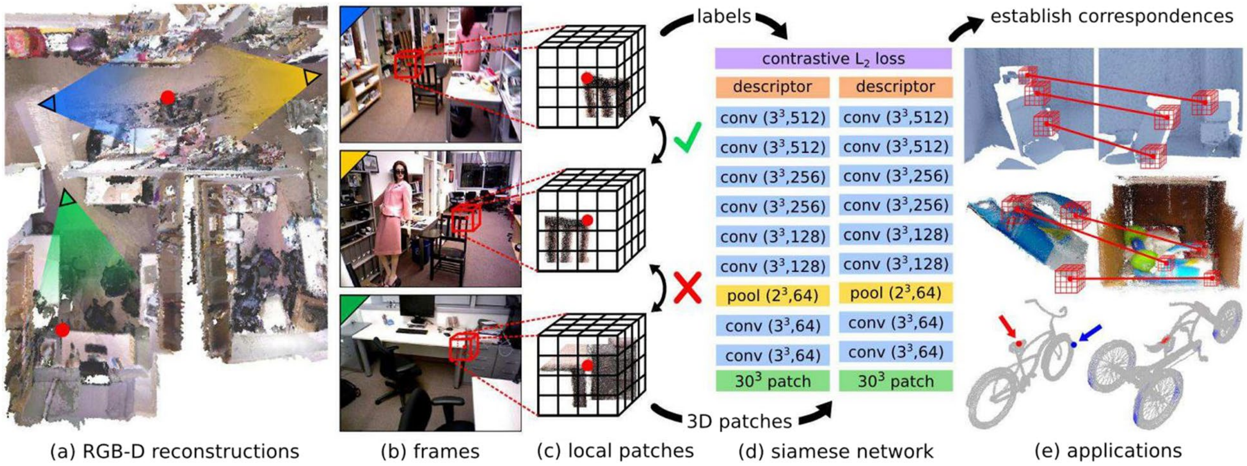

Fig. 11 Schematic overview of the 3DMatch algorithm

point cloud registration methods can be categorized into two classes: pre deep learning and post deep learning.

The pre deep learning methods mainly use optimization-based strategies. One typical example is ICP [61], which iterates the transformation estimation and correspondence estimation. The vanilla ICP is sensitive to noise and outliers. Many variations [62–68] are proposed to tackle the limitations of ICP. GO-ICP[62] utilizes a branch-and-bound (BnB) scheme that searches the entire 3D motion space SE(3) and develops the ICP to obtain global optimal solution. TEASER, introduced in the paper by Yang et al. (2020) [64], utilizes Truncated Least Squares cost and separates the estimation of scale, rotation, and translation. It has demonstrated robustness against 99% outliers. With advancements in 3D sensor technology, a new subfield called cross-source point cloud registration has emerged to handle data from different domains, which exhibit greater variations. Huang et al. [66–68] have proposed several methods to address the challenges of cross-source registration by enhancing the Iterative Closest Point (ICP) algorithm and achieving high accuracy in handling cross-source point cloud registration.

The post deep learning methods leverage the strong ability of deep neural networks to optimize the transformation. The post deep learning methods [57, 69–74] can be further divided into two categories: correspondence-based and correspondence-free methods. The main idea of correspondence-based methods is to replace the feature extractor with deep learning features and uses the learned features to extract correspondences. These correspondences are utilized to estimate the transformation with RANSAC or SVD. The typical example of correspondence-based methods is 3DMatch [57], which leverages a 3DCNN to extract the deep features and uses a RANSAC to estimate the transformation. The key objective of correspondence-free methods is to develop an end-to-end neural network to directly estimate the transformation. The typical examples are DGR [69] and FMR [75]. DGR [69] develops a neural network to estimate the descriptor and the weights of weighted SVD. Then, the transformation is optimized with a closed-form solution. FMR [75] estimate the feature alignment error of two point clouds. Then, the error is backbprogated to estimate the transformation by optimizing a LM-based registration algorithm. Obviously, these correspondence-free methods combine the deep learning with the conventional registration theoretical framework.

The recent registration methods [73, 74] follow this line and propose advanced feature extraction module to enhance the association of both given point clouds, as shown in Fig. 12.

