



Deep learning based computer vision under the prism of 3D point clouds: a systematic review

Kyriaki A. Tychola¹ · Eleni Vrochidou¹ · George A. Papakostas¹

Accepted: 14 December 2023
© The Author(s) 2024

Abstract

Point clouds consist of 3D data points and are among the most considerable data formats for 3D representations. Their popularity is due to their broad application areas, such as robotics and autonomous driving, and their employment in basic 3D vision tasks such as segmentation, classification, and detection. However, processing point clouds is challenging compared to other visual forms such as images, mainly due to their unstructured nature. Deep learning (DL) has been established as a powerful tool for data processing, reporting remarkable performance enhancements compared to traditional methods for all basic 2D vision tasks. However new challenges are emerging when it comes to processing unstructured 3D point clouds. This work aims to guide future research by providing a systematic review of DL on 3D point clouds, holistically covering all 3D vision tasks. 3D technologies of point cloud formation are reviewed and compared to each other. The application of DL methods for point cloud processing is discussed, and state-of-the-art models' performances are compared focusing on challenges and solutions. Moreover, in this work the most popular 3D point cloud benchmark datasets are summarized based on their task-oriented applications, aiming to highlight existing constraints and to comparatively evaluate them. Future research directions and upcoming trends are also highlighted.

Keywords Point cloud · LiDAR · RGB-D · Deep learning · Computer vision · 3D data · Review

1 Introduction

Point clouds constitute an alternative data format for 3D scenes' representation. Their popularity is attributed to the increasing availability of point cloud capturing devices and the wide range of their application in various scientific fields [1, 2]. Different sensing devices are currently available, accompanied by algorithms, for the detailed acquisition of point clouds in a wide range of computational and economical costs. Note that Apple has included Light Detection And Ranging (LiDAR) capabilities in its latest hardware to help the camera autofocus faster and capture details even in low-lighting conditions. A point cloud consists of thousands of unorganized colored 3D points that identify objects' shapes. Each point is denoted by a set of three Cartesian coordinates (X, Y, Z), providing at the same time additional information, such as intensity or reflectance, when active sensors are used

to generate it, geometric information, scale, as well as distance and speed estimations [3]. Point cloud representations allow for adaptive storing space and imagining details of varying levels by controlling the number of points based on the desired density [4]. This flexible control stems from the unstructured nature of point clouds, lacking a strict topology and thus enabling their easy formatting to properly adapt to any real-time application. However, these same advantages denote substantial challenges related to point cloud management, associated with data sparsity, unstructured nature, uneven distributions, redundant data, modeling errors, and noise artifacts.

Point clouds are generated by 3D laser scanners, referring mainly to LiDAR technology [5] and Red Green Blue-Depth (RGB-D) cameras [6] with different resolutions and sensor restrictions, or by photogrammetry software [7]. Each laser scan measurement is represented by a point, while all scans register to form the entire scene. Point clouds having temporal dimensions are referred to as *dynamics* and consist of a sequence of static point clouds. Dynamic point clouds can be generated at speed by mounting sensors on mobile mapping

✉ George A. Papakostas
gpapak@cs.ihu.gr

¹ MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece

devices, i.e., ground vehicles or Unmanned Aerial Vehicle (UAVs) [8].

Point clouds are employed in a great variety of applications, such as 3D object recognition [9], robotics [10] for simultaneous localization and mapping (SLAM) [11], odometry (visual odometry and LiDAR odometry), autonomous driving [12], change detection [13], remote sensing [14], medical treatment [15], image matching [16], shape analysis [17], etc. To enrich the quality of low-density point clouds, up-sampling is performed by combining point clouds with the corresponding 2D images of objects [18]. This combination may seem ideal, especially for objects' representation that their original 2D images are available and can be used as input data. However, the fact that 2D and 3D images obtain different characteristics makes it challenging; 3D point clouds are unstructured and have thousands of points with geometry and attribute information, while 2D images are in a limited structured grid shape. Moreover, a point cloud denotes the 3D external surface of objects, in contrast to 2D images which project the 3D world on a 2D plane. These inherent differences need to be considered when developing point cloud processing methods [19]. Thereby, from point cloud formation and processing, several challenges are emerging; robust and accurate methods as well as efficient algorithms need to be considered.

Deep learning (DL) methods are proven powerful tools for data processing in computer vision due to their capability for automatic feature extraction and high reported performance. For this reason, combined with the simultaneous development of powerful Graphics Processing Units (GPUs) and the existence of suitable training datasets, DL has been adapted for point cloud processing and analysis for all popular 3D vision tasks: semantic segmentation [20], classification [3], object detection [21], as well as 3D registration [22], completion [23] and compression [24, 25]. It was in 2017 when, for the first time, PointNet [26], a deep network, was introduced directly to sets of points, without any pre-processing or conversion to other forms, followed by PointNet++ [27] to resolve drawbacks of PointNet and form the basis for upcoming deep networks. These works set the start of a new 3D point cloud processing era. In the following years, related research focused on point cloud generation, processing, and the description of specific datasets for various applications [28–30]. A recent method proposed a new point cloud re-identification network (PointReIDNet) consisting of a global semantic module and a local feature extraction module, able to decrease the 3D shape representation parameters from 2.3 M to 0.35 M [31]. However, DL application on raw 3D point clouds remains challenging; the limited scale of existing datasets, the high dimensionality and the irregular nature of unstructured 3D point clouds pose the basic limitations in the utilization of DL methods for the direct processing of point clouds. In recent years, there has

been a plethora of available point cloud datasets derived from various sensors, such as Structure from Motion (SfM), RGB-D cameras, and LiDAR systems. Many existing available datasets include real single-sensor data such as Argoverse [32], real multi-sensor data such as KITTI [33], and synthetic data such as Apollo [34]. Available benchmark datasets, however, decrease as their size and complexity increase, consist of real or virtual scenes, and focus on different tasks. Yet, the existence of large-scale multi-sensory datasets is crucial for DL applications, that need great amounts of ground truth labels for training deep networks.

To this end, this study aims to provide an exhausting overview and present the current status of DL methods on 3D point cloud processing. This work covers a wide range of aspects, summarized in the following distinct points: (1) a comparison of existing point clouds acquisition technologies, (2) a holistic review of all 3D point cloud related vision tasks, (3) the presentation of available point cloud datasets, (4) the presentation of emerging challenges by using point clouds in DL applications, in contrast to the use of other image data formats, e.g., 2D images, (5) proposed solutions to face these challenges and (6) future research directions. The review is based on a holistic taxonomical classification of DL methods for 3D point clouds as illustrated in Fig. 1. This work to the best of the authors' knowledge is the first to particularly focus on DL algorithms for all basic 3D point cloud related tasks, including classification, segmentation, detection and tracking, registration, completion and compression, as well as the first work that integrates all aforementioned research aspects. These tasks are the most commonly addressed in research and applications related to 3D computer vision and point cloud processing, contributing towards extracting meaningful information from point cloud data. Their selection is based on their significance in applications employing 3D point clouds, such as robotics and autonomous vehicles, as defined from the investigation of the high-frequency terms used in deep learning on point clouds based on relevant papers of the examined literature. This work comprises a Systematic literature review (SLR) that aims to cover a wide range of 3D point clouds related aspects, focusing on all fundamental concepts and tasks, so as to provide a complete road map for people newly introduced to this research field.

The rest of this paper is organized as follows. Section 2 provides the motivation and the contributions of this work. Section 3 presents the research strategy followed in this work. Section 4 describes point cloud formation technologies, including a brief comparison between them. Section 5 focuses on the relationship between computer vision and point clouds, presenting DL methods on 3D point clouds, and various challenges. Section 6 reviews the literature on point cloud learnable methods and techniques for main vision tasks. Section 7 summarizes available point cloud datasets.

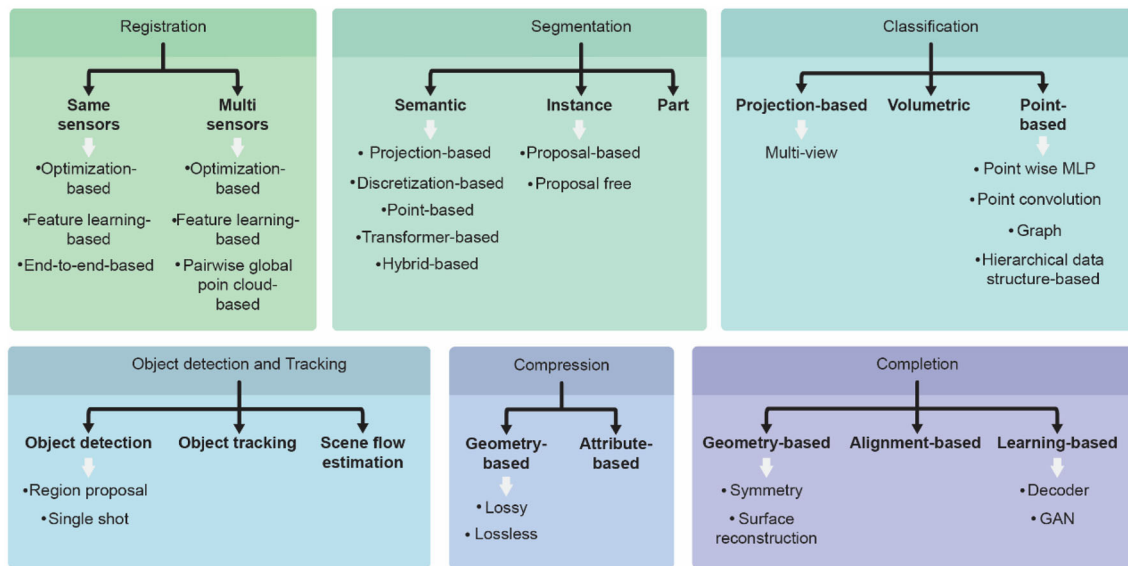


Fig. 1 Organization of 3D point clouds DL methods

Section 8 includes an exhausting discussion based on the research findings and provides future research directions. Finally, Sect. 9 concludes the paper.

2 Motivation and contribution

During the last decade, sensory 3D point cloud data acquisition has increased, allowing users to visualize highly detailed and realistic scenes easily, as well as to manipulate, explore and analyze them to the extent needed for various tasks, identifying potential issues, and concluding to better decisions. However, since these data are complex and large-scale for their manipulation, more robust and efficient methods are required. At the same time, DL methods are established as a powerful tool for point cloud processing and analysis. Researchers are turning to the investigation of robust and efficient DL algorithms for point cloud data inputs for various vision tasks, and simultaneously various point cloud datasets are developing.

Several similar review articles can be found in the recent literature, however, there is a lack of a complete investigation of DL on point clouds. Guo et al. [35] in their review, cover only three major tasks, i.e., classification, segmentation, and object detection and tracking. The authors focus on the comprehensive comparison of existing DL methods on several datasets, providing evaluation results for all corresponding tasks. Wang et al. [36] review urban reconstruction algorithms and evaluate their performance in the context of architectural modeling, focusing on LiDAR capturing technologies. In [37], technical developments of RGB-D sensors

and consequent data processing methods to handle various challenges, such as missing depth, are reviewed, while [38] deals with novel developments in high resolution synthetic aperture radar (SAR) interferometry. Ahmed et al. [39] provide a comprehensive overview of various 3D representations, discuss DL methods for each representation and compare algorithms based on certain datasets. Liu et al. [40] focus on feature learning methods for point clouds and analyze their advantages and disadvantages, including the three basic vision tasks and corresponding datasets. In [41], Vinodkumar et al., present a review of DL-based tasks for 3D point clouds, including segmentation, detection, and classification. Evaluation performance is reported, as well as the used datasets. Ioannidou et al. [42] survey methods that apply DL on 3D data and classify them according to the way the input data is treated before being inserted into the DL models. Camuffo et al. [43] review DL-based semantic scene understanding, compression, and completion, introducing a new taxonomy classification based on the characteristics of the acquisition setup and the data peculiarities. Bello et al. [44] provide a review of DL-based classification segmentation and detection, including popular benchmark point cloud datasets. Xiao et al. [45] focus on unsupervised point cloud representation learning using DL. In general, more review articles published over the years, yet they focus only on specific tasks, such as registration [46], classification [3, 47], completion [23], compression [48], and segmentation [20]. In [44], Bello et al. compile a review for DL on 3D point clouds, focusing on DL state-of-the-art approaches for raw point cloud data.

Table 1 includes the basic features of the aforementioned related works regarding DL methods on point clouds. There

Table 1 Comparative table of the characteristics of present work (Ours) versus related works from the literature

Characteristics		Point cloud and deep learning methods											Ours
		[35] 2021	[36] 2018	[37] 2019	[38] 2009	[39] 2019	[40] 2019	[41] 2023	[42] 2018	[43] 2022	[44] 2020	[45] 2023	
Point cloud generation technologies	LiDAR	×	✓	×	×	✓	✓	✓	✓	✓	×	×	✓
	RGB-D	×	×	✓	×	✓	✓	×	×	✓	×	×	✓
	Radar	×	×	×	✓	✓	×	×	×	✓	×	×	✓
Tasks	Registration	×	×	✓	×	×	×	×	×	×	×	×	✓
	Segmentation	✓	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓
	Classification	✓	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓
	Detection	✓	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓
	Tracking	✓	×	✓	×	×	×	×	×	×	×	×	✓
	Compression	×	×	×	×	✓	×	×	×	✓	×	×	✓
	Completion	×	×	×	×	✓	×	×	×	✓	×	×	✓
	Datasets	×	✓	×	✓	×	✓	×	✓	✓	✓	✓	✓
Other visual data	2D images	×	×	×	×	×	×	×	×	×	×	×	✓
	Depth images	×	×	×	×	×	×	×	×	×	×	×	✓
Methods	Deep learning	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓
	Traditional	✓	×	✓	×	×	✓	✓	×	×	×	×	✓

are also other proposed surveys about point clouds in the literature not included in Table 1 since their index is far non-comparable to the proposed study. Such indicative works include the survey implemented by Xiao et al. [49] focusing on label-efficient learning of point clouds, the survey of Li et al. [50] for DL for scene flow estimation on point clouds, the review of Grill et al. [51] for point cloud segmentation and classification algorithms, and the review on DL-based semantic segmentation for point clouds of Zhang et al. [20]. Therefore, the most contextual similar works were considered at this point, aiming to comparatively highlight the contribution of the present review work (Ours) versus previous ones.

According to Table 1, the present work aims to fill the identified research gap, by providing: an overview of the main 3D point cloud generation technologies and comparing the quality of point clouds among different acquisition sensors; discussing all DL-based 3D vision tasks; highlighting challenges and constraints from earlier traditional methods; reporting solutions to face the challenges stemming from DL methods; providing the corresponding datasets for each task; comparing point cloud data to other visual data forms; providing deeper insights and underlining differences, advantages, and drawbacks of each modality; providing a critical evaluation of point clouds' utilization for different tasks and datasets; and, finally, suggesting future research directions and trends in the field.

To the best of the authors' knowledge, this review is the first to holistically cover DL-based tasks, including segmentation, classification, detection and tracking, registration, completion, and compression, and to combine point cloud fundamentals, DL research advances on point clouds for all tasks, challenges, solutions, datasets, and future research directions, as opposed to already existing reviews. Performance comparison results of DL algorithms on 3D point cloud processing tasks can be found in [35, 43, 44].

