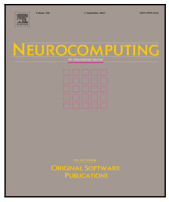
Neurocomputing 608 (2024) 128436

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey Paper

Deep learning for 3D object recognition: A survey

A.A.M. Muzahid [a,*], Hua Han [a], Yujin Zhang [a], Dawei Li [b], Yuhe Zhang [c], Junaid Jamshid [d], Ferdous Sohel [a,e]

[a] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, 201620, China
[b] College of Information Sciences and Technology, Donghua University, Shanghai, 201620, China
[c] School of Information Science and Technology, Northwest University, Xi'an, 710127, China
[d] School of Communications and Information Engineering, Shanghai University, Shanghai, 200444, China
[e] School of Information Technology, Murdoch University, Murdoch WA 6150, Australia

ARTICLE INFO

Communicated by B. Fan

Keywords:
3D shape analysis
3D object classification
Point cloud classification
Volumetric CNN

ABSTRACT

With the growing availability of extensive 3D datasets and the rapid progress in computational power, deep learning (DL) has emerged as a highly promising approach for learning from 3D data, addressing critical tasks like object detection, segmentation, and recognition. Despite the unique challenges in processing geometry data with deep neural networks, recent advancements in DL for 3D object recognition have shown remarkable success, with various methods proposed to tackle different issues. This paper aims to stimulate future research by providing a comprehensive review of recent progress in DL techniques for 3D object recognition, which are systematically categorized based on their learning behavior. We discuss the advantages, limitations, and application of each approach, highlighting their performance in 3D object classification on benchmark datasets such as ModelNet, ScanObjectNN, and Sydney Urban Object. The survey offers insightful observations and inspires future research directions.

1. Introduction

The 3D shape is one of the most crucial features of objects. Object recognition aims to develop systems to detect, and recognize the contents in a scene. Therefore, 3D object recognition using DL refers to categorizing or labeling three-dimensional objects based on their features and attributes using DL networks. It involves training DL models, typically a convolutional neural network (CNN) or a variant specialized for 3D data, to automatically learn patterns and representations from 3D object data. This enables accurate classification of objects into predefined categories or classes. The DL model learns from a large dataset of labeled 3D objects, and it can recognize and categorize unseen or new objects based on the training it received. The automatic 3D object recognition method is a vital step forward in the true intelligence of smart computer vision, and it has numerous real-world applications, including robot navigation and autonomous driving, and augmented reality [1–9]. The availability of advanced 3D scanning technology [10,11], the growing applications of 3D shape analysis, advancements in DL, availability of 3D datasets and benchmarks, technological advances in computing, and cross-disciplinary collaboration have collectively fueled the popularity and interest in 3D shape analysis in recent years [12–17]. The field of 3D computer vision has witnessed significant advancements, especially in the use of different sensor modalities [18] such as LiDAR, camera images, and radar. These advancements have paved the way for more accurate, efficient, and robust systems. By combining information from multiple sensors, DL models can leverage the complementary strengths of each modality, leading to a more reliable and comprehensive understanding of the environment [19,20]. However, understanding 3D data is still an open research problem. Traditional approaches to 3D object recognition relied on handcrafted features, such as local descriptors or shape-based representations. However, these approaches often struggled with handling variations in viewpoints, poses, scales, and occlusions [21–25]. In this context, DL algorithms have emerged as a promising solution. Inspired by the success of DL algorithms in image processing [26,27], research on 3D object recognition has evolved from the use of conventional techniques to DL methods [5,28,29]. By leveraging the power of these neural networks, recently, the performance of 3D object recognition has improved significantly by using DL-based methods, such as Auto Encoder (AE) [30], Deep Boltzmann Machine (DBM) [31], Deep Belief Network (DBF) [32,33] and CNN [34,35]. Feature extraction by DL models (e.g., CNN frameworks) can learn the

* Corresponding author.
E-mail addresses: muzahid@sues.edu.cn (A.A.M. Muzahid), 2070967@mail.dhu.edu.cn (H. Han), yjzhang@sues.edu.cn (Y. Zhang), daweili@dhu.edu.cn (D. Li), zhangyuhe0601@nwu.edu.cn (Y. Zhang), junaid20860155@shu.edu.cn (J. Jamshid), f.sohel@murdoch.edu.au (F. Sohel).

https://doi.org/10.1016/j.neucom.2024.128436
Received 8 February 2024; Received in revised form 21 May 2024; Accepted 19 August 2024
Available online 22 August 2024

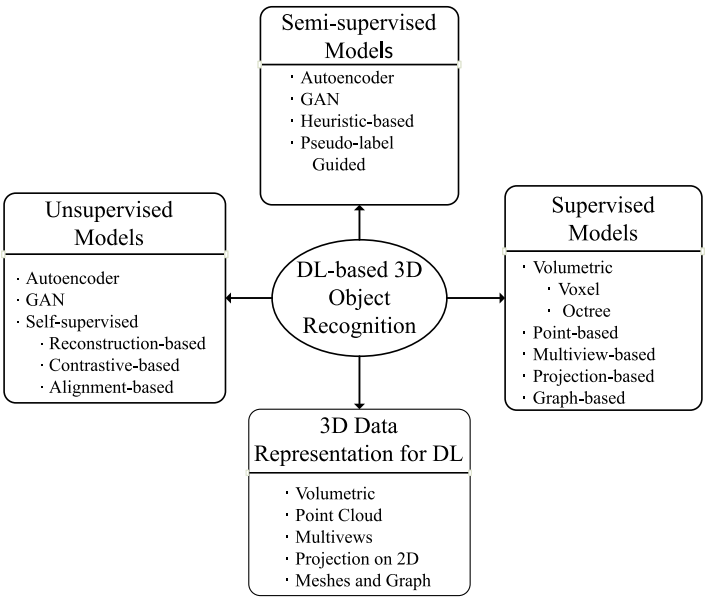


Fig. 1. A taxonomy of DL-based 3D object recognition.

distribution of 3D data better, and they provide effective results when compared with traditional methods based on handcrafted features [5]. DL models not only outperform traditional methods in terms of accuracy, but they also offer better generalization and robustness. Although DL can leverage the complementary strengths of multimodal data [36], exploring the interaction between point clouds and images has posed a challenge. However, recent advancements in graph matching-based feature alignment strategies, which exploit both graph and self-attention feature alignment, have provided a more reliable and comprehensive understanding of the environment [19,20]. These strategies address the unnatural interaction between point clouds and images, improving performance in 3D object detection and classification tasks.

Yet, DL in 3D faces several major challenges, such as small 3D datasets, complex representations of 3D objects, high demand for computing resources, and increased computational costs with high dimensionality. Consequently, several survey articles on DL on 3D data for several tasks have been published in the last few years [18,25,37–43]. However, this survey aims to provide a comprehensive focus specifically on DL methods employed for 3D object classification task only, utilizing a variety of representations for 3D data. We thoroughly explore these methods and summarize their potential applications in the context of object recognition. Supervised methods for 3D object recognition are categorized based on the various data representations and processing techniques they use, such as volumetric and multiview methods. In contrast, semi-supervised and unsupervised methods are categorized based on the different approaches and objectives they use to learn representations from labeled and unlabeled data, respectively, such as autoencoders and contrastive-based techniques. A new taxonomy of this survey is presented in Fig. 1.

In comparison to the existing literature, this paper makes the following significant contributions:

- To the best of our knowledge, this is the first comprehensive survey that thoroughly covers state-of-the-art (SOTA) DL techniques for 3D object recognition, systematically categorizing them based on their learning behavior across a wide range of 3D data representations.
- This paper distinguishes itself from existing reviews [5,37,40,42,44] by specifically focusing on the applications of DL in 3D object recognition, rather than encompassing all other tasks within 3D shape analysis.
- Thorough comparisons of existing methods on publicly available datasets are provided comprehensively, accompanied by concise summaries and insightful discussions. This survey also highlights current challenges and lays the groundwork for future research and development in DL-based 3D object recognition.

The rest of the paper is organized as follows: Section 2 presents the background and previous review of works on 3D object recognition. Section 3 presents the most popular categories of 3D data representation methods for 3D object recognition. Section 4 discusses traditional methods, and Section 5 surveys recent advances in deep learning methods for 3D object recognition. Section 6 provides the possible future research direction for this field. Finally, we conclude this survey in Section 7.

2. Background and related works

This section briefly discusses the recent reviews on 3D shape analysis, including object recognition. The pipeline of 3D object recognition using DL methods can be divided into two broad parts: (i) 3D data representation, and (ii) the design of DL networks. A 3D object can be represented differently, e.g., point clouds, volumetric (voxel and octree), depth images, meshes, and multiview. Ioannidou et al. [5] summarized traditional and early descriptor-based DL approaches for various 3D computer vision tasks, leveraging handcrafted features of 3D data. Carvalho et al. [39] reviewed earlier methods (2006–2016) for 3D object recognition and classification tasks. This work, while overlapping with previous surveys [39]. However, DL on 3D data became popular after publicly releasing the ModelNet [33] dataset in 2015. Ahmed et al. [37] subsequently reviewed advanced DL methods for different 3D data types, categorizing them into Euclidean and non-Euclidean representations for diverse tasks. While this survey [37] offers high-quality content of advanced DL for learning 3D data in computer vision. Several surveys and review articles on DL techniques for 3D point clouds [25,40,42,44,45] and multiview [41] data, respectively, were subsequently published, each focusing on a specific type of 3D data. Recently, Gezawa et al. [46] reviewed DL methods for 3D object retrieval and classification tasks. However, they [46] covered articles published up to 2020, and performance comparisons of DL models were limited to supervised learning.

This survey paper presents a new taxonomy that aims to provide a comprehensive overview of the advancements in 3D object recognition

using DL techniques. The objective is to present a critical analysis, highlight limitations, and suggest future research directions. An extensive literature review was conducted using prominent databases, including Web of Science and Scopus. The paper focuses on the common 3D data representations used in DL: volumetrics, point clouds, multi-view, projections, meshes, and graphs. Each representation is discussed in terms of its characteristics, advantages, and limitations. The paper categorizes DL methods for 3D object recognition into three main types: (i) Supervised, (ii) Semi-Supervised, and (iii) Unsupervised learning. For each category, the performance in terms of mean accuracy and model parameters reported in the corresponding articles is compared. The paper also identifies and reports the limitations of each method at the end of each section and subsection, respectively. Based on the analysis and insights gained from the survey, the paper identifies several promising future research directions. However, the survey's scope is limited to the classification of isolated 3D objects. Additionally, the survey does not delve into the practical implementation challenges and computational requirements.

2.1. Datasets

The results and efficacy of DL models have been significantly affected by the size of the datasets. The growth of DL algorithms for 3D computer vision led to many 3D datasets to analyze 3D objects. A set of popular 3D datasets for 3D object classification are enlisted together in Table 1. Two types of 3D datasets are available for 3D object classification, including synthetic datasets [33] and real-world datasets [47–49]. 3D objects in synthetic datasets are complete without occlusion and background noises. On the other hand, 3D objects in real-world datasets may be contaminated with background noise with occlusion. ModelNet40[1] and ModelNet10 [33] are widely used 3D shape classification datasets, containing synthetic CAD models across 40 and 10 categories, respectively, facilitating evaluation of DL models in this domain. ModelNet40 comprises 12,311 CAD models across 40 classes, with 9,843 models for training and 2,468 for testing, while ModelNet10 is a subset containing 4,899 objects from popular categories within ModelNet40. ScanObjectNN[2] [48] is a real-world point cloud dataset comprising 15,000 objects categorized into 15 classes used for 3D object recognition in indoor environments. The Sydney Urban Objects[3] [47] dataset consists of common urban road objects scanned with a Velodyne HDL-64E LIDAR, facilitating object detection and classification in outdoor settings. These datasets provide real-world object data for indoor and outdoor contexts, respectively. The Omni3D dataset is a recently introduced benchmark public dataset [50] for 3D object detection and recognition. It is designed to address the limitations of existing datasets by providing a larger number of object categories, finer-grained annotations, and a wide range of sensor modalities. The dataset contains 3 million samples across 98 categories of objects, which is about 20 times larger than existing benchmarks. It is a valuable resource for evaluating algorithms and advancing 3D object detection and recognition research. Another benchmark, robust datasets for 3D objects, have also been introduced lately [51] with systematic categorization into weather, sensor, motion, object, and alignment levels. These datasets consider real-world driving scenarios and include 27 distinct corruptions. Each corruption has five severities, resulting in 135 distinct corruptions in total. To evaluate the performance of DL models in challenging scenarios, these corruptions were applied to typical autonomous driving datasets such as KITTI [52], nuScenes [53], and Waymo [54], creating three corruption robustness benchmarks: KITTI-C, nuScenes-C, and Waymo-C [51]. These benchmarks provide valuable resources for researchers in the field of 3D vision to investigate more challenging tasks and improve the performance of DL models in handling 3D shapes. The number of 3D samples and the form of 3D data might be varied for different datasets. The attributes of those datasets are summarized in Table 1.

2.1.1. Dataset augmentation

DL requires to be trained with a large dataset [55,56] in contrast, current public datasets of 3D objects are smaller (i.e., ModelNet40 contains 12K+ samples) when compared with image datasets such as ImageNet contains 14M samples. To increase the number of 3D training samples, it is a common practice to apply geometric transformations, including rotation, scaling, translation, flipping and jittering, and points dropout [57–59]. Therefore, the most common practice to increase the number of 3D samples is to rotate each object 360 degrees. Then, multiple shapes of a 3D object are captured from its different viewpoints by placing virtual scanners at a certain interval of angles [58,60,61]. For example, capturing shapes at every 30 degrees can generate twelve new samples from a single 3D object. These newly augmented samples help improve the network's performance by learning them from different angles. The rotation-based augmentation can be defined by Eq. (1) [62]. The world coordinates (x, y, z), projected into a 3D camera frame (x_c, y_c, z_c) by applying matrix rotation (**R**-rotation matrix) and translation (t-translation vector) operation. The f is the focal length, and λ refers to the skewness between the horizontal and vertical axis of the camera in the Intrinsic matrix, **K**.

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \underbrace{\begin{bmatrix} f & \gamma & w/2 \\ 0 & f & h/2 \\ 0 & 0 & 1 \end{bmatrix}}_{\text{intrinsic } [\mathbf{K}]} \underbrace{\begin{bmatrix} r_1 & r_2 & r_3 & t_1 \\ r_4 & r_5 & r_6 & t_2 \\ r_7 & r_8 & r_9 & t_3 \end{bmatrix}}_{\text{rotation|translation } [\mathbf{R} \mid \mathbf{t}]} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{1}$$

This traditional augmentation method considerably expands the dataset by adding additional information. However, it does not generate a new variant of samples, and not all views effectively learn new features [63]. Recently, generating synthesis 3D shapes using conditional GAN (CGAN) was introduced to augment the dataset for training the DL framework on 3D object classification tasks [63,64]. To improve model generalization and address the risk of inconsistency during training, a synchronous data augmentation (SDA) strategy has been introduced for 3D object detection tasks [65]. This strategy involves performing synchronous augmentation on corresponding image data and point clouds, maintaining correspondence between modalities. By avoiding inconsistency issues and increasing dataset diversity through various transformations, SDA improves model accuracy, robustness, and adaptability to different scenarios and environments.

2.2. Evaluation metrics for classification task

In general, the numbers of samples across different classes are not necessarily equal in the datasets. Therefore, both instance accuracy (*Inst. Acc.*) and average class accuracy (*Class Acc.*) are measured to evaluate the performance of classification methods (Eq. (2), (3)) [66], where TP_i, TN_i, P_i and N_i are the number samples of true positive, true negative, positive and negative outcomes corresponding to the ith categories respectively, and C is the total number of categories in the sample. In addition, the $F1$ score is a metric used to evaluate a classification model's performance. It considers precision and recall to provide a more balanced measure of a model's accuracy (Eq. (4)). To compare the performance, the *Inst. Acc.* and the $F1$ score of DL methods on 3D object classification tasks is reported in this survey from the corresponding papers.

$$Inst.\ Acc. = \frac{\sum_{i=1}^{C} TP_i + TN_i}{\sum_{i=1}^{C} P_i + N_i}. \tag{2}$$

[1] https://modelnet.cs.princeton.edu/
[2] https://hkust-vgd.github.io/scanobjectnn/
[3] https://www.acfr.usyd.edu.au/papers/SydneyUrbanObjectsDataset.shtml

Table 1
A list of popular 3D datasets for 3D object analysis using DL methods.

Datasets	Year	Data type	No. samples/scenes	Training/Test	Classes
Omni3D [50]	2023	RGB	3M	175K/39K	98
ScanObjectNN [48]	2019	Point clouds	2902	2321/581	15
ScanNet [49]	2017	RGB-D	12 283	9677/2606	17
ShapeNet [55]	2015	Mesh	51 190	Not split	55
ModelNet40 [33]	2015	Mesh	12 311	9843/2468	40
ModelNet10 [33]	2015	Mesh	4899	3991/605	10
Sydney Urban Objects [47]	2013	Point clouds	588	Not split	14

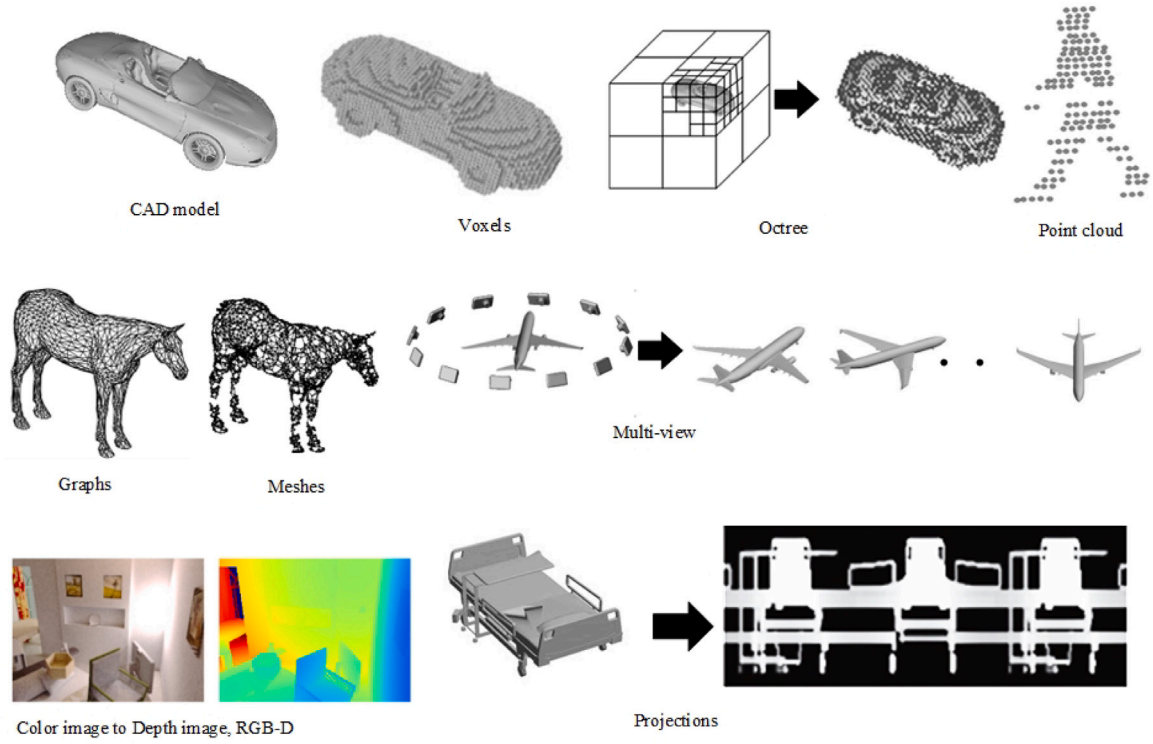


Fig. 2. Various 3D data representation for DL-based 3D shape analysis. Figures are adopted from Voxel and Octree [61], Point cloud [47], Graphs and Meshes [67] and Depth image [68], respectively.

$$Class\ Acc. = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i + TN_i}{P_i + N_i}. \tag{3}$$

$$F1 - score = 2\frac{(Precision * Recall)}{(Precision + Recall)}, \tag{4}$$

where,

$Precision = TP/(TP + FP)$, and $Recall = TP/(TP + FN)$

3. Data representations methods for 3D object recognition

Raw 3D data can be obtained directly using several low-cost scanning devices that have made it easier to generate or capture 3D objects in the real environment [10,69]. Microsoft Kinect [70] device can be used to scan objects into a 3D point cloud or RGB-D. In addition, 3D data is inherently high dimensional, making it challenging to apply DL algorithms directly. Transforming 3D data into fixed-sized representations (e.g., voxel grids or point clouds) enables consistent input sizes for DL models, reducing computational complexity and improving efficiency. Different representations can enable higher levels of abstraction and generalization. By transforming 3D data into a more abstract form (e.g., learned embeddings or encoded representations), DL models can detect underlying patterns and generalize across different 3D objects or scenes efficiently. The popular representation of 3D data used for 3D object recognition tasks using DL methods can be classified into point clouds, volumetric, multiview, projection, meshes, and graphs as shown in Fig. 2.

3.1. Point clouds

In 3D space, a point cloud is represented by a collection of individual points, each defined by X, Y, and Z coordinates. These coordinates determine the position of each point in the 3D Cartesian coordinate system. Point clouds can be visualized as dots or markers in a 3D space, with each dot representing a point in the cloud. The density and distribution of the points in the point cloud can provide insights into the shape, structure, and characteristics of the objects or environment being represented. Point clouds are commonly obtained through sensor technologies, such as LiDAR photogrammetry and depth cameras (e.g., Microsoft Kinect [70]). Point cloud representation has become very popular for the analysis of 3D objects using DL algorithms for several reasons. One reason would be that generating point clouds has become easier and more accessible [71–76]. Another reason would be to implement advanced DL and computer vision algorithms for point cloud processing as they make extracting valuable information from the data easier. Learning features from point clouds using DL algorithms can be divided into raw-point-based methods [1,72,77,78], and K-dimensional tree (Kd) [79–81] based methods. The detail of point-based

methods is discussed in Section 5.1.2. Increasing the number of points provides more detailed information on 3D shapes. The input points are usually sampled with 2048 and 1024 points, respectively, to use in the DL model [77,78].