5 Benchmark

We present the basic information of popular benchmarks in 3D classification, segmentation, detection, tracking, matching and registration. These benchmarks are widely used by both industrial and academic communities.

ModelNet40 [76]. It is a synthetic object-level benchmark and consists of 12,311 CAD models. The quantity of object categories is 40. It includes common objects such as chairs, tables, cars, *etc*. These CAD models are captured from different angles and orientations. In ModelNet40, 9,843 samples are used for training while the rest 2,468 samples are chosen for testing.

ShapeNet [77]. It is a widely used dataset for 3D shape analysis. It covers a wide range of object categories and is comprised of approximately 51,300 unique 3D shapes from over 55 different categories.

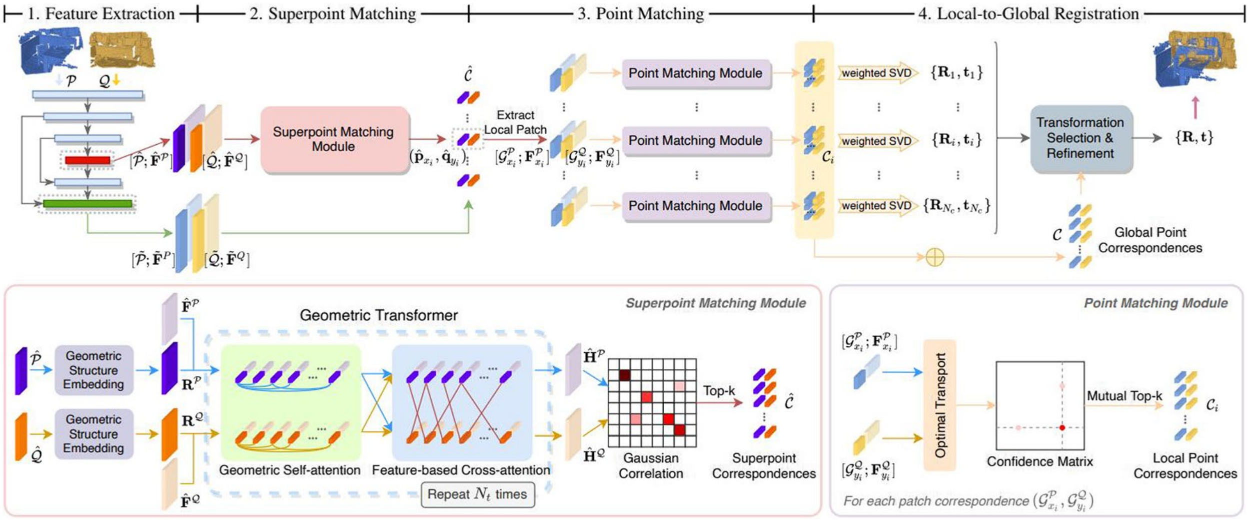
1. Feature Extraction
2. Superpoint Matching
3. Point Matching
4. Local-to-Global Registration
Superpoint Matching Module
Extract Local Patch
Point Matching Module
weighted SVD
Transformation Selection & Refinement
Global Point Correspondences
Superpoint Matching Module
Geometric Transformer
Geometric Structure Embedding
Geometric Self-attention
Feature-based Cross-attention
Repeat N_t times
Gaussian Correlation
Top-k
Superpoint Correspondences
Point Matching Module
Optimal Transport
Confidence Matrix
Mutual Top-k
Local Point Correspondences
For each patch correspondence $(\mathcal{G}^{\mathcal{P}}_{x_i}, \mathcal{G}^{\mathcal{Q}}_{y_i})$

Fig. 12 Schematic overview of the GeoTransformer algorithm

ScanNetV2 [78]. The ScanNetV2 dataset is an indoor segmentation benchmark and provides 1,201 and 312 scenes for training and validation, respectively. The total number of categories is 20.

S3DIS [79]. The S3DIS dataset for semantic scene parsing consists of 271 rooms in six areas from three different buildings, and the number of categories is 13. Following a common protocol, area 5 is withheld during training and used for testing.

SUN RGB-D [80]. The SUN RGB-D dataset contains over 10,000 RGB-D images captured from different indoor scenes, covering a wide range of object categories and scene types.