3 Research methodology

Within the context of this work, a systematic literature review took place by using the Kitchenham approach [52] to identify the status of research in DL on 3D point clouds, based on six basic research questions:

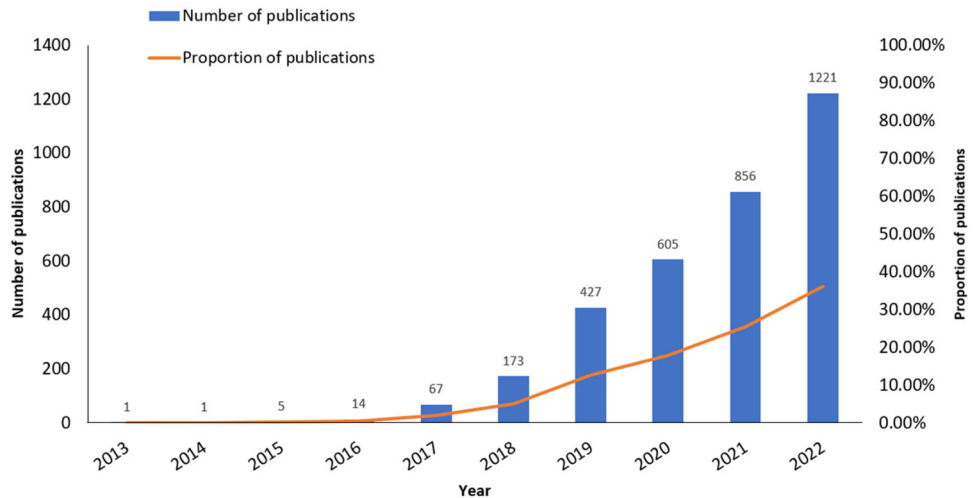
RQ1: *What are the challenges regarding point cloud data processing?*

RQ2: *What are the challenges that DL models face with 3D point cloud data?*

RQ3: *What is the status of 3D point cloud datasets for DL-based applications?*

RQ4: *In which applications does it make sense to apply point clouds?*

RQ5: *To what extent do different sensors affect the point cloud resolution?*

Fig. 2 Number of publications per year (2013–2022)**Fig. 3** Cloud map of high-frequency terms used in deep learning on point clouds based on paper keywords

RQ6: *Under what conditions does the use of point clouds provide benefits against 2D images?*

We performed a search of peer-reviewed journal publications in the Scopus database using the query “(TITLE-ABS-KEY (point AND cloud) AND TITLE-ABS-KEY (deep AND learning)) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ch")) AND (LIMIT-TO (LANGUAGE, "English")) AND (EXCLUDE (PUBYEAR, 2023))”. The process returned 3370 documents. Figure 2 summarizes the number and the proportion of total published works on the subject per year from 2013 to 2022, to illustrate a full period of 10 years. References from 2023 were also considered in this work, as indicated in the query above; however, they were not illustrated in the graph as related research in 2023 is ongoing. Although DL methods have been applied to point clouds only just in the last decade, the even increasing number of publications, arithmetically and proportionally, shows an overall upward trend, indicating the significance of this research topic. Figure 3 illustrates a tag of high-frequency used keyword terms in DL on point clouds literature, based on their

occurrence. The font size indicates the frequency of the used terms based on the keywords of the papers. As it can be observed, most of the literature focuses on deep learning networks. In addition, segmentation classification and object detection tasks, are the most used for various applications.

4 Point cloud essentials

This section summarizes the form and characteristics of 3D point clouds. Working principles of the main technologies for point cloud data acquisition and generation are also discussed. The section concludes with the comparison of point clouds with alternative visual data formats.

4.1 Point cloud data acquisition and generation technologies

Point clouds can be captured with either laser scanners or photogrammetry. Currently, there are several frequently used techniques able to generate a 3D point cloud using 3D laser scanners. Point clouds quality depends on the technology that is adopted for its acquisition since each technology has its own features and peculiarities.

4.1.1 LiDARs

LiDAR is an active remote sensing technology that employs a laser beam to sense objects through ultraviolet visible or near-infrared sources and measures the distance between an object and the scanner. This is achieved through multiple light waves (pulses) that scan the scene from side to side. LiDAR technology can cover large areas from the ground and above, when mounted on aerial vehicles, at flight height between 100–1000 m, while the angle scan is from 40° to 75° maximum with rhythm 20–40 Hz. A typical range of pulse

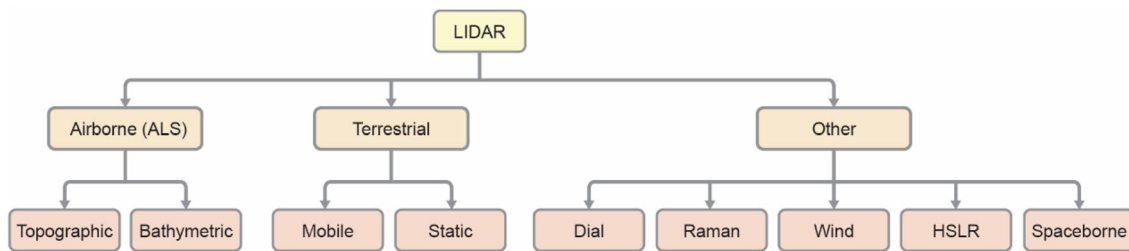
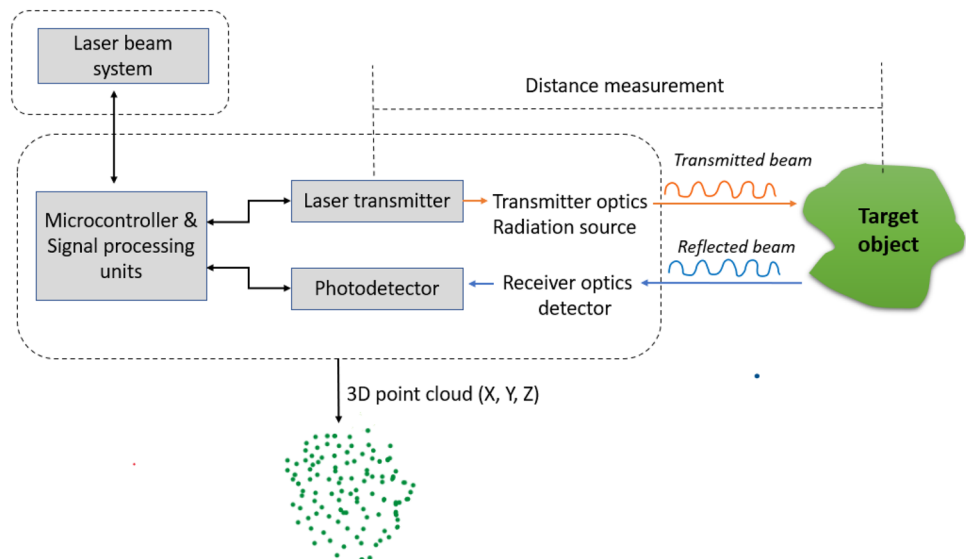


Fig. 4 LiDARs taxonomy

Fig. 5 The basic LiDAR unit



is 10 ns with repetition 5–33 kHz-max50Khz and frequency 10 kHz, i.e., 10,000 points per second [53]. Information of distance and direction are recorded to generate a point in 3D space and the differences in the pulse, return times, and wavelengths are used to generate the 3D representation of the target and calculate the exact distance from the objects [54, 55].

LiDARs can be classified based on their functionality and their inherent characteristics in three broader categories. Based on their functionality, they can be divided into Airborne (ALS) and terrestrial LiDAR. ALS LiDARs are mounted on aerial vehicles and can be further classified as topographic, to monitor the topography in terms of geomorphology, and bathymetric, to measure the depth of water and locate objects in the bottom of water bodies, e.g., oceans, lakes, etc. Terrestrial LiDARs are mounted on stable places, e.g., a tripod, or on moving vehicles, and can be classified as static, when it is portable and located at fixed points, or mobile when it is mounted on moving platforms. A third category, includes all other LiDAR types designated for special applications, including Differential Absorption LiDAR (DIAL) for sensing the ozone, Raman LiDAR for monitoring water vapor and aerosol, Wind LiDAR to measure wind data, Spaceborne LiDAR for out-of-space detection and tracking,

and airborne High Spectral Resolution LiDAR (HSRL) for aerosols and clouds characterization. Figure 4 illustrates the classification of LiDARs.

A laser mapping LiDAR system comprises (1) the LiDAR unit itself, which emits rapid pulses of infrared laser light to scan the scene, (2) a Global Positioning System (GPS), (3) an inertial measurements unit (IMU) and (4) a computer for controlling the system and storing the data. GPS and IMU combination allows identifying accurately the location of the laser at the time, at which the corresponding pulse is transmitted, while by using another GPS the ground truth is measured. IMU is responsible for accurate elevation calculations using orientation to accurately determine the actual position of the pulse on the ground. Figure 5 shows a typical LiDAR unit. The unit consists of a laser rangefinder and a scan system. The rangefinder system includes a laser transmitter, photodetector, optics and microcontroller, and signal processing electronics. Different azimuths and vertical angles of laser beams are steered from the scan system. The operation of a typical LiDAR is based on the scanning of its field of view with one or more laser beams, via a beam steering system which is produced by a laser diode with modulated amplitude, emitting at near-infrared wavelength. Laser beams are reflected from the environment backwards to the scanner.

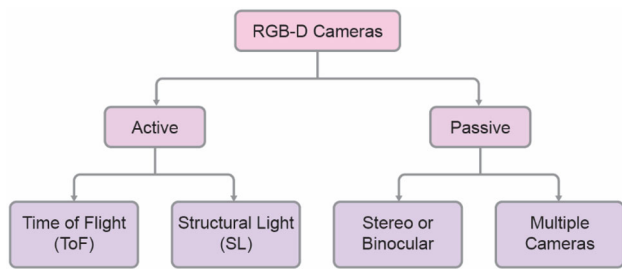


Fig. 6 RGB-D taxonomy

The returning signal is sensed by the photodetector. The signal is filtered by fast electronics and the difference between transmitted and received signal is estimated. The exact range is calculated based on this difference by the sensor model. Signal processing is used to compensate for differences in variations of the transmitted and reflected signals because of surface materials. The outputs of LiDAR are 3D point clouds corresponding to scanned environments, and intensities corresponding to the reflected laser energies [56].

4.1.2 RGB-D cameras

RGB-D cameras are a type of depth camera able to provide both depth and color data from their field of view, towards a point cloud generation in real-time [37, 57]. Microsoft Kinect, as the first RGB-D sensor commercially released, paved the way for range sensing technologies to flood the market and promote research, providing cheap and powerful tools for static and dynamic scene reconstruction.

RGB-D images can be captured with either active or passive sensing. Passive ranging is feasible due to the input combination of two (stereo or binocular) or multiple cameras (monochrome or color). For estimating the depth of a scene, the triangulation process is employed [58]. Active sensing can be classified in structured light (SL) and time of flight (ToF) cameras. SL techniques refer to the process of projecting a distinctive pattern in the scene and therefore, adding known features to enable feature matching and compute the depth even for areas in the image that lack discriminative features. ToF cameras emit a pulse of light and estimate distance by the round-trip time. In both cases, depth information is retrievable through a depth map/image acquired from infrared measurements.

Figure 6 illustrates the taxonomy of the RGB-D cameras, while Fig. 7 shows the typical workflow of point cloud generation using an RGB-D camera. Color and depth data are captured concurrently by the different sensor types. The color images are transformed, while infrared images lead to 3D mapping. Then, the camera's position and orientation are determined relative to the desired object (target) and the pose is estimated from 2D images using pixel correspondence and

3D object points [59]. The intrinsic parameters of the camera contribute to pixel-by-pixel point cloud projection from the depth images to 3D points. In the next step, the camera is calibrated to correct possible errors, while in real-time applications it is calibrated to obtain the desired coordinate system. Thereupon the features are extracted, and homologous points are detected between previous and current frames at each given time and matched. Then, a low-resolution sparse point cloud is generated. The local coordinates of the point cloud are converted into a global coordinate system with the aim of co-linearity equations. From the sparse cloud arises a denser point cloud whose density is based on the frame's number; moreover, when the depth map is combined with color information, the point cloud obtains color [60].

RGB-D video allows capturing active depth when the sensor is moving in a static scene. By fusing the captured frames, the scene's reconstruction is possible. Multiple RGB-D cameras could also be employed to enable dynamic scene reconstruction. The recent advancements in DL made monocular depth estimation also possible [61]. Prior information, such as relations between geometric structures, is used to conclude from a single image into depth information. Depending on the used ground image, monocular DL-based depth estimation can be classified into supervised [62], unsupervised [63] and semi-supervised [64]. DL models for monocular depth estimation are usually jointly trained in the framework of other basic tasks, such as segmentation, therefore depth estimation is not examined separately in this work as an independent task.

4.1.3 Radars

Synthetic Aperture Radar (SAR) is a significant active microwave imaging sensor [65]. A SAR point cloud generation system processes SAR data acquired from multiple spatially separated SAR apertures so as to calculate the exact 3D positions of all scatterers in the image scene. Aperture is the opening used to collect the reflected energy and form an image. Interferometric Synthetic Aperture Radar (InSAR) is a geodetic radar technique for remote sensing applications generating maps of deformation on surfaces or digital elevations by comparing two or more SAR images.

SAR techniques stand out for their simple design process, their flexibility to change any scanning scheme, and the high computation efficiency for processing. However, data acquisition is generally slow, many antenna pairs or scan positions are required and are more suitable for stationary or slowly moving targets. It should be noted here that mm-Wave radars in general, e.g. Multi-input Multi-output (MIMO) [66] can be used for (range, azimuth and elevation) point cloud generation to detect moving targets. Point clouds generated by mm-Wave radars are attracting growing attention from academia and industry [67] due to their excellent

Fig. 7 A typical process of point cloud generation

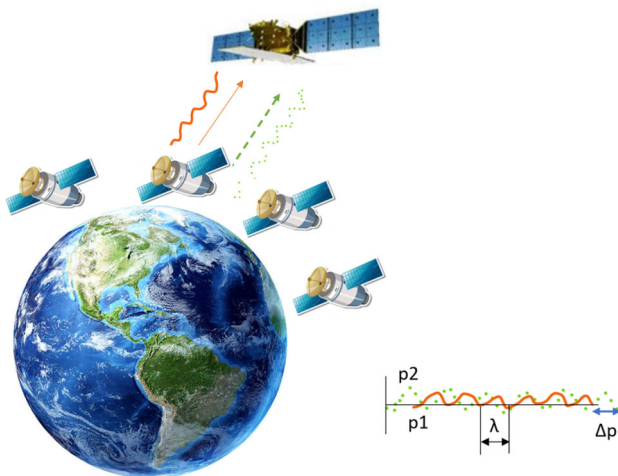
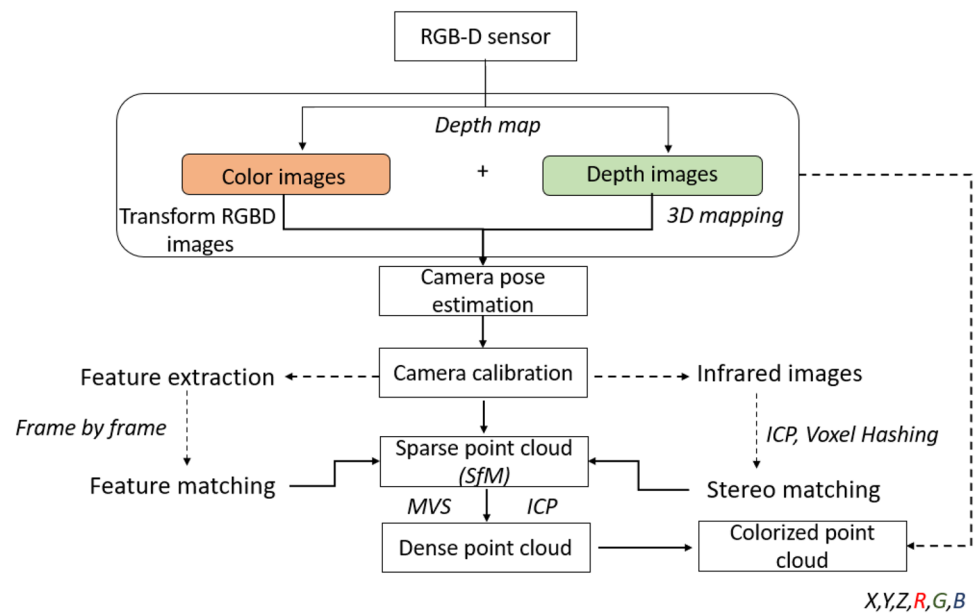


Fig. 8 The basic operational principles of SAR

performance and capabilities. Frequency-Modulated Continuous Wave Radars (FMCW) are another category of radar sensors that radiate continuous transmission power. FMCWs can alter their operating frequency during the measurement, offering more robust sensing [68].