However, processing raw point clouds is a challenging task because of unordered data structure, missing points, lack of connectivity information, noise, and other constraints from the environment and acquisition setup [55,56].

3.2. Volumetric

To solve the computational challenges of unordered point clouds using DL, volumetric representation is widely used to represent spars [14,56,82] 3D data on a regular grid. To transform volumetric data, a point cloud of a 3D shape is inserted into a volume, and then the points inside the volume are encoded corresponding to the direction of x, y, and z as a set, where the value of v is the property of data at a 3D location [83]. The value of v is either 0 or 1, which refers to the background and object, respectively. Voxels and octree are two major types of volumetric representations of 3D data used in DL [31,58]. Voxels are utilized to model 3D data throughout a scene in 3D space. The distribution of points on a 3D voxel grid is a popular method of volumetric representation to describe a geometric shape using a probability distribution function. The data tensor of the voxel grid is encoded into binary data. A binary tensor 1 and 0 is used to identify whether the voxel is in occupied or free space [33]. By categorizing the occupied voxels as visible, occluded, or self-occluded, it is also possible to encode viewpoint information about the 3D object. In addition, voxel representation with an occupancy grid provides richer information as it contains the information of the unknown space together in addition to occupied and free space [57]. Voxel can be treated as the resolution of the 3D shape or the size of the volume. Unlike the image pixel ratio, the number of voxels at the 3D axis is always equal, e.g., $128 \times 128 \times 128$. The size of grids with $32 \times 32 \times 32$ voxels and $30 \times 30 \times 30$ voxels, respectively, are commonly used to DL as an input sample of 3D objects. The dimensions of voxel grids with a size of 64^3 would be the same as the dimensions of images with a resolution of 256×256 pixels [33,58], respectively. Voxel-based representation depicts the scene's occupied and unoccupied areas, requiring a larger memory space and increasing the computational footprint. Therefore, this method is not always effective in computing high-resolution 3D data.

To minimize the computational and memory requirements of 3D CNNS, Octree is an alternative volumetric representation that is more efficient in analyzing 3D shapes using DL algorithms. Octrees are often used to divide the 3D spaces into eight octants recursively [84,85]. Octree follows the hierarchical data structure, which is similar to quadtree [86,87], where the subdivision process of octrees is applied both inside and outside of the object in the cube. The number of octree partitions or the generation of octrees is continued until the required depth of the octree is reached. The octree at the 4th, 5th, 6th, and 7th depths can be compared with voxel resolutions of 16^3, 32^3, 64^3, and 128^3, respectively [84]. Only non-empty nodes are considered when encoding information from octrees. Therefore, less memory is required to store information than a voxel representation. Since octree representation can extract the fine structure of a 3D shape, it provides richer information than voxel [61]. For the 3D object classification challenge, octree representations allow deeper networks with high resolutions and faster computational speed.

However, neither voxel nor octree representations capture the intrinsic characteristics or the smoothness of the surface of the 3D object, and implementing octree data requires a different DL kernel [61,84,88].

3.3. Multiview

Multiview representation of a 3D object is a set of 2D images from various points of view as shown in Fig. 2. For computer vision applications, the surface information of a 3D object provides very high-level features that can be generated with multiple 2D views of a 3D object. In general, several virtual cameras are set up around the object, and photos are taken by rotating the 3D object uniformly. This rotation can be applied both horizontally and vertically around the model. Multiview representation became popular as previously published many successful DL models in image processing can be used directly to solve the problem of 3D object recognition [3,34,89–91]. In addition, the multiview representation of 3D data enables the learning of multiple features by aggregating multiple views to recognize the shape accurately [88,92–95].

However, multiview approaches heavily rely on the availability of multiple viewpoints of an object during training. Consequently, they may struggle to generalize to novel viewpoints not seen during training. For example, a few views might fail to characterize the full geometry of the 3D shape, and too many views result in an unnecessary computing overhead.

3.4. Data projections

3D data projection is a technique to visualize a 3D shape on a 2D surface. Projecting 3D data into 2D space is a straightforward way to apply conventional DL methods on projected 2D image data of 3D shapes. When complex 3D data is projected into a simpler 2D space, some of the important characteristics of the original 3D geometry are preserved. This representation greatly solves the view-shifting problem while the 3D shape is rotated using the multiview method [96]. 3D data projections into cylindrical [96] and spherical domains [97]are a common practice to extract features of 3D shapes into 2D grids. Multiple projections are generally employed to capture both depth and surface information of the 3D shapes and their elements. These projections assist in making the projected data invariant to rotations along the primary axis of the projection. Projecting 3D shapes into 2D images with panoramic views and shooting multiple images from different view angles by avoiding image overlap are popular methods of 3D data projections [96,98–103]. Information loss is the major limitation of this method, so it cannot be used for some complicated 3D vision applications, e.g., dense correspondence [104]. However, occlusions or missing data can occur, especially when dealing with complex or cluttered scenes. These conditions can introduce noise or incomplete information into the data projection representation, potentially affecting classification accuracy.

3.5. Meshes and graph

One of the most common ways to represent 3D shapes is with 3D meshes. A 3D mesh model is a computer-aided design that consists of a set of polygons. The elements of meshes consist of a collection of vertices, edges, and faces. ModelNet10 and ModelNet40 [33] datasets of 3D shapes are the most popular SOTA datasets for 3D object recognition, where all samples are presented with 3D meshes. However, it is challenging to learn 3D meshes for DL models because of irregular representations [105]. However, 3D meshes can be easily converted to point clouds, volumetric, or other DL-supported representations, which can be input directly into DL models. The most common practice of using 3D meshes in DL is to represent graph-structured data where the vertices of the mesh are presented as the nodes of the graph [106–110]. To compute irregular points, the graph is used to relate neighboring points in geometric space and compute through graph convolution to learn the point clouds representation of a 3D shape [109]. While graphs can capture the connectivity information between vertices and edges of a 3D object, they may not fully represent the topological

Table 2
A summary of 3D data representations for the 3D object recognition task.

3D representations	Data acquisition method	Related publications	Advantages	Limitations
Point clouds	LiDAR, optical scanning	[1,78,97,111]	Contains both intrinsic and extrinsic features.	Computation is challenging because of the irregular distribution of points [112,113].
Voxels	Data Transformation	[114–116]	Represent full geometry of a 3D shape into regular grid-style data.	Computational cost is high as it includes both occupied and non-occupied spaces [40,117].
Octree	Data Transformation	[63,90,104]	Represent the full geometry with fine details of a 3D shape and require low computational cost.	Implementation is complex and cannot preserve intrinsic features.
RGB-D	Kinect Sensor	[49,106,118, 119]	Effective representation to learn 3D data from a noisy environment.	2D images with depth information are limited to learning the full geometry of a 3D object [119].
Mesh (CAD)	Computer-aided	[55,107]	Provide rich surface information of a 3D shape	Learning geometric features is difficult from complex mesh data.
Multiview	Virtual Cameras	[88,108,120]	The straightforward way to learn complex 3D objects using a 2D DL model.	Cannot represent the geometric information and the number of view selections is an open issue [121,122].
Graphs	3D meshes to graph-structured data	[109,110, 123]	Graph-based DL models can extract geometric information and can be used for mesh data.	Requires different DL kernels to implement [124].
3D Projection	Projecting 3D into 2D	[96,98]	Projected data is invariant to rotations.	Cannot preserve geometric information [104].

characteristics of the object. More complex topological features like handles, tunnels, or voids may be difficult to capture accurately in a graph representation. Graph representations may struggle to handle scalability issues when dealing with large-scale 3D objects.

Moreover, as the object's size grows, it becomes harder to process the graph structure efficiently. It is worth mentioning that while graph representations have limitations, they also offer advantages, such as efficient graph algorithms, easy traversal, and graph-based computations for certain types of analysis on 3D objects.

3.6. RGB-D

RGB-D Data presents 2.5D information which contains the depth map and RGB color information of 2D images (see Fig. 2). Recently, RGB-D representation of 3D objects has become popular in analyzing 3D shapes because of the availability of low-cost RGB-D scanners such as Microsoft's Kinect [70]. Since RGB-D data is simple but very effective in representing a 3D shape. It has become popular to analyze 3D data using DL algorithms for understanding 3D scenes and 3D object detection from noises and occlusion backgrounds. In addition, there are more public RGB-D datasets available compared to 3D datasets of point clouds or meshes such as SUN RGB-D [123], H3D Waymo Open, and nuScenes [40].

However, RGB-D sensors typically have a limited range for capturing depth information. Consequently, the analysis of 3D shapes might be compromised under certain lighting environments. RGB-D representations are also sensitive to dynamic changes in the environment, such as moving objects or variations in scene lighting. Therefore, it is important to consider the potential impact of environmental dynamics when using RGB-D data for 3D shape analysis.

3.7. Summary of 3D data representation methods

Processing or extracting features of 3D objects using DL algorithms is required to represent the 3D data into compatible shapes for computing. Apart from being inexpensive, the 3D point cloud is also a rich representation that defines surface and geometric information to represent a 3D object. Although multiview and data projection representations have become popular for analyzing 3D data in DL, they suffer from losing geometrical information. Voxel and octree (volumetric) are very precise representations to compute structured grid data of a 3D object in a volume. However, they are required to pre-process from point clouds or mesh data, which is time-consuming. The graph-based DL model has recently received attention in 3D object recognition tasks. Graphs concisely represent complex 3D shapes by describing their connectivity relationships between individual vertices and edges. This allows for efficient storage and processing of shape data. Table 2 summarizes 3D data representation methods widely used for 3D object recognition tasks.

4. Traditional methods for 3D object recognition

Traditional methods of 3D object recognition, before the advent of DL, often relied on handcrafted features [23,125,126] and conventional machine learning (ML) techniques [127]. Handcrafted 3D object recognition methods refer to approaches where features and descriptors are manually designed or engineered to recognize objects in three-dimensional space. These methods rely on domain knowledge and expertise to extract relevant information from the data, which can be divided into global feature-based methods and local feature-based methods [23]. Global feature-based methods aim to characterize the entire object as a single feature vector, capturing the overall shape and appearance. These methods typically involve extracting global descriptors from the 3D shape, such as histograms of shape distributions [128,129], moments, Spin Images, Extended Gaussian Images, 3D Shape Context, or Fourier descriptors [130]. These descriptors provide a compact representation of the entire object, which was used for 3D object classification. A histogram of the geometric features was one of the popular global feature-based methods. It involves computing distances between points or angles between vectors. By comparing these histograms, shapes can be matched and recognized [131]. However, global features may not capture fine-grained local details of objects, which can be crucial for distinguishing between objects with similar global shapes.

On the other hand, local feature-based 3D object recognition methods are techniques used to identify and classify objects based on distinctive local features extracted from the object's surface [131]. Local features are inherently robust to variations in viewpoint, scale, and illumination, making them suitable for handling objects with diverse appearances and configurations [132]. These features can include points, edges, corners, or other distinctive patterns. Common algorithms for feature extraction include SIFT (Scale-Invariant Feature Transform), SURF (Speeded Up Robust Features), ORB (Oriented FAST and Rotated BRIEF), and others [133–135]. Once the local features

are detected, they are described in a way that makes them robust to changes in viewpoint, scale, lighting conditions, and noise. This description allows for matching features across different views of the same object. Then the extracted features from the query object are compared with the features extracted from the objects in the scene. The goal is to find correspondences between the features of the query object and those in the scene [136]. Finally, based on the matches and pose estimation, the query object is recognized and classified within the 3D scene. Local feature-based methods offer superior performance in 3D object classification by leveraging robust, fine-grained representations of object geometry and appearance. Their ability to handle variations, selective attention to informative regions, and efficient matching capabilities make them a preferred choice for many computer vision applications [23]. However, they may require significant computational resources and may struggle with certain types of transformations or highly cluttered scenes [132].

In the following, machine learning (ML) algorithms played a crucial role by providing the capability to learn from data, generalize knowledge, and adapt to diverse scenarios in improving recognition accuracy and efficiency [137]. To improve the performance, a structured and informative representation of objects is extracted by local feature-based methods in 3D space. Once the local features are extracted and represented, traditional ML algorithms such as Support Vector Machines (SVM), k-nearest Neighbors (k-NN), Decision Trees, or Random Forests are employed for classification and recognition tasks [138,139]. These algorithms learn patterns from the extracted features and make predictions based on learned models [127,140]. While handcrafted-based traditional ML methods have been valuable in the early stages of 3D object recognition research, their limitations in scalability, generalization, and adaptability have driven the shift towards more data-driven and learning-based approaches, such as DL and neural networks [42].

Limitation. Both global and local feature-based methods for 3D object recognition come with their limitations [131].

- Noise or occluded regions can distort global features, leading to recognition errors. It may also not sufficiently distinguish between similar objects or handle intra-class variation well.
- Some global feature extraction algorithms can be computationally expensive, especially when dealing with large-scale 3D datasets or complex object shapes. This can hinder real-time performance in applications such as robotics or augmented reality.
- Similar local features may exist in different objects, leading to ambiguity in matching. Local features may also struggle with deformations that alter the object's surface characteristics
- Traditional ML methods heavily rely on handcrafted features designed by experts. This process is time-consuming and requires domain knowledge, making it less scalable and adaptable to new datasets or environments.

These limitations highlight the challenges faced by both approaches in achieving robust and accurate 3D object recognition across various scenarios. Despite these drawbacks, global feature-based methods remain valuable in certain applications where holistic shape representations are sufficient or where computational resources are limited. However, they are often complemented by local feature-based methods to improve robustness and accuracy in 3D object recognition tasks. In addition, local feature-based methods assist in employing traditional ML algorithms in 3D object recognition tasks by providing a meaningful representation of objects' surface characteristics [127,137].

5. DL methods for 3D object recognition

The 3D object recognition problem has been researched for the last few decades [4]. Still, DL-based techniques have been applied widely for a few years after releasing the Princeton ModelNet dataset in 2015 [5,33,141]. DL approaches can be categorized into two types based on how they are used: (i) Discriminative (i.e., CNNs), and (ii) Generative (Auto Encoders). Discriminative models compute the probability of an output for a certain given input. Meanwhile, generative models [101] calculate the joint probability distribution of the input–output. Several 3D convolutional, pooling, and fully connected layers are typically used in building deep convolutional neural networks (CNNs). Based on the input data type of the neural network, existing 3D object classification methods can be categorized into different groups, including volumetric-base, point-based, multiview-based, projection-based, and graph-based methods. This section presents DL methods according to their learning methods, including supervised, semi-supervised, and unsupervised methods. Supervised learning models give high accuracy to increase classification accuracy, and the volume of research using supervised models is the largest. In contrast, semi-supervised and unsupervised learning methods have been considered recently for 3D object recognition due to the lack of labeled training samples. A list of early approaches (2015–2019) for 3D object classification can be found on the leaderboard of the ModelNet webpage.[4]

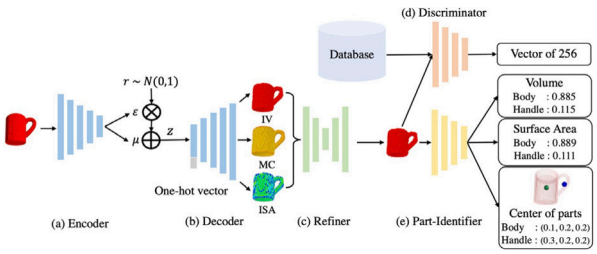


Fig. 3. PG-Net architecture for robust 3D object classification using volumetric representation [116].

5.1. Supervised DL models for 3D object classification

Supervised learning is defined by its use of labeled datasets to train the DL model to learn how to recognize future inputs of unlabeled data. The raw 3D data captured by scanning devices come in different forms. The raw 3D data are further transformed into a supportive input data format of the particular DL framework for the 3D object classification task. Some SOTA DL methods are discussed below, followed by supervised learning strategies.

5.1.1. Volumetric methods

Generally, volumetric methods involving two key operations, offline (preprocessing) and online (learning), refer to voxelizing a point cloud model into 3D grids and learning features by applying a 3D convolution neural network (CNN) on volumetric representation, respectively. Fig. 3 shows a volumetric model that extracts surface and volumetric representation using a single descriptor [116]. However, applying 3D CNN learning on volumetric data was previously described as a "nightmare" because of computational intractability [57]. Wu et al. introduced 3D ShapeNet [55], pioneering work for volumetric methods. 3D ShapeNet introduced 3D CNN on a 3D voxel grid for the first time in 3D shape classification. The idea was used for a similar level of detail at an image resolution of 165 × 165 pixels, which can be compared with a volume of 30 × 30 × 30 voxels. They also maintain a similar computational cost. Deep Belief Network [DBN] was extended to Convolutional DBN (CDBN) to reduce the model parameters. SVM was used to classify the object using features from the 5th layer of the ShapeNet. Classification performance was evaluated on their new 3D datasets, ModelNet40 and ModelNet10 [33], respectively. These datasets became the SOTA

[4] https://modelnet.cs.princeton.edu/

datasets for 3D object recognition in the following. Maturana et al. introduced VoxNet [57], which also used a voxel grid but could process multiple sources of point clouds (from both LiDAR and RGB-D), including meshes. To find a gap between volumetric and multiview CNNs, Qi et al. [117] investigated three volumetric frameworks, including auxiliary training by slicing volume into subvolumes, 3D to 2D projection using network layers incorporating an elongated anisotropic kernel, and multi-orientation pooling by aggregating information from multiple views of a 3D object, respectively. Inspired by this, several subvolume and auxiliary learning approaches were introduced to improve the learning of volumetric data and increase the classification accuracy, respectively, [35,60,84,142].

Hybrid Model. To improve the generalization ability of DL network, Cai et al. [143] introduced a hybrid parallel network to fuse multiple features from depth projection views of a 3D shape at the front, top, and side views, respectively. This parallel network was designed with four branches of networks, three branches were used to learn the projected images from three views, and one branch was used to learn voxels. Their findings depict that the combined view-based model performed better than a single-view one. Inspired by this, Gezawa et al. [144] introduced a new hybrid model to extract the local geometric feature by integrating voxel and point cloud data. This hybrid model better learned a higher-order local approximation function by combining two different types of 3D data while keeping a fixed number of points in each grid cell. It outperformed the SOTA volumetric methods on ModelNet40 and ModelNet10 datasets, except the VRN-ensemble method [101]. Liu et al. [145] proposed VB-Net by exploiting a broad learning system (BLS) to improve robustness. VB-Net [145] adopted a pre-trained VoxNet [57] to extract features from voxels, and then BLS was applied to create robust feature representation. The performance of VB-Net [145] on 3D shape classification was evaluated through a series of experiments to check its robustness. Gaussian noises were added to the 3D point clouds to evaluate classification performance over noisy input. In the following, the noisy point cloud models were voxelized accordingly. Experimental results show that the performance of VB-NET was slightly worse than noiseless inputs, but other SOTA methods such as VoxNet [57] were affected largely by noisy input. However, the experimental results of VB-Net revealed a new finding that the classification accuracy is improved by increasing voxel resolution and nodes and/or feature nodes [145].

Rotation Invariant Model. Evaluation of the robust performance of DL models is another issue for the real-time 3D object recognition task because most of the public datasets are built with noiseless 3D objects [48,55]. However, scanning environments and registration 3D models might have several kinds of noises. The performance of the 3D object recognition task might be affected by noisy input. It is easy to get noises to 3D models while scanning objects in natural environments or working on point cloud registration. To obtain a robust representation of 3D data, Mukhaimar et al. [102] proposed a rotation invariant model, RSCNN. This model applied a light spherical convolutional neural network framework on voxel grids. However, the voxelization process of the point clouds was followed by a spherical harmonics approach to perform robustness to noise and uncertainty in point cloud data. Spherical CNN was used to learn features on the voxel grid of concentric spheres over the unit ball. Generally, a spherical CNN is rotation equivariant, which helps improve classification performance. In addition, RSCNN incorporated the convolution operations in the Fourier domain just to avoid the inverse transformation, which helped to improve its robustness to data inaccuracies such as noise and outliers [102].

To improve the classification accuracy, dataset augmentation by rotation is a common practice to create multiple copies of each sample due to the lack of labeled 3D samples [57,81]. Therefore, PC-GAN [63] introduced a new concept of data augmentation using a conditional GAN model [101] to produce new samples of each category of 3D object in ModelNet40 [33]. With this new augmentation method, PC-GAN [63] outperformed the SOTA methods on ModelNet40 and ModelNet10 datasets. The major advantage of PC-GAN was that it augments the dataset with new varieties of existing 3D samples, which helps the DL model to learn new features from different varieties of 3D objects.

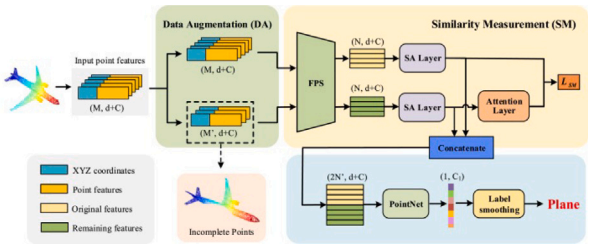


Fig. 4. IPC-Net framework for complete and incomplete point cloud classification [150].

Octree-based Methods. To reduce the computational footprints, OctNet [148] is one of the earliest DL models that takes octree representation of 3D data to classify their labels. Instead of using a single unbalanced octree, OctNet used several shallow octrees by restricting the maximal depth of an octree to a small number. A set of such shallow octrees was encoded using a bit string representation. The input resolution of $265 \times 256 \times 256$ was used to train the OctNet using a GPU of 12 GB memory only. In the following, O-CNN [61] and adaptive O-CNN [88] were introduced for the 3D object classification task. O-CNN was followed by the regular hierarchically octree partitions. It experimented with several sets of input regulations to experience the performance of O-CNN under different resolutions. In contrast, adaptive octree [88] introduced a patch-guided portioning strategy, and the computation starts from the leaf octants simultaneously at different octree levels. To better capture the sparse geometric features from both a full and partial part of the 3D object, MS-DNN [84] introduced sub-volume and multiscale residual learning strategies on octree data. Compared to the voxel method, MS-DNN performed seven times faster since octree only exploits nonempty nodes to store data. Motivated by these, octree has also been exploited significantly in 3D shape analysis [149] because of its efficient computational cost and low memory space requirement. A comprehensive comparison among volumetric methods is presented in Table 3.