SemanticKITTI [81]. It is a popular outdoor LiDAR segmentation dataset and is comprised of 22 point cloud sequences, where sequences 00 to 10, 08, and 11 to 21 are used for training, validation and testing, respectively. The total number of categories is 19. It only provides front-view images for multi-modal fusion.

KITTI [82]. It is well-known autonomous driving benchmark for 3D detection. It contains 7,481 training samples and 7,518 testing samples. Average precision (AP) on easy, moderate and hard levels is the primary evaluation criterion.

nuScenes [83]. It is comprised of 1,000 scenes with approximately 20 seconds for each scene. The key frames are labeled at 2 Hz. Each sample contains six multi-view RGB images captured from different cameras and one point cloud frame collected by a 16-beam LiDAR. For the 3D detection task, the number of annotated 3D boxes and categories is 1.4M and 10, respectively.

Waymo [84]. It is one of the largest autonomous driving benchmarks, with 798, 202 and 150 sequences chosen for training, validation and testing, respectively. Each sequence has around 200 frames and there are five RGB images accompanying each point cloud frame. The standard metrics for evaluating detectors are average precision (AP), average precision weighted by heading (APH) on LEVEL 1 (L1) and LEVEL 2 (L2) difficulty levels.

ONCE [85]. It is a large-scale autonomous driving dataset that contains 1 million LiDAR frames and 7 million camera images. 15 K scenes are fully annotated with 5 classes, including car, bus, truck, pedestrian and cyclist.

3DMatch [57]. It is a widely used indoor point cloud dataset to evaluate the performance of registration algorithms. The dataset consists of 62 indoor scenes, which are captured by a RGB-D sensor. The 54 scenes are used for training and 8 scenes for testing. The ground-truth transformation is estimated by the RGB-D reconstruction pipeline.

KITTI [82], nuScenes [83], and Waymo Open Dataset (WOD) [84] are the three most influential benchmarks for BEV-based 3D perception. KITTI is a well-known benchmark for 3D perception. It consists of 3712, 3769, and 7518 samples for training, validation, and testing, respectively. It provides both 2D and 3D annotations for cars, pedestrians, and cyclists. The detection is divided into three levels, i.e., easy, moderate, and hard, based on the size of detected objects, occlusion, and truncation levels. NuScenes contains 1000 scenes with durations of 20 seconds for each scene. Each frame contains six calibrated images covering the 360-degree horizontal FOV, making nuScenes one of the most commonly used datasets for vision-based BEV perception algorithms. WOD is a large-scale autonomous driving dataset with 798 sequences, 202 sequences, and 150 sequences for training, validation, and testing, respectively. Apart from the above-mentioned three datasets, more benchmarks such as Argoverse, H3D, and Lyft L5 can also be used for BEV-based perception.

6 Evaluation metric

For classification tasks, accuracy is the major metric to evaluate the performance of classifiers. For 3D object detection, mean average precision (mAP) and nuScenes detection scores (NDS) [83] are often employed as evaluation criterion. The most commonly used criterion is which is the area under the precision-recall curve.

Specifically, to calculate AP, we first use Intersection-over-Union (IoU) to measure the distance between the predictions and labels. The definition of IoU between prediction A and label B is given below: $\mathbf{IoU}(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$. If the IoU between prediction and its label is larger than a pre-defined value, the prediction is seen as True Positive (TP). Otherwise, it is treated as False Positive (FP). Then, Precision and Recall can be computed based on TP, FP, and FN: $\mathbf{Precision} = \frac{TP}{TP+FP}, \mathbf{Recall} = \frac{TP}{TP+FN}$. Here, FN denotes False Negative. And AP is calculated using the interpolated precision values: $\mathbf{AP} = \frac{1}{|R|}\sum_{r \in R} p_{interp}(r)$, where R is the set of all recall positions, $p_{interp}(.)$ is the interpolation function, defined as: $p_{interp}(r) = \max_{r':r' \geq r} p(r')$, and the mean average precision (mAP) is the average of APs of different classes or difficulty levels.

For LiDAR semantic segmentation tasks, we adopt the Intersection-over-Union (IoU), mean of Intersection-over-Union (mIoU), mean of class-wise accuracy (mAcc), and overall point-wise accuracy (OA) as the evaluation criterion. Since the number of points may vary significantly among different classes, we can also use the inverse class frequency scores to reweight the calculation of mIoU.