Figure 8 illustrates the operational principles of SAR; p_1 and p_2 are the phases of two reflected signals, λ refers to the wavelength of a signal and Δp is the displacement (between two different phases).

4.1.4 Photogrammetry

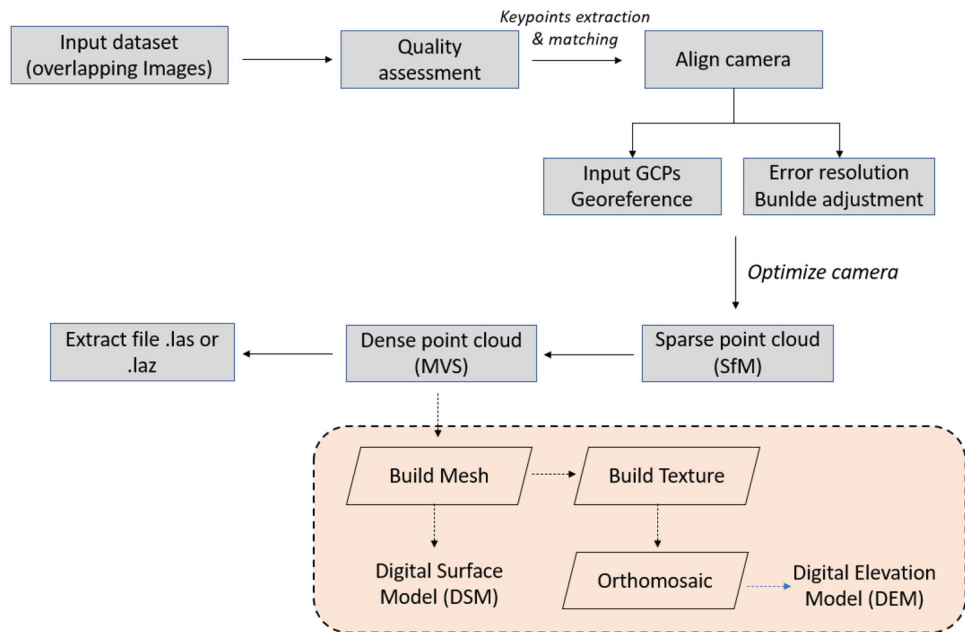
Photogrammetry is an alternative method to generate 3D models, by using photographs instead of light to collect data,

and methods from optics and projective geometry [69]. Photogrammetry needs a conventional camera to capture the images, a computer, and specialized software to create the 3D representation of the objects. Photogrammetry, the same as laser scanning, can be terrestrial, based on photos taken from the ground, or aerial, based on photos taken from an aerial vehicle with a mounted camera.

The most common aerial vehicles for point cloud capturing, are the Unmanned Aerial Vehicles (UAVs), namely drones. A UAV is an aircraft without a pilot that has an assistive onboard system that is controlled remotely or autonomously [70]. Various categories of drones differ in terms of flexibility, accuracy, weight, and performance in altering weather conditions. A general categorization is to classify them according to their flight mechanism into Multi-Rotor Fixed-Wing and Hybrid-Wing drones [71]. The choice depends on the intended use and application requirements. Their detecting and surveying systems usually incorporate high resolution visual cameras, RADAR and LiDAR. This technology is the most popular due to the handy, low acquisition and operating cost for point cloud generation through both commercial and open-source photogrammetric software packages [72]. The primary data are obtained by the sensors mounted on UAVs (vertical and oblique overlapping images) and then the point cloud data are extracted after the data post-processing.

Figure 9 shows the data post-processing procedure for the point cloud generation. Images are inserted into the software, the homologous points are detected and matched, and images are aligned and oriented. In the next step, input Ground Control Points (GCPs), i.e., markers with known coordinates, are defined to geo-reference. At this stage, error resolution

Fig. 9 Point cloud generation through 3D model reconstruction. The context in pink color complements the whole process



is also computed. A sparse point cloud is consequently generated using Structure from Motion (SfM), followed by a denser cloud that is created by Multi-view Stereo (MVS), with a metric value. In this phase, for each image, the corresponding depth map is calculated. Dense point cloud can be extracted and stored in .las or .laz file format for further processing. Moreover, a mosaic (orthomosaic) arises when a dense point cloud is converted to mesh with texture. Finally, from the orthomosaic, the Digital Surface Model (DSM) and Digital Elevation Model (DEM) are also extracted.

4.1.5 Comparison and evaluation of different point cloud acquisition and generation technologies

The main advantages of LiDAR technology are high accuracy, fast data acquisition, fast processing time, automated procedures therefore independent from human interventions, independent functionality from bad weather conditions, independent from lighting conditions, e.g., sun inclination as well as during night-time. Since point clouds have high data density, they can be used as input data to create several elevation models, such as DSM, Digital Terrain Model (DTM) and DEM. DSM is a digital representation of the heights of the surface of earth, including man-made structures and above-ground features. DTM is a bare-earth topographic representation of earth's surface, while DEM is a superset of DSM and DTM. However, disadvantages of LiDAR technology also exist. It functions better for static objects, and for moving objects it needs to be combined with other technologies, e.g., camera, GPS, IMU, to establish a complete mapping system. Even though it can penetrate dense foliage, as the rays of light, it cannot penetrate very dense structures.

Finally, it has high operational costs due to costly equipment and the need for experienced operators able to interpret and analyze the captured data. Additional limitations are accuracy problems caused by reflective surfaces; in extreme weather conditions data collection can be interrupted; high dependency of its accuracy on the quality and calibration of the scanning system, the GPS, and IMU components.

RGB-D technology has advantages such as affordable acquisition, computational cheap 3D reconstruction methods, and low power consumption translated to high autonomy. However, in many cases, final images may comprise missing values, translated to holes, which must be filled, or depth maps of low resolutions, which must be up-sampled. Moreover, disadvantages at sensory level are observed. RGB-D sensors may fail to capture objects and surfaces with reflections, transparencies, absorptive materials, motion blurred, noisy characteristics and errors can be displayed (systematic and random) due to strong light and to their limited scanning speed. Additionally, low-cost RGB-D sensors cannot provide high quality data. In this case, a metrological analysis of their performance needs to be considered. Moreover, the detection of point correspondences between two cameras during triangulation with passive RGB-D sensors is also challenging since it needs adequate local intensities and variations of colors in images. Therefore, passive sensor data can provide accurate depth information only in rich textured areas within a scene. For areas with less information (fewer features), active sensors provide better depth measurements. Moreover, since with ToF cameras depth is estimated by the round-trip time of emitted light, measurements are not affected at all by the lack of features on the scene.

By conducting a direct comparison between point cloud generation techniques from LiDAR data and image data (photogrammetry), a set of definitive conclusions emerge. Point cloud quality from LiDAR (aerial and terrestrial) depends on scan frequency, point density and flying height. Point cloud quality from images using SfM is affected by the ground sample distance (GSD), flight altitude and image content. Moreover, the point clouds created from LiDAR are denser (2–100 ppsm) than that from images (1 ppsm or less) [73]. By using UAV platforms, a 3D point cloud can be created with photogrammetry only when a second homologous point is found in another image or overlapping images. The main difference, however, that distinguishes photogrammetry from LiDAR is color, since photogrammetry results in a colored point cloud. Yet, LiDAR point clouds can be more accurate, as already said, due to the fact that LiDAR emits light and reflects features (ground or surfaces), thus, the scene's texture does not affect the modelling. Furthermore, in LiDAR each reflected point has a coordinate location (X, Y, Z) without having to find a second or third point in overlapping images. Due to the total amount of points sprayed at once, the LiDAR laser can penetrate below heavily vegetated areas and provide more accurate surface models. From the photorealism aspect, photogrammetry provides photorealistic mapping (orthomosaics, point clouds, textured mesh), while LIDAR provides a sparse laser point cloud which is colored based on the intensity of reflection; yet, it is without contextual detail [74].

Compared to LiDAR, SAR tomography (TomoSAR) offers moderate accuracy on the order of 1 m, as it is reconstructed from spaceborne data. In contrast, ALS LiDARs provide much higher accuracy on the order of 0.1 m [75]. TomoSAR focuses on different objects than LiDAR due to its coherent imaging nature and side-looking geometry system. It can provide rich information and high-resolution reconstructions in complex scenes, such as buildings, by leveraging multiple viewing angles [76]. The combination of LiDAR and SAR sensors can provide 4D information from space [77]. However, TomoSAR does have some drawbacks, such as its limited orbit spread, the small image number, and multiple scattering, which can lead to location errors and outliers. Another advantage of LiDAR over Radar is the difference in wavelength; the lower wavelength in LiDAR enables the identification of extremely small objects, such as cloud particles. Additionally, it's important to note that LiDAR performance declines in bad weather conditions, while radars can function effectively regardless of weather conditions. Finally, Radars are more robust to weather changes and possess day and night operational capabilities.



Fig. 10 Point cloud visual representation. An example from an archaeological site

4.2 Point cloud formats

A point cloud is sparse, noisy, irregular, and represents objects' shape, size, position and orientation in a scene. The term "cloud" refers to its collection of unorganized points and spatial coherence. However, it has unsharp boundaries, and consists of numerous and scattered points described by 3D coordinates (X, Y, Z) and attributes, such as intensity, while they can also contain additional information, e.g., color. In the case of different sensory combinations, a point cloud can also provide additional multispectral or thermal information. Figure 10 shows a point cloud sample from an archaeological site.

A variety of file formats for point cloud data storage is currently available. The two main categories of point cloud files are ASCII (XYZ, OBJ, PTX, and ASC) and binary (FLS, PCD, and LAS) or both binary and ASCII (e.g., PLY, FBX, and E57). The format selection depends on the data acquisition source and the intended use, e.g., for data meant to be saved for a long time, the best packing format is in ASCII file.

It should be noted that when dealing with point cloud processing using DL models, the input data can either be in its raw or transformed into a more easily handled data structure that suits the requirements of the DL model architecture. Commonly, used structures are volumetric [78], shell (or boundary), and depth maps.

4.3 Comparison of 3D point clouds with other visual data forms

Nowadays, there exist various representation types of the physical world, including 2D images, orthomosaics, depth images, meshes and 3D point clouds. While humans can perceive and understand any kind of representation through vision, the understanding of scenes in the computer vision field is achieved mainly by 2D images and 3D point clouds, used differently due to their distinct characteristics. Therefore, different visual forms are employed for different problems, due to their inherent differences. 2D images are

presented in a regular grid, i.e., an RGB pixel array, while 3D point clouds consist of thousands of points where each point is encoded with spatial coordinates (X, Y, Z), including other information as well. Moreover, 2D images are captured by light rays using a lens and are projections of the 3D world on 2D planes, whereas 3D point clouds represent surfaces, are sparse and contain outliers.

RGB-D images combine four channels of which the three channels include the color (RGB) and the fourth channel represents the depth. In depth images each pixel describes the distance between the object (target) and the image plane [19]. Comparing 3D point clouds and depth maps, one could say that they have different goals and purposes. More specifically, a 3D point cloud has an irregular shape form, whereas a depth map conveys information about the distance. In terms of viewpoint, from a point cloud, it is visible each point used to create the image, while a depth map provides a view of the data points from a particular angle [79]. From the dimension aspect, point cloud images are visible into three axes (X, Y, Z), unlike depth maps which present information only from Z-axis. A depth image, if compared to a flat image, is more accurate and provides additional elements around and behind the target. Point clouds generated from images obtained from UAV platforms vary in terms of quality, outliers, and holes. Their quality depends on the spatial resolution of the images, which is affected by several parameters, such as the flying height, sensor characteristics, and weather conditions, as already mentioned.

Nowadays, there are cases, where 3D point clouds are complemented by 2D images and depth data in various applications, towards a better understanding of a scene. Recently, researchers have developed algorithms and applied learnable approaches using either LiDAR point clouds combined with digital images as input data [80], aerial images (orthophotos) fused with airborne LiDAR point clouds [81], LiDAR and depth data combinations [82]. However, images may present limitations as opposed to point clouds due to sun angle and viewing geometry, occlusions shadows, lack of texture, illumination, atmospheric conditions reflections, and image displacement in areas with steep terrain [83].

5 Computer vision and point cloud processing

Computer vision (CV) was developed in the late 1960s aiming to simulate the human visual system and through automated tasks to achieve, from images or videos, a high-level understanding. This was achieved by information extraction related to their structure. In the next decades, many algorithms and mathematical models have been applied to object and shape representation from various cues, such as shading, texture, and contours [84]. In the last decades, the need for 3D

reconstruction and visualization of the real-world, including camera calibration, led to optimization methods and multi-view stereo techniques. However, images are limited from spectral characterization, sampling effectiveness, measurement accuracy, and operating conditions and the natural data process in raw form is also limited due to the parameters' sensitivity, the algorithms' strength, and the results' accuracy [85].

To tackle these issues, classic machine learning methods were applied to various 3D point cloud related applications. Considering the rapid computer vision evolution, the needs, and requirements of high-precision data for real-world recording and modelling are increasing. By using 2D images the latter cannot be achieved, since 2D images do not provide depth and position information that are essential for advanced applications, e.g., robotics and autonomous driving. At the same time the enhancement of technologies for 3D geospatial acquisition of data from various 3D sensors brought to the fore a plethora of computer vision applications providing new data formats such as point clouds, for the rich representation of the scenes [86].