Limitation. The voxel-based DL models provide encouraging performance, although they increase computational cost cubically. In addition, volumetric representation cannot preserve the primitive features of the original 3D models because the transformation of point clouds to voxels discards some information.

5.1.2. Point-based methods

Point cloud representations are popular for their ease of 3D shape representation and their direct usage in CNNs as inputs. An example of the point-based method is shown in Fig. 4. In the top panel of Fig. 4, we can observe the data augmentation and similarity measurement modules, both incorporating an attention mechanism. Moving to the bottom panel, the left side demonstrates the visualization outcome of the point cloud after undergoing random erasing. On the right side, we can inspect the enhanced classification network's structure [150]. To learn irregular points data, PointNet [78] is one of the earliest approaches proposed to address the problems of 3D object classification and segmentation tasks. PointNet directly inputs the pointset to a multilayer perceptron (MLP) network and outputs the object class label. However, PointNet does not capture the local structure induced by the pointwise feature extraction method. This problem was addressed in PointNet++ [70] by sampling a set of points in a hierarchical order. Both PointNet [78] and PointNet++ [70] learn features with several MLP (Multi-Layer Perceptions) layers. In the following, many point-based methods have been introduced with MLP because of its powerful

Table 3
A comparison of classification results by different volumetric methods. '#Params' refers to the network parameters, 'MN40 and MN10' represent ModelNet40 and ModelNet10, respectively. The symbol '-' implies that the results are unavailable. Results have been reported from corresponding articles accordingly.

Methods	Year	Input	#Params (M)	Accuracy (Inst Acc %)		
				M40	M10	Sydney (F1)
Hybrid Representation [144]	2022	Voxels, 32^3	8M	88.20	93.40	–
RSCNN [115]	2022	64*64*7	–	82.20	–	–
Hybrid parallel network [143]	2021	Voxel, projection	–	92.80	95.60	–
PG-Net [116]	2021	Voxels, 40^3	–	89.10	95.60	–
VB-Net [145]	2020	Voxels, 32^3	–	83.99	90.80	–
MV-DNN [61]	2020	Octree, 128^3	6.2M	92.93	–	–
LP-3DCNN [146]	2019	Voxels, 32^3	2M	92.10	94.40	–
NormalNet [59]	2019	Voxels, 30^3	6.5	91.90	93.10	74
3DWINN [147]	2019	Voxel, 32^3	–	–	93.60	–
Adaptive-OCNN [90]	2018	Octree, 32^3	–	90.50	–	–
OctNet [104]	2017	Octree, 128^3	–	86.50	90.90	–
O-CNN [63]	2017	Octree, 64^3	–	90.60	–	–
Aniprobing [117]	2016	Voxels, 30^3	–	89.90	–	–
Voxception [114]	2016	Voxels, 32^3	18M	90.60	93.28	–
VRN-ENSEMBLE [114]	2016	Voxels, 32^3	90M	**95.54**	**97.10**	–
3D ShapeNet [55]	2015	Voxels, 30^3	38M	83.00	83.50	83
VoxNet [58]	2015	Voxels, 30^3	0.9M	85.90	92.00	72

mapping ability [151–153]. To learn point cloud directly from LiDAR, Ben-Shabat et al. [154] proposed a hybrid point cloud representation, namely 3D modified Fisher Vectors (3DmFV). This method efficiently combines discrete point structures with a continuous generalization of Fisher Vectors using CNN for 3D point cloud classification and part segmentation. To recognize multiple objects in a scene, Gao et al. [155] proposed a new approach using the Euclidean distance clustering segmentation method for learning point cloud data. The main idea was to separate multiple point cloud objects in the context of different texture and color information in a scene into a single and then train a DL network to learn an individual point cloud sample for classification tasks. To capture the fine-grained contextual information, Liu et al. introduced Point2Sequence [156] using attention-based sequence to sequence network. It extracted features from point clouds by aggregating multi-scale features from each local region of 3D objects with attention. To learn contextual shape-aware information the Relation-Shape Convolutions Neural Network (RS-CNN) [157] was introduced for point cloud classification and segmentation problems. RS-CNN [157] was developed with shared MLP layers and 2D grid convolution layers for relation reasoning. To extract local features directly from disordered point clouds, PointFusionNet [158] exploited a feature-fusion strategy based on MLP. PointFusionNet [158] introduced Feature Fusion Convolution (FFC) and Global Relationship Reasoning Modules (GRM) for fusing the point-wise features and developing a global map of local features, respectively. Followed by [158], Qiu et al. [159] introduced a Dense-Resolution Network (DRNet) to learn local features from point clouds in multiple resolutions. DRNet was equipped with its newly proposed local neighborhood searching and error-minimizing modules. In the following, GBNet [160] was introduced to learn the geometric features of point clouds. The main contribution of this work was to improve learning of the geometric feature representation of point clouds by a geometric back-projection network. The idea of an error-correcting feedback structure was used to design a back-projection CNN. It introduced a channel-wise affinity attention module to refine the raw features of point clouds. However, point-based DL models always suffer from arbitrary rotations, so the classification accuracy will decrease once rotation-invariant is not maintained.

Rotation-invariant Methods. To tackle arbitrary rotations issue, several rotation-invariant models have been proposed using spherical coordinates to represent rotation-invariant features of point clouds [111,161–165]. To extract invariant properties to both global shape transformations and to local rotation, SPHNet [111] framework was proposed as a spherical harmonics kernel to rigid motions using point-based convolutions. Rao et al. [161] introduced spherical fractal convolutional neural networks to learn point clouds robustly. This framework was resilient against rotations and perturbations for point cloud recognition. To improve the generalization ability of DL to arbitrary orientations, Li et al. [162] introduced a new low-level rotation-invariant point cloud representation in spherical coordinates, which encodes both local and global features. These rotation invariant features are learned by a deep hierarchical network with their newly designed region relation convolution layers and MLP. Although the classification accuracy of this method [162] was not the highest, it is still comparable with the SOTA methods. In addition, it provided the most stable performance with no drop in accuracy while handling inputs at arbitrary rotations. Meng et al. [165] introduced a residual transformer network, Res-TNet, with a spatial transformer for learning non-alignment points. The pose alignment was applied using a transformation matrix before feature extraction, and residual learning was adopted to improve classification accuracy.

Occlusion and truncation are general problems while scanning point clouds in real-time, which might affect classification accuracy. To solve this problem, several approaches have been introduced to learn features from incomplete point clouds [150,166,167]. IPC-Net [150] utilized a random erasing-based data augmentation method where complete and augmented incomplete models of point clouds were used to train the network. To understand incomplete point clouds, Zhang et al. [166] introduced the PointSetVoting framework for learning partial point clouds by the latent feature encoding methods. The effectiveness of this voting strategy was demonstrated in its ability to handle partial observations. Each vote was created as a distribution in the latent space, enabling diverse predictions. Through the proposed training strategy on complete point clouds, the resulting model exhibits robust performance when dealing with partial observations during testing. This approach minimizes the requirement of collecting extensive partial point cloud datasets, thus reducing associated costs.

Robust Methods. Robustness is another important property that is very demanding for many safety-critical applications (i.e., autonomous driving) [112]. However, numerous studies revealed that the general classifier models are vulnerable to noisy input data or adversarial attacks while data points are modified, added, or deleted [113,168–172]. Liu et al. proposed PointGuard [173], a robust deep 3D model that worked against the point modification attack. PointGuard [173] generated several subsamples of point clouds. These subsamples contained a random subset of the points from the original point cloud. The majority votes among the labels of the subsamples predicted the label of the original point clouds. However, point clouds are somewhat different from rasterized data (voxels or pixels), which are formed in an irregular spatial fashion. Still, the main benefit of their coordinates is that they can be used straightforwardly to CNNs as inputs. A comparison of classification results of point-based methods is provided in Table 4.

Table 4
A comparison of classification results of point-based methods. '#Params' refers to the network parameters, 'MN40 and ScanObj' represents for Modelnet40 and ScanObjectNN datasets, respectively. Results have been reported from corresponding articles accordingly. The symbol '-' implies the results are unavailable.

Methods	Year	Input/Specialty	#Params(M)	*Inst. Acc.* (%)	
				M40	ScanObj
PointCAT [174]	2023	points,Robust	–	91.33	–
AdaptConv [175]	2023	2Kpoints,CNN	2.13M	**93.9**	82.1
IPC-Net [150]	2023	1K points,Incomplete point clouds	–	93.7	86.7
XPCC [176]	2023	1K points,KP-CNN	–	92.18	–
SPRIN [97]	2022	Points,Rotation Invariant	–	86.1	–
Snowpoints [1]	2022	1K points,CNN	0.12M	92.7	82.3
SC-CNN [124]	2022	1k points,CNN	–	93.80	**96.4**
GBNet [160]	2022	1K points, CNN	8.39M	93.8	80.5
PointFusionNet [158]	2021	1K points,CNN	1.4M	93.8	–
PRANet [177]	2021	2K points, MLP	–	93.7	82.1
DRNet [159]	2021	1K point,MLPs	–	93.1	80.3
SPHNet [111]	2021	2K Points,Rotation Invariant	–	87.7	–
Res-TNet [165]	2021	1K points, Rotation Invariant	–	71.2	–
PointSetVoting [166]	2021	1K points,MLP	–	91.4	–
RS-CNN [157]	2020	1K points,MLP& Conv	–	93.6	–
DenX-Conv [178]	2020	1K points,KNN& CNN	1M	92.5	–
Gao et al. [155]	2020	1K points,MLP & Conv	–	89.70	–
PIRIN [179]	2020	points, Rotation Invariant	0.44M	70.35	–
Point2Sequence [156]	2019	1K points,MLP	–	92.6	–
DGCNN [180]	2019	1K points,CNN	1.48M	92.2	78.1
SO-Net [181]	2018	2K points,KNN	–	90.9	–
SpiderCNN [182]	2018	1K points,CNN	–	92.4	–
PointCNN [183]	2018	1K points,CNN	–	92.2	78.5
PointNet [78]	2017	1K points,MLP	–	89.2	63.4
PointNet++ [70]	2017	MLP,1K points	–	90.7	77.9

Limitation. Due to the advancement of 3D scanning tools, it is very easy and fast to develop a point cloud model. However, it fails to preserve the surface information of the object in the real environment because of the unordered point sampling. In addition, the occlusion and truncation may destroy the completeness of objects, which also affects the classification performance. Therefore, incomplete point cloud classification needs to be investigated extensively [45].

5.1.3. Multiview-based methods

DL method on multiview representation is another popular method for shape classification. MVCNN [89] is a pioneering work of multiview-based methods for 3D object recognition. The main idea of MVCNN was to learn a 3D shape from a collection of its rendered views on 2D images. Multiple vanilla-type CNNs were used as descriptors to learn each view independently and produce high-level features by aggregating multiple view-wise representations. A max-pooling layer combined those high-level features into a global descriptor to predict the object labels. This method can adopt contemporary successful DL models in 2D computer vision, which can be used directly as backbone descriptors for learning a 3D object. Inspired by MVCNN [89], the multiview techniques have become popular and developed rapidly in the past few years. The max-pooling layer in MVCNN discards information by retaining only maximum values from each view, which results in an information loss. This characteristic of multiview methods limits their use in real-time applications such as robot navigation. In addition, selecting the number of views is also an open research question for multiview-based methods. To choose the next views of a 3D shape, Chen et al. [122] introduced a recurrent attention framework, VERAM. The key idea of this method was to exploit how humans can form a hypothesis about the category of a 3D shape without observing all views and how they move for the next viewpoint to narrow down the uncertainty. Inspired by this idea, VERAM exploited LSTM [184] for recurrent subnetwork. However, AlexNet [26] and ResNet [185] can be used in VERAM as the observation subnetwork. A similar approach of combining CNN and LSTM was proposed by Ma et al. [186] to produce correlative information from multiple views. To investigate how to jointly realize the selection of representative views and the measurement of similarity using multiview, MVSG-DNN [93] introduced the multiview saliency-guided method to explore the discriminative information of multiview sequences without specific viewpoint settings. Initially, they employed a multiview projection module to capture multiple views, then computed the visual saliency of individual views for learning visual context in the Saliency LSTM. It adaptively selects the perspective views, then, LSTM was used as the visual descriptor to perform the classification task. Inspired by this, Nie et al. [187] also utilized a similar approach to find discriminative views rather than taking all views equally. They exploited two modules, including view-attention pooling and LSTM, for computing the dynamic weighted sum of multiview features to CNN and developing visual saliency for aggregating multiview information, respectively. Finally, they employed a fusion module to fuse features from multiview CNN and employed LSTM to perform the classification task. Instead of using LSTM and the pooling method, Liu et al. [121] introduced the channel attention mechanism (CAM) and context information fusion module (CFM) to extract view-wise semantic information and aggregate multiview features. CAM and CFM consist of convolution modules to exploit the multiview context information and fuse context information to the classifier module. To address the imbalance of local features by appearing repeated parts, Huang et al. [188] investigated the representation of 3D features using 2D projections, and they found that different views of the same object have different features that contribute to the classification task.

However, not all views contain effective features, and they introduced a view-based weight network to identify only "good views" of a 3D object by assigning different weights to different projections. They used ResNet to extract features from multiview, and then view-based weight pooling was used to perceive good view images. The final classification task was performed using a voting scheme for multichannel integrated classifiers consisting of ELM, KNN, SVM, and RF. Nie et al. [189] introduced a hybrid Multimodal Information Fusion Network (MMFN) that fused features from MVCNN [89] and PointNet [77]. However, MMFN [188] incorporated two MVCNN [95] networks to learn features from multiview and panorama views of 3D objects, respectively. Kanezaki et al. [120] introduced RotationNet for joint object classification and pose estimation. RotationNet is a differentiable multi-layer neural network that can predict object labels by inputting a partial set of multiview images. This method is very effective for practical usage when it is mostly not allowed to scan an object in 3D.

Table 5
A comparison of classification results by different multiview-based methods. '#Params' refers to the network parameters in Million (M), 'MN40 and MN10' represent ModelNet40 and ModelNet10, respectively. Results have been reported from corresponding articles accordingly. The symbol '-' implies the results are unavailable.

Methods	Year	No. of Views	Pre-trained	#Params	*Inst. Acc.* (%)	
					M40	M10
Xu et al. [66]	2023	12	–	–	95.62	97.79
AMHN [191]	2023	12	–	–	**97.86**	–
MVDAN [192]	2022	12	ImageNet	–	96.3	–
Huang et al. [190]	2022	12	ImageNet	–	94.41	–
MVMSAN [92]	2022	20	ImageNet	–	96.96	–
FSDCNet [193]	2022	12	–	–	95.3	–
RAP-EDIR [194]	2022	12	–	–	–	95.2
Chu et al. [108]	2022	20	–	13.45M	97.49	**99.34**
HMVCM [195]	2021	12	ImageNet	25.6M	94.57	95.7
GA-MVCNN [196]	2021	12	–	–	96.2	–
MVSG-DNN [93]	2020	12	–	62.6M	92.3	94.0
Nie et al. [187]	2020	12	ImageNet	–	93.02	–
SCFN [121]	2020	12	ImageNet	–	93.1	94.1
VWN [188]	2020	12	ImageNet	–	93.82	95.16
MMFN [189]	2020	Fusion	ImageNet	–	94.00	–
VERAM [122]	2019	12	ImageNet	12.8M	92.1	93.7
Chao Ma et al. [186]	2019	12	ImageNet	–	91.05	95.47
RotationNet [34]	2019	20	ImageNet	102.M	97.4	98.5
GVCNN [197]	2018	8	ImageNet	–	93.1	–
MVCNN [89]	2015	80	ImageNet	138M	90.1	–

RotationNet achieves the SOTA performance on 3D object classification tasks. However, it cannot distinguish effective views from multiple views, so processing all views increases the computational footprint. In the following, Chu et al. [108] extended the idea of RotationNet [120] and introduced an independent viewpoint features extraction method to decrease the number of viewpoint selections. Zeng et al. [190] introduced a novel graph attention-based view selection module for multiview representation where MVCNN [95] extracted features from multiview images. Then, the extracted features were aggregated by the graph attention module to produce global features. Recently, Xu et al. [66] introduced another view-independent approach by computing inter-view relations via LSTM. The feature extraction method using view-relation constrained works in feature selection and feature aggregation tasks, respectively. Although important view-selection methods are very efficient for multiview observation, not all views but front views generally provide more important clues in 3D object recognition. Thus, the selection of views and content level should be investigated further. Generally, multiview-based methods benefit from leveraging techniques developed in 2D image processing and CNNs. The rich body of research and pre-trained models for multiview images can be readily applied, facilitating feature extraction and representation learning. Thus, the ability of multiview-based DL methods to harness multiple perspectives of 3D objects enables them to achieve better classification performance than other methods. A comparison of classification results by different multiview-based methods is provided in Table 5.

Limitation. Although multiview methods are dominating on 3D object classification tasks, however, some issues need to be investigated. For example, the number of view selections is still an open issue of multiview networks. Recently, some view selection methods [93,121, 122,186,187] have been introduced; however, initially, all views are taken into consideration to select some representative views of a 3D object. Further, a few selective views are used in the inference stage. Therefore, this method is computationally inefficient. In addition, it will suffer problems for 3D objects in occlusion. Several approaches were initiated to tackle this problem, but there are still several challenges [93] selecting of good views that need to be investigated further. Multiview approaches mainly rely on learned representations from the training set, which may not fully generalize to diverse datasets or novel object classes. As a result, their performance might be restricted when applied to different domains or categories of objects. These limitations highlight some considerations one should consider when utilizing multiview representations for 3D object classification.

5.1.4. Projection-based methods

3D data projection is another representation of 3D data where some key properties of the original 3D raw data are projected into 2D space. Shi et al. [96] proposed DeepPano for 3D object recognition and retrieval tasks. A cylindrical 2D projection around the principal direction of the 3D object was made to extract 2D panoramic views. Cao et al. [98] proposed two complementary projections on a spherical domain, producing cylindrical patches. The first projection stores depth variations and contour information encoded by the second projection on different angles. The proposed 2D CNN inputs a set of cylindrical patches as features of a 3D object. The trained model was applied for 3D object classifications. The major advantage of this projection method is using 2D DL directly for 3D applications [100]. The performance of this method highly depends on the type of projection method being used.

Limitation. This kind of representation is not optimal for 3D vision tasks because of information loss in the projection [104]. When 3D data is projected onto 2D, depth information is lost, making it difficult to estimate the true 3D structure of objects accurately. Ambiguous silhouettes can also lead to misclassification, as different objects can have similar projected silhouettes. Occlusion and self-occlusion further complicate recognition by obstructing parts of objects. Additionally, 2D projections are sensitive to viewpoint changes, requiring prior knowledge of an object's appearance from multiple viewpoints. To address these limitations, researchers are exploring alternative approaches that directly leverage 3D data, such as depth sensors or multiview images, to improve the accuracy and robustness of 3D object recognition.

5.1.5. Graph-based methods

Graph-based convolutional neural networks (CNNs) have gained attention for their ability to process data with irregular or non-Euclidean structures, such as point cloud data. Graph-based methods define the operation of neural networks (e.g., convolution, pooling) on graphical data. In such cases, each point in the point clouds is considered a vertex of the graph, with the connection between nodes as edges. An example of a graphical model is shown in Fig. 5. To aggregate multiview features while considering the interconnections among nodes in the graph, Wei et al. [123] proposed a novel view-based graph network, view-GCN. Multiple views were used as graph nodes to build a graph convolutional neural network. The view-GCN was followed by ResNet-18 architecture, which used images as input. To extract the global structure of the point cloud, Wan et al. [110] introduced a 3D classifier by applying deep graph convolutions to Reeb graphs. Reeb graphs provide very compact and informative information on raw point clouds.

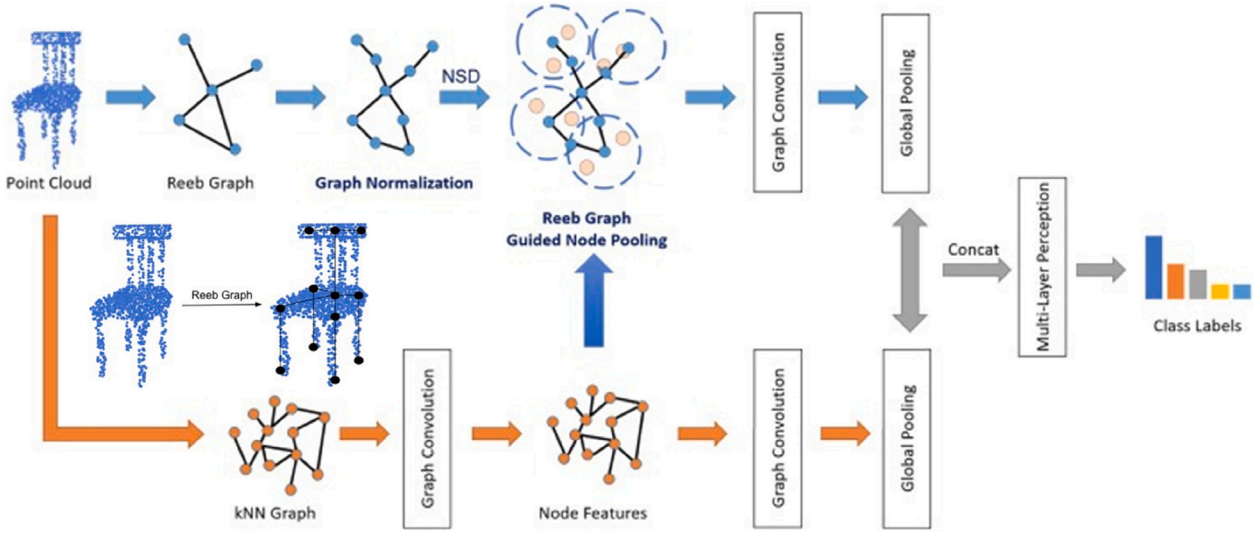


Fig. 5. An example of 3D object classification based on graph convolution [110]. The Reeb graph serves as a simplified representation of a manifold, capturing its skeletal form. This graph effectively encodes the overall structure of the point cloud data.