For the matching and registration tasks, five metrics are usually adopted. (1) Registration Recall (RR) refers to the proportion of point cloud pairs that meet the accuracy threshold. (2) Rotation Error (RE) is the average angular deviation between the estimated rotation and the ground truth rotation. (3) Translation Error (TE) is the average discrepancy between the estimated translation and the ground truth translation.. (4) The F1-score (F1) is calculated using the formula: $\text{F1} = \frac{2TP}{2TP+FN+FP}$, where TP represents the true positive, FN represents the false negative, and FP represents the false positive. The F1-score is utilized to assess the balance between precision and recall, measuring their stability. (5) Inlier Recall (IR) quantifies the proportion of estimated correspondences that have residuals below a specified threshold (e.g., 0.1m) based on the ground-truth transformation. The first three metrics focus on assessing registration accuracy, while the remaining metrics aim to evaluate the ability to reject outliers.

7 Future work

In this section, we list several promising research directions that are worth exploring in the 3D domain.

Large Language Models. The emergence of Large Language Models (LLMs) [86, 87] is undoubtedly the milestone event in the deep learning field. Considering that LLMs are trained on a large corpus of textual information, they typically embrace rich world knowledge and information. How to incorporate these powerful LLMs into the 3D field is also a hot topic worth exploring. Uni3D-LLM [88] makes the first attempt to use LLMs to process 3D perception and generation tasks in a unified manner. FrozenCLIP [89] employs the frozen CLIP model to perform LiDAR-based scene understanding.

Knowledge Transfer from 2D to 3D. PointCLIP [20] makes the first attempt to introduce the strong vision-language model, *i.e.* CLIP [19], for the point cloud understanding tasks and achieves appealing zero-shot classification and segmentation performance in indoor benchmarks. Since the amount of training samples in 2D and NLP domains is much larger than the 3D domain, how to effectively leverage the precious knowledge hidden in these domains is also an interesting research direction. Effective multi-modal and cross-modal knowledge distillation algorithms [90–93] are needed to better utilize the abundant information hidden in 2D images and videos.

Synthetic Data. Manually collecting and annotating 3D data is extremely expensive. To relieve the heavy reliance on the large-scale training data, using synthetic data is undoubtedly a promising direction. Compared to the real data, the expense of collecting synthetic data and making annotations is much cheaper. Therefore, how to effectively utilize a large amount of valuable synthetic samples is an important direction to explore. However, since synthetic data may differ from the real data in many aspects, such as texture, lighting, material, physical constraints, models trained on synthetic data may witness significant performance drops when directly deployed in the real-world applications. Techniques such as domain adaptation, neural rendering, data augmentation, are required to relieve the gap between synthetic and real samples [94, 95].

Foundation Models. Note that 2D foundation models, such as the SAM series [96], have reshaped the 2D vision field and greatly facilitate many downstream tasks and applications. However, due to the lack of large-scale 3D benchmarks, the 3D foundation models have not appeared. The building of the 3D foundation models is inevitably the urgent task to date as they can greatly reduce the design and deployment cost of many 3D tasks, paving the way for the Artificial General Intelligence (AGI) 3D era. To build such a foundation model in the 3D domain, one can either train a powerful model on a large quantity of 3D samples, or take foundation models from other fields as the backbone and employ adapters to re-train or fine-tune such model on the provided samples.

Network Design. For 2D domains, networks such as ResNet [97], EfficientNet [98], MobileNet [99], have become the de facto architectures for many downstream applications. However, for the 3D field, there is no such network series that can be applicable to various tasks. Since sparse convolution is adept at capturing local information while self attention excels at grasping global relationship, combining the strengths of both operators is straightforward, constituting the hybrid 3D networks. Drawing inspirations from the Inception Transformer [100], designing hybrid networks that fully leverage the strengths of both operations is a promising research direction. We can also follow the pipeline of [39] and propose novel modules to relieve the shortcomings of the current model.

8 Conclusion

In this survey, we summarize the recent advances in the 3D field, including the main-stream pre-training strategies, 3D perception tasks, benchmarks as well as the evaluation metrics. We also point out several promising research directions of the 3D field. We hope this survey can lay the foundation for both academic and industrial communities and inspire more fundamental works.