Point cloud processing for information extraction is a complex and challenging task due to its unordered structure and different sizes, which depend on the recorded scene. Moreover, matching between scenes is not feasible due to the lack of neighboring. However, traditional machine learning methods for the processing of point clouds depend on handcrafted features and specifically designed optimization methods. Point cloud features of static properties are invariant to transformations, therefore, application-oriented optimization methods need to be developed in each case, and generalization cannot be achieved [87]. Therefore, the need for developing enhanced and more efficient methods to process point cloud data is apparent. DL methods can automatically learn discriminative features, have proven their effectiveness, and therefore have been also adapted to point cloud processing. Recently, researchers and industrial organizations have employed DL techniques to handle point clouds. In deep learning, the features are learned automatically based on artificial neural networks during the training process. However, used methods depend on the application and the computer vision task and pose many challenges.

In the next subsection, the basic challenges of point cloud processing are reviewed, while in the following section, task-oriented challenges of DL methods for point cloud processing are analyzed.

5.1 A brief review of DL-based point cloud processing and corresponding challenges

Currently, the 3D representation of scenes via point clouds is promoted by a variety of different advanced sensors. Yet, data

does not contain topology, and connectivity, including occlusions, can be affected by illumination, objects' motion, noise of sensors, and sources of external radiation. The latter issues can lead to wrong coordinates' estimation and thus, point clouds can be sparse due to the mostly concentrated points around key visual features, appearing holes and missing data due to unsampled areas around smooth regions [88, 89]. This data sparsity and uneven points' distribution can get worse depending on the quality of the acquisition device or in cases of specific sensors, such as ToF sensors where occlusions and hidden surfaces deteriorate the generated point cloud.

Redundancy of data is another critical issue on point clouds. Point cloud representations can be highly redundant, compared to meshes representations, especially on planar surfaces. The latter results in large files that cannot be shared or stored. In such cases, the point cloud needs to be efficiently organized to provide good representations that could be feasibly processed. Finally, the last basic challenge in point cloud processing is the existence of noise in the resulting models. Illumination, radiation, motion blur, sensory noises, etc. can severely affect the point cloud, deriving false estimations of surfaces and flying pixels. These kinds of artifacts can be observed on vision-based acquisition devices, as well as in all devices where environmental changes can alter the quality of the derived data, e.g., in FMCW radar sensors.

Addressing such issues with traditional methods can lead to increased memory costs [89]. Computer vision can offer more powerful tools and point cloud processing techniques. DL methods are capable of confronting the limitations of traditional computer vision solutions using deep denoising [90], volumetric multi-view, and point-based methods [91]. In addition, DL methods for feature learning can be applied pointwise, such as Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) as the PointNet family [92], or on graph and hierarchical data structures [93] by either converting the point cloud into other formats or directly on the raw data [94]. Challenges related to resolution were faced by super-resolution methods aiming to upscale low-resolution representations [95]. The challenge imposed by the great number of points of a point cloud, in regular conditions, is handled by finding similarities or dissimilarities i.e., comparing corresponding pairs of pixels like in the case of images; however, on point clouds, there is no 3D dissimilarity measurement. The latter was tackled by using supervised, unsupervised and autoencoder methods [96–98]. Finally, additional methods have been developed focusing on capturing local structures and providing richer representations through sampling, grouping, and mapping functions [98, 99].

It should be noted that all aforementioned challenges are more general; specific challenges emerge when DL methods are used in different tasks, as reviewed in the following section.

6 DL-based computer vision tasks using point clouds

In this section, the main vision tasks are classified into six categories: *registration*, *segmentation*, *classification*, *3D object detection and tracking*, *compression* and *completion*. The advantages, disadvantages and challenges of using point clouds in each task are discussed separately, aiming to deliver an in-depth understanding of the impact of DL on point clouds and the extent to which various point cloud management challenges have already been addressed. This analysis is significant as it can highlight research gaps, current trends, and future research directions in the field.

DL-based methods for geometric data pose challenges in terms of performing convolution. Numerous DL model architectures have been proposed aiming to learn geometric features from point clouds and implementing main DL operations on 3D points. The main idea is to interpret point clouds locally as structured data by considering each point concerning its neighboring points or achieving a learning process that remains invariant to the order of the point cloud. Based on this, feature learning on point clouds is classified according to Liu et al. [40] in *raw point-based* methods, where DL models directly use raw point cloud data and *k-dimensional tree* (Kd-tree) methods, where the point cloud is transformed into another representation before being inserted into the DL models.

Table 2 summarizes the main vision tasks as defined in this work, providing optical examples, definitions for every task as well as information about the type of input/output data.




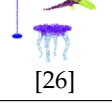
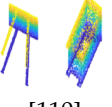
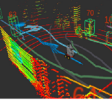


6.1 Registration

In registration, two-point clouds that are acquired from different angle views of the same scene are aligned through a rigid transformation, aiming to obtain a common coordinate system [46].

When point cloud data is captured by the same sensor, at different times, it may contain noise and outliers, while being partially overlapped due to varying viewpoints. In cases of cross-sensory data acquisition, different scales are introduced due to the physical metrics, making rigid motion prediction for aligning one point cloud into another challenging. Nowadays various challenges are being addressed by DL methods, including partial point cloud registration. To classify DL methods for the registration task, we can first distinguish them based on the origin of the data: *same sensor* or *multi sensors* [100]. Figure 11 illustrates the classification of DL methods for the registration task.

Methods for the same sensors are based on optimization, feature learning and end-to-end learning. Optimization methods utilize techniques such as Iterative Closest Point (ICP)

Table 2 Taxonomy of main 3D point cloud vision tasks

Tasks	Examples	Definition	Input	Output
Registration	 [108]	Finding a rigid transformation for aligning two-point clouds	Two 3D point clouds	Point clouds union
Segmentation	 [99]	Semantic: point cloud classification into multiple homogeneous regions with the same properties (scene level)	One 3D point cloud	Class prediction for every point in the cloud
	 [109]	Instance: point assignment to each object instance and predict its semantic label (object level)		
	 [26]	Part: data point classification where a group represents a physical part of an object (part level)		
Classification	 [110]	Categorization of the points by a set of geometric attributes to a predefined set of classes (e.g., vegetation, ground, roofs, etc.)		Prediction of classes
Object detection and Tracking	 [111]	Identification and Localization of objects in a sequence of images or video		Objects inside the bounding boxes and class prediction
Compression	 [3]	Volumetric visual data compression (geometry and attributes)		3D point clouds
Completion	 [112]	Shape generation and estimation, appearance of real objects derived from a partial point cloud		

[101], graphs [102], Gaussian mixture models (GMM) [103] and semi-definite registration [104]. One of the key advantages of this category is the presence of rigorous mathematical theories that guarantee their convergence. Additionally, these methods do not require training data and can generalize well to unknown scenes. To address challenges like noise, outliers' density variations, and partial overlap, optimization methods are employed; nonetheless, the computation cost is increased.

Feature learning methods are used for accurate correspondence estimation, including learning on both volumetric

and point cloud data. Volumetric methods involve converting point clouds into 3D volumetric data and then utilizing a Neural Network (NN). However, they require a large Graphic Process Unit (GPU) memory and are sensitive to rotation variations. Some representative algorithms in this category are PPFNet [105], SiamesePointNet [106] and deep closest point (DCP) [107]. These methods offer robust and accurate registration using a simple Random Sample Consensus (RANSAC) iterative algorithm. Nonetheless, certain issues persist, such as the necessity for large training data and poor registration performance in unknown scenes.

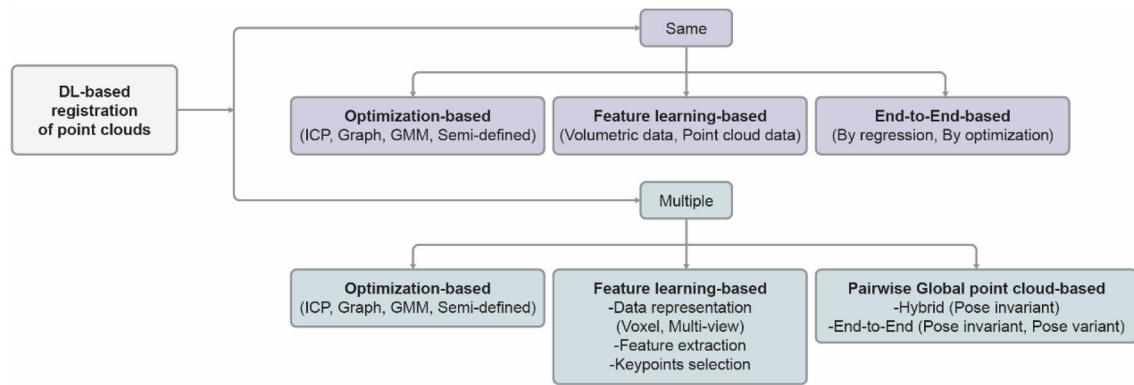


Fig. 11 Classification of DL methods for registration of point clouds

In End-to-end learning methods, two-point clouds are inserted, and a transformation matrix that aligns them is obtained as output [113]. These methods encompass registration by regression and optimization. Regression transforms the registration problem into a regression task [114], combining the conventional optimization theories with deep neural networks (DNN), resulting in improved accuracy compared to previous methods. Algorithms for both regression and optimization have also been developed. End-to-end methods are developed especially for the task, and the optimization of the NN depends on the specific objective. Note as drawbacks that DL regression methods are considered “black boxes”, the coordinate measurements based on Euclidean space are sensitive to noise and density differences, and feature-metric registration focuses on local structure information.

Point cloud registration methods based on cross-sensors face more challenges compared to single-sensor cases, requiring the use of advanced registration frameworks. The benefits of cross-sensor methods include leveraging advantages from combining different sensors, providing the best information for augmented reality applications; however, they also suffer from limitations such as lower accuracy and higher computational cost. These methods can be categorized into *optimization-based*, *feature learning-based* and *pairwise global point cloud-based*. Optimization-based methods aim to estimate the transformation matrix using optimization techniques [115] or deep networks [116]. While these methods are similar to the same sensor approaches, the computational cost problem remains an issue, and their performance with different datasets can be problematic. DL methods offer models focusing on various aspects of registration, including feature extraction and key point selection [117], key point detector [118], and the entire registration process embedded in a DL network [119]. Pairwise global point cloud-based methods [120] consist of hybrid methods that exploit pose-invariant features and feature descriptors for local features’ extraction. Additionally, there are End-to-End methods, which comprise both pose-invariant and

pose-variant feature methods, with pose-invariant methods excelling [121]. Recently, a probability driven approach for point cloud registration has been proposed [122], outperforming state-of-the-art registration methods on registration accuracy.

6.1.1 Comparative discussion on image-based registration

When it comes to images, image registration involves aligning multitemporal and multimodal images, as well as images from different viewpoints. Image registration methods aim to address specific challenges such as finding similarity measurements, especially for multimodal images, reducing the computational cost, particularly in real-time applications, improving quality of images and handling deformations. Moreover, traditional methods often suffer from good generalization and usually converge to local minima [46]. To address these challenges, DL models, such as CNN, RNN, Autoencoder, Reinforcement Learning (RL), Generative Adversarial Network (GAN), as well as regular intensity-based similarity metrics like sum-of-square distance (SSD) and mean square distance (MSD), have been extended to tackle the geometric computer vision task of registration [46]. However, when dealing with multimodal images, the results were found to be poor. To address the metric problem, handcrafted descriptors were applied, but these descriptors were error-prone, and deep similarity metric methods slowed down the registration process. Additionally, the image alignment’s quality directly impacted the accuracy of the models. To tackle accuracy problems, special data augmentation techniques were proposed [123]. Despite presenting satisfactory results, these techniques posed difficulties in optimization and did not reduce the computational cost. DL methods mainly focused on rigid registration, since non-rigid registration models involved high dimensionality and non-linearity. With DL methods, an improvement of 20%–30% was observed [124] compared to traditional methods.

The most significant shortcoming of DL lies in the limitation of the transformation model from high to low dimensionality [125]. The high dimensionality of the output parametric space, coupled with the scarcity of datasets for training, containing ground truth transformations, and the challenges of regularization in predicted transformations are tackled through supervised transformation prediction and the use of data augmentation methods [126]. However, these approaches insert additional errors, like the bias of unrealistic artificial transformations and shifts of image domain between the testing and training phases. Additional problems arise when the transformation fails to captivate the wide range of variations found in real image registration scenarios, leading to potential mismatches between image pairs. To address this issue, transformation generation models [127] are employed, and to overcome the scarcity of training datasets, unsupervised transformation prediction is applied [128].

6.2 Segmentation

Point cloud segmentation is utilized for scene understanding and to determine the shape, size, and other assets of objects in 3D data [129]. During segmentation, a point cloud is divided into different segments (subsets) with identical attributes; in other words, points are clustered based on similar characteristics into homogenous regions. Segmentation is an essential task in 3D point cloud processing since it is the first step for detecting objects in a scene that cannot be directly discerned from a raw point cloud directly [130].

Three types of segmentation can be distinguished: *semantic*, *instance*, and *part* segmentation. In semantic segmentation, objects are grouped into predefined categories. Instance segmentation is a specialized form of semantic segmentation that detects instances of objects with the same semantic meaning and defines their boundaries. Object part segmentation addresses the challenge of providing pixel-level semantic annotations that imply fine-grained object parts, instead of just object labels. Semantic, instance and part segmentation are applied at scene, object, and part levels, respectively. All forms of segmentation present challenges related to comprehending details of the global geometric structure for every point, defining surface descriptors that describe the object's parts, as well as developing robust algorithms to compute these features [131]. Segmentation methods learn point distribution patterns from annotated datasets and make predictions. Previous traditional segmentation methods have encountered challenges in defining feature calculation units and developing suitable feature descriptors for classifier training. However, handcrafted features act as a limitation factor for the generalization performance of algorithms in complex scenes. In contrast to traditional machine learning methods, DL methods address the aforementioned challenges by employing DNN training to encode point features

and make predictions, or to design effective backbones, leveraging the unique characteristics of point clouds.

The classes of segmentation methods are illustrated in Fig. 12. Semantic segmentation includes *projection-based* methods, further categorized into *multi-view*, *spherical* and *cylindrical* methods. Other methods involve *discretization-based* methods (Dense or Sparse), *point-based* methods (Point-wise multi-layer perceptron (MPL), Point convolution, or RNN, Graph), *Transformer-based* and *hybrid-based* methods. Instance segmentation methods are categorized in proposal and proposal-free methods, while the last category is part segmentation. In their simplest form, these methods often apply pre-trained CNN models, e.g., AlexNet, VGG, GoogLeNet, etc., on various image datasets. In what follows, each category of Fig. 13 is examined separately.

6.2.1 Semantic segmentation-projection-based methods

In projection-based methods, point clouds are projected into 2D images. These methods are efficient in terms of computational complexity and can result in improvement of performance for various 3D tasks by capturing several views of the area of interest. Predictions are then made based on the outputs, either through fusion or majority voting. However, multi-view segmentation methods are easily affected by viewpoint selection and occlusions, and they do not exploit geometric and structural information due to information loss [132]. On the other hand, spherical methods achieve fast and accurate segmentation, making them suitable even for the segmentation of LiDAR point clouds in real-time [133]. Semantic labels of 2D range images are assigned to 3D point clouds to enhance the discretization of errors and improve the quality of outputs. Spherical projection retains more information compared to multi-view methods, making it suitable for labeling LiDAR point clouds. Nevertheless, discretization errors and occlusion issues persist. Methods based on cylindrical coordinates have recently proven to be really effective in representing LiDAR point clouds for various tasks [134]. Despite their sparsity and density effectiveness, they still encounter noise issues [135].