Table 6
A comparison of classification results of projection-based and graph-based methods. '#Params' refers to the network parameters, 'MN40 and MN10' represent ModelNet40 and ModelNet10, respectively. Results have been reported from corresponding articles accordingly. The symbol '-' implies the results are unavailable.

Model type	Methods	Year	Input size	#Params	*Inst. Acc.* (%)	
					M40	M10
Projection	Spherical Projections [98]	2017	Image	–	**94.24**	–
	DeepPano [96]	2015	Image	–	2.54	88.66
Graph-based models	GN-CNN [198]	2023	1k points	–	93.00	95.9
	MVDVAGCN [199]	2023	1k points	–	96.44	–
	Wei at al. [200]	2023	Image	–	**97.60**	–
	SACNN [201]	2023	1k points	–	93.20	–
	IPC-Net [150]	2023	1k points	–	93.70	–
	DRGCNN [109]	2022	Images	–	91.00	–
	IGCN [202]	2022	1k points	0.57M	90.10	–
	Reeb graph [110]	2021	1k points	–	89.10	–
	SPH3D-GCN [203]	2021	10k points	0.8M	92.10	–
	Guo et al. [204]	2021	2k points	–	93.60	–
	GFA-Net [74]	2021	1k points	–	93.20	–
	Wang et al. [205]	2021	1k points	–	92.90	–
	CVA [206]	2021	1k points	–	93.70	–
	DGCNN [207]	2021	1k points	1.84M	94.00	–
	View-GCN [123]	2020	Images	73.4M	96.50	96.4
	Grid-GCN [208]	2020	1kpoints	–	91.80	–
	DGNN [180]	2019	2k points	–	90.70	–
	DPAM [209]	2019	1k points	–	91.90	94.6
	Momenet [151]	2019	1k points	–	92.40	–
	PointGCN [210]	2018	2k points	–	89.50	91.90
	KCNet [211]	2018	2k points	0.9M	91.00	94.40
	ECC [212]	2017	1k points	–	87.40	0.80

Reeb graph features were normalized and fused with a KNN graph model for 3D object classification. To exploit geometric relationships in 3D data, Lei et al. [203] proposed a graph neural network, SPH3D-GCN, for learning point clouds. It utilized a spherical kernel to maintain translation-invariant and asymmetry properties. Similar to [203], Meng et al. [201] proposed self-augmented CNNs (SCNN), that aggregated dynamic graph structures to encoded local features from the point cloud without destroying its intrinsic structure. The graph structures were different across the self-augmented CNNs and their parameters were updated adaptively to learn point clouds' local and global features. This method can alleviate the translation-invariant problem with robust performance on 3D object classification. To ensure the transformation invariance of the global feature, Guo et al. [204] introduced an adaptive feature fusion module with dynamic graph CNN. It utilized three modules: local feature extractor, global feature extractor, and adaptive fusion. Motivated by this, GFA-Net [74] introduced a novel feature aggregation module to learn geometric features between points by integrating them in Euclidean space. Cui et al. [207] proposed DGCNN to learn both extrinsic and intrinsic features of point clouds using a geometric attention module in dynamic graph CNNs. It can learn geometric level relations as attentional weights and update dynamically through attentional EdgeConv.

To address the challenge of losing geometric information due to quantization, SC-CNN [124] introduced a new type of depth-wise separable convolution to extract positional and feature relationships between center and neighboring points. This network learned both surface and fine-grained shape information. Instead of using a random multiview of point clouds, Yue et al. [109] introduced DRGCNN, which employs the adaptive selection of local region features by graph CNN and attention mechanism. To reduce the complexity of CNN operations, IGCN [202] utilized interpolation graph CNNs to learn unordered point clouds. In addition, IGCN reduced the model parameter to 0.57M with

Table 7
A summary of supervised methods for 3D object classification task.

Supervised methods	Advantages	Limitations	Applications
Point-based Methods	Directly process raw 3D point data without requiring complex preprocessing or conversion, allowing for more accurate and efficient classification of irregular and detailed geometric structures [174,175].	Variations in point density and distribution can affect classification accuracy, making it challenging to maintain consistent performance across different resolutions and noise levels [112,113].	Aids in city planning and infrastructure maintenance include classifying urban environments, including buildings, and roads, from 3D point cloud data captured by drones or LiDAR scanners [213,214].
Volumetric Methods	Uniformly represent and process the entire 3D structure, enabling comprehensive feature extraction and robust classification results [59,116].	High computational and memory demands due to the need to process dense 3D grids [40,117].	Extensively used in medical imaging to classify and segment 3D medical scans such as MRI and CT images [215]. In autonomous driving, classify objects such as pedestrians, vehicles, and obstacles [216].
Multiview-based Methods	Integrate multiple 2D views for comprehensive feature capture and improve classification accuracy [217].	Aggregating and aligning multiple 2D views to form a comprehensive 3D representation is complex [121,122].	Enhance user interaction and experience in Virtual reality by using multiple viewpoints [41].
Projection-based Methods	Simplify processing and enhances computational efficiency by utilizing well-established 2D CNNs [218].	Spatial and depth information is lost when converting 3D data into 2D representations, which can reduce classification performance [104].	Simplify object recognition and manipulation, in robotics by analyzing 3D data projected onto 2D planes [218].
Graph-based Methods	Effectively capture complex relationships and geometric structures, improving accuracy and robustness [205].	High computational complexity and difficulty in scaling to large datasets [124].	In social robotics, these methods classify and understand human gestures by representing 3D skeletons as graphs, enhancing human–robot interaction [219].

a competitive performance with other SOTA methods. To learn deep-level visual features of point clouds, MVDVAGCN [199] introduced a new idea of multiview deep visual adaptive graph CNNs to learn deep-level abstract features by forming the optimal view from the multiview correlation. Sun et al. [198] proposed a novel data-driven model, GN-CNN to tackle the learning difficulties for unlabeled 3D point cloud samples. It utilized a GAN to generate prior knowledge to perform a warm startup pre-training for GN-CNN. The key idea of this approach was to focus on the visual appearance of the point clouds instead of their exact positions. In summary, the advantages of graph-based DL for 3D object classification lie in their ability to preserve structure, handle irregular data, model relationships, interactions, scalability, and effective integration with DL techniques. These advantages make graph-based methods a promising approach for 3D object classification tasks. A list of graph-based methods is presented in Table 6.

Limitation. Although graph-based methods have attracted more attention in the last few years, the performance of current graph-based methods is still worse than multiview-based methods on 3D object classification tasks. Traditional CNNs possess translation invariance, which allows them to recognize patterns regardless of their position within the input data. However, graph-based CNNs lack this translation invariance since the graph structure depends on node connectivity. Consequently, graph-based CNNs may struggle to detect patterns in different spatial locations within the graph. Interpreting the decisions made by a graph-based CNN can be challenging due to the lack of visual representation typical in grid-based CNNs. Understanding how specific nodes or edges contribute to the model's predictions is not as straightforward, making it difficult to examine the internal workings of the network.

5.1.6. Summary of supervised learning

A comparison of supervised 3D object classification methods is listed based on different categories of 3D data representations and chronologically by year of publication in Tables 4, 3, 5, and 6, respectively. The ModelNet40 and ModelNet10 [33] are the most used datasets to compare the 3D object classification performance. Finding representative views of 3D objects and model generalization are the key problems in 3D object analysis. We have summarized the advantages, limitations, and possible applications of supervised methods for 3D object classification in Table 7.

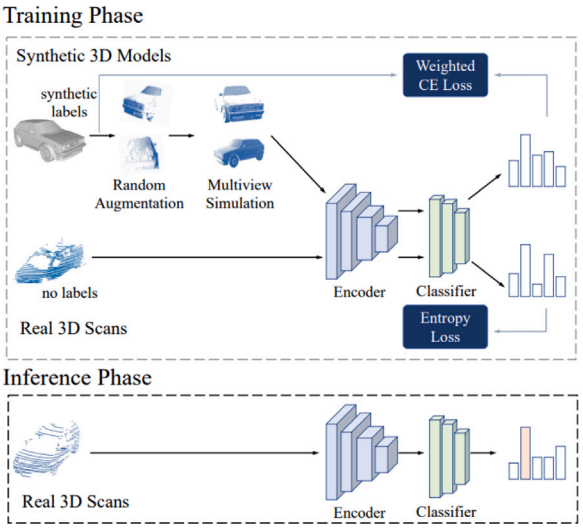


Fig. 6. The point-syn2real framework employed semi-supervised learning for point cloud classification using multiview representation [220].

5.2. Semi-supervised DL models for 3D object classification

Semi-supervised learning is a combination of supervised and unsupervised learning techniques. In this approach, a small portion of the labeled data and a larger amount of unlabeled data are provided. The labeled data is used to guide and improve the learning process, while the unlabeled data helps to capture the underlying distribution and structure of the data. The semi-supervised learning method is a good alternative to supervised learning when a large collection of label data is not readily available to train a DL model [221,222]. Taking inspiration from these methods, multiple semi-supervised techniques were developed for 3D shape classification by utilizing various DL variations such as autoencoders, GANs, and other similar methodologies.

5.2.1. Autoencoder-based methods

To utilize autoencoders for semi-supervised 3D object classification, a large amount of unlabeled 3D object data is used to pre-train the autoencoder. The model is trained to reconstruct the input data

accurately. The encoder part of the pre-trained autoencoder is used to extract the latent representation from both labeled and unlabeled data [223,224]. This latent representation captures important features of the objects. The encoder is combined with a classification layer, and the model is fine-tuned using the limited labeled data. The entire network is trained to minimize the classification loss by adjusting the encoder and classifier parameters. By adapting the pre-trained autoencoder to the semi-supervised 3D object classification task, the model can learn rich and discriminative latent representations from the unlabeled data. This enables the network to generalize to the labeled data more effectively and potentially improve classification performance.

To tackle the difficulties faced in unsupervised learning on point clouds, Yang et al. proposed FoldingNet using a graph-based encoder [225] for 3D object classification. It was trained on the ShapeNet dataset in a supervised fashion. Then, this trained model was used for 3D object classification on the ModelNet10 dataset using a linear SVM. FoldingNet performs well while using a small number of training samples. It achieves classification accuracy of more than 55% and close to 85% while using 1% and 20% of labeled data from the ModelNet10 dataset, respectively. Inspired by this, Wang et al. [220] proposed a semi-supervised cross-domain learning technique in generalizing across different domains to enhance the generalizability of the Model. This approach enables the Model to transfer the knowledge from synthetic 3D point clouds to real-world data obtained through LiDAR scanners. It follows RotationNet [34] to prepare the multiview representation of 3D data. Fig. 6 shows this semi-supervised framework was trained using synthesized data through automated simulation. This training enables the model to learn and acquire knowledge, which can then be utilized for making accurate inferences on real data [220]. Zdobylak et al. [226] introduced a new algorithm for a semi-supervised approach by combining triplet learning with a specific autoencoding structure trained within a comprehensive framework. The autoencoder was designed to represent point cloud data in a bottleneck space efficiently and was trained using unsupervised techniques. Incorporating the triplet model can effectively train a strong data representation using supervised data.

Limitation. Auto-encoder based methods can struggle to generalize well to unseen object classes or instances, as the learned latent representations may not capture the full complexity and diversity of 3D shapes. In addition, the learned latent representations in auto-encoder based methods can be difficult to interpret, which can make it challenging to understand the underlying features and patterns that the model is using for classification.

5.2.2. GAN-based methods

GAN-based DL methods in semi-supervised 3D object classification employ adversarial training to generate realistic objects and learn discriminative representations. By leveraging the generator and discriminator networks, these methods can enhance the classification performance using limited labeled data. To address the rotation issue in point cloud data and enhance the model's recognition capability, 3D-SsGAN [227] integrated spatial transformation networks into the discriminator of GAN. In this approach, 3D-SsGAN utilized a semi-supervised GAN to automatically enable the model to identify the most significant feature regions during training. To enhance the representation of a model's visual attributes for Metaverse applications, Sun et al. [198] introduced GN-CNN to generate prior knowledge of an unlabeled 3D point cloud using GAN. It can achieve competitive performance with limited supervision and classification accuracy of 91.80% when training size was reduced to 30%.

Limitation. Evaluating the performance of GAN-based semi-supervised methods for 3D object classification can be challenging. Traditional metrics may not fully capture the quality and diversity of generated samples, making it difficult to assess the model's performance objectively.

5.2.3. Heuristic-based methods

To tackle the difficulties associated with comprehending the unlabeled Semantic 3D point cloud, Shi et al. [228] introduced a framework for Open-Set semi-supervised learning by exploiting heuristic learning techniques. The heuristic approach is a learning-based paradigm that defines a meta-objective, usually as the loss on a separate labeled dataset. In open-set semi-supervised learning, the goal is to learn per-sample weights on unlabeled data in such a way that the model trained with these weighted losses minimizes the meta-objective. This entire problem can be reformulated as a bi-level optimization problem. In this context, the upper level involves optimizing the meta-objective, while the lower level involves optimizing the model parameters based on the weighted loss function. This formulation allows for the incorporation of both labeled and unlabeled data in training, with the aim of improving the model's performance on unseen data. This framework employs an unlabeled point cloud with a selective sample weighting technique. The research revealed that this Open-Set semi-supervised learning approach can significantly enhance the learning performance for vast quantities of unlabeled data.

Limitation. Heuristic-based methods can be sensitive to noise, missing data, or irregularities in the 3D object representations, which are common in real-world scenarios. This can limit their robustness and applicability.

5.2.4. Pseudo-label guided methods

A pseudo-label generated semi-supervised method is a technique used for training DL models when labeled data is limited. In this approach, the model first trains on the available labeled data. Then, it uses this initial model to predict labels for the unlabeled data points. These predicted labels are known as pseudo-labels. The model is then retrained using both the labeled data and the pseudo-labeled data to improve its performance. Inspired by this, Deng et al. [229] proposed a framework called WSC-Net based on weakly supervised learning to overcome the challenges of a real-world point cloud cluttered with background. This framework comprises two stages. In the first stage, a novel semi-supervised PointRGCN generates pseudo-foreground-background labels for each partially labeled object in the real-world point cloud. This allows for additional supervision using limited labeled points. In the second stage, a weakly supervised real-world point cloud classification network is utilized to classify the point cloud. A unique multi-task noise-robust loss function is introduced to ensure accurate training of the final classification model. This loss function is designed to handle shape-level, point-level, sparse, and noisy pseudo labels. Incorporating pseudo-label generation and noisy label learning improves overall accuracy by 0.7% compared to the baseline approach. Recently, He et al. [230] introduced a novel semi-supervised model for point cloud classification tasks. This method involves unsupervised learning in point cloud classification, where it simultaneously handles representation embedding and pseudo-classification tasks. Unlabeled samples are utilized to generate both hard and soft pseudo labels through a shared classifier. These labels guide the semi-supervised contrastive learning process, enhancing classification accuracy. The framework integrates components such as a point cloud encoder, projector, classifier, momentum encoder, and momentum projector. These components enable semi-supervised learning with guidance from hard–soft labels, facilitating adaptive interaction between supervised and unsupervised branches via representational distribution alignment. The classification accuracy by this method on ModelNet40 was 75.3%, 81.8%, 83.2%, and 87.9% when using data labels of 5%, 10%, 20%, and 40%, respectively. A list of semi-supervised methods is presented in Table 8.

Limitation. Pseudo-labeling involves assigning labels to the unlabeled data based on predictions made by a classifier trained on the labeled data. The accuracy of the pseudo-labeling process heavily influences the quality of the semi-supervised classification. If the pseudo-labels are incorrect or noisy, it can negatively impact the performance of the classifier.

Table 8
A comparison of classification results of semi-supervised methods. 'AEC' refers to Autoencoder, '#Params' denotes the network parameters in a million (M), and 'MN40' represents for Modelnet40. The symbol '-' implies the results are unavailable. Results have been reported from corresponding articles accordingly, but those marked with '*' have been reported from [228].

Type	Methods	Year	Input	Pretrained	#Params(M)	*Inst. Acc.* (%)		
						M40	ScanNet	ShapeNet
AEC	Point-Syn2Real [220]	2023	Multiview	M40	–	–	59.13	–
	Point-Syn2Real [220]	2023	Multiview	ShapeNet	–	–	63.8	–
	Zdobylak et al. [226]	2020	2K points	–	–	–	–	73.4
	FoldingNet(20%) [225]	2018	Voxel, 32^3	M40	–	85.0	–	–
GAN	DG-CNN [198]	2023	1k point-Graph	M40	1.8M	**91.8**	–	–
	3D-SsGAN [227]	2019	Points	M40	–	84.6	–	–
Heuristic-based	*ReBo [228]	2022	1Kpoints	–	–	85.5	–	–
	*LTWA [231]	2022	1Kpoints	–	–	84.9	–	–
	*OP-Match [232]	2021	1Kpoints	–	–	84.6	–	–
	*Multi-OS [233]	2020	1Kpoints	–	–	83.6	–	–
	*DS3L [234]	2017	1Kpoints	–	–	84.4	–	–
Pseudo-label	He et al.(40%) [230]	2024	1K points	M40	–	87.9	–	**83.9**
	WSC-Net [229]	2022	2Kpoints	–	–	–	**76.6**	–

Table 9
A summary of semi-supervised methods for 3D object classification task.

Sem-supervised methods	Advantages	Limitations	Applications
Autoencoder-based Methods	Autoencoders can leverage large amounts of unlabeled data to learn meaningful features, enhancing performance with limited labeled data. They can also be trained to denoise input data, improving classification robustness and reliability [225].	The learned feature representations by autoencoders can be difficult to interpret, making it hard to understand classification decisions. Additionally, the quality of these representations depends on the quantity and quality of the unlabeled data [238].	In manufacturing, detecting and classifying defects in 3D printed or machined parts to ensure quality control. Classifying and reconstructing 3D models of artifacts and archaeological findings for preservation and study [238].
GAN-based Methods	GAN-based methods have several advantages over autoencoder-based methods. They generate diverse synthetic examples, improve the generalization of classification models, and help learn robust representations resilient to noise and adversarial attacks, enhancing model reliability [227].	GAN-based models can be vulnerable to adversarial attacks, where slight perturbations to the input data can lead to incorrect classifications. This vulnerability poses a risk in applications where robustness and reliability are critical [239].	GAN-based methods often outperform autoencoder-based models in applications where generating high-quality, realistic samples is crucial. They are useful for simulating diverse driving scenarios and augmenting training datasets, thereby improving the robustness and generalization of object classification models [240].
Heuristic-based Methods	Heuristic methods leverage domain-specific knowledge, improving classification accuracy in specialized domains by tailoring solutions to unique 3D object characteristics [228]	Heuristic methods may struggle to adapt to complex or evolving datasets and may require manual adjustment when faced with new or diverse types of 3D objects [241].	Heuristic methods can classify components of 3D building models to aid in construction planning, monitoring, and facility management, ensuring that models conform to standards and specifications [241].
Pseudo-label guided Methods	Enhance the effectiveness of semi-supervised learning in 3D object classification by assisting in domain adaptation, managing class imbalance, and improving model performance and generalization [229].	Pseudo-labels generated by an initial model can be noisy, especially if the model is not well-trained, leading to poor learning outcomes [242].	Pseudo-label guided methods enhance object classification for improving robot navigation, path planning, and autonomous vehicle perception in complex environments [242].

5.2.5. Summary of semi-supervised learning

The major advantage of this model is that they do not require human annotation explicitly to train it. The limited availability of labeled 3D object datasets hinders the model's ability to learn comprehensive and diverse representations. In addition, heuristic-based semi-supervised methods can be more robust to certain types of data irregularities in comparison to autoencoders and GAN, as they rely on domain-specific rules and feature engineering. Table 8 presents a comparison of semi-supervised 3D object classification methods. Semi-supervised learning aims to leverage labeled and unlabeled data for training, obtaining labeled data for 3D objects can be expensive and time-consuming. In semi-supervised autoencoders, generative, and descriptive DL models are commonly used [225,235–237] to perform robustness. The major limitation of semi-supervised 3D object classification is the scarcity of labeled data [231]. Semi-supervised learning in 3D object classification provides practical benefits in various fields where labeled data is limited but unlabeled 3D data is abundant. Table 9 summarize key advantages, limitations, and possible applications of semi-supervised methods for 3D object classification.

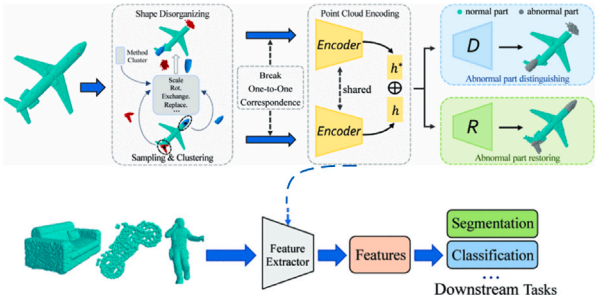


Fig. 7. A point cloud encoding network for unsupervised point cloud understanding [243].

5.3. Unsupervised DL models for 3D object classification

Unsupervised learning is a type of ML where an algorithm learns patterns and relationships within a dataset without being given explicit

labels or targets. It explores the inherent structure or distribution of the data to find meaningful patterns, clusters, or representations. Unsupervised representation learning can be broadly classified into generative and discriminative approaches. Generative approaches generally utilize auto-encoding [225,244,245] and adversarial learning [237,246–248]. On the other hand, discriminative approaches, including self-supervised learning [62,249], generate discriminative labels without supervision to learn representation learning.