Abbreviations

LLM	Large language model
spconv	Sparse convolution
MLP	Multi-layer perceptron
FPS	Farthest point sampling
SoTA	State-of-the-art
ICP	Iterative closest point
IoU	Intersection-over-union
Acc	mean of class-wise accuracy
NDS	NuScenes detection scores
TP	True positive
FP	False positive
OA	Overall point-wise accuracy
mAP	Mean average precision
mIoU	Mean of intersection-over-union
RR	Registration recall
RE	Rotation error
TE	Translation error
AGI	Artificial general intelligence

Acknowledgements

None.

Authors' contributions

All authors contributed to the study conception and design. Yuenan drafted the work. Tong He was in charge of the pre-training section. Yuenan and Xiaoshui were in charge of the downstream task sections. Prof. Wanli and Shixiang critically revised the work.

Funding

No funding was received to assist with the preparation of this manuscript.

Availability of data and materials

All data generated or analysed during this study are included in this manuscript.

Code availability

Not applicable.

Declarations

Competing interesst

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this manuscript.

Received: 17 March 2024 Revised: 13 June 2024 Accepted: 1 July 2024
Published online: 23 July 2024

References

1. Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, M. Bennamoun, Deep learning for 3d point clouds: A survey. IEEE Trans. Pattern Anal. Mach. Intell. **43**(12), 4338–4364 (2020)
2. Z. Shi, S. Peng, Y. Xu, A. Geiger, Y. Liao, Y. Shen, Deep generative models on 3d representations: A survey (2022). arXiv preprint arXiv:2210.15663
3. C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation. in *Proceedings of the IEEEs conference on computer vision and pattern recognition* (IEEE, USA, 2017), pp. 652–660
4. C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inf. Process. Syst. **30** (2017)
5. S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2019), pp. 770–779
6. Y. Zhou, O. Tuzel, Voxelnet: End-to-end learning for point cloud based 3d object detection. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2018), pp. 4490–4499
7. Y. Yan, Y. Mao, B. Li, Second: Sparsely embedded convolutional detection. Sensors **18**(10), 3337 (2018)
8. B. Graham, M. Engelcke, L. Van Der Maaten, 3d semantic segmentation with submanifold sparse convolutional networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2018), pp. 9224–9232
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
10. S. Xie, J. Gu, D. Guo, C.R. Qi, L. Guibas, O. Litany, Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. in *ECCV* (Springer, Germany, 2020)
11. J. Hou, B. Graham, M. Nießner, S. Xie, Exploring data-efficient 3d scene understanding with contrastive scene contexts. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 15587–1559
12. H. Yang, T. He, J. Liu, H. Chen, B. Wu, B. Lin, X. He, W. Ouyang, Gd-mae: Generative decoder for mae pre-training on lidar point clouds. in *CVPR* (IEEE, USA, 2023)
13. J. Hou, X. Dai, Z. He, A. Dai, M. Nießner, Mask3d: Pre-training 2d vision transformers by learning masked 3d priors. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 13510–13519
14. D. Huang, S. Peng, T. He, H. Yang, X. Zhou, W. Ouyang, Ponder: Point cloud pre-training via neural rendering. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2023), pp. 16089–16098