6.2.2 Semantic segmentation-discretization-based methods

Discretization-based methods transform a point cloud into a discrete representation structure either dense, referring to voxels or octrees, or sparse, referring to permutohedral lattices. Then, dense or sparse convolution can be easily employed. In dense methods the space taken by point clouds is divided into volumetric occupancy grids and all points that belong to the same cell are assigned to the same label. Subsequently, predictions are made for each voxel center using a convolutional architecture. Previous methods voxelized the

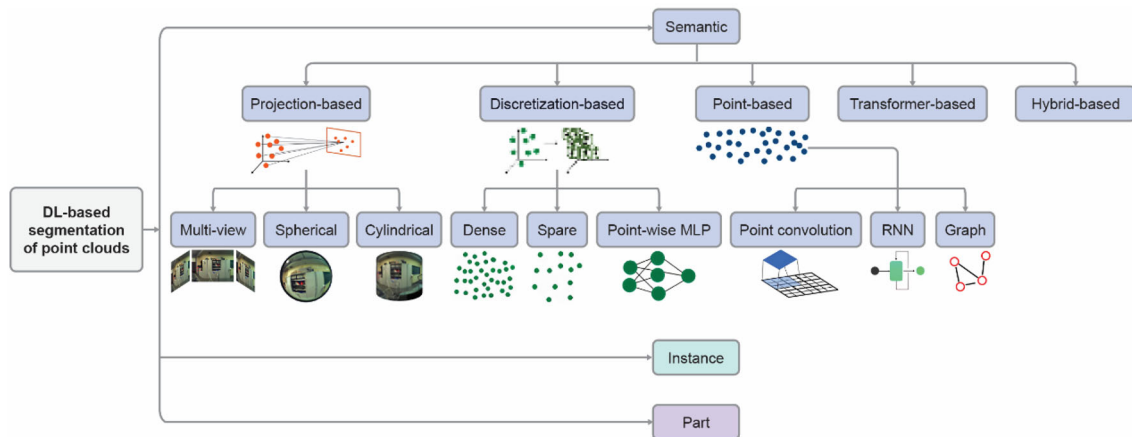
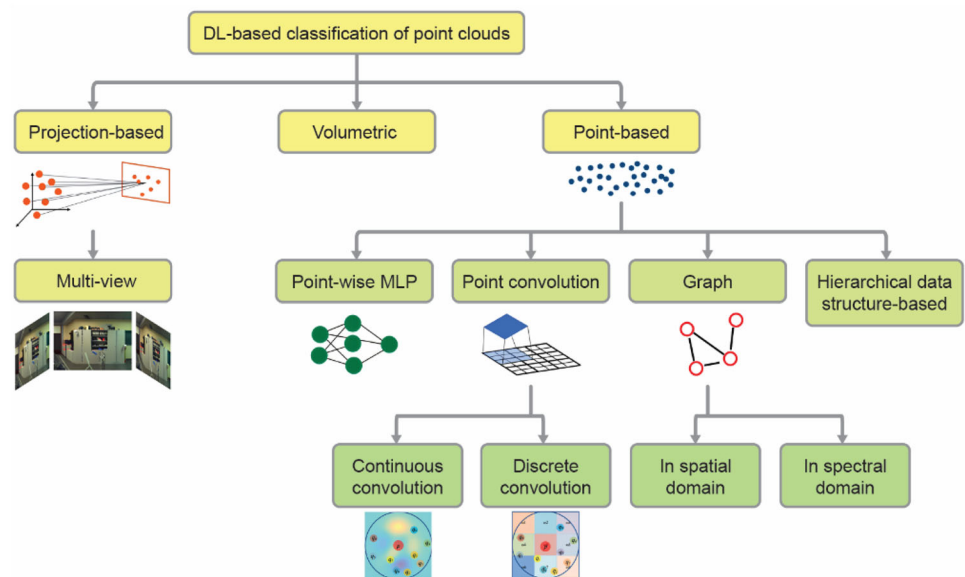


Fig. 12 Categorization of DL-based segmentation tasks on point clouds

Fig. 13 Categorization of DL-based classification tasks for point clouds



point clouds as dense grids; however, these methods were obstructed by the granularity of the voxels and the boundary artifacts due to the partitioning of the point cloud. In practice, there is no selection of a suitable grid resolution. To address these issues, advanced methods exploit the scalability of fully CNNs, allowing them to handle even large-scale point clouds [136]. Moreover, these methods can train volumetric networks with different spatial sizes point clouds. The latter can lead to high computational costs due to the high resolution and the loss of details. To mitigate this, trilinear interpolation models that can learn automatically or 3D convolution filters have been used [137]. The natural sparsity of point cloud models results in a relatively small number of filled cells in volumetric representations. To resolve this, sparse convolutional networks have been proposed [138]. These networks reduce memory and computational costs by limiting the output of convolutions to be related solely to

occupied voxels. In this way, these methods can process efficiently high-dimensional and spatially sparse data.

6.2.3 Semantic segmentation-Point-based methods

Point-based networks operate on unstructured point clouds, avoiding some limitations posed by previous methods, such as projection and discretization. In this category, point cloud data processing is conducted directly. These methods are divided into point-wise MLP, point convolution, RNN based, and graph-based methods.

Point-wise MLP methods utilize the joint MLP as the main unit of their network due to its superior effectiveness. However, the extracted features by these methods may not fully describe the local geometry and common interactions between points. To address this, various networks have been

proposed, involving attention aggregation, neighboring feature pooling, and local–global feature concatenation [139], which enable better local structure learning and wider point capturing. Point convolution methods apply specific 3D convolution operators tailored to continuous or discrete point clouds. The definition of the 3D continuous convolution kernels is done on a continuous space, where the weights for neighboring points are associated to spatial distribution. In discrete point clouds, CNNs are specified on regular grids, and the neighboring points' weights are associated with the offsets regarding the center point [140]. RNN networks can model the interdependency of acquired point cloud at different times. PointRNN leveraged this idea [141], while other solutions have been proposed by combining CNNs and recurrent architectures [142], capturing inherent context features from point clouds [143], exploring several RNN architectures [144], and using dynamic models [145]. Additionally, graph methods have also been developed for capturing the shape and geometric structure of 3D point clouds [146].

6.2.4 Transformer-based methods

Transformer-based methods are decoder–encoder structures that consist of input embedding, positional (order) encoding, and self-attention, enabling the learning context and tracking relationships in sequential data. Particularly, self-attention plays the most important role as it generates sophisticated input attention features, according to the global context, and consequently the output attention also learns the global context [147]. These methods are suitable for processing of point clouds due to their natural independence of the input order. In these frameworks, Natural Language Processing (NLP) methods provide better performance than CNN, allowing for parallel processing and are much faster than any other model with similar performance [148].

6.2.5 Hybrid-based methods

Hybrid-based methods are popular and involve the utilization of over-segmentation or point cloud segmentation algorithms [146] as a pre-segmentation stage to reduce the data volume. However, reducing the amount of data may lead to a slight loss of accuracy. Moreover, additional methods are learning multi-modal features from 3D scans and leverage all available information. For example, 3D-multi-view networks combining RGB and geometric features [149], 3D CNN stream and a back-projection layer to learn 2D embeddings and 3D geometric features [150], or a unified point-based framework for learning 2D textural appearance, 3D structures and global context features from point clouds. These networks are directly applied to extract local geometric features and global context from a sparse sampling of point sets without voxelization. In contrast, other techniques like

Multi-view PointNet (MVPNet) [151] combine appearance features from 2D multi-view images and spatial geometric features in the canonical point cloud space.

6.2.6 Instance segmentation

Instance segmentation focuses on distinguishing points of different semantic meanings and separating instances accordingly. It combines the advantages of semantic segmentation and object detection; however, it requires more accuracy and granularity due to points, compared to semantic segmentation methods. Instance segmentation presents some significant challenges, such as difficulty in segmenting smaller objects, dealing with occlusions, inaccurate depth estimation, and handling of aerial images. Existing DL-based instance segmentation methods can be categorized into *proposal* and *proposal-free* approaches [35].

Proposal methods transform the problem of instance segmentation in 3D object detection and prediction of instance mask [152]. For this purpose, several methods have been introduced [153, 154], with Generative Shape Proposal Network (GSPN) [150] being the first reported approach. However, these techniques are computationally expensive, require substantial memory and rely on large amounts of data, presenting challenges in their implementation. Proposal-free methods [155, 156] consider instance segmentation as a successive step of clustering at a pixel level to generate instances after semantic segmentation, without involving any object detection module. Existing methods assume that points within the same instances can have alike features, thus focusing on discriminative feature learning and grouping of points. Group Proposal Network (SGPN) was the first proposal-free method [157] reported in the literature. Proposal-free methods do not require computationally costly region-proposal components. However, they exhibit lower objectiveness in instance segmentation since they do not clearly identify boundaries of objects. Essentially, these methods rely on grouping/clustering techniques at a pixel level to generate instances, covering potential gaps through the use of semantic segmentation methods.

6.2.7 Part segmentation

In part segmentation, semantic annotations indicate fine-grained object parts at the pixel level, rather than just object labels. The difficulties in this task are related to 3D shapes. For instance, parts of shapes having the same semantic label exhibit big geometric variations and ambiguity, and the total parts having the same semantic meaning can differ significantly. These challenges have been partially faced by volumetric CNNs [158], Synchronized Spectral CNN [159], Shape Fully Convolutional Networks [160], and part decomposition networks [161], which have reported improvements

in part segmentation outcomes. However, they have stated sensitivity to initial parameters and limitations in learning local features.

6.2.8 Semantic segmentation

As a general conclusion, DL-based methods for semantic segmentation on point clouds offer numerous benefits, even for very large-scale point clouds. Instance segmentation requires more discriminative features, while the combination of semantic and instance segmentation can enable label prediction simultaneously [162]. Such an approach can be particularly useful for still images displaying many overlapping objects in a scene, as it allows models to be better trained in real-world scenarios, effectively handling dense objects and significant overlaps between them [163]. Furthermore, image segmentation in various fields, such as medicine, highlights the need for large-scale annotated 3D image datasets, which can be challenging to create, compared to datasets in lower dimensional counterparts.

6.2.9 Comparative discussion on image-based segmentation

Image segmentation methods on 2D data have been developed using interpretable deep models [164], weakly-supervised and unsupervised learning [165], unsupervised learning [166], self-supervised learning [167] and Reinforcement Learning [168]. However, challenges regarding the kind of information used with interpretable models, their behavior, dynamics and the efficiency related to accuracy and computational cost remain. Image segmentation algorithms depend on the spatial properties of image intensity values. However, these intensities are not purely quantitative and can be influenced by a variety of factors, including hardware, protocols, and noise. Researchers have made attempts to address these challenges by using traditional methods [169]. Unfortunately, these methods have had limited success, as they often require manual interventions for abnormal cases and lack the necessary robustness to handle sensitive input data effectively. To overcome these limitations, more advanced approaches have been developed, such as U-net [170] and DeepLab [171].

These methods leverage large amounts of image data to enhance performance, increase robustness, and obtain more reliable estimates. Additionally, they help mitigate the computational cost by combining differently constructed architectures in various applications, particularly in the field of medicine. Methods based on RGB images in general, lack information to achieve semantic segmentation of complex scenes. It should be noted that RGB-D semantic segmentation providing additional depth information, was concluded to reach to better segmentation results [172].

6.3 Classification

Point cloud classification refers to the assignment of pre-defined category labels to groups of points within a point cloud, determining which points belong to which objects. In the past, methods for point cloud classification relied on handcrafted features and traditional classifiers for point cloud preprocessing, as well as machine learning techniques, like unsupervised, supervised, or a combination of them [173]. However, unsupervised methods were limited by their dependency on thresholds, leading to poor adaptability. Supervised methods struggled to learn high-level features, making it challenging to achieve significant improvements in classification accuracy. Although combining these methods improved classification accuracy to some extent, they still inherited certain limitations [174]. Nowadays, DL methods have proven their powerful capabilities for representation learning directly from the data. The latter has led to significant advancements in the field of 3D point cloud classification.

DL-based methods for classification of point clouds can be divided in *projection-based*, *volumetric-based* and *point-based* methods according to the different input data formats used by the neural networks, as illustrated in Fig. 13. It should be noted that many of these classification methods share, to some extent, similar concepts with segmentation methods.

6.3.1 Projection-based methods

Multi-view methods [175] involve projecting 3D shapes into multiple views and extracting view-wise features, which are then fused to achieve precise shape classification. However, these methods often encounter information loss. One of the main challenges for these methods lies in the way to combine the several view-wise features in a discriminative global representation. As a result, several methods have been suggested aiming towards improving the accuracy of recognition [176, 177].

6.3.2 Volumetric-based methods

Volumetric-based methods voxelize point clouds into 3D grids and utilize 3D CNN for shape classification [178]. The main challenge with this approach is the scaling of dense 3D data, as both memory footprint and computations increase exponentially with the resolution. To address these concerns, researchers have introduced OctNet [137], which reduces the computational and memory costs, as well as the runtime, especially for high-resolution point clouds. However, despite these efforts, volumetric-based methods are not appropriate for handling large-scale point clouds because of the persistently high computational cost, which has not yet been efficiently resolved.

6.3.3 Point-based methods

Point-based methods are applied directly for processing raw points without voxelization or projection, enabling high precision and efficiency due to the irregularity of the distribution of point clouds and the scenes' complexity [179]. These methods encompass point-wise MLP, convolution, graph, and hierarchical data structure-based approaches. It is noted that point-wise MLP, convolution and graph methods share similarities with the segmentation methods, albeit with some variations.

Point-wise MLP methods aggregate global features using a symmetric aggregation function. However, applying DL methods for images directly to a 3D point cloud is challenging because of their irregular data nature. Unlike images where kernels are described on a 2D grid structure, designing point clouds convolutional kernels is complex due to their irregularity. Numerous methods have been developed based on point convolutional kernels, which can be categorized into *continuous* and *discrete* convolution methods. Continuous methods define a convolutional kernel in a continuous space, where the neighboring points' weights are determined based on their spatial distribution regarding the center point. These methods can be translated as a weighted sum over a given subset [180, 181]. 3D discrete methods describe convolutional kernels on conventional grids, and the neighboring points' weights are determined based on the offsets from the center point [182]. In graph NNs each point is treated as a vertex in a graph, and directed edges are generated for the graph based on the neighboring points of each vertex. Then, feature learning is applied in either the spectral or spatial domain for effectively capturing the local structure data of point clouds [183]. However, the receptive field size of many graph NNs is often not enough for capturing comprehensive contextual information. Graph-based methods in the spatial domain define operations like convolution and pooling, while convolutions are defined as spectral filtering, implemented through signals' multiplication on the graph with eigenvectors of the graph Laplacian matrix. On the other hand, hierarchical methods employ networks constructed using different hierarchical data structures, wherein the learning of point features is done hierarchically from leaves to the root node of a tree structure [184]. Recently, a unified representation of image, text, and 3D point cloud was introduced, namely ULIP, pre-trained by using object triplets from all three modalities [185]. The ULIP reported state-of-the-art performances in standard 3D classification and zero-shot 3D classification, bringing multi-modal point cloud classification in the forefront of point cloud related research. A 3D point cloud classification method based on dynamic coverage of local area was presented in [186], introducing a new type of convolution to aggregate local features. For point cloud classification and segmentation, it was also proposed a new

space-cover CNN (SC-CNN) [187], towards implementing a depth-wise separable convolution to the point cloud using a space-cover operator. The latter approach was proven capable of better perceiving the shape information of point clouds and improving the robustness of the DL model.