5.3.1. Autoencoder-based methods

The encoder-based models capture meaningful and discriminative features in the latent space, allowing for effective clustering or classification of 3D objects. By training the encoder on unlabeled data, these models can learn representations that capture the inherent structure and patterns within the data, enabling unsupervised classification without the need for labeled examples. Sharma et al. [141] proposed a CNN-based volumetric auto-encoder, VConv-DAE, to reconstruct 3D shapes from noisy 3D data. It achieves competitive results for the unsupervised 3D object classification task. A new autoencoder called FoldingNet [225] was introduced to obtain a codeword capable of representing high-dimensional embedding point clouds. This approach replaced the traditional fully-connected decoder with a folding-based decoder. To attain robustness to adversarial attacks for volumetric data, Huang et al. proposed 3DWINN [141] followed by Wasserstein Introspective Neural Networks, which consist of volumetric generative and discriminative networks for modeling 3D objects and 3D scenes, respectively. 3DWINN achieved a competitive performance for unsupervised classification. MAP-VAE [244] employed involved making predictions by dividing the point cloud into two halves (front and back) at various angles. These predictions were combined with global self-supervision techniques to capture the point cloud's geometry and structure. Despite the potential advancements shown by these methods, the issue of sampling variation still poses a considerable challenge when representing decoder output as point clouds. This is because the network needs to encapsulate the geometric aspects and deficiencies arising from the specific point cloud sampling in the latent code [250]. To acquire robust shape representations through unsupervised learning, Chen et al. [243] introduced an autoEncoder-based framework for unsupervised Point Cloud Understanding. The main idea of this method is to learn strong representations of 3D shapes by destroying local parts and restoring incomplete parts to normal for the classification task (see Fig. 7). Encoder-based models may struggle to accurately assign objects to specific classes when there is ambiguity or overlap in the learned representations.

Limitation. Encoder-based unsupervised methods primarily focus on learning low-level features and representations from raw 3D data. While they may capture geometric and structural properties of objects, they often lack semantic understanding. This means that they may struggle to differentiate between objects of similar shapes or classify objects based on their functional or categorical properties.

5.3.2. GAN-based methods

Unsupervised methods for learning 3D features [237,251,252] do not usually perform as well as supervised ones [23,205]. However, GAN-based models draw much attention to the analysis of 3D objects, including 3D object generation, reconstruction, and classification [101,237,248,253]. 3D-GAN [237] generates 3D objects from a probabilistic space incorporating volumetric CNN. To evaluate the generated objects, the learned model performed the 3D object classification task on the ModelNet dataset [33] in an unsupervised fashion. 3D-GAN achieves better classification accuracy than other unsupervised methods. Khan et al. [248] proposed an unsupervised primitive 3D GAN to represent 3D objects which was trained with a single-layer CNN for classification. The primitive approach of 3D-GAN also achieved a competitive performance for 3D object classification on the ModelNet dataset. A multiview-based View Inter-Prediction GAN (VIP-GAN) [254] was introduced to solve the multiview inter-prediction task by training a Recursive Neural Network (RNN). In general, multiview-based networks [185,255] achieve better performance on 3D object classifications, and this VIP-GAN also achieved the highest accuracy among the unsupervised methods. However, GANs are prone to mode collapse, where the generator fails to generate diverse and representative samples. In 3D object classification, mode collapse can result in limited variations in the generated objects, leading to inadequate coverage of the underlying object distribution. This narrow variety of generated objects can hinder the ability of the GAN-based model to learn a comprehensive representation space.

Limitation. GANs are prone to the problem of mode collapse, where the generator learns to produce only a limited set of samples, leading to reduced diversity in the generated 3D objects. This can negatively impact the classifier's ability to learn robust features.

5.3.3. Self-supervised methods

To learn the implicit representation of 3D data without manual annotation, self-supervised (SSL) learning has gained popularity in unsupervised learning because of its ability to leverage unlabeled data effectively, learn generalized representations, reduce reliance on labeled data, and facilitate transfer learning for downstream tasks [76,256,257]. SSL is an ML technique in which the machine predicts and learns from different parts of its input for any observed component [258]. SSL can be divided into two categories: pretext tasks and downstream tasks. Pretext tasks involve training an AI system in SSL to acquire valuable interpretations of unorganized data [259]. These acquired interpretations can later serve as input for downstream tasks, such as supervised learning or reinforcement learning. Based on this pretext, SSL can be broadly categorized into reconstruction-based [258], contrast-based [256], and alignment-based [62] methods for unsupervised 3D object classification.

Reconstruction-based Methods. The key idea of reconstruction-based models is that objects from the same class should have consistent and similar reconstructions. Therefore, after training the model, objects from different classes can be reconstructed and evaluated based on their similarity to the reconstructed objects of known classes. To exploit the implicit surface representation of point cloud, IAE [250] introduced an implicit autoEncoder that encodes both geometric and information of the discrete sampling. This model specifically addressed the sampling variation problem of point cloud data by replacing the traditional decoder with an implicit decoder, which reconstructs the discrete point cloud in a continuous representation. To address the challenge of location information leakage and uneven information density of point cloud data, Point-MAE [258] divided point cloud into irregular point patches and randomly masked them at a high ratio. Then, a standard transformer framework adopted an asymmetric encoder–decoder design where the last layer of this model adopts a regular prediction head. This model was pre-trained on the ShapeNet dataset. Point-MAE achieved (93.8%) the SOTA performance on the ModelNet40 dataset compared to other self-supervised methods. To enhance the capture of local geometric information, Zhang et al. [260] introduced Mask Surfel Prediction framework, MaskSurf. This approach simultaneously estimates the surfel position (points) and per-surfel orientation (normals). A two-head pre-training paradigm in MaskSurf has been validated to yield more effective representations than relying solely on reconstruction as a pretext. The major limitation of reconstruction-based methods is that they focus more on capturing low-level or pixel-level details, which may not always be meaningful or relevant for higher-level tasks. This can limit the ability of the learned representations to generalize well to complex and abstract concepts, as they may not capture the underlying semantic or structural information in the data.

Contrast-based Methods. Contrast-based models for unsupervised 3D object classification work by leveraging contrastive learning to

learn discriminative representations of objects without requiring explicit class labels. These models enable extracting meaningful features that can be used for classification tasks. To overcome to learn rotation-invariant representation, [261] introduced a decoder-free representation learning method by maximizing mutual information between objects and their transformations. To fasten the convergence of the model, Wang et al. [262] introduced a self-supervised model using a mask transformer and contrastive learning. The masking in the transformer can effectively learn the masked parts, and contrastive learning maximizes the mutual information between the input and learned high-level features, respectively. Following [262], Shao et al. [256] introduced an unsupervised framework using contrastive learning concepts for 3D Intracranial aneurysms classification to improve unsupervised learning. This method incorporated a dual-branch encoder model to augment the input point cloud by utilizing CNN for the representation vector and MLP for abstraction levels. In the following, the outputs of each encoder were used as input for unsupervised tasks. The experimental results show that this dual-branch contrastive framework performs better than general contrastive learning methods. For contrastive learning, the mutual information (MI) between clean samples or adversarial samples (X, Y) and classification tasks, T can be formulated as [256]:

$$MI(\mathrm{T}, X) = \sum_{X,Y} p(X,Y) \log \frac{p(X,Y)}{p(X)p(Y)}, \tag{5}$$

$$MI\left(\mathrm{T}, X_{adv}\right) = \sum_{X_{adv},Y} p\left(X_{adv}, \widetilde{Y}\right) \log \frac{p\left(X_{adv}, \widetilde{Y}\right)}{p\left(X_{adv}\right) p(\widetilde{Y})}. \tag{6}$$

To improve the generalization performance of 3D DL, Sun et al. [249] introduced adversarial point cloud classification using an encoder framework using semi-supervised learning. The adversarial output of the encoder was used as input for the classification framework. This approach performed better adversarial classification accuracy by incorporating DGCNN [180] model. To construct a stable and invariant point cloud representation, AFSEL utilized [263] a fusion network by applying data-level augmentation and feature enhancement together in an unsupervised manner. However, contrastive-based self-supervised learning may struggle with capturing high-dimensional or fine-grained structures in the data, as the similarity or dissimilarity relationships are typically based on global features or representations. This can limit the discriminative power of the learned representations, especially for tasks that require fine-grained details. In addition, this approach involves computing pairwise similarities, which can pose challenges for scaling up to larger datasets or real-time applications with limited computational resources.

Alignment-based Methods. Point cloud representation exhibits inherent invariance to transformations such as time flow, spatial motion, and multiview photography. Leveraging this characteristic, alignment-based methods have been developed to learn implicit embeddings of point clouds. These methods preserve the coherence of point features through techniques like spatiotemporal consistency, multiview alignment, and multimodal fusion [264]. To address the intricate nature of 3D scene understanding, STRL [265] utilized the rich spatio-temporal cues from the point cloud to learn the invariant representation self-supervised. It takes two temporally correlated point clouds as input and applies spatial augmentation by changing local geometry to learn a better spatial structure representation of point clouds. To extract a robust representation of point clouds in occlusion, Wang et al. introduced OcCo [76] to generate mask point clouds through view-point occlusion and reconstruct the occluded object using an encoder–decoder framework. This approach enhances accuracy in few-shot learning scenarios and boosts overall generalization accuracy in fully-supervised tasks. Tran et al. [264] introduced a knowledge distillation method to establish a global loss that promotes similarity in feature distribution between images and point clouds to enhance the regularization of self-supervision. However, Alignment-based approaches may struggle to generalize well to unseen or novel data that deviates significantly from the training distribution. In addition, these methods often require significant computational resources and time for aligning and processing the data, making them less practical for real-time or resource-constrained applications. Table 10 shows a comparative classification result of several semi-supervised and unsupervised models for 3D classification tasks.

Limitation. The pretext tasks used in self-supervised learning are often designed for specific applications or datasets, which can limit the generalization of the learned representations to other 3D object classification tasks. However, the latent representations learned through reconstruction-based methods [260] may not be optimally suited for the end task of 3D object classification, as the model may prioritize accurate reconstruction over learning discriminative features

5.3.4. Summary of unsupervised learning

The comparison of unsupervised 3D object classification methods is provided chronologically in Table 10. In unsupervised 3D object classification, contrastive learning generally outperforms autoencoders and GANs due to its ability to learn robust and discriminative features, better scalability, effective generalization, and more stable training processes [262,267]. Table 11 summarizes key advantages, limitations, and possible applications of unsupervised methods for 3D object classification.

6. Future research direction

In this section, we have discussed future research directions and identified potential areas for further advancements in DL-based 3D object recognition as follows:

1. ***Learning with Limited Data.*** Developing DL techniques for learning from limited annotated 3D data, such as few-shot or zero-shot learning approaches, can help address the challenge of data scarcity and enable recognition of rare or novel object categories.
2. ***Domain Adaptation and Transfer Learning.*** Domain adaptation for 3D object classification is a challenging task, especially due to the differences in data distributions between source and target domains. Techniques such as meta-learning approaches [273] can adapt quickly to new domains by training on various source domains and their corresponding adaptation strategies. Domain-adaptive normalization methods can adjust their parameters adaptively based on the input domain [220]. Ensemble methods can combine multiple adaptation strategies [189], while semantic segmentation methods can aid in aligning object semantics across domains. Additionally, self-supervised learning can help the model learn transferable features that are robust to domain shifts, thereby improving domain adaptation for 3D object classification [264].
3. ***Robust 3D object recognition.*** The robustness of 3D classification models is underexplored, with current methods vulnerable to noise, incomplete data, and adversarial attacks [24,59,102,113,145,169,173,274–276]. These models often fail when data points are modified, added, or deleted, highlighting the need for more robust real-time, partial 3D object learning models. Additionally, 3D object recognition is challenged by sensitivity to orientation and the difficulty of identifying representative viewpoints for daily objects. Developing rotation-invariant representations and establishing benchmarking protocols are crucial for improving reliability and performance in 3D object classification. Future robust, rotation-invariant 3D object classification models may include quaternion-based convolutions, attention mechanisms, adversarial training, self-supervised learning, and meta-learning for improved adaptability and resilience.

Table 10
A comparison of classification results of unsupervised methods. 'AEC' refers to Autoencoder, '#Params' denotes the network parameters in a million (M), and ''MN40 and MN10' represent ModelNet40 and ModelNet10, respectively. Results have been reported from corresponding articles accordingly. The symbol '-' implies the results are unavailable.

Type	Methods	Year	Input	#Params (M)	*Inst. Acc.* (%)	
					M40	M10
Autoencoder	IAE [250]	2023	1K points	1.8M	93.7	–
	Transformer-OcCo [76]	2022	1K points	1.8M	92.1	–
	OcCo [62]	2021	1K points	1.8M	93.0	–
	Chen et al. [266]	2021	1K points	89.28K	90.4	–
	MAP-VAE [244]	2019	Points	–	90.15	94.82
	3DWINN [141]	2019	Voxel, 32^3	–	–	91.9
	FoldingNet [225]	2018	2k points	–	88.4	94.4
	Vconv-DAE [245]	2016	Voxel, 24^3	–	75.5	80.5
	PG-Net [116]	2021	Voxel, 32^3	–	89.10	**95.6**
GAN	PointDist [236]	2020	Points	–	84.70	–
	VIP-GAN [247]	2019	Points	–	**91.98**	94.05
	3D-GAN [237]	2017	Voxel, 32^3	–	83.3	91.0
	Primitive GAN [248]	2019	Voxel, 50^3	–	86.4	92.2
	3D-DescripNet [253]	2018	Voxel, 32^3	–	83.8	92.4
	3D-GAN [248]	2019	Voxel, 50^3	22.53M	84.5	91.2
Self-supervised	Wang et al. [262]	2023	1K points	–	93.0	–
	Jiang et al. [267]	2023	1K points	–	90.32	95.09
	Shao et al. [256]	2023	2K points	–	90.79	95.01
	Sun et al. [249]	2023	1K points	1.8M	**94.20**	–
	Mltiview Rendering [264]	2022	Hybrid	–	91.70	–
	Point-MAE [258]	2022	1K points	–	93.80	–
	Point-BERT [76]	2022	1k points	22.1M	93.20	–
	CrossNet [268]	2022	1K points	–	93.40	–
	MaskSurf [260]	2022	–	–	92.96	–
	STRL [265]	2021	1K points	–	90.9	–
	SG-NET [269]	2021	–	–	90.9	–
	Chen et al. [266]	2021	1K points	–	90.36	–
	Info3D [261]	2020	1K point	–	89.8	–
	PointOE [259]	2020	Points	–	90.75	–
	ACD [235]	2020	Points	–	89.80	–
	PointGrow [270]	2020	Points	0.25M	85.80	–
	Multitask [254]	2019	Graph	–	89.10	–

Table 11
A summary of unsupervised methods for 3D object classification task.

Unsupervised methods	Advantages	Limitations	Applications
Autoencoder-based Methods	Powerful for unsupervised 3D object classification, offering efficient feature learning, robustness, and flexibility without the need for labeled data [141].	Autoencoders might suffer from mode collapse, where they fail to capture the full diversity of the input data [238].	Autonomous vehicles, robotics, medical imaging, AR/VR, geospatial analysis, manufacturing, cultural heritage, and entertainment [238].
GAN-based Methods	GAN-based methods provide high-quality feature learning, robust data augmentation, improved generalization, effective anomaly detection, efficient transfer learning, and broad applicability across different domains [254] .	GAN-based methods in unsupervised 3D object classification face limitations like training instability, high computational demands, imperfect data generation, evaluation challenges, mode collapse, data dependency, complex architecture design, and limited interpretability [239].	Autonomous driving, medical imaging, and augmented reality, where high-quality 3D object classification is essential [240].
Self-supervised methods	Self-supervised methods in unsupervised 3D object classification enhance feature learning, generalization, versatility, and robustness while reducing dependency on labeled data by using reconstruction, contrastive learning, and data augmentations [250,256,267,271]	They face limitations including complex training, reconstruction errors, challenges in contrastive learning, reliance on augmentation quality, scalability issues, evaluation difficulties, domain-specific adaptation needs, and limited interpretability [272].	Autonomous vehicles, robotics, medical imaging, AR/VR, geospatial analysis, manufacturing, cultural heritage preservation, and entertainment, enhancing object recognition, scene understanding, and interaction with 3D environments [272].

4. ***Real-World Deployment Considerations.*** To enhance the real-world deployment of 3D object classification systems, it is crucial to address challenges in data collection, processing, model performance, and usability. This can be achieved by integrating data from multiple sensors like LIDAR and RGB-D cameras to enrich input data, incorporating attention mechanisms to focus on relevant 3D data features, and adapting EfficientNet for 3D data to balance accuracy and computational efficiency. Additionally, employing advanced neural architectures such as sparse convolutions for efficient data processing, graph neural networks for modeling relationships in point clouds, and hybrid models combining CNNs and Transformers can significantly improve performance. Implicit neural representations like Neural Radiance Fields (NeRF) further support real-time 3D object classification.
5. ***Lifelong and Incremental Learning.*** Exploring methods that can continuously learn and adapt over time, incrementally incorporating new object categories or adapting to concept drift, can enable lifelong learning in 3D object recognition systems.
6. ***Object recognition from 3D videos.*** With the advancement of 3D imaging and display technologies, nowadays 3D movies are very popular. Since it contains spatial and temporal information, object recognition from 3D videos could be a new window for analyzing 3D objects in cluttered scenes or occlusion. DL for 3D

object recognition in videos may include leveraging spatiotemporal CNNs to capture both spatial and temporal information, incorporating RNNs or attention mechanisms for better temporal dynamics, and utilizing self-supervised learning for improved generalization. Additionally, using graph neural networks to model spatial relationships and exploring novel architectures like capsule networks or neural architecture search (NAS) can enhance efficiency and performance in 3D object recognition tasks.

7. ***Non-rigid 3D object recognition.*** Non-rigid object recognition poses several challenges due to the deformable nature of the objects. Non-rigid objects can undergo significant shape variation, high intra-class variability, and viewpoint variation, which make it challenging to establish correspondences between different instances of the same object. Future DL methods for non-rigid 3D object recognition include using deformable convolutional neural networks (DCNNs) to adapt to varying shapes, dynamic graph CNNs (DGCNNs) to capture local and global geometric features, and spatio-temporal graph neural networks (ST-GNNs) to understand non-rigid deformations over time in dynamic sequences.

7. Conclusion

This paper provides a comprehensive survey of state-of-the-art DL techniques for 3D object recognition, exploring various approaches based on different types of 3D data representations from 2015 to 2023. We present a detailed taxonomy, summarizing and analyzing the strengths, limitations, and possible applications of each method. By examining the performance of these DL techniques on 3D benchmark datasets, we highlight key challenges in the field and propose potential future research directions.

CRediT authorship contribution statement

A.A.M. Muzahid: Writing – original draft, Investigation, Conceptualization. **Hua Han:** Funding acquisition, Formal analysis. **Yujin Zhang:** Investigation, Funding acquisition. **Dawei Li:** Writing – review & editing. **Yuhe Zhang:** Resources, Investigation. **Junaid Jamshid:** Resources, Investigation. **Ferdous Sohel:** Writing – review & editing, Supervision.

Declaration of competing interest

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 paper.

Data availability

The data that has been used is confidential.

Acknowledgments

This work is partly supported by the Natural Science Foundation of Shanghai, China (Grant No. 22ZR1426200 and 17ZR1411900) and the National Natural Science Foundation of China (Grant No. 62103257).

References

[1] Z. Xin, H. Wang, J. Zhang, Snowpoints: Lightweight neural network for point cloud classification, Comput. Electr. Eng. 104 (2022) 108463, http://dx.doi.org/10.1016/j.compeleceng.2022.108463.

[2] J. Zhang, Z. Sun, J. Sun, 3-DFineRec: Fine-grained recognition for small-scale objects in 3-D point cloud scenes, IEEE Trans. Instrum. Meas. 71 (2022) 1–11, http://dx.doi.org/10.1109/TIM.2021.3139685.

[3] Z. Yu, P. Tiwari, L. Hou, L. Li, W. Li, L. Jiang, X. Ning, MV-ReID: 3D multi-view transformation network for occluded person re-identification, Knowl.-Based Syst. 283 (2024) 111200, http://dx.doi.org/10.1016/j.knosys.2023.111200.

[4] G.J. Mamic, M. Bennamoun, Review of 3D object representation techniques for automatic object recognition, in: K.N. Ngan, T. Sikora, M.-T. Sun (Eds.), in: Visual Communications and Image Processing 2000, vol. 4067, SPIE, International Society for Optics and Photonics, 2000, pp. 1185–1197, http://dx.doi.org/10.1117/12.386708.

[5] A. Ioannidou, E. Chatzilari, S. Nikolopoulos, I. Kompatsiaris, Deep learning advances in computer vision with 3D data: A survey, ACM Comput. Surv. 50 (2) (2017) http://dx.doi.org/10.1145/3042064.

[6] L. Wang, Z. Song, X. Zhang, C. Wang, G. Zhang, L. Zhu, J. Li, H. Liu, SAT-GCN: Self-attention graph convolutional network-based 3D object detection for autonomous driving, Knowl.-Based Syst. 259 (2023) 110080, http://dx.doi.org/10.1016/j.knosys.2022.110080.

[7] V.R. Kumar, C. Eising, C. Witt, S.K. Yogamani, Surround-view fisheye camera perception for automated driving: Overview, survey & challenges, IEEE Trans. Intell. Transp. Syst. 24 (4) (2023) 3638–3659, http://dx.doi.org/10.1109/TITS.2023.3235057.

[8] X. Zhang, L. Wang, J. Chen, C. Fang, L. Yang, Z. Song, G. Yang, Y. Wang, X. Zhang, J. Li, Z. Li, Q. Yang, Z. Zhang, S.S. Ge, Dual radar: A multi-modal dataset with dual 4D radar for autonomous driving, 2023, arXiv:2310.07602.

[9] S.Y. Alaba, J.E. Ball, A survey on deep-learning-based LiDAR 3D object detection for autonomous driving, Sensors 22 (24) (2022) http://dx.doi.org/10.3390/s22249577.

[10] W. Jang, M. Park, E. Kim, Real-time driving scene understanding via efficient 3-D LIDAR processing, IEEE Trans. Instrum. Meas. 71 (2022) 1–14, http://dx.doi.org/10.1109/TIM.2022.3197771.