15. H. Zhu, H. Yang, X. Wu, D. Huang, S. Zhang, X. He, T. He, H. Zhao, C. Shen, Y. Qiao et al., Ponderv2: Pave the way for 3d foundataion model with a universal pre-training paradigm (2023). arXiv preprint arXiv:2310.08586
16. A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding (2018). arXiv preprint arXiv:1807.03748
17. M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, R. Rodrigo, 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 (CVPR)* (IEEE, USA, 2022), pp. 9902–9912
18. C. Sautier, G. Puy, S. Gidaris, A. Boulch, A. Bursuc, R. Marlet, Image-to-lidar self-supervised distillation for autonomous driving data. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, USA, 2022), pp. 9891–9901
19. A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., Learning transferable visual models from natural language supervision. in *International conference on machine learning,* (PMLR, USA, 2021), pp. 8748–8763
20. R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, H. Li, Pointclip: Point cloud understanding by clip. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2022), pp. 8552–8562
21. X. Zhu, R. Zhang, B. He, Z. Guo, Z. Zeng, Z. Qin, S. Zhang and P. Gao, Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2023), pp. 2639–2650
22. L. Nunes, R. Marcuzzi, X. Chen, J. Behley, C. Stachniss, SegContrast: 3D Point Cloud Feature Representation Learning through Self-supervised Segment Discrimination. IEEE Robot. Autom. Lett. (RA-L) **7**(2), 2116–2123 (2022). https://doi.org/10.1109/LRA.2022.3142440
23. Y. Pang, W. Wang, F.E. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning. in *ECCV,* (Springer, Germany, 2022), pp. 604–621
24. X. Ma, C. Qin, H. You, H. Ran, Y. Fu, Rethinking network design and local geometry in point cloud: A simple residual mlp framework (2022). arXiv preprint arXiv:2202.07123
25. X. Wu, Y. Lao, L. Jiang, X. Liu, H. Zhao, Point transformer v2: Grouped vector attention and partition-based pooling. Adv. Neural Inf. Process. Syst. **35**, 33330–33342 (2022)
26. T. Xiang, C. Zhang, Y. Song, J. Yu, W. Cai, Walk in the cloud: Learning curves for point clouds shape analysis. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2021), pp. 915–924
27. G. Qian, Y. Li, H. Peng, J. Mai, H. Hammoud, M. Elhoseiny, B. Ghanem, Pointnext: Revisiting pointnet++ with improved training and scaling strategies. Adv. Neural Inf. Process. Syst. **35**, 23192–23204 (2022)
28. Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, J.M. Solomon, Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. (tog) **38**(5), 1–12 (2019)
29. C. Choy, J. Gwak, S. Savarese, 4d spatio-temporal convnets: Minkowski convolutional neural networks. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2019), pp. 3075–3084
30. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation. in *MICCAI,* (Springer, Germany, 2015), pp. 234–241
31. H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, S. Han, Searching efficient 3d architectures with sparse point-voxel convolution. in *European conference on computer vision*, (Springer, Germany, 2020), pp. 685–702
32. X. Zhu, H. Zhou, T. Wang, F. Hong, Y. Ma, W. Li, H. Li, D. Lin, Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2021), pp. 9939–9948
33. X. Lai, Y. Chen, F. Lu, J. Liu, J. Jia, Spherical transformer for lidar-based 3d recognition. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 17545–17555
34. A. Ando, S. Gidaris, A. Bursuc, G. Puy, A. Boulch, R. Marlet, Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 5240–5250
35. L. Kong, Y. Liu, R. Chen, Y. Ma, X. Zhu, Y. Li, Y. Hou, Y. Qiao, Z. Liu, Rethinking range view representation for lidar segmentation. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2023), pp. 228–240
36. J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, S. Pu, Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2021), pp. 16024–16033
37. Y. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li et al., Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2023), pp. 21662–21673
38. J. Li, H. Dai, H. Han, Y. Ding, Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 21694–21704
39. X. Wu, L. Jiang, P.S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, H. Zhao, Point transformer v3: Simpler, faster, stronger. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2024), pp. 4840–4851
40. S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 10529–10538
41. Y. Bai, B. Fei, Y. Liu, T. Ma, Y. Hou, B. Shi, Y. Li, Rangeperception: Taming lidar range view for efficient and accurate 3d object detection. Adv. Neural Inf. Process. Syst. **36** (2024)
42. Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, X. Zhu, Vision-centric bev perception: A survey (2022). arXiv preprint arXiv:2208.02797
43. Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, J. Dai, Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. in *European conference on computer vision*, (Springer, Germany, 2022), pp. 1–18
44. S. Vora, A.H. Lang, B. Helou, O. Beijbom, Pointpainting: Sequential fusion for 3d object detection. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 4604–4612
45. C. Wang, C. Ma, M. Zhu, X. Yang, Pointaugmenting: Cross-modal augmentation for 3d object detection. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2021), pp. 11794–11803
46. X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao et al., Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 17524–17534
47. C.R. Qi, Y. Zhou, M. Najibi, P. Sun, K. Vo, B. Deng, D. Anguelov, Offboard 3d object detection from point cloud sequences. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2021), pp. 6134–6144
48. T. Ma, X. Yang, H. Zhou, X. Li, B. Shi, J. Liu, Y. Yang, Z. Liu, L. He, Y. Qiao and Y. Li, Detzero: Rethinking offboard 3d object detection with long-term sequential point clouds. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE, USA, 2023), pp. 6736–6747
49. L. Vacchetti, V. Lepetit, P. Fua, Stable real-time 3d tracking using online and offline information. IEEE Trans. Pattern Anal. Mach. Intell. **26**(10), 1385–1391 (2004)
50. T.X. Xu, Y.C. Guo, Y.K. Lai, S.H. Zhang, Cxtrack: Improving 3d point cloud tracking with contextual information. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 1084–1093
51. T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2021), pp. 11784–11793
52. R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (fpfh) for 3d registration. in *2009 IEEE international conference on robotics and automation,* (IEEE, USA, 2009), pp. 3212–3217
53. W. Wohlkinger, M. Vincze, Ensemble of shape functions for 3d object classification. in *2011 IEEE international conference on robotics and biomimetics,* (IEEE, USA, 2011), pp. 2987–2992
54. D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**, 91–110 (2004)