6.3.4 Comparative discussion on image-based classification

Classification of 2D images using DL refers to the training process of a model to classify the images into predetermined classes. CNNs are widely employed in image classification due to their ability to capture and learn hierarchical features. Due to their structure, their powerful feature learning abilities, as well as the availability of GPU computing, outperform in most cases the traditional machine learning techniques. Therefore, CNNs have reported significant performances in various large-scale identification computer vision tasks. Despite their great achievements, DL models for 2D image classification still face challenges to tackle, such as insufficient data because due to the fact that DL models require large amounts of labeled data for effective training, overfitting issues especially when the model is complex having a large number of hyperparameters, resulting in capturing noise instead of general patterns. Addressing these challenges requires a combination of advanced algorithmic improvements, data curation strategies and deployment of robust image classification models that could be applied to a wider range of cases.

6.4 Object detection and tracking

Object detection and tracking task also involves 3D scene flow estimation. Given the arbitrary nature of point cloud data, the aim of object detection is the identification and localization of instances of predefined categories, providing their geometric 3D location, orientation, and semantic instance label. This information is embodied by a bounding box encompassing the target, indicating the object's center position, and orientation size [188, 189]. Figure 14 shows the categorization of object detection and tracking methods.

6.4.1 3D object detection

Object detection finds applications in various real-world scenarios, including autonomous driving, surveillance, transportation, scene analysis from drones, and robotic vision [40]. However, specific challenges arise in this task, such as simultaneous classification and localization, real-time processing, handling multiple spatial scales and aspect ratios, dealing with multiple feature maps, limited data, and addressing class imbalances [190]. To address the issue of simultaneous classification and localization, specific methods use multi-task loss functions, penalizing misclassifications

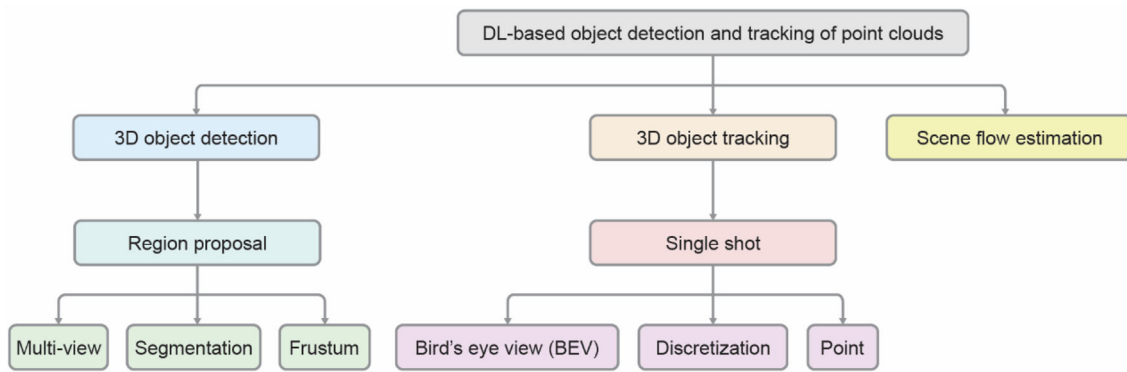


Fig. 14 Classification of DL-based object detection and tracking tasks for 3D point clouds

and localization errors. Convolutional Neural Networks are employed to handle false classifications and misalignment of bounding boxes. For real-time detection, speed methods like Yolo or Faster R-CNN are applied to mitigate the problem to a certain extent. It is worth to note that maintaining real-time speed (e.g., at least 24 fps) can be challenging when processing video shoots continuously. Currently, YOLO v3 offers object detection at multiple scales. However, there are still challenges that need improvement, such as achieving real-time detection with high-level classification and localization accuracy, as well as ensuring continuity in video tracking between frames, rather than processing them separately [191]. In addition, during object tracking, several challenges arise, including smooth object motion without abrupt changes, dealing with sudden and gradual changes in both object and scene backgrounds, maintaining camera stability, handling varying numbers and sizes of objects, and addressing occlusion limitations [192]. These challenges are tackled by using region proposal [193, 194] and single-shot proposal methods [195]. Region proposal methods suggest possible regions (proposals) that may contain objects and subsequently the extraction of region-wise features towards determining the class label of each proposal.

Region proposals can be achieved through multi-view, segmentation and frustum methods, while single-shot methods include Bird's Eye View (BEV), discretization and point-based methods. Multi-view methods aim to obtain 3D rotated boxes by fusing proposal-wise features from different view maps [196, 197], but they often suffer from high computational costs. To address this, researchers have developed various methods to effectively fuse data from alternative modalities, enabling robust representation extraction from the input data [198–200]. Such methods exhibit superior object recall rates and are more appropriate for complex scenes with strong occlusions and packed objects, when compared to the previous approaches.

Frustum methods leverage 2D object detectors for generating 2D potential regions of objects and derive a 3D frustum

proposal for each region. However, the performance of these methods is restricted by their reliance on 2D image detectors [201]. Single-shot methods predict directly the class probability and regress the 3D bounding box of objects by utilizing a single-stage network. These methods do not need of the generation of region proposal or post-processing, allowing them to operate at high speed.

Based on the type of input data, single-shot methods are further categorized into *BEV*, *discretization* and *point-based* methods. BEV methods use representations as their input for estimating the heading angles and location of objects. The reported generalization performance to point clouds with different densities was poor. However, this issue was resolved by using normalization maps that take into account the variations between different LiDARs [202]. Discretization methods transform point clouds into regular discrete representations and employ CNNs for predicting all classes and objects' 3D boxes. The challenge with these methods lies in the requirement of significant computation resources because of 3D convolutions, and the data sparsity. For addressing this issue, a voting scheme has been applied to cover each non-empty voxel, resulting in complexity of computations that was analogous to the number of occupied voxels. Additionally, other methods have been developed towards saving memory and accelerating computation by fully utilizing the sparsity of voxels. Point-based methods use directly raw point clouds as inputs. While these methods can be time-consuming, they can be mitigated using fusion sampling strategies [193]. While saliency perception could aid segmentation, localization and detection tasks, it should be noted that relevant research on 3D point clouds is limited. In [203], salient detection is performed by employing principal component analysis in a sigma-set feature space, achieving high performances without using topological information.

6.4.2 3D object tracking

The object tracking task uses point clouds sequences along with the object's location in the first point cloud towards estimating the location of the object in subsequent frames [204, 205]. This category draws inspiration from object tracking algorithms used with 2D images [206]. Object tracking exploits the strong geometric information found in point clouds, enabling it to beat challenges related to image-based tracking, occlusions, illuminations, and scale variations [207].

6.4.3 Scene flow estimation

Scene flow estimation of 3D point clouds uses two point clouds to describe how each point moves as it traverses through a scene, with the aim to learn valuable insights from a point clouds sequence. Scene flow estimation encounters several challenges, such as vectors that may deviate from the ground truth, handling deformable objects, achieving accuracy in rigid dynamic scenes, managing computational costs, and processing sequential point clouds. Although these challenges have been addressed by various methods, their performance is still limited because of the small scale of available datasets [208].

6.4.4 Comparative discussion on image-based object detection and tracking

In 2D image object detection and tracking, the goal is to determine the existence of objects from given categories in an image and localize their spatial positions using bounding boxes [209]. In this context, the recognition aspect involves finding an appropriate feature space and similarity metric. On the other hand, 2D visual object tracking faces several challenges, including tracking a target image in each video frame, localizing its Region-Of-Interest (ROI) and detecting the object over the video frames, while dealing with issues like object deformation, blur, abrupt object motion, noise in image sequences, changes in scene illumination, object sizes, object-to-object and object-to-scene occlusions, non-rigid object structures, camera movements, and real time processing requirements. For both single and multi-object cases, additional challenges arise, such as the existence of artificial or sunlight, varying weather conditions, different times of day, shadows on the ground, reflections, and occlusions. From the object representation perspective, the problems center around extracting the minimum amount of information from the object, such as color, intensity, feature points, and spatialized color histograms. To tackle these challenges, several methods have been developed [210].

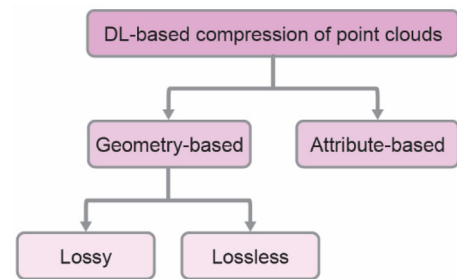


Fig. 15 Categorization of DL-based compression tasks for point clouds

6.5 Compression

Since the point cloud data acquired from various 3D technologies comprises of a huge number of points, there is a need to efficiently transfer and store this data. This is accomplished through the compression task. Point clouds consist of geometry information, which indicates the position of the points and attributes that provide additional details for each point. This information can be coded jointly or separately, using either lossless or lossy computer techniques [211].

Challenges encountered in the compression task include computational costs, support for incrementally acquired data, local decompression, geometry resolution, and managing the points' number. The geometry of objects is sometimes based on sparse signals, resulting in information distributed across an irregular sparse domain with irregular neighborhoods. As a result, compression may lead to feature loss, making object identification difficult. For this reason, DL methods based on coarse-to-fine geometry are employed to tackle these problems. Besides, DL methods are further differentiated according to the encoding domain and prior information. Figure 15 illustrates the categories of DL methods used for point cloud compression tasks. Furthermore, voxel methods based on encoding domain use denser point clouds compared to point methods, while methods that require prior information are divided into unstructured and LiDAR sensor data utilization.

6.5.1 Geometry-based methods

Geometry-based methods for point cloud compression can be categorized into *lossy* and *lossless* approaches. Lossy methods utilize CNNs based on autoencoders architecture for transforming the point cloud into a latent space of a lower dimension and then reconstructing an output similar to the input, treating the point cloud as a binary signal on a voxels grid [211, 212]. The decoding process is considered a problem of binary classification in a voxel grid, even though point clouds are sparse, resulting in class imbalance [213]. This is resolved by using focal loss methods. However, challenges arise with spatiotemporal complexities and

filling certain gaps, which are tackled through block partitioning and sparse convolutions, and by applying multiscale approaches, respectively. Nevertheless, these methods may degrade the sparsity of point clouds. Another challenge is the decoding mismatch related to DNN, which is addressed using adaptive thresholding approaches to maintain density consistency between training and tests datasets, and encoding thresholds to ensure the reconstructed point cloud retains the same number of points as the original [214]. Additionally, to reduce training time, a sequential training scheme has been proposed [215].

Lossless methods aim to enhance the prediction occupancy probabilities by utilizing entropy models. Different frameworks have been proposed, focusing on the improvement of octree coding, applying networks based on entropy models and continuous convolutions, and leveraging already decoded frames used to dynamic point clouds [216, 217]. These techniques provide limited information; deep CNN with masked convolutions is applied in these cases to improve the performance on sparse regions. However, in such methods, the sequential dependency increases the complexity, which can be addressed using multiscale approaches [218]. Signal point-based approaches such as PointNet for geometric compression, suggest improved schemes based on adaptive octree partitioning and clustering NN architectures, which integrate novel neural graph sampling modules, point-based NNs specifically designed for LiDAR point cloud compression, and deconvolution operators to compress point cloud geometry [219]. Encoding domain approaches are more suitable for dense and sparse point clouds since they depend on the number of points, unlike voxel approaches that depend on the voxel grid dimension (precision). In addition, prior information should be considered regarding the used data acquisition sensor. For example, LiDAR sensors use spinning mechanisms, and a spherical coordinate system is used to model the point cloud data, while data from camera arrays are represented as several 2D images, and RGB data are stored on a single 2D images. For this reason, compression tasks may vary depending on the data acquisition source.

6.5.2 Attribute-based methods

Compressing attributes in 3D point clouds involves compressing information such as colors or normal directions. A significant challenge in this category is the difficulty of dealing with the geometrical irregularity of point clouds, which is influenced by the attributes, making modeling and prediction challenging. Irregularity problems are resolved by assuming that a point cloud is a sampling of a 2D manifold with a 2D parameterization, and attributes are projected onto a 2D image [220]. Attributes can be compressed by either image compression methods or can be mapped directly on a voxel

grid. In some cases, geometry and attributes are compressed to define a voxel grid. To deal with the irregular geometry caused by attributes mapped onto a 2D plane, CNNs are used for compressing attributes on a voxel grid or directly to define convolutions on the points. In real-time applications, data is usually transmitted via restricted bandwidth networks, necessitating rate reduction through approaches focused on compression and data structure for mapping environments. Although real-time processing rates are efficient, many methods do not support incremental compression and local decompression, leading to increased computational costs as the actual time does not match the compression time. Research results have reported compression times of the order of hours [221].

6.5.3 Comparative discussion on image-based detection and tracking

When comparing point cloud compression methods with corresponding methods for 2D images, the compression system depends on both a compressor and a decompressor, considering that an image is a two-dimensional signal composed of binary numbers [220]. Traditional methods aimed to improve compression ratio and image quality, as well as the encoder structure. However, these frameworks contain multiple sub-modules and many parameters, which limit the optimization space. In contrast, DL methods rely on the network's characteristics, updating parameters through automatic feature extraction during learning, and extracting patterns from the input images by adjusting weights [222]. To improve compression performance and address challenges such as complex unknown correlations between pixels, progressively increasing compression ratio, and iteratively consuming execution time during encoding and decoding, unsupervised learning based on Auto Encoder (AE) networks have been used [223]. Problems such as image reconstruction and diversity of original images are tackled using CNNs, RNNs and GANs. Unlike CNNs and RNNs which complete the compression by extracting feature information, GANs generate artificial data and the network parameters are optimized [224]. Therefore, the image is reconstructed based on the image coding information. However, network computation and parameters can be large. For optimal performance, polymerization schemes are applied, and prediction generation is achieved by using the original image instead of the residual signal [225].

6.6 Completion

Point clouds suffer from missing points due to peculiarities of the scanned object, specular reflection, signal absorption caused by the surface material of the object, occlusions, as well as blind spots. Additionally, the limited stability of the

3D scanner may cause topology errors that influence the quality of point clouds, and thus, 3D models' reconstruction, extraction of local spatial information, and successive processing [23]. On top of these issues, denoising, smoothing, fusion, and computational costs, pose additional challenges in the completion task.

Traditional methods based on geometry and alignment, utilize the geometric attributes of the objects [226] and retrieve the complete structure from a database [227]. However, they often struggle to robustly generalize to complex 3D surfaces with large missing parts. Other methods focused on using a 3D voxel grid [228], but they faced limitations due to computational costs, which increase cubically with the shape resolution. Nowadays, to resolve such challenges, DL methods are applied, which can be divided into geometric, alignment and learning-based, as illustrated in Fig. 16.