[11] L. Yang, B. Li, W. Li, H. Brand, B. Jiang, J. Xiao, Concrete defects inspection and 3D mapping using CityFlyer quadrotor robot, IEEE/CAA J. Autom. Sin. 7 (4) (2020) 991–1002, http://dx.doi.org/10.1109/JAS.2020.1003234.

[12] M. Rezaei, M. Azarmi, F.M.P. Mir, 3D-net: Monocular 3D object recognition for traffic monitoring, Expert Syst. Appl. 227 (2023) 120253, http://dx.doi.org/10.1016/j.eswa.2023.120253.

[13] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D.L. Rus, S. Han, BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation, in: 2023 IEEE International Conference on Robotics and Automation, ICRA, 2023, pp. 2774–2781, http://dx.doi.org/10.1109/ICRA48891.2023.10160968.

[14] Z. Song, G. Zhang, J. Xie, L. Liu, C. Jia, S. Xu, Z. Wang, VoxelNextFusion: A simple, unified, and effective voxel fusion framework for multimodal 3-D object detection, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12, http://dx.doi.org/10.1109/TGRS.2023.3331893.

[15] L. Yang, X. Zhang, J. Li, L. Wang, C. Zhang, L. Ju, Z. Li, Y. Shen, SGV3D:Towards scenario generalization for vision-based roadside 3D object detection, 2024, arXiv:2401.16110.

[16] L. Yang, J. Yu, X. Zhang, J. Li, L. Wang, Y. Huang, C. Zhang, H. Wang, Y. Li, MonoGAE: Roadside monocular 3D object detection with ground-aware embeddings, 2023, arXiv:2310.00400.

[17] L. Wang, X. Zhang, F. Zhao, C. Wu, Y. Wang, Z. Song, L. Yang, J. Li, H. Liu, Fuzzy-NMS: Improving 3D object detection with fuzzy classification in NMS, 2023, arXiv:2310.13951.

[18] L. Wang, X. Zhang, Z. Song, J. Bi, G. Zhang, H. Wei, L. Tang, L. Yang, J. Li, C. Jia, L. Zhao, Multi-modal 3D object detection in autonomous driving: A survey and taxonomy, IEEE Trans. Intell. Veh. 8 (7) (2023) 3781–3798, http://dx.doi.org/10.1109/TIV.2023.3264658.

[19] Z. Song, C. Jia, L. Yang, H. Wei, L. Liu, GraphAlign++: An accurate feature alignment by graph matching for multi-modal 3D object detection, IEEE Trans. Circuits Syst. Video Technol. 34 (4) (2024) 2619–2632, http://dx.doi.org/10.1109/TCSVT.2023.3306361.

[20] Z. Song, H. Wei, L. Bai, L. Yang, C. Jia, GraphAlign: Enhancing accurate feature alignment by graph matching for multi-modal 3D object detection, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 3335–3346, http://dx.doi.org/10.1109/ICCV51070.2023.00311.

[21] J. Gu, H. Hu, H. Li, Local robust sparse representation for face recognition with single sample per person, IEEE/CAA J. Autom. Sin. 5 (2) (2018) 547–554, http://dx.doi.org/10.1109/JAS.2017.7510658.

[22] S. Biasotti, A. Cerri, M. Aono, A.B. Hamza, V. Garro, A. Giachetti, D. Giorgi, A. Godil, C. Li, C. Sanada, et al., Retrieval and classification methods for textured 3D models: A comparative study, Vis. Comput. 32 (2016) 217–241.

[23] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, 3D object recognition in cluttered scenes with local surface features: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 36 (11) (2014) 2270–2287, http://dx.doi.org/10.1109/TPAMI.2014.2316828.

[24] B. Drost, M. Ulrich, N. Navab, S. Ilic, Model globally, match locally: Efficient and robust 3D object recognition, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 998–1005, http://dx.doi.org/10.1109/CVPR.2010.5540108.

[25] W. Liu, J. Sun, W. Li, T. Hu, P. Wang, Deep learning on point clouds and its application: A survey, Sensors 19 (19) (2019) http://dx.doi.org/10.3390/s19194188.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, vol. 25, 2012.

[27] W. Wei, T. Can, W. Xin, L. Yanhong, H. Yongle, L. Ji, et al., Image object recognition via deep feature-based adaptive joint sparse representation, Comput. Intell. Neurosci. 2019 (2019).

[28] J.-C. Su, M. Gadelha, R. Wang, S. Maji, A deeper look at 3D shape classifiers, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.

[29] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, et al., Deep learning for computer vision: A brief review, Comput. Intell. Neurosci. 2018 (2018).

[30] Z. Zhu, X. Wang, S. Bai, C. Yao, X. Bai, Deep learning representation using autoencoder for 3D shape retrieval, Neurocomputing 204 (2016) 41–50, http://dx.doi.org/10.1016/j.neucom.2015.08.127, Big Learning in Social Media Analytics.

[31] B. Leng, X. Zhang, M. Yao, Z. Xiong, A 3D model recognition mechanism based on deep Boltzmann machines, Neurocomputing 151 (2015) 593–602, http://dx.doi.org/10.1016/j.neucom.2014.06.084.

[32] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.

[33] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D ShapeNets: A deep representation for volumetric shapes, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015, pp. 1912–1920, http://dx.doi.org/10.1109/CVPR.2015.7298801.

[34] A. Kanezaki, Y. Matsushita, Y. Nishida, RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5010–5019, http://dx.doi.org/10.1109/CVPR.2018.00526.

[35] S. Zhi, Y. Liu, X. Li, Y. Guo, Toward real-time 3D object recognition: A lightweight volumetric CNN framework using multitask learning, Comput. Graph. 71 (2018) 199–207, http://dx.doi.org/10.1016/j.cag.2017.10.007.

[36] Z. Song, G. Zhang, L. Liu, L. Yang, S. Xu, C. Jia, F. Jia, L. Wang, RoboFusion: Towards robust multi-modal 3D object detection via SAM, 2024, arXiv:2401.03907.

[37] E. Ahmed, A. Saint, A.E.R. Shabayek, K. Cherenkova, R. Das, G. Gusev, D. Aouada, B. Ottersten, Deep learning advances on different 3D data representations: A survey, 1, 2018, arXiv preprint arXiv:1808.01462.

[38] W.G. Hatcher, W. Yu, A survey of deep learning: Platforms, applications and emerging research trends, IEEE Access 6 (2018) 24411–24432, http://dx.doi.org/10.1109/ACCESS.2018.2830661.

[39] L. Carvalho, A. von Wangenheim, 3D object recognition and classification: a systematic literature review, Pattern Anal. Appl. 22 (2019) 1243–1292, http://dx.doi.org/10.1007/s00371-023-02921-y.

[40] 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) (2021) 4338–4364, http://dx.doi.org/10.1109/TPAMI.2020.3005434.

[41] S. Qi, X. Ning, G. Yang, L. Zhang, P. Long, W. Cai, W. Li, Review of multi-view 3D object recognition methods based on deep learning, Displays 69 (2021) 102053.

[42] P.K. Vinodkumar, D. Karabulut, E. Avots, C. Ozcinar, G. Anbarjafari, A survey on deep learning based segmentation, detection and classification for 3D point clouds, Entropy 25 (4) (2023) http://dx.doi.org/10.3390/e25040635.

[43] Z. Song, L. Liu, F. Jia, Y. Luo, G. Zhang, L. Yang, L. Wang, C. Jia, Robustness-aware 3D object detection in autonomous driving: A review and outlook, 2024, arXiv:2401.06542.

[44] A. Hazer, R. Yildirim, Deep learning based point cloud processing techniques, IEEE Access 10 (2022) 127237–127283, http://dx.doi.org/10.1109/ACCESS.2022.3226211.

[45] S.A. Bello, S. Yu, C. Wang, J.M. Adam, J. Li, Review: Deep learning on 3D point clouds, Remote Sens. 12 (11) (2020) http://dx.doi.org/10.3390/rs12111729.

[46] A.S. Gezawa, Y. Zhang, Q. Wang, L. Yunqi, A review on deep learning approaches for 3D data representations in retrieval and classifications, IEEE Access 8 (2020) 57566–57593, http://dx.doi.org/10.1109/ACCESS.2020.2982196.

[47] M. De Deuge, A. Quadros, C. Hung, B. Douillard, Unsupervised feature learning for classification of outdoor 3D scans, in: Australasian Conference on Robitics and Automation, vol. 2, University of New South Wales Kensington, Australia, 2013, pp. 1–9.

[48] M.A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, S.-K. Yeung, Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 1588–1597, http://dx.doi.org/10.1109/ICCV.2019.00167.

[49] A. Dai, A.X. Chang, M. Savva, M. Halber, T. Funkhouser, M. Nießner, ScanNet: Richly-annotated 3D reconstructions of indoor scenes, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 2432–2443, http://dx.doi.org/10.1109/CVPR.2017.261.

[50] G. Brazil, A. Kumar, J. Straub, N. Ravi, J. Johnson, G. Gkioxari, Omni3D: A large benchmark and model for 3D object detection in the wild, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 13154–13164, http://dx.doi.org/10.1109/CVPR52729.2023.01264.

[51] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, J. Zhu, Benchmarking robustness of 3D object detection to common corruptions in autonomous driving, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 1022–1032, http://dx.doi.org/10.1109/CVPR52729.2023.00105.

[52] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3354–3361, http://dx.doi.org/10.1109/CVPR.2012.6248074.

[53] 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: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 11618–11628, http://dx.doi.org/10.1109/CVPR42600.2020.01164.

[54] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, Z. Chen, D. Anguelov, Scalability in perception for autonomous driving: Waymo open dataset, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 2443–2451, http://dx.doi.org/10.1109/CVPR42600.2020.00252.

[55] 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.

[56] M. Alam, M. Samad, L. Vidyaratne, A. Glandon, K. Iftekharuddin, Survey on deep neural networks in speech and vision systems, Neurocomputing 417 (2020) 302–321, http://dx.doi.org/10.1016/j.neucom.2020.07.053.

[57] B. Akay, D. Karaboga, R. Akay, A comprehensive survey on optimizing deep learning models by metaheuristics, Artif. Intell. Rev. (2022) 1–66.

[58] D. Maturana, S. Scherer, VoxNet: A 3D convolutional neural network for real-time object recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 922–928, http://dx.doi.org/10.1109/IROS.2015.7353481.

[59] C. Wang, M. Cheng, F. Sohel, M. Bennamoun, J. Li, NormalNet: A voxel-based CNN for 3D object classification and retrieval, Neurocomputing 323 (2019) 139–147, http://dx.doi.org/10.1016/j.neucom.2018.09.075.

[60] X. Yan, C. Zheng, Z. Li, S. Wang, S. Cui, Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5589–5598.

[61] A.A.M. Muzahid, W. Wan, F. Sohel, N.U. Khan, O.D. Cervantes Villagómez, H. Ullah, 3D object classification using a volumetric deep neural network: An efficient octree guided auxiliary learning approach, IEEE Access 8 (2020) 23802–23816, http://dx.doi.org/10.1109/ACCESS.2020.2968506.

[62] H. Wang, Q. Liu, X. Yue, J. Lasenby, M.J. Kusner, Unsupervised point cloud pre-training via occlusion completion, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9762–9772, http://dx.doi.org/10.1109/ICCV48922.2021.00964.

[63] P.-S. Wang, Y. Liu, Y.-X. Guo, C.-Y. Sun, X. Tong, O-CNN: Octree-based convolutional neural networks for 3D shape analysis, ACM Trans. Graph. 36 (4) (2017) http://dx.doi.org/10.1145/3072959.3073608.

[64] A. Muzahid, W. Wanggen, F. Sohel, M. Bennamoun, L. Hou, H. Ullah, Progressive conditional GAN-based augmentation for 3D object recognition, Neurocomputing 460 (2021) 20–30, http://dx.doi.org/10.1016/j.neucom.2021.06.091.

[65] J. Bi, H. Wei, G. Zhang, K. Yang, Z. Song, Dyfusion: Cross-attention 3D object detection with dynamic fusion, IEEE Lat. Am. Trans. 22 (2) (2024) 106–112, http://dx.doi.org/10.1109/TLA.2024.10412035.

[66] R. Xu, Q. Mi, W. Ma, H. Zha, View-relation constrained global representation learning for multi-view-based 3D object recognition, Appl. Intell. 53 (7) (2023) 7741–7750.

[67] A.M. Bronstein, M.M. Bronstein, R. Kimmel, A. Bronstein, M. Bronstein, R. Kimmel, Discrete geometry, Numer. Geom. Non-Rigid Shapes (2009) 41–65.

[68] J. McCormac, A. Handa, S. Leutenegger, A.J. Davison, SceneNet RGB-D: Can 5M synthetic images beat generic ImageNet pre-training on indoor segmentation? in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 2697–2706, http://dx.doi.org/10.1109/ICCV.2017.292.

[69] Z. Bi, L. Wang, Advances in 3D data acquisition and processing for industrial applications, Robot. Comput.-Integr. Manuf. 26 (5) (2010) 403–413, http://dx.doi.org/10.1016/j.rcim.2010.03.003.

[70] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems, vol. 30, 2017.

[71] J. Han, L. Shao, D. Xu, J. Shotton, Enhanced computer vision with microsoft kinect sensor: A review, IEEE Trans. Cybern. 43 (5) (2013) 1318–1334, http://dx.doi.org/10.1109/TCYB.2013.2265378.

[72] G. Wang, L. Wu, Y. Hu, M. Song, Point cloud simplification algorithm based on the feature of adaptive curvature entropy, Meas. Sci. Technol. 32 (6) (2021) 065004.

[73] X. Zou, K. Li, Y. Li, W. Wei, C. Chen, Multi-task Y-shaped graph neural network for point cloud learning in autonomous driving, IEEE Trans. Intell. Transp. Syst. 23 (7) (2022) 9568–9579, http://dx.doi.org/10.1109/TITS.2022.3150155.

[74] Y. Wang, C. Yue, X. Tang, A geometry feature aggregation method for Point-Cloud classification and segmentation, IEEE Access 9 (2021) 140504–140511, http://dx.doi.org/10.1109/ACCESS.2021.3119622.

[75] A. Akhtar, Z. Li, G.V.d. Auwera, L. Li, J. Chen, PU-dense: Sparse tensor-based point cloud geometry upsampling, IEEE Trans. Image Process. 31 (2022) 4133–4148, http://dx.doi.org/10.1109/TIP.2022.3180904.

[76] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, J. Lu, Point-BERT: Pre-training 3D point cloud transformers with masked point modeling, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 19291–19300, http://dx.doi.org/10.1109/CVPR52688.2022.01871.

[77] Y. Cui, Y. An, W. Sun, H. Hu, X. Song, Lightweight attention module for deep learning on classification and segmentation of 3-D point clouds, IEEE Trans. Instrum. Meas. 70 (2021) 1–12, http://dx.doi.org/10.1109/TIM.2020.3013081.

[78] R.Q. Charles, H. Su, M. Kaichun, L.J. Guibas, PointNet: Deep learning on point sets for 3D classification and segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 77–85, http://dx.doi.org/10.1109/CVPR.2017.16.

[79] R. Klokov, V. Lempitsky, Escape from cells: Deep kd-networks for the recognition of 3D point cloud models, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 863–872, http://dx.doi.org/10.1109/ICCV.2017.99.

[80] R. Zhang, G. Li, W. Wiedemann, C. Holst, KdO-net: Towards improving the efficiency of deep convolutional neural networks applied in the 3D pairwise point feature matching, Remote Sens. 14 (12) (2022) http://dx.doi.org/10.3390/rs14122883.

[81] W. Zeng, T. Gevers, 3DContextNet: Kd tree guided hierarchical learning of point clouds using local and global contextual cues, in: Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018.

[82] Z. Song, H. Wei, C. Jia, Y. Xia, X. Li, C. Zhang, VP-net: Voxels as points for 3-D object detection, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12, http://dx.doi.org/10.1109/TGRS.2023.3271020.

[83] A.E. Kaufman, 43 - Volume visualization in medicine, in: I.N. Bankman (Ed.), Handbook of Medical Imaging, in: Biomedical Engineering, Academic Press, San Diego, 2000, pp. 713–730, http://dx.doi.org/10.1016/B978-012077790-7/50050-3.

[84] A. Muzahid, W. Wan, L. Hou, A new volumetric CNN for 3D object classification based on joint multiscale feature and subvolume supervised learning approaches, Comput. Intell. Neurosci. 2020 (2020) 1–17.

[85] P.-S. Wang, Y. Liu, X. Tong, Dual octree graph networks for learning adaptive volumetric shape representations, ACM Trans. Graph. 41 (4) (2022) http://dx.doi.org/10.1145/3528223.3530087.

[86] F. Tian, Y. Gao, Z. Fang, Y. Fang, J. Gu, H. Fujita, J.-N. Hwang, Depth estimation using a self-supervised network based on cross-layer feature fusion and the quadtree constraint, IEEE Trans. Circuits Syst. Video Technol. 32 (4) (2022) 1751–1766, http://dx.doi.org/10.1109/TCSVT.2021.3080928.

[87] F. Jaillet, C. Lobos, Fast quadtree/octree adaptive meshing and re-meshing with linear mixed elements, Eng. Comput. 38 (4) (2022) 3399–3416.

[88] Z. Liu, Y. Zhang, J. Gao, S. Wang, VFMVAC: View-filtering-based multi-view aggregating convolution for 3D shape recognition and retrieval, Pattern Recognit. 129 (2022) 108774, http://dx.doi.org/10.1016/j.patcog.2022.108774.

[89] A.-A. Liu, N. Hu, D. Song, F.-B. Guo, H.-Y. Zhou, T. Hao, Multi-view hierarchical fusion network for 3D object retrieval and classification, IEEE Access 7 (2019) 153021–153030, http://dx.doi.org/10.1109/ACCESS.2019.2947245.

[90] P.-S. Wang, C.-Y. Sun, Y. Liu, X. Tong, Adaptive O-CNN: A patch-based deep representation of 3D shapes, ACM Trans. Graph. 37 (6) (2018) 1–11.

[91] Z. Yu, L. Li, J. Xie, C. Wang, W. Li, X. Ning, Pedestrian 3D shape understanding for person re-identification via multi-view learning, IEEE Trans. Circuits Syst. Video Technol. (2024) http://dx.doi.org/10.1109/TCSVT.2024.3358850, 1–1.

[92] W. Wang, X. Wang, G. Chen, H. Zhou, Multi-view SoftPool attention convolutional networks for 3D model classification, Front. Neurorobotics 16 (2022) http://dx.doi.org/10.3389/fnbot.2022.1029968.

[93] H.-Y. Zhou, A.-A. Liu, W.-Z. Nie, J. Nie, Multi-view saliency guided deep neural network for 3-D object retrieval and classification, IEEE Trans. Multimed. 22 (6) (2020) 1496–1506, http://dx.doi.org/10.1109/TMM.2019.2943740.

[94] Y. Wang, W. Zhong, H. Su, F. Zheng, Y. Pang, H. Wen, K. Cai, An improved MVCNN for 3D shape recognition, in: 2021 IEEE International Conference on Emergency Science and Information Technology, ICESIT, 2021, pp. 469–472, http://dx.doi.org/10.1109/ICESIT53460.2021.9696941.

[95] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 945–953.

[96] B. Shi, S. Bai, Z. Zhou, X. Bai, DeepPano: Deep panoramic representation for 3-D shape recognition, IEEE Signal Process. Lett. 22 (12) (2015) 2339–2343, http://dx.doi.org/10.1109/LSP.2015.2480802.

[97] Y. You, Y. Lou, R. Shi, Q. Liu, Y.-W. Tai, L. Ma, W. Wang, C. Lu, PRIN/SPRIN: on extracting point-wise rotation invariant features, IEEE Trans. Pattern Anal. Mach. Intell. 44 (12) (2022) 9489–9502, http://dx.doi.org/10.1109/TPAMI.2021.3130590.

[98] Z. Cao, Q. Huang, R. Karthik, 3D object classification via spherical projections, in: 2017 International Conference on 3D Vision, 3DV, 2017, pp. 566–574, http://dx.doi.org/10.1109/3DV.2017.00070.

[99] M. Yavartanoo, E.Y. Kim, K.M. Lee, SPNet: Deep 3D object classification and retrieval using stereographic projection, in: C. Jawahar, H. Li, G. Mori, K. Schindler (Eds.), Computer Vision – ACCV 2018, Springer International Publishing, Cham, 2019, pp. 691–706.

[100] Z. Xie, K. Xu, W. Shan, L. Liu, Y. Xiong, H. Huang, Projective feature learning for 3D shapes with multi-view depth images, in: Computer Graphics Forum, vol. 34, Wiley Online Library, 2015, pp. 1–11.

[101] P. Papadakis, I. Pratikakis, T. Theoharis, S. Perantonis, PANORAMA: A 3D shape descriptor based on panoramic views for unsupervised 3D object retrieval, Int. J. Comput. Vis. 89 (2010) 177–192.

[102] S. heum Kim, Y. Hwang, I.S. Kweon, Category-specific upright orientation estimation for 3D model classification and retrieval, Image Vis. Comput. 96 (2020) 103900, http://dx.doi.org/10.1016/j.imavis.2020.103900.

[103] Q.T. Chiem, M. Lech, R.H. Wilkinson, A hybrid two-stage 3D object recognition from orthogonal projections, in: 2019 13th International Conference on Signal Processing and Communication Systems, ICSPCS, 2019, pp. 1–5, http://dx.doi.org/10.1109/ICSPCS47537.2019.9008740.

[104] G. Riegler, A.O. Ulusoy, A. Geiger, OctNet: Learning deep 3D representations at high resolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6620–6629, http://dx.doi.org/10.1109/CVPR.2017.701.

[105] J. Hong, K. Kim, H. Lee, Faster dynamic graph CNN: Faster deep learning on 3D point cloud data, IEEE Access 8 (2020) 190529–190538, http://dx.doi.org/10.1109/ACCESS.2020.3023423.

[106] M. Sohrabi Nasrabadi, R. Safabakhsh, 3D object recognition with a linear time-varying system of overlay layers, IET Comput. Vis. 15 (5) (2021) 380–391.