55. R.B. Rusu, Z.C. Marton, N. Blodow, M. Beetz, Learning informative point classes for the acquisition of object model maps. in *2008 10th International Conference on Control, Automation, Robotics and Vision,* (IEEE, USA, 2008), pp. 643–650
56. X. Huang, G. Mei, J. Zhang, R. Abbas, A comprehensive survey on point cloud registration (2021). arXiv preprint arXiv:2103.02690
57. A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao, T. Funkhouser, 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2017), pp. 1802–1811
58. G. Riegler, A. Osman Ulusoy, A. Geiger, Octnet: Learning deep 3d representations at high resolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2017), pp. 3577–3586
59. C. Choy, J. Park, V. Koltun, Fully convolutional geometric features. in *Proceedings of the IEEE/CVF international conference on computer vision* (IEEE, USA, 2019), pp. 8958–8966
60. X. Huang, W. Qu, Y. Zuo, Y. Fang, X. Zhao, Imfnet: Interpretable multimodal fusion for point cloud registration. IEEE Robot. Autom. Lett. **7**(4), 12323–12330 (2022)
61. P.J. Besl, N.D. McKay, Method for registration of 3-d shapes. in *Sensor fusion IV: control paradigms and data structures*, vol. 1611 (Spie, USA, 1992), pp. 586–606
62. J. Yang, H. Li, D. Campbell, Y. Jia, Go-icp: A globally optimal solution to 3d icp point-set registration. IEEE Trans. Pattern Anal. Mach. Intell. **38**(11), 2241–2254 (2015)
63. Q.Y. Zhou, J. Park, V. Koltun, Fast global registration. in *ECCV 2016,* (Springer, Germany, 2016), pp. 766–782
64. H. Yang, J. Shi, L. Carlone, Teaser: Fast and certifiable point cloud registration. IEEE Trans. Robot. **37**(2), 314–333 (2020)
65. A. Myronenko, X. Song, Point set registration: Coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2262–2275 (2010)
66. X. Huang, J. Zhang, L. Fan, Q. Wu, C. Yuan, A systematic approach for cross-source point cloud registration by preserving macro and micro structures. IEEE Trans. Image Process. **26**(7), 3261–3276 (2017)
67. X. Huang, J. Zhang, Q. Wu, L. Fan, C. Yuan, A coarse-to-fine algorithm for matching and registration in 3d cross-source point clouds. IEEE Trans. Circ. Syst. Video Technol. **28**(10), 2965–2977 (2017)
68. X. Huang, G. Mei and J. Zhang, Cross-source point cloud registration: Challenges, progress and prospects. Neurocomputing. **548**, 126383 (2023)
69. C. Choy, W. Dong, V. Koltun, Deep global registration. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 2514–2523
70. X. Huang, S. Li, Y. Zuo, Y. Fang, J. Zhang, X. Zhao, Unsupervised point cloud registration by learning unified gaussian mixture models. IEEE Robot. Autom. Lett. **7**(3), 7028–7035 (2022)
71. X. Huang, Y. Wang, S. Li, G. Mei, Z. Xu, Y. Wang, J. Zhang, M. Bennamoun, Robust real-world point cloud registration by inlier detection. Comp. Vision Image Underst. **224**, 103556 (2022)
72. G. Mei, H. Tang, X. Huang, W. Wang, J. Liu, J. Zhang, L. Van Gool, Q. Wu, Unsupervised deep probabilistic approach for partial point cloud registration. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 13611–13620
73. Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, K. Xu, Geometric transformer for fast and robust point cloud registration. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2022), pp. 11143–11152
74. S. Ao, Q. Hu, H. Wang, K. Xu, Y. Guo, Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 1255–1264
75. X. Huang, G. Mei, J. Zhang, Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 11366–11374
76. Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3d shapenets: A deep representation for volumetric shapes. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2015), pp. 1912–1920
77. A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su et al., Shapenet: An information-rich 3d model repository (2015). arXiv preprint arXiv:1512.03012