6.6.1 Geometry-based methods

Methods based on geometry aim predicting hidden shape parts of objects directly from the seen shape parts using former geometric assumptions [229]. The shapes are reconstructed from partial input through interpolation techniques [230] like Laplacian smoothing [231] and Poisson surface reconstruction [232], without the need for external data. Other techniques focus on detecting consistencies in the structures of models and repeating them towards predicting missing data according to identified symmetry axes. Such methodologies are inferring missing data straight from the region of observation, providing notable outcomes. Yet, they require hand-crafted geometric consistencies that can be defined in advance for certain kind of models and are employed only for models that have a finite level of incompleteness [233].

6.6.2 Alignment-based methods

Methods based on the alignment either match the partial input and substitute it with a model from a database, or multiple input parts are fit and combined to obtain the full surface [234]. Alternative methodologies utilize synthesized models after deformation [235] or non-3D geometric primitives, like planes and quadrics instead of 3D shapes from the database [236]. Such methodologies can be applied to many different types of models and to variable levels of incompleteness. However, they are computationally demanding during construction of the database and inference optimization, and they exhibit sensitivity to noise.

6.6.3 Learning-based methods

Learning-based methods construct a parameterized model towards learning a mapping between the two feature spaces

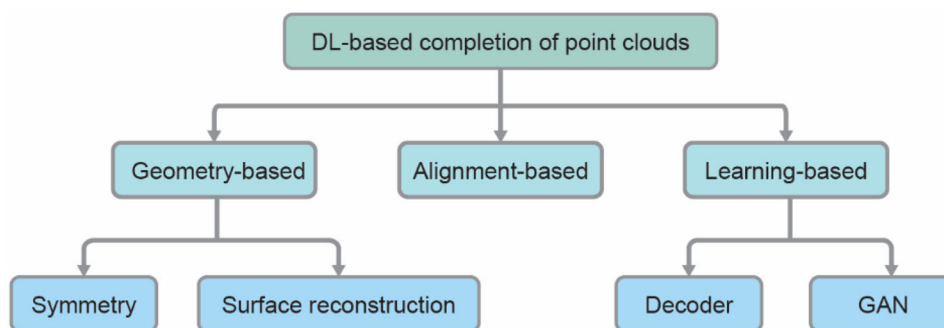
of the complete and incomplete point cloud, typically using encoder-decoder NNs. For the shape representation, most models use voxels, which are intuitive and appropriate for 3D convolution. To maintain local geometric details, models are employed on direct point sets. Due to the fact that points and voxels are mono-modality inputs, it is challenging to achieve accuracy in mapping between a complete and incomplete point cloud. Hence these methods are considered efficient only on small-scale incompleteness of shapes or objects. To resolve this, coarse-to-fine strategies are applied [237, 238] towards preserving the observed geometric details from the local features in incomplete inputs, and voxel methods are used for the completion process [239]. Other methods focus on unpaired shape completion techniques because the paired ground truth of real scans is difficult to obtain [240].

Concluding, the existing completion networks are still inferior to the maintenance of the details, especially for thin structures, where they cannot generalize sufficiently and it is difficult to extend to scene completion. Among CNNs for compression and completion tasks, there is a great parallelism between them, because models require meaningful features to be extracted from the input towards the original shape reconstruction.

Image completion involves completing missing regions in an image according to the available visual data, including eliminating unwanted noise and blur while preserving the biggest part of image details. Completion and image enhancement can occur either simultaneously or separately. Completion challenges such as the poor-quality images, mainly in real-world applications due to missing and masked image areas. The plausible of completions during the transition between known and unknown regions, the removal of unwanted objects and the generation of occluded regions for 3D reconstruction, are additional challenges, addressed by using GANs [241], while residual learning techniques [242] are used to address enhancement challenges. Despite their effectiveness, these methods may struggle to deal with big incomplete regions in images, especially when the corrupted regions are big or unrelated to the visual data. They might also fail in cases where the original image lacks sufficient data for completion or when the models are unsuitable for handling noisy images.

In real-time applications, several challenges persist when image samples are required to be included in the training data. Then, masks on the corrupted regions are necessary during training, and context encoders usually lead to blurry and noisy results [243]. Denoising approaches can only improve the picture's clarity but cannot fully recover it if the management of texture and structure is unsatisfactory [244]. Enhancement models applied to very smooth regions (e.g., clear sky) may be susceptible to the halo effect, which can be addressed by iteratively optimizing the pixel gradient in edge transitional regions [245]. The issue of generating plausible

Fig. 16 Categorization of DL-based completion tasks for point clouds



completion results during the transition between known and unknown region methods, is confronted with the following handling [246, 247]: building a contextual attention architecture, employing efficient loss functions for generating a more representative content in the incomplete region, using partial convolutions to focus on the unidentified region, and ensuring structural consistency to achieve continuity. Furthermore, many methods rely on GANs to achieve better details. However, GANs may present structure constraints since they completely rely on the training image for generating semantically relevant structures and texture confidence [248]. To tackle the removal of unwanted objects or the generation of occluded regions, patch-based image synthesis approaches were proposed [249]. These methods have limited effectiveness as they require high-level scene recognition, completed textured patterns, and object and scene anatomy understanding. Other methods focused on better controlling the completion behavior of networks and the computational costs, yet, without serious improvements in large regions [250].

6.6.4 Comparative discussion on image-based completion

Completion of 2D images using DL involves training models to fill in missing or damaged parts of an image, creating a visually coherent and realistic result. This process is particularly useful in applications where images are incomplete or have regions that need restoration, such as in image editing, restoration, or inpainting. Commonly used architectures include CNNs, which are well-suited for completion due to their capability to capture hierarchical features, GAN models to generate realistic and high-resolution completions, or AE where the input image is encoded, and missing parts of 2D images are completed. However, there are challenges such as the model's understanding of the context of the image, the striving for realistic and visually consistent completions that could blend smoothly with the existing image, and the need for computational resources, especially in the case of large images or complex architectures.

7 Point cloud datasets

In recent years, many point cloud datasets obtained from various sensors and comprising virtual or real scenes, both indoor and outdoor, have been published, catering to different tasks. This section provides an overview of the most popular datasets associated with various tasks. These datasets are created mainly by industries and university communities, and they play a crucial role in DL-based applications since they offer a substantial amount of ground truth labels for network training and serve as the basic benchmarks for comparing and evaluating different methods. However, the disposal of benchmark datasets diminishes as their complexity and size increase, mainly due to the diverse platforms and acquisition methods used to collect point cloud data [44].

DL approaches require large datasets, but the availability and size of data can vary [3]. Some datasets can be utilized for multiple tasks, while others are specific to particular applications. Datasets can be classified based on the types of data they provide. Classification datasets typically include textured surfaces that can be synthetically generated using CAD tools or actual datasets directly captured by 3D sensors like Kinect or LiDAR. Real data, though valuable, often have small sizes due to the challenges involved in obtaining them [251]; their acquisition can be expensive in terms of both money and time, and they may suffer from noise and occlusions. On the other hand, synthetic data are available in huge amounts without occlusions and background; however, they have limited generalization abilities [252]. For 3D object detection and tracking, datasets are derived mainly from indoor [253] and outdoor [254] scenes. The data can be obtained by converting dense depth information, sampling from 3D meshes, or designing them with separated objects spatially. Point cloud segmentation datasets come from various sensors, such as MLS, ALS, static TLS, RGB-D, and other 3D scanners, and developing robust algorithms is crucial in this context. Completion task datasets include both artificial and real-world data, and for them to be effective, the dataset needs to be rich and diverse. The creation of a big and efficient dataset requires significant manpower, material, and financial resources.

Table 3 Properties of indicative well-known 3D point cloud benchmark datasets

Dataset	Year	Type	Origin	Description	Reg	Segm	Class	Detect & Track	Compr	Compl
ApolloScape [255]	2018	Real	LiDAR	over 140,000 video frames, 35 semantic classes, 28 instance classes				✓		
Argoverse [32]	2019	Real	LiDAR	360-degree images from 7 cameras				✓		
Audi Autonomous Driving Dataset (A2D2) [255]	2020	Real	6 cameras + LiDAR	41,277 non-sequential video frames, 38 classes		✓		✓		
DBNet [256]	2018	Real	LiDAR	1,000 km driving data				✓		
iQmulus [257]	2015	Real	mobile laser scanning (MLS) LiDAR	300 M points, 50 classes, 10 scans (outdoor roadway level)		✓				
KITTI [258]	2012	Real	RGB + LiDAR	Outdoor, 15 K frames, 93 thousand depth maps, 22 scenes, 8 classes	✓	✓	✓	✓	✓	✓
Matterport3D [33]	2017	Real	RGBD	194,400 images	✓	✓	✓			
Nuage de Points et Modlisation 3D (NPM3D) [259]	2017	Real	LiDAR	1,431 M points data, 5 labeled point cloud classes		✓	✓			
nuScenes [253]	2020	Real	LiDAR	Outdoor, 1000 scenes, 31 classes, 10 annotated frames		✓		✓		
Oakland [260]	2009	Real	LiDAR	1.6 M points, 17 scans, 44 classes		✓				
Paris-Lille-3D [259]	2018	Real	MLS LiDAR	50 classes, 143 M (outdoor Roadway level)		✓	✓			
Paris-rue-Madame [261]	2017	Real	MLS LiDAR	143 M points, 50 classes (outdoor Roadway level)		✓				

Table 3 (continued)

Dataset	Year	Type	Origin	Description	Reg	Segm	Class	Detect & Track	Compr	Compl
ScanNet [251]	2017	Real	Occipital structure sensor	1,513 scenes, 21 categories (Indoor level)		✓	✓	✓	✓	
ScanObjectNN [262]	2019	Real	Structure sensor, CAD	2,902 samples, 15 indoor classes, 15,000 frames			✓			
SceneNN [263]	2016	Real	RGB-D	Over 100 indoor scenes		✓				
Semantic3D [264]	2017	Real	Terrestrial laser scanning (TLS) LiDAR	8 classes, 3 billion points		✓				
SemanticKITTI [265]	2019	Real	MLS LiDAR	28 classes, 43,000 labels, 4,549 M points (outdoor roadway level)		✓			✓	
SUN3D [266]	2015	Real	RGB-D	700 classes, 272 scans, 41 objects		✓		✓		
Sydney Urban Objects [267]	2013	Real	LiDAR	2.3 M manual points and 26 classes			✓			
3DMatch [268]	2017	Synthetic	RGB-D	62 indoor scenes from existing data	✓					
Apollo [34]	2018	Synthetic/Real	CAD/RGB + LiDAR	5,277 images				✓		
Completion3D [269]	2019	Synthetic	CAD	8 classes, 30,958 models, partial and ground truth point clouds with 2,048 points each						✓
Multisensor Indoor Mapping and Positioning (MIMAP) [270]	2020	Synthetic	MLS LiDAR	dense laser scanning point cloud for indoor mapping and positioning			✓			
Multi-View Partial (MVP) [271]	2021	Synthetic	CAD	16 classes, over 100,000 high-quality scans						✓
New York University Depth Dataset v2 (NYUDv2) [272]	2013	Synthetic	RGB-D	Indoor, 40 classes		✓				✓

Table 3 (continued)

Dataset	Year	Type	Origin	Description	Reg	Segm	Class	Detect & Track	Compr	Compl
PartNet [273]	2019	Synthetic	CAD	Indoor, 570 k part instances, 24 classes (object level)		✓				
ShapeNet [274]	2015	Synthetic	CAD	300 M models, 53 classes, 51,190 samples (object level)		✓	✓			✓
Stanford 2D-3D-Semantics [275]	2017	Synthetic	RGB + Matterport 3D camera	over 70,496 images, 6 large scale indoor areas covering 271 rooms, 13 classes		✓				
Stanford Large-scale 3D Indoor Spaces Dataset (S3DIS) [276]	2017	Synthetic	RGB + Matterport 3D camera	272 scenes, 13 classes, 273 M points (indoor scene level)		✓		✓		
SynLiDAR [277]	2021	Synthetic	LiDAR	13 sequences of LiDAR point cloud 20 k scans (over 19 billion points and 32 classes)		✓				
SynthCity [278]	2019	Synthetic	MLS LiDAR	367.9 M points, 9 classes		✓				
Wuhan University TLS [279]	2018	Synthetic	TLS LiDAR	Outdoor, comprises 115 scans, over 1,740 million 3D points, 11 classes	✓					
ModelNet 40 [178]	2015	Synthetic	CAD	40 classes, 12,311 models			✓			✓
McGill Benchmark [280]	2008	Synthetic	CAD	456 samples of 19 classes			✓			
Sydney Urban Objects [267]	2013	Real	LiDAR	588 samples of 4 classes			✓			
ModelNet10 [178]	2015	Synthetic	CAD	4,899 samples of 10 classes			✓			
SUN RGB-D [252]	2015	Real	RGB-D	Indoor, 47 scenes of 32 classes, 5 K annotated frames and 65 K 3D boxes				✓		

Table 3 (continued)

Dataset	Year	Type	Origin	Description	Reg	Segm	Class	Detect & Track	Compr	Compl
ScanNetV2 [251]	2018	Real	RGB-D	Includes mesh sensors, 1.5 K scenes of 28 classes of annotated voxelized objects				✓		
H3D [281]	2019	Real	LiDAR	Includes RBD sensory data, 160 scenes and 8 classes, 27 K frames of urban driving				✓		
Lyft L5 [282]	2019	Real	LiDAR	Includes RBD sensory data, 366 scenes and 9 classes, 46 K annotated frames of urban driving				✓		
A*3D [283]	2019	Real	LiDAR	Includes RBD sensory data of 7 classes, 39 K frames, 230 K boxes of urban driving				✓		
Waymo Open [284]	2020	Real	LiDAR	Includes RBD sensory data, 1 K scenes of 4 classes, 200 K frames, 12 M boxes of urban driving				✓		
ISPRS [285]	2012	Real	ALS LiDAR	1.2 M points, 9 classes		✓				
Toronto 3D [286]	2020	Real	MLS LiDAR	78.3 M points, 8 and 9 classes, 4 scans		✓				
DALES [287]	2020	Real	ALS LiDAR	550 M points, 8 and 9 classes, 40 scans		✓				
ONCE [288]	2021	Real	LiDAR	230 K scenes, 1 M images, 12 M 3D boxes, 4 classes for autonomous driving			✓			

Table 3 (continued)

Dataset	Year	Type	Origin	Description	Reg	Segm	Class	Detect & Track	Compr	Compl
MVPNet [289]	2023	Real	RGB camera (multi-view images)	3D object cloud dataset of 87,200 samples from 150 categories with class labels derived by dense reconstruction on MVIImgNet			✓			

Table 3 indicatively summarizes well-known benchmark datasets, providing information about the type of point cloud data, their origin, a brief description, and the application tasks for which they have been used, based on the reviewed literature.

8 Discussion and future directions

In this section, the questions that guided the research can be answered, valuable conclusions can be drawn, and future research directions can be suggested. Therefore, based on the conducted research, it can be concluded that point clouds are used in a wide variety of applications in industry and academia, whenever an accurate 3D representation of objects or surroundings is essential. Such applications include 3D modeling, mapping, robotics and autonomous systems, scene reconstruction, medical imaging, virtual and augmented reality, gaming applications, and more (answer to RQ4: *In which applications does it make sense to apply point clouds?*).