[107] A.A.M. Muzahid, W. Wan, F. Sohel, L. Wu, L. Hou, CurveNet: Curvature-based multitask learning deep networks for 3D object recognition, IEEE/CAA J. Autom. Sin. 8 (6) (2021) 1177–1187, http://dx.doi.org/10.1109/JAS.2020.1003324.

[108] H. Chu, C. Le, R. Wang, X. Li, H. Ma, Learning representative viewpoints in 3D shape recognition, Vis. Comput. (2022) 1–16.

[109] C. Yue, Y. Wang, X. Tang, Q. Chen, DRGCNN: Dynamic region graph convolutional neural network for point clouds, Expert Syst. Appl. 205 (2022) 117663, http://dx.doi.org/10.1016/j.eswa.2022.117663.

[110] W. Wang, Y. You, W. Liu, C. Lu, Point cloud classification with deep normalized reeb graph convolution, Image Vis. Comput. 106 (2021) 104092, http://dx.doi.org/10.1016/j.imavis.2020.104092.

[111] A. Poulenard, M.-J. Rakotosaona, Y. Ponty, M. Ovsjanikov, Effective rotation-invariant point CNN with spherical harmonics kernels, in: 2019 International Conference on 3D Vision, 3DV, 2019, pp. 47–56, http://dx.doi.org/10.1109/3DV.2019.00015.

[112] D. Liu, C. Chen, C. Xu, Q. Cai, L. Chu, F. Wen, R. Qiu, A robust and reliable point cloud recognition network under rigid transformation, IEEE Trans. Instrum. Meas. 71 (2022) 1–13, http://dx.doi.org/10.1109/TIM.2022.3142077.

[113] Y. Zhao, Y. Wu, C. Chen, A. Lim, On isometry robustness of deep 3D point cloud models under adversarial attacks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 1198–1207, http://dx.doi.org/10.1109/CVPR42600.2020.00128.

[114] A. Brock, T. Lim, J.M. Ritchie, N. Weston, Generative and discriminative voxel modeling with convolutional neural networks, 2016, arXiv preprint arXiv:1608.04236.

[115] A. Mukhaimar, R. Tennakoon, C.Y. Lai, R. Hoseinnezhad, A. Bab-Hadiashar, Robust object classification approach using spherical harmonics, IEEE Access 10 (2022) 21541–21553, http://dx.doi.org/10.1109/ACCESS.2022.3151350.

[116] S. Kim, H. gun Chi, K. Ramani, Object synthesis by learning part geometry with surface and volumetric representations, Comput. Aided Des. 130 (2021) 102932, http://dx.doi.org/10.1016/j.cad.2020.102932.

[117] C.R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, L.J. Guibas, Volumetric and multi-view CNNs for object classification on 3D data, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 5648–5656, http://dx.doi.org/10.1109/CVPR.2016.609.

[118] H. Wang, W. Li, J. Kim, Q. Wang, Attention-guided RGB-D fusion network for category-level 6D object pose estimation, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2022, pp. 10651–10658, http://dx.doi.org/10.1109/IROS47612.2022.9981242.

[119] L. Huang, B. Zhang, Z. Guo, Y. Xiao, Z. Cao, J. Yuan, Survey on depth and RGB image-based 3D hand shape and pose estimation, Virtual Real. Intell. Hardw. 3 (3) (2021) 207–234, http://dx.doi.org/10.1016/j.vrih.2021.05.002, Hand and gesture.

[120] A. Kanezaki, Y. Matsushita, Y. Nishida, RotationNet for joint object categorization and unsupervised pose estimation from multi-view images, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2021) 269–283, http://dx.doi.org/10.1109/TPAMI.2019.2922640.

[121] A.-A. Liu, F.-B. Guo, H.-Y. Zhou, W.-H. Li, D. Song, Semantic and context information fusion network for view-based 3D model classification and retrieval, IEEE Access 8 (2020) 155939–155950, http://dx.doi.org/10.1109/ACCESS.2020.3018875.

[122] S. Chen, L. Zheng, Y. Zhang, Z. Sun, K. Xu, VERAM: View-enhanced recurrent attention model for 3D shape classification, IEEE Trans. Vis. Comput. Graphics 25 (12) (2019) 3244–3257, http://dx.doi.org/10.1109/TVCG.2018.2866793.

[123] X. Wei, R. Yu, J. Sun, View-GCN: View-based graph convolutional network for 3D shape analysis, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 1847–1856, http://dx.doi.org/10.1109/CVPR42600.2020.00192.

[124] C. Wang, X. Ning, L. Sun, L. Zhang, W. Li, X. Bai, Learning discriminative features by covering local geometric space for point cloud analysis, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15, http://dx.doi.org/10.1109/TGRS.2022.3170493.

[125] B. Leng, C. Du, S. Guo, X. Zhang, Z. Xiong, A powerful 3D model classification mechanism based on fusing multi-graph, Neurocomputing 168 (2015) 761–769, http://dx.doi.org/10.1016/j.neucom.2015.05.048.

[126] A.J. Rodríguez-Sánchez, S. Szedmak, J. Piater, Scurv: A 3D descriptor for object classification, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 1320–1327, http://dx.doi.org/10.1109/IROS.2015.7353539.

[127] F. Chen, R. Ji, L. Cao, Multimodal learning for view-based 3D object classification, Neurocomputing 195 (2016) 23–29, http://dx.doi.org/10.1016/j.neucom.2015.09.120, Learning for Medical Imaging.

[128] S. Salti, F. Tombari, L. Di Stefano, SHOT: Unique signatures of histograms for surface and texture description, Comput. Vis. Image Underst. 125 (2014) 251–264, http://dx.doi.org/10.1016/j.cviu.2014.04.011.

[129] A.F. Sheta, A. Baareh, M. Al-Batah, 3D object recognition using fuzzy mathematical modeling of 2D images, in: 2012 International Conference on Multimedia Computing and Systems, 2012, pp. 278–283, http://dx.doi.org/10.1109/ICMCS.2012.6320118.

[130] Z. Gao, Z. Yu, X. Pang, A compact shape descriptor for triangular surface meshes, Comput. Aided Des. 53 (2014) 62–69, http://dx.doi.org/10.1016/j.cad.2014.03.008.

[131] R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin, Shape distributions, ACM Trans. Graph. 21 (4) (2002) 807–832, http://dx.doi.org/10.1145/571647.571648.

[132] Y. Lei, M. Bennamoun, M. Hayat, Y. Guo, An efficient 3D face recognition approach using local geometrical signatures, Pattern Recognit. 47 (2) (2014) 509–524, http://dx.doi.org/10.1016/j.patcog.2013.07.018.

[133] A. Frome, D. Huber, R. Kolluri, T. Bülow, J. Malik, Recognizing objects in range Data Using Regional point descriptors, in: T. Pajdla, J. Matas (Eds.), Computer Vision - ECCV 2004, Springer, Berlin, Heidelberg, 2004, pp. 224–237.

[134] N. Bayramoglu, A.A. Alatan, Shape index SIFT:Range image recognition using local features, in: 2010 20th International Conference on Pattern Recognition, 2010, pp. 352–355, http://dx.doi.org/10.1109/ICPR.2010.95.

[135] S. Filipe, L.A. Alexandre, A comparative evaluation of 3D keypoint detectors in a RGB-D object dataset, in: 2014 International Conference on Computer Vision Theory and Applications, Vol. 1, VISAPP, 2014, pp. 476–483.

[136] S. Salti, F. Tombari, L.D. Stefano, A performance evaluation of 3D keypoint detectors, in: 2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 2011, pp. 236–243, http://dx.doi.org/10.1109/3DIMPVT.2011.37.

[137] H. Zhang, H. Su, Wrapped phase based SVM method for 3D object recognition, in: 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009, pp. 206–209, http://dx.doi.org/10.1109/ICCSIT.2009.5234564.

[138] H. Chen, B. Bhanu, Efficient recognition of highly similar 3D objects in range images, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 172–179, http://dx.doi.org/10.1109/TPAMI.2008.176.

[139] M. Pontil, A. Verri, Support vector machines for 3D object recognition, IEEE Trans. Pattern Anal. Mach. Intell. 20 (6) (1998) 637–646, http://dx.doi.org/10.1109/34.683777.

[140] W. Li, P. Dong, B. Xiao, L. Zhou, Object recognition based on the region of interest and optimal bag of words model, Neurocomputing 172 (2016) 271–280, http://dx.doi.org/10.1016/j.neucom.2015.01.083.

[141] W. Huang, B. Lai, W. Xu, Z. Tu, 3D volumetric modeling with introspective neural networks, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 (01) (2019) 8481–8488, http://dx.doi.org/10.1609/aaai.v33i01.33018481.

[142] N. Sedaghat, M. Zolfaghari, E. Amiri, T. Brox, Orientation-boosted voxel nets for 3D object recognition, 2017, arXiv:1604.03351.

[143] W. Cai, D. Liu, X. Ning, C. Wang, G. Xie, Voxel-based three-view hybrid parallel network for 3D object classification, Displays 69 (2021) 102076.

[144] A.S. Gezawa, Z.A. Bello, Q. Wang, L. Yunqi, A voxelized point clouds representation for object classification and segmentation on 3D data, J. Supercomput. 78 (1) (2022) 1479–1500.

[145] Z. Liu, W. Song, Y. Tian, S. Ji, Y. Sung, L. Wen, T. Zhang, L. Song, A. Gozho, VB-net: Voxel-based broad learning network for 3D object classification, Appl. Sci. 10 (19) (2020) http://dx.doi.org/10.3390/app10196735.

[146] S. Kumawat, S. Raman, LP-3DCNN: Unveiling local phase in 3D convolutional neural networks, 2019, arXiv:1904.03498.

[147] W. Huang, B. Lai, W. Xu, Z. Tu, 3D volumetric modeling with introspective neural networks, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 (01) (2019) 8481–8488, http://dx.doi.org/10.1609/aaai.v33i01.33018481.

[148] A. Sinha, J. Bai, K. Ramani, Deep learning 3D shape surfaces using geometry images, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VI 14, Springer, 2016, pp. 223–240.

[149] P.-S. Wang, Y.-Q. Yang, Q.-F. Zou, Z. Wu, Y. Liu, X. Tong, Unsupervised 3D learning for shape analysis via multiresolution instance discrimination, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 2773–2781, http://dx.doi.org/10.1609/aaai.v35i4.16382.

[150] Y. He, Z. Zhang, Z. Wang, Y. Luo, L. Su, W. Li, P. Wang, W. Zhang, IPC-net: Incomplete point cloud classification network based on data augmentation and similarity measurement, J. Vis. Commun. Image Represent. 91 (2023) 103769, http://dx.doi.org/10.1016/j.jvcir.2023.103769.

[151] M. Joseph-Rivlin, A. Zvirin, R. Kimmel, Momen(e)t: Flavor the moments in learning to classify shapes, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, 2019, pp. 4085–4094, http://dx.doi.org/10.1109/ICCVW.2019.00503.

[152] H. Zhao, L. Jiang, C.-W. Fu, J. Jia, PointWeb: Enhancing local neighborhood features for point cloud processing, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 5560–5568, http://dx.doi.org/10.1109/CVPR.2019.00571.

[153] C. Wang, B. Samari, K. Siddiqi, Local spectral graph convolution for point set feature learning, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 56–71.

[154] Y. Ben-Shabat, M. Lindenbaum, A. Fischer, 3DmFV: Three-dimensional point cloud classification in real-time using convolutional neural networks, IEEE Robot. Autom. Lett. 3 (4) (2018) 3145–3152, http://dx.doi.org/10.1109/LRA.2018.2850061.

[155] R. Gao, X. Li, J. Zhang, Recognition of point sets objects in realistic scenes, Mob. Inf. Syst. 2020 (2020) 1–13.

[156] X. Liu, Z. Han, Y.-S. Liu, M. Zwicker, Point2sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8778–8785.

[157] Y. Liu, B. Fan, S. Xiang, C. Pan, Relation-shape convolutional neural network for point cloud analysis, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 8887–8896, http://dx.doi.org/10.1109/CVPR.2019.00910.

[158] P. Liang, Z. Fang, B. Huang, H. Zhou, X. Tang, C. Zhong, PointFusionNet: Point feature fusion network for 3D point clouds analysis, Appl. Intell. 51 (2021) 2063–2076.

[159] S. Qiu, S. Anwar, N. Barnes, Dense-resolution network for point cloud classification and segmentation, in: 2021 IEEE Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 3812–3821, http://dx.doi.org/10.1109/WACV48630.2021.00386.

[160] S. Qiu, S. Anwar, N. Barnes, Geometric back-projection network for point cloud classification, IEEE Trans. Multimed. 24 (2022) 1943–1955, http://dx.doi.org/10.1109/TMM.2021.3074240.

[161] Y. Rao, J. Lu, J. Zhou, Spherical fractal convolutional neural networks for point cloud recognition, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 452–460, http://dx.doi.org/10.1109/CVPR.2019.00054.

[162] X. Li, R. Li, G. Chen, C.-W. Fu, D. Cohen-Or, P.-A. Heng, A rotation-invariant framework for deep point cloud analysis, IEEE Trans. Vis. Comput. Graphics 28 (12) (2022) 4503–4514, http://dx.doi.org/10.1109/TVCG.2021.3092570.

[163] X. Sun, Z. Lian, J. Xiao, Srinet: Learning strictly rotation-invariant representations for point cloud classification and segmentation, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 980–988, http://dx.doi.org/10.1145/3343031.3351042.

[164] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, P. Riley, Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, 2018, arXiv:1802.08219.

[165] S. Meng, D. Liang, Y. Li, Residual transformer network for 3D objects classification, in: 2021 International Wireless Communications and Mobile Computing, IWCMC, 2021, pp. 1175–1179, http://dx.doi.org/10.1109/IWCMC51323.2021.9498886.

[166] J. Zhang, W. Chen, Y. Wang, R. Vasudevan, M. Johnson-Roberson, Point set voting for partial point cloud analysis, IEEE Robot. Autom. Lett. 6 (2) (2021) 596–603, http://dx.doi.org/10.1109/LRA.2020.3048658.

[167] X. Ma, X. Li, J. Song, Point cloud completion network applied to vehicle data, Sensors 22 (19) (2022) http://dx.doi.org/10.3390/s22197346.

[168] C. Xiang, C.R. Qi, B. Li, Generating 3D adversarial point clouds, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 9128–9136, http://dx.doi.org/10.1109/CVPR.2019.00935.

[169] M. Wicker, M. Kwiatkowska, Robustness of 3D deep learning in an adversarial setting, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 11759–11767, http://dx.doi.org/10.1109/CVPR.2019.01204.

[170] H. Zhou, D. Chen, J. Liao, K. Chen, X. Dong, K. Liu, W. Zhang, G. Hua, N. Yu, LG-GAN: Label guided adversarial network for flexible targeted attack of point cloud based deep networks, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 10353–10362, http://dx.doi.org/10.1109/CVPR42600.2020.01037.

[171] I. Lang, U. Kotlicki, S. Avidan, Geometric adversarial attacks and defenses on 3D point clouds, in: 2021 International Conference on 3D Vision, 3DV, 2021, pp. 1196–1205, http://dx.doi.org/10.1109/3DV53792.2021.00127.

[172] C. Ma, W. Meng, B. Wu, S. Xu, X. Zhang, Efficient joint gradient based attack against SOR defense for 3D point cloud classification, in: Proceedings of the 28th ACM International Conference on Multimedia, MM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1819–1827, http://dx.doi.org/10.1145/3394171.3413875.

[173] H. Liu, J. Jia, N.Z. Gong, PointGuard: Provably robust 3D point cloud classification, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 6182–6191, http://dx.doi.org/10.1109/CVPR46437.2021.00612.

[174] Q. Huang, X. Dong, D. Chen, H. Zhou, W. Zhang, K. Zhang, G. Hua, N. Yu, PointCAT: Contrastive adversarial training for robust point cloud recognition, 1, 2023, arXiv:2209.07788.

[175] J. Chen, Y. Zhang, F. Ma, Z. Tan, EB-LG module for 3D point cloud classification and segmentation, IEEE Robot. Autom. Lett. 8 (1) (2023) 160–167, http://dx.doi.org/10.1109/LRA.2022.3223558.

[176] N.I. Arnold, P. Angelov, P.M. Atkinson, An improved explainable point cloud classifier (XPCC), IEEE Trans. Artif. Intell. 4 (1) (2023) 71–80, http://dx.doi.org/10.1109/TAI.2022.3150647.

[177] S. Cheng, X. Chen, X. He, Z. Liu, X. Bai, PRA-net: Point relation-aware network for 3D point cloud analysis, IEEE Trans. Image Process. 30 (2021) 4436–4448, http://dx.doi.org/10.1109/TIP.2021.3072214.

[178] J. Lee, S.-U. Cheon, J. Yang, Connectivity-based convolutional neural network for classifying point clouds, Pattern Recognit. 112 (2021) 107708, http://dx.doi.org/10.1016/j.patcog.2020.107708.

[179] Y. You, Y. Lou, Q. Liu, Y.-W. Tai, L. Ma, C. Lu, W. Wang, Pointwise rotation-invariant network with adaptive sampling and 3D spherical voxel convolution, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12717–12724.

[180] 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) (2019) 1–12.

[181] J. Li, B.M. Chen, G.H. Lee, SO-net: Self-organizing network for point cloud analysis, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 9397–9406, http://dx.doi.org/10.1109/CVPR.2018.00979.

[182] Y. Xu, T. Fan, M. Xu, L. Zeng, Y. Qiao, Spidercnn: Deep learning on point sets with parameterized convolutional filters, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham, 2018, pp. 90–105.

[183] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, B. Chen, Pointcnn: Convolution on x-transformed points, in: Advances in Neural Information Processing Systems, vol. 31, 2018.

[184] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735, arXiv:https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf.

[185] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[186] C. Ma, Y. Guo, J. Yang, W. An, Learning multi-view representation with LSTM for 3-D shape recognition and retrieval, IEEE Trans. Multimed. 21 (5) (2019) 1169–1182, http://dx.doi.org/10.1109/TMM.2018.2875512.

[187] W. Nie, L. Qu, M. Ren, Q. Liang, Y. Su, Y. Li, H. Jin, Two-stream network based on visual saliency sharing for 3D model recognition, IEEE Access 8 (2020) 5979–5989, http://dx.doi.org/10.1109/ACCESS.2019.2963511.

[188] Q. Huang, Y. Wang, Z. Yin, View-based weight network for 3D object recognition, Image Vis. Comput. 93 (2020) 103828, http://dx.doi.org/10.1016/j.imavis.2019.11.006.

[189] W. Nie, Q. Liang, Y. Wang, X. Wei, Y. Su, MMFN: Multimodal information fusion networks for 3D model classification and retrieval, ACM Trans. Multimed. Comput. Commun. Appl. 16 (4) (2020) http://dx.doi.org/10.1145/3410439.

[190] J. Huang, W. Yan, T. Li, S. Liu, G. Li, Learning the global descriptor for 3-D object recognition based on multiple views decomposition, IEEE Trans. Multimed. 24 (2022) 188–201, http://dx.doi.org/10.1109/TMM.2020.3047762.

[191] L. Nong, J. Peng, W. Zhang, J. Lin, H. Qiu, J. Wang, Adaptive multi-hypergraph convolutional networks for 3D object classification, IEEE Trans. Multimed. 25 (2023) 4842–4855, http://dx.doi.org/10.1109/TMM.2022.3183388.

[192] W. Wang, Y. Cai, T. Wang, Multi-view dual attention network for 3D object recognition, Neural Comput. Appl. 34 (4) (2022) 3201–3212.

[193] W. Wang, H. Zhou, G. Chen, X. Wang, Fusion of a static and dynamic convolutional neural network for multiview 3D point cloud classification, Remote Sens. 14 (9) (2022) http://dx.doi.org/10.3390/rs14091996.

[194] X. Jin, D. Li, Rotation prediction based representative view locating framework for 3D object recognition, Comput. Aided Des. 150 (2022) 103279, http://dx.doi.org/10.1016/j.cad.2022.103279.

[195] A.-A. Liu, H. Zhou, W. Nie, Z. Liu, W. Liu, H. Xie, Z. Mao, X. Li, D. Song, Hierarchical multi-view context modelling for 3D object classification and retrieval, Inform. Sci. 547 (2021) 984–995, http://dx.doi.org/10.1016/j.ins.2020.09.057.

[196] H. Zeng, T. Zhao, R. Cheng, F. Wang, J. Liu, Hierarchical graph attention based multi-view convolutional neural network for 3D object recognition, IEEE Access 9 (2021) 33323–33335, http://dx.doi.org/10.1109/ACCESS.2021.3059853.

[197] Y. Feng, Z. Zhang, X. Zhao, R. Ji, Y. Gao, GVCNN: Group-view convolutional neural networks for 3D shape recognition, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 264–272, http://dx.doi.org/10.1109/CVPR.2018.00035.

[198] Q. Sun, Y. Xu, Y. Sun, C. Yao, J.S.A. Lee, K. Chen, GN-CNN: A point cloud analysis method for metaverse applications, Electronics 12 (2) (2023) http://dx.doi.org/10.3390/electronics12020273.

[199] H. Fan, Y. Zhao, G. Su, T. Zhao, S. Jin, The multi-view deep visual adaptive graph convolution network and its application in point cloud., Trait. Signal 40 (1) (2023).

[200] X. Wei, R. Yu, J. Sun, Learning view-based graph convolutional network for multi-view 3D shape analysis, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2023) 7525–7541, http://dx.doi.org/10.1109/TPAMI.2022.3221785.

[201] X. Meng, X. Lu, H. Ye, B. Yang, F. Cao, A new self-augment CNN for 3D point cloud classification and segmentation, Int. J. Mach. Learn. Cybern. (2023) 1–12.

[202] Y. Liu, L. Yao, B. Li, C. Sammut, X. Chang, Interpolation graph convolutional network for 3D point cloud analysis, Int. J. Intell. Syst. 37 (12) (2022) 12283–12304.

[203] H. Lei, N. Akhtar, A. Mian, Spherical kernel for efficient graph convolution on 3D point clouds, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2021) 3664–3680, http://dx.doi.org/10.1109/TPAMI.2020.2983410.