78. A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, Scannet: Richly-annotated 3d reconstructions of indoor scenes. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2017), pp. 5828–5839
79. I. Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3d semantic parsing of large-scale indoor spaces. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2016), pp. 1534–1543
80. S. Song, S.P. Lichtenberg, J. Xiao, Sun rgb-d: A rgb-d scene understanding benchmark suite. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2015), pp. 567–576
81. J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, J. Gall, Semantickitti: A dataset for semantic scene understanding of lidar sequences. in *Proceedings of the IEEE/CVF international conference on computer vision* (IEEE, USA, 2019), pp. 9297–9307
82. A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The kitti dataset. Int. J. Robot. Res. **32**(11), 1231–1237 (2013)
83. H. Caesar, V. Bankiti, A.H. Lang, S. Vora, V.E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 11621–11631
84. P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine et al., Scalability in perception for autonomous driving: Waymo open dataset. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (IEEE, USA, 2020), pp. 2446–2454
85. J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li et al., One million scenes for autonomous driving: Once dataset (2021). arXiv preprint arXiv:2106.11037
86. S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y. Li, S. Lundberg et al., Sparks of artificial general intelligence: Early experiments with gpt-4 (2023). arXiv preprint arXiv:2303.12712
87. T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners. Adv. Neural Inf. Process. Syst. **33**, 1877–1901 (2020)
88. D. Liu, X. Huang, Y. Hou, Z. Wang, Z. Yin, Y. Gong, P. Gao, W. Ouyang, Uni3d-llm: Unifying point cloud perception, generation and editing with large language models (2024). arXiv preprint arXiv:2402.03327
89. X. Huang, Z. Huang, S. Li, W. Qu, T. He, Y. Hou, Y. Zuo, W. Ouyang, Frozen clip transformer is an efficient point cloud encoder. in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38 (AAAI Press, USA, 2024), pp. 2382–2390
90. R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, W. Wang, Clip2scene: Towards label-efficient 3d scene understanding by clip. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 7020–7030
91. M. Klingner, S. Borse, V.R. Kumar, B. Rezaei, V. Narayanan, S. Yogamani, F. Porikli, X3kd: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2023), pp. 13343–13353
92. Y. Hou, X. Zhu, Y. Ma, C.C. Loy, Y. Li, Point-to-voxel knowledge distillation for lidar semantic segmentation. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, USA, 2022), pp. 8479–8488
93. X. Xing, Z. Chen, Y. Hou, Y. Yuan, Gradient modulated contrastive distillation of low-rank multi-modal knowledge for disease diagnosis. Med. Image Anal. **88**, 102874 (2023)
94. B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis. in *European Conference on Computer Vision* (Springer, Germany, 2020)
95. J. Liu, X. Huang, T. Huang, L. Chen, Y. Hou, S. Tang, Z. Liu, W. Ouyang, W. Zuo, J. Jiang et al., A comprehensive survey on 3d content generation (2024). arXiv preprint arXiv:2402.01166
96. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo et al., Segment anything. in *Proceedings*

of the IEEE/CVF International Conference on Computer Vision (IEEE, USA, 2023), pp. 4015–4026

97. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* (IEEE, USA, 2016), pp. 770–778
98. M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks. in *International conference on machine learning,* (PMLR, USA, 2019), pp. 6105–6114
99. A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017). arXiv preprint arXiv:1704.04861
100. C. Si, W. Yu, P. Zhou, Y. Zhou, X. Wang, S. Yan, Inception transformer. Adv. Neural Inf. Process. Syst. **35**, 23495–23509 (2022)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.