[204] R. Guo, Y. Zhou, J. Zhao, Y. Man, M. Liu, R. Yao, B. Liu, Point cloud classification by dynamic graph CNN with adaptive feature fusion, IET Comput. Vis. 15 (3) (2021) 235–244.

[205] L. Wang, J. Li, D. Fan, A graphical convolutional network-based method for 3D point cloud classification, in: 2021 33rd Chinese Control and Decision Conference, CCDC, 2021, pp. 1686–1691, http://dx.doi.org/10.1109/CCDC52312.2021.9601582.

[206] W. Jakub, N. Patryk, S. Rafał, W. Adam, CVA-GNN: Convolutional vicinity aggregation graph neural network for point cloud classification, in: 2021 International Joint Conference on Neural Networks, IJCNN, 2021, pp. 1–8, http://dx.doi.org/10.1109/IJCNN52387.2021.9533545.

[207] Y. Cui, X. Liu, H. Liu, J. Zhang, A. Zare, B. Fan, Geometric attentional dynamic graph convolutional neural networks for point cloud analysis, Neurocomputing 432 (2021) 300–310.

[208] Q. Xu, X. Sun, C.-Y. Wu, P. Wang, U. Neumann, Grid-GCN for fast and scalable point cloud learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 5660–5669, http://dx.doi.org/10.1109/CVPR42600.2020.00570.

[209] J. Liu, B. Ni, C. Li, J. Yang, Q. Tian, Dynamic points agglomeration for hierarchical point sets learning, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 7545–7554, http://dx.doi.org/10.1109/ICCV.2019.00764.

[210] Y. Zhang, M. Rabbat, A graph-CNN for 3D point cloud classification, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 6279–6283, http://dx.doi.org/10.1109/ICASSP.2018.8462291.

[211] Y. Shen, C. Feng, Y. Yang, D. Tian, Mining point cloud local structures by kernel correlation and graph pooling, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4548–4557, http://dx.doi.org/10.1109/CVPR.2018.00478.

[212] M. Simonovsky, N. Komodakis, Dynamic edge-conditioned filters in convolutional neural networks on graphs, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 29–38, http://dx.doi.org/10.1109/CVPR.2017.11.

[213] X.-F. Xing, M.-A. Mostafavi, S.H. Chavoshi, A knowledge base for automatic feature recognition from point clouds in an urban scene, ISPRS Int. J. Geo-Inf. 7 (1) (2018) http://dx.doi.org/10.3390/ijgi7010028.

[214] J. Park, Y. Cho, Y.-S. Shin, Nonparametric background model-based LiDAR SLAM in highly dynamic urban environments, IEEE Trans. Intell. Transp. Syst. 23 (12) (2022) 24190–24205, http://dx.doi.org/10.1109/TITS.2022.3204917.

[215] B.U. Mahmud, G.Y. Hong, A.A. Mamun, E.P. Ping, Q. Wu, Deep learning-based segmentation of 3D volumetric image and microstructural analysis, Sensors 23 (5) (2023) http://dx.doi.org/10.3390/s23052640.

[216] R. Qian, X. Lai, X. Li, 3D object detection for autonomous driving: A survey, Pattern Recognit. 130 (2022) 108796, http://dx.doi.org/10.1016/j.patcog.2022.108796.

[217] J. Jiang, Z. Liu, J. Li, J. Tu, L. Li, J. Yao, iMVS: Integrating multi-view information on multiple scales for 3D object recognition, J. Vis. Commun. Image Represent. 101 (2024) 104175, http://dx.doi.org/10.1016/j.jvcir.2024.104175.

[218] G. Yang, S. Mentasti, M. Bersani, Y. Wang, F. Braghin, F. Cheli, Lidar point-cloud processing based on projection methods: a comparison, in: 2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive, (AEIT AUTOMOTIVE), 2020, pp. 1–6, http://dx.doi.org/10.23919/AEITAUTOMOTIVE50086.2020.9307387.

[219] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, L. Lin, Graph convolutional neural network for human action recognition: A comprehensive survey, IEEE Trans. Artif. Intell. 2 (2) (2021) 128–145, http://dx.doi.org/10.1109/TAI.2021.3076974.

[220] Z. Wang, R. Arablouei, J. Liu, P. Borges, G. Bishop-Hurley, N. Heaney, PointSyn2Real: Semi-supervised synthetic-to-real cross-domain learning for object classification in 3D point clouds, in: 2023 IEEE International Conference on Multimedia and Expo, ICME, 2023, pp. 1481–1486, http://dx.doi.org/10.1109/ICME55011.2023.00256.

[221] D.P. Kingma, D.J. Rezende, S. Mohamed, M. Welling, Semi-supervised learning with deep generative models, 2014, arXiv:1406.5298.

[222] E. Denton, S. Gross, R. Fergus, Semi-supervised learning with context-conditional generative adversarial networks, 2016, arXiv:1611.06430.

[223] M.E. Abbasnejad, A. Dick, A. van den Hengel, Infinite variational autoencoder for semi-supervised learning, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 781–790, http://dx.doi.org/10.1109/CVPR.2017.90.

[224] A. Abdulaziz, J. Zhou, A. Di Fulvio, Y. Altmann, S. McLaughlin, Semi-supervised Gaussian mixture variational autoencoder for pulse shape discrimination, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2022, pp. 3538–3542, http://dx.doi.org/10.1109/ICASSP43922.2022.9747313.

[225] Y. Yang, C. Feng, Y. Shen, D. Tian, FoldingNet: Point cloud auto-encoder via deep grid deformation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 206–215, http://dx.doi.org/10.1109/CVPR.2018.00029.

[226] A. Zdobylak, M. Zieba, Semi-supervised representation learning for 3D point clouds, in: N.T. Nguyen, K. Jearanaitanakij, A. Selamat, B. Trawiński, S. Chittayasothorn (Eds.), Intelligent Information and Database Systems, Springer International Publishing, Cham, 2020, pp. 480–491.

[227] H. Wang, S. Zhang, Y. Zhang, Z. Liu, Semi-supervised generative adversarial model for 3D recognition, in: 2019 IEEE International Conferences on Ubiquitous Computing & Communications (IUCC) and Data Science and Computational Intelligence (DSCI) and Smart Computing, Networking and Services (SmartCNS), 2019, pp. 381–385, http://dx.doi.org/10.1109/IUCC/DSCI/SmartCNS.2019.00090.

[228] X. Shi, X. Xu, W. Zhang, X. Zhu, C.S. Foo, K. Jia, Open-set semi-supervised learning for 3D point cloud understanding, in: 2022 26th International Conference on Pattern Recognition, ICPR, 2022, pp. 5045–5051, http://dx.doi.org/10.1109/ICPR56361.2022.9956506.

[229] A. Deng, Y. Wu, P. Zhang, Z. Lu, W. Li, Z. Su, A weakly supervised framework for real-world point cloud classification, Comput. Graph. (2022).

[230] Y. He, G. Hu, S. Yu, Hard-soft pseudo labels guided semi-supervised learning for point cloud classification, IEEE Signal Process. Lett. 31 (2024) 1059–1063, http://dx.doi.org/10.1109/LSP.2024.3386115.

[231] Z. Ren, R. Yeh, A. Schwing, Not all unlabeled data are equal: Learning to weight data in semi-supervised learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), in: Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 21786–21797.

[232] K. Saito, D. Kim, K. Saenko, OpenMatch: Open-set semi-supervised learning with open-set consistency regularization, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), in: Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 25956–25967.

[233] Q. Yu, D. Ikami, G. Irie, K. Aizawa, Multi-task curriculum framework for open-set semi-supervised learning, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 438–454.

[234] L.-Z. Guo, Z.-Y. Zhang, Y. Jiang, Y.-F. Li, Z.-H. Zhou, Safe deep semi-supervised learning for unseen-class unlabeled data, in: Proceedings of the 37th International Conference on Machine Learning, ICML '20, JMLR.org, 2020, pp. 3897–3906.

[235] M. Gadelha, A. RoyChowdhury, G. Sharma, E. Kalogerakis, L. Cao, E. Learned-Miller, R. Wang, S. Maji, Label-efficient learning on point clouds using approximate convex decompositions, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 473–491.

[236] Y. Shi, M. Xu, S. Yuan, Y. Fang, Unsupervised deep shape descriptor with point distribution learning, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 9350–9359, http://dx.doi.org/10.1109/CVPR42600.2020.00937.

[237] J. Wu, C. Zhang, T. Xue, W.T. Freeman, J.B. Tenenbaum, Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling, 2017, arXiv:1610.07584.

[238] P. Li, Y. Pei, J. Li, A comprehensive survey on design and application of autoencoder in deep learning, Appl. Soft Comput. 138 (2023) 110176, http://dx.doi.org/10.1016/j.asoc.2023.110176.

[239] D. Saxena, J. Cao, Generative adversarial networks (GANs): Challenges, solutions, and future directions, ACM Comput. Surv. 54 (3) (2021) http://dx.doi.org/10.1145/3446374.

[240] C.-C. Hsu, L.-W. Kang, S.-Y. Chen, I.-S. Wang, C.-H. Hong, C.-Y. Chang, Deep learning-based vehicle trajectory prediction based on generative adversarial network for autonomous driving applications, Multimedia Tools Appl. 82 (7) (2023) 10763–10780, http://dx.doi.org/10.1007/s11042-022-13742-x.

[241] Y. Zou, H. Sun, C. Fang, J. Liu, Z. Zhang, Deep learning framework testing via hierarchical and heuristic model generation, J. Syst. Softw. 201 (2023) 111681, http://dx.doi.org/10.1016/j.jss.2023.111681.

[242] W. Liu, H. Wang, H. Luo, K. Zhang, J. Lu, Z. Xiong, Pseudo-label growth dictionary pair learning for crowd counting, Appl. Intell. 51 (12) (2021) 8913–8927, http://dx.doi.org/10.1007/s10489-021-02274-w.

[243] Y. Chen, J. Liu, B. Ni, H. Wang, J. Yang, N. Liu, T. Li, Q. Tian, Shape self-correction for unsupervised point cloud understanding, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 8362–8371, http://dx.doi.org/10.1109/ICCV48922.2021.00827.

[244] Z. Han, X. Wang, Y.-S. Liu, M. Zwicker, Multi-angle point cloud-VAE: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 10441–10450, http://dx.doi.org/10.1109/ICCV.2019.01054.

[245] A. Sharma, O. Grau, M. Fritz, Vconv-DAE: Deep volumetric shape learning without object labels, in: G. Hua, H. Jégou (Eds.), Computer Vision – ECCV 2016 Workshops, Springer International Publishing, Cham, 2016, pp. 236–250.

[246] C.-L. Li, M. Zaheer, Y. Zhang, B. Barnabas Poczosand Ruslan Salakhutdinov, Point cloud GAN, 1, 2018, arXiv preprint arXiv:1810.05795v1.

[247] Z. Han, M. Shang, Y.-S. Liu, M. Zwicker, View inter-prediction GAN: Unsupervised representation learning for 3D shapes by learning global shape memories to support local view predictions, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 (01) (2019) 8376–8384, http://dx.doi.org/10.1609/aaai.v33i01.33018376.

[248] S.H. Khan, Y. Guo, M. Hayat, N. Barnes, Unsupervised primitive discovery for improved 3D generative modeling, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2019, pp. 9731–9740, http://dx.doi.org/10.1109/CVPR.2019.00997.

[249] N. Sun, B. Jin, J. Guo, J. Zheng, D. Shao, J. Zhang, 3D point cloud adversarial sample classification algorithm based on self-supervised learning and information gain, IEEE Access 11 (2023) 119544–119552, http://dx.doi.org/10.1109/ACCESS.2023.3326990.

[250] S. Yan, Z. Yang, H. Li, C. Song, L. Guan, H. Kang, G. Hua, Q. Huang, Implicit autoencoder for point-cloud self-supervised representation learning, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, 2023, pp. 14484–14496, http://dx.doi.org/10.1109/ICCV51070.2023.01336.

[251] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, C.L.P. Chen, Unsupervised learning of 3-D local features from raw voxels based on a novel permutation voxelization strategy, IEEE Trans. Cybern. 49 (2) (2019) 481–494, http://dx.doi.org/10.1109/TCYB.2017.2778764.

[252] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, C.L.P. Chen, Mesh convolutional restricted Boltzmann machines for unsupervised learning of features with structure preservation on 3-D meshes, IEEE Trans. Neural Netw. Learn. Syst. 28 (10) (2017) 2268–2281, http://dx.doi.org/10.1109/TNNLS.2016.2582532.

[253] J. Xie, Z. Zheng, R. Gao, W. Wang, S.-C. Zhu, Y.N. Wu, Learning descriptor networks for 3D shape synthesis and analysis, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8629–8638, http://dx.doi.org/10.1109/CVPR.2018.00900.

[254] K. Hassani, M. Haley, Unsupervised multi-task feature learning on point clouds, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 8159–8170, http://dx.doi.org/10.1109/ICCV.2019.00825.

[255] J. Jiang, D. Bao, Z. Chen, X. Zhao, Y. Gao, MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval, in: 2019Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8513–8520, http://dx.doi.org/10.1609/aaai.v33i01.33018513.

[256] D. Shao, X. Lu, X. Liu, 3D intracranial aneurysm classification and segmentation via unsupervised dual-branch learning, IEEE J. Biomed. Health Inf. 27 (4) (2023) 1770–1779, http://dx.doi.org/10.1109/JBHI.2022.3180326.

[257] C. Zeng, W. Wang, A. Nguyen, J. Xiao, Y. Yue, Self-supervised learning for point cloud data: A survey, Expert Syst. Appl. 237 (2024) 121354, http://dx.doi.org/10.1016/j.eswa.2023.121354.

[258] Y. Pang, W. Wang, F.E.H. Tay, W. Liu, Y. Tian, L. Yuan, Masked autoencoders for point cloud self-supervised learning, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022, Springer Nature, Switzerland, Cham, 2022, pp. 604–621.

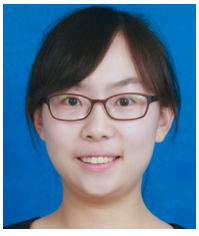
[259] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, V.G. Kim, Self-supervised learning of point clouds via orientation estimation, in: 2020 International Conference on 3D Vision, 3DV, 2020, pp. 1018–1028, http://dx.doi.org/10.1109/3DV50981.2020.00112.
[260] Y. Zhang, J. Lin, C. He, Y. Chen, K. Jia, L. Zhang, Masked surfel prediction for self-supervised point cloud learning, 2022, arXiv:2207.03111.
[261] A. Sanghi, Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 626–642.
[262] D. Wang, Z.-X. Yang, Self-supervised point cloud understanding via mask transformer and contrastive learning, IEEE Robot. Autom. Lett. 8 (1) (2023) 184–191, http://dx.doi.org/10.1109/LRA.2022.3224370.
[263] Z. Lu, Y. Dai, W. Li, Z. Su, Joint data and feature augmentation for self-supervised representation learning on point clouds, Graph. Models 129 (2023) 101188, http://dx.doi.org/10.1016/j.gmod.2023.101188.
[264] B. Tran, B.-S. Hua, A.T. Tran, M. Hoai, Self-supervised learning with multi-view rendering for 3D point cloud analysis, in: L. Wang, J. Gall, T.-J. Chin, I. Sato, R. Chellappa (Eds.), Computer Vision – ACCV 2022, Springer Nature, Switzerland, Cham, 2023, pp. 413–431.
[265] S. Huang, Y. Xie, S.-C. Zhu, Y. Zhu, Spatio-temporal self-supervised representation learning for 3D point clouds, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 6515–6525, http://dx.doi.org/10.1109/ICCV48922.2021.00647.
[266] H. Chen, S. Luo, X. Gao, W. Hu, Unsupervised learning of geometric sampling invariant representations for 3D point clouds, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW, 2021, pp. 893–903, http://dx.doi.org/10.1109/ICCVW54120.2021.00105.
[267] J. Jiang, X. Lu, W. Ouyang, M. Wang, Unsupervised contrastive learning with simple transformation for 3D point cloud data, Vis. Comput. (2023) http://dx.doi.org/10.1007/s00371-023-02921-y.
[268] Y. Wu, J. Liu, M. Gong, P. Gong, X. Fan, A.K. Qin, Q. Miao, W. Ma, Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding, IEEE Trans. Multimed. 26 (2024) 1626–1638, http://dx.doi.org/10.1109/TMM.2023.3284591.
[269] Q. Wu, J. Wan, A.B. Chan, Progressive unsupervised learning for visual object tracking, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2021, pp. 2992–3001, http://dx.doi.org/10.1109/CVPR46437.2021.00301.
[270] Y. Sun, Y. Wang, Z. Liu, J.E. Siegel, S.E. Sarma, PointGrow: Autoregressively learned point cloud generation with self-attention, in: 2020 IEEE Winter Conference on Applications of Computer Vision, WACV, 2020, pp. 61–70, http://dx.doi.org/10.1109/WACV45572.2020.9093430.
[271] Y. Zhang, C. Zhou, D. Huang, STAL3D: Unsupervised domain adaptation for 3D object detection via collaborating self-training and adversarial learning, IEEE Trans. Intell. Veh. (2024) 1–12, http://dx.doi.org/10.1109/TIV.2024.3397194.
[272] Y. Zhang, M. Li, Y. Xie, C. Li, C. Wang, Z. Zhang, Y. Qu, Self-supervised exclusive learning for 3D segmentation with cross-modal unsupervised domain adaptation, MM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 3338–3346, http://dx.doi.org/10.1145/3503161.3547987.
[273] F. Xu, J. Chen, Y. Shi, T. Ruan, Q. Wu, X. Zhang, 3D meta-classification: A meta-learning approach for selecting 3D point-cloud classification algorithm, Inform. Sci. 662 (2024) 120272, http://dx.doi.org/10.1016/j.ins.2024.120272.
[274] L. Wu, K. Zhong, Z. Li, M. Zhou, H. Hu, C. Wang, Y. Shi, PPTFH: Robust local descriptor based on point-pair transformation features for 3D surface matching, Sensors 21 (9) (2021) http://dx.doi.org/10.3390/s21093229.
[275] H. Naderi, I.V. Bajić, Adversarial attacks and defenses on 3D point cloud classification: A survey, IEEE Access 11 (2023) 144274–144295, http://dx.doi.org/10.1109/ACCESS.2023.3345000.
[276] L. Pan, X. Chen, Z. Cai, J. Zhang, H. Zhao, S. Yi, Z. Liu, Variational relational point completion network for robust 3D classification, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 11340–11351, http://dx.doi.org/10.1109/TPAMI.2023.3268305.

A.A.M. Muzahid is a lecturer at the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science (SUES), China. He received his Ph.D. in Communication and Information Systems from Shanghai University, Shanghai, China in 2021. He received an M.E in Communication and Information Engineering from Chongqing University of Posts and Telecommunications, Chongqing, China in 2016 and a B.Sc. in Electronics and Telecommunications Engineering from Daffodil International University, Dhaka, Bangladesh in 2011. He received the "Chinese Govt. Outstanding International Student Scholarship Award 2020" and "Shanghai University Public-spirited Star Award 2019". His current research interests include Machine Learning, Pattern Recognition, 3D vision, particularly on 3D feature learning, 3D object recognition, 3D modeling, scene understanding, and AR and VR.

Han Hua received her B.S. degree in Communication Engineering from Hangzhou Dianzi University, Hangzhou, Zhejiang, China in 2007, Ph.D. degree in Pattern Recognition and Intelligent System from Donghua University, Shanghai, China in 2012. She joined the New Jersey Institute of Technology (NJIT), Newark, NJ in 2016 as a Research Scholar of Electrical and Computer Engineering. She is a Full Professor at the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. Her research interests include multi-target tracking, pedestrian re-identification, pattern recognition, and machine learning. She has published over 50 papers and received 10 patents.

Yujin Zhang received a Ph.D. degree in communication and information systems from Shanghai Jiao Tong University, Shanghai, China, in 2014. He is currently an Associate Professor with the School of Electronic and Electrical Engineering at Shanghai University of Engineering Science, and an Opening Project Researcher with the Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai Jiao Tong University. His research interests include multimedia forensics, signal processing, artificial intelligence, and pattern recognition.

Dawei Li received the Bachelor of Engineering degree in Automation in 2006 from Tongji University, Shanghai, China. In 2013, he received a Ph.D. in Control Theory and Control Engineering from Tongji University, Shanghai, China. During 2013–2015, he was a postdoc at the Department of Computer Sciences and Technology, Tongji University, Shanghai, China. He is now working as an associate professor at the College of Information Sciences and Technology, Donghua University, Shanghai, China. He is currently the secretary general of IEEE CIS's Shanghai Chapter. From 2009–2010, he was a visiting researcher at Michigan State University. His current research interests include plant phenotyping, point cloud processing, and artificial intelligence.



Yuhe Zhang received a B.S. degree in software engineering and a Ph.D. degree in computer applied technology from Northwest University of China in 2012 and 2017. From July 2017 to July 2020, she was a lecturer at the School of Information Science and Technology, Northwest University of China. Since August 2020, she has been an Associate Professor at the School of Information Science and Technology, Northwest University of China. Her research interests include computer graphics, image processing, intelligent information processing, and the digital restoration of cultural heritage.

Junaid Jamshid received the B.E (electronics) and M.S (telecommunication) degrees from the Iqra University, Pakistan in 2010 and 2016, respectively. Currently, he is pursuing his Ph.D. degree with the School of information and Communication Engineering at Shanghai University, China. His Ph.D. studies are focusing on computer vision and autonomous robots. His research interests include 3D reconstruction and DL.

Ferdous Sohel received a Ph.D. degree from Monash University, Australia. He is currently a professor of information technology at Murdoch University, Australia. He worked as a Research Assistant Professor at the University of Western Australia. His research interests include computer vision, machine learning, pattern recognition, and digital agriculture. He is an Associate Editor of IEEE Transactions on Multimedia, IEEE Signal Processing Letters, and Computers and Electronics in Agriculture. He is a member of the Australian Computer Society and a senior member of the IEEE.