Despite the rapid evolution of point cloud acquisition technologies, several important issues continue to emerge. These issues are mainly related to point clouds' nature, such as being unordered, unorganized, irregular, and sparse. In a point cloud, there is no explicit connectivity between the points, as each point is independently scanned. Consequently, the distance between adjacent points is not always stable resulting in potential information loss in the object's representation. Another challenge lies in efficiently storing large point clouds in a permutation-invariant file format. LiDAR technology, in particular, can generate massive file sizes, posing storage and processing problems. Furthermore, the data quality directly depends on the quality of the sensors used; higher quality data often requires more expensive sensors. In general, the resolution of point clouds is directly depending on the quality of acquisition sensors. Resolution implies

the level of detail and precision of representation of a 3D scene's geometry. Therefore, different sensors can affect the point cloud resolution, based on their specifications and their manufacturing technology (answer to RQ5: *To what extent do different sensors affect the point cloud resolution?*). LiDAR resolution can be affected by the laser pulse density, the angular resolution and the scanning range. SL sensors are affected by the pattern's complexity, while the quality of ToF sensors depends on the capturing frame rate, the high accuracy and low noise levels, towards contributing to precise depth measurements.

Same for RGB-D cameras, the depth sensor resolution is the most important factor in providing higher-resolution point clouds. Moreover, the proper calibration of sensors can improve the quality of captured data. It should be noted, however, that while the quality of sensors holds a substantial role, other factors may influence the point cloud resolution. Such factors are the ambient light and the reflectance of surfaces, that can lead to inaccurate depth measurements. Therefore, the selection of the proper sensor should be based on the requirements of the application, so as the captured point cloud to be characterized by sufficient precision and the desired level of detail. Post-processing tasks such as denoising, filtering, registration, etc., can also enhance the overall captured quality of a point cloud.

In addition to data-related challenges, issues also arise in point cloud processing and analysis. Developing user interfaces able to manage and visualize complex 3D data is a challenge, as existing interfaces are typically designed for 2D data and may not be well-suited for handling point clouds effectively. Adjustments are considered necessary since the use of 3D point clouds could provide substantial benefit over the use of 2D images in applications where a more accurate representation of surroundings is crucial (answer to RQ6: *Under what conditions does the use of point clouds provide benefits against 2D images?*). The ability of point clouds

to capture accurate depth information, spatial geometry of objects and precise measurements of the surrounding environment, is beneficial to a multitude of applications where a simple 3D representation of objects is not enough; true 3D representations are essential for in-depth analysis of the environment towards the transition to actions, for autonomous interaction of systems and reliable decision-making.

The implementation of suitable algorithms for point cloud processing is crucial for achieving effective results. Nowadays, DL models have demonstrated remarkable performance in various tasks, thanks to the convolution operation, although conventionally they perform on regular, structured, and ordered data, such as 2D images. DL models can learn automatically more distinct and robust feature representations that contain particularly symmetric and repetitive elements, poor geometric features, and limited overlaps.

Based on the up-to-date adaptation of DL on 3D point clouds, the following conclusions can be drawn regarding the issues that emerge from the point cloud data (answer to RQ1: *What are the challenges regarding point cloud data processing?*):

- i. Sensors inherently contain various types of noise, which can lead to disturbances and outliers in the point cloud data.
- ii. Point density in point clouds can be highly diversified, presenting challenges in data processing and analysis.
- iii. Reflective intensity varies depending on the distance between the target and LiDAR sensors, affecting the quality of the captured data.
- iv. Incompleteness in point clouds may occur due to occlusion between the target object and cluttered background, leading to confusion of categories in tasks such as segmentation.
- v. Handling big data in point clouds requires intense processing and can result in significant computational costs.

These challenges underscore the importance of developing robust DL-based computer vision algorithms and methodologies that can effectively handle the complexities and irregularities in 3D point cloud data, enabling more accurate and efficient processing and analysis.

Regarding DL models applied to 3D point clouds, the following issues are noted (answer to RQ2: *What are the challenges that DL models face with 3D point cloud data?*):

- i. *Permutation and orientation invariants*: DL models need to handle the unordered and unoriented nature of point cloud data, which presents challenges in establishing consistent correspondences between points and ensuring robustness to different point cloud representations.

- ii. *Rigid transformation challenges*: Point clouds may undergo rigid transformations, such as translation and rotation, which can affect the performance of DL models. Addressing this issue requires developing models that can effectively handle and generalize to various transformations.
- iii. *Accuracy dependence on data quality and scene variation*: The accuracy of DL models for point clouds is strongly influenced by the quality of the input data and the variability of scenes. High-quality data and diverse scene representations are essential to achieve robust and reliable performance.

Overcoming these challenges is vital to harness the full potential of DL models for point cloud processing and analysis, enabling a wide range of applications in various fields.

Although, there are public standard benchmark datasets for various tasks, which have proven to be effective in evaluating the performance of DL models on point cloud processing, several issues regarding datasets need to be highlighted (answer to RQ3: *What is the status of 3D point cloud datasets for DL-based applications?*):

- i. *Shift towards Point Cloud Semantic Segmentation (PCSS)*: Since 2009, many datasets have been labeled for Point Cloud Semantic Segmentation (PCSS) rather than Point Cloud Classification (PCS), limiting the availability of diverse and comprehensive labeled datasets for classification tasks.
- ii. *Limitations of certain datasets*: Some datasets, such as Oakland outdoor MLS dataset, the Sydney Urban Objects MLS dataset, the Paris-rue-Madame MLS dataset, and the IQmulus MLS dataset, may not provide sufficient object representations and labeled points for certain tasks.
- iii. *Challenges in labeling*: Datasets like KITTI and NYUv2 contain more objects and points but may not directly provide labeled point clouds, requiring additional processing for certain applications.
- iv. *Diverse measurement ranges and scenes*: Datasets like Wuhan University TLS (Whu-TLS) cover a wide range of scenes with variations in environmental and geometric shapes, but overlapping and low-density adjacent point clouds pose challenges.
- v. *Weak geometric features in repetitive structures*: Some datasets contain particularly symmetric and repetitive elements, leading to weak geometric features, especially in scenes with periodic changes due to moving objects.
- vi. *Symmetric structures and mirror reflections*: The presence of symmetric structures and mirror reflections, including virtual points, further complicates point cloud analysis.

Addressing these dataset-related issues will be crucial for advancing the development and evaluation of DL models in point cloud processing and analysis tasks. Ensuring diverse, comprehensive, and well-labeled datasets will aid in enhancing the accuracy and generalization capabilities of DL models in handling different point cloud scenarios.

In order to unlock the potential of point cloud data, efficient processing solutions are essential. Artificial intelligence (AI) and automation can play a significant role in simplifying tasks and improving overall processing. Novel algorithms tailored for specific point cloud tasks and exploration of cloud-computing processing are crucial for enhancing efficiency. Cloud computing offers the advantage of simultaneous processing of a vast number of scans, thus accelerating various algorithms.

Considering the challenges faced in point cloud processing, future research can be directed towards the following four key areas. Table 4 summarizes the four defined research areas and the potential research suggestions as derived from this review.

8.1 Advancements in 3D technologies for point cloud generation

- 3D point cloud generation technologies present some technical limitations in the operation of sensors. For this purpose, there is a crucial need for the development of new sensory mechanisms, particularly for LiDARs, to enhance radiation sources and integrate multi-sensory systems for efficient data fusion.
- There is a significant demand to cover large-scale scenes, yet point clouds contain a global fine-scale. To effectively handle the big data of 3D point clouds, a promising approach would involve the integration of various technologies such as edge computing, artificial intelligence, and deep learning. This combination can lead to more efficient processing and analysis of point cloud data at scale.

8.2 Point cloud data management

- Due to the irregular and disordered nature of the point clouds, many methods initially voxelize them before further processing. However, in the cases of multimodal data from sources like LiDAR, RGB-D camera or Radar, voxelization can lead to information loss and increased computational complexity, especially in complex scenes. Therefore, the development of attention models that focus on efficient feature extraction and fusion would be highly beneficial.
- As point clouds continue to grow in size with the advancement of 3D technologies, the need for larger storage space

becomes increasingly critical. Hence, a future direction should focus on the development of updated mechanisms that efficiently support the handling of 3D point cloud big data.

- The utilization of cloud capabilities for point cloud processing is essential to consider. Cloud computing can offer advantages such as 5G access capabilities and dynamic parallelization mechanisms, which can significantly enhance the processing speed and efficiency of point clouds. Leveraging cloud resources for point cloud analysis can enable faster and more scalable computations, making it possible to handle large datasets and complex tasks with greater ease and effectiveness.
- Since point clouds and voxels are often sparse and irregular, sparse convolutions are commonly applied. However, these algorithms can lead to higher GPU memory consumption, posing challenges for efficient execution. Therefore, there is a need to emphasize the design of more efficient hardware, including RAM and processors, to better handle the demands of sparse convolutions. Additionally, the software should be optimized to match the capabilities of the hardware, ensuring that the system can take full advantage of its resources for faster and more effective point cloud processing. By optimizing both hardware and software, more efficient and powerful solutions can be developed for handling point cloud data.

8.3 Deep learning models for point clouds

- The development of DL networks should focus on object-oriented point cloud big data, leveraging the power of artificial intelligence to transform point clouds directly for object classification and boundary extraction. This approach aims to move away from dealing with point clouds point-by-point, enabling more efficient and accurate processing of complex 3D data.
- The development of robust techniques is crucial to provide enhanced precision, fast processing speed, and guaranteed accuracy. These techniques should be capable of handling the challenges posed by point cloud data, such as noise, occlusions, and irregularity while ensuring high levels of accuracy in various tasks.
- The development of DL models capable of supporting different types of data, and providing adequate generalization is essential. Efforts should be made to design more competent and efficient pre-trained DL models specifically tailored for real-life point cloud scenes. Additionally, exploring various backbones and architectures could facilitate knowledge transfer between different datasets and tasks. As a future research direction, it would be beneficial to focus on developing a unified backbone that can be

Table 4 Defined research areas on point clouds and potential future research directions

Research areas			
3D technologies for point cloud generation	Point cloud data management	Deep learning models for point clouds	Point cloud datasets
Development of new sensory mechanisms	Attention models for feature extraction and fusion	Object-oriented point cloud big data models	Realistic diverse, and annotated datasets
Integration of multi-sensory systems for efficient data fusion	Updated mechanisms to support 3D point cloud big data	Development of robust techniques	Synthetic realistic datasets using GAN
Integration of various technologies (edge computing, AI, DL)	Leveraging cloud resources for point cloud processing	Models capable of supporting different types of data	Inclusion of metadata
	Design of more efficient hardware	Efficient unified backbone for point cloud processing tasks	High-quality pre-training datasets of object- and scene-level data
	Optimized software to match the capabilities of the hardware	More accurate registration algorithms for multitemporal point clouds	Unified evaluation standard for indoor scenes
		Interpretable algorithms	Ground truth information datasets
		Efficient metrics to consider perception and downstream tasks	
		Self-supervised and transfer learning	
		Models to handle noise, missing points, occlusions	
		Few-shot and zero-shot learning techniques	
		Real-time processing DL architectures	
		Integration with other modalities for complementary information to DL models	
		Privacy-preserving techniques	

adapted and fine-tuned for various point cloud processing tasks, ultimately enhancing the versatility and effectiveness of DL-based approaches.

- The development of more accurate registration algorithms, particularly for multitemporal point clouds, is crucial for real-life applications that require ground truth accuracy. Future research should focus on creating registration techniques capable of effectively combining multiple data sets, even when acquired from different sensory devices. This approach would ensure robust and reliable registration for

various applications involving point clouds from diverse sources.

- Since DL object detection models often suffer from challenges related to interpretability, such as occlusion and noise, their general behavior is often characterized as a black box. In future research, it would be essential to focus on designing DL models that offer stronger interpretability, particularly for applications where understanding the model's decision-making process is critical. By enhancing interpretability, we can gain better insights into how

these models function and improve their reliability and trustworthiness in real-world scenarios.

- In applications like autonomous driving, the development of efficient metrics necessitates models capable of considering additional parameters, such as perception and downstream tasks. These two aspects are closely correlated, and incorporating them into the evaluation metrics would provide a more comprehensive understanding of the overall system's performance. By capturing the interplay between perception and downstream tasks, we can better assess the effectiveness of the models and ensure their suitability for real-world applications.
- Self-supervised learning, generative models, and transfer learning should be enhanced towards advancing this field.
- One important future direction is the improvement of the robustness of deep learning models to noise and incomplete data in point clouds. Point clouds obtained from real-world sensors often contain noise, missing points, or occlusions. Developing techniques that can handle such challenges and effectively utilize imperfect data is crucial for real-world applications.
- Exploration of few-shot and zero-shot learning techniques for point clouds, that can train models to generalize to new object categories or tasks with limited or no labeled data. Developing methods that can effectively leverage prior knowledge, as well as transfer learning, to new scenarios can greatly enhance the applicability of deep learning models for point clouds.
- Real-time processing of point clouds is essential for many applications, such as autonomous driving. Future research can focus on developing efficient deep learning architectures and algorithms that can handle large-scale point clouds in real-time.
- Point clouds are often captured along with other sensor modalities, such as images, depth maps, or semantic information. Integration with other modalities can provide complementary information towards improving the performance of deep learning models.
- Point cloud data are nowadays established in various applications, thus, privacy and security preservation become crucial, directing research to privacy-preserving techniques for deep learning on point clouds.

8.4 Point cloud datasets

- Further evolution of sensors to generate data based on realistic environments, such as the SynthCity dataset, is a promising future direction. Realistic and diverse datasets, and more specifically annotated datasets, are crucial for training and evaluating DL models, especially in complex scenarios. By creating more sophisticated datasets that

closely resemble real-world environments, we can push the boundaries of point cloud processing and drive advancements in the field.

- Synthetic realistic datasets using GAN would be useful. In addition, some datasets are used only in specific tasks, such as downstream tasks at the object level for synthetic objects (ModelNet40), real object classification (ScanObjectNN), few-shot classification (Few-shot ModelNet40), and synthetic object segmentation (ShapeNet), and it is difficult to generalize to other datasets. Hence, it would be useful to enhance data acquisition to help the pre-trained models transfer to real scenes.
- Another important direction would be to focus on the inclusion of metadata, such as class labels, timestamps, and geospatial coordinates, in addition to other information like color, intensity, and multispectral bands. Integrating such contextual data with point clouds can provide valuable insights and enhance the performance of DL models for various tasks.
- The sheer volume and diversity of point cloud data pose significant challenges for various tasks. To address this, there is a need to develop high-quality pre-training datasets, focusing on both object-level and scene-level data. These datasets can serve as a foundation for training DL models with a better understanding of the underlying structures and patterns in point clouds.
- When it comes to indoor scene detection and segmentation, there is a lack of a unified model and common framework. Similarly, for outdoor scenes, the inherent diversity in scenes and weather conditions makes it challenging to establish a standardized evaluation method that can provide unbiased outcomes, unlike the more controlled environments of indoor scenes. Therefore, it is essential to focus on developing a unified evaluation standard specifically tailored to indoor scenes to facilitate fair and objective comparisons between different models and methods. Such an evaluation standard will help researchers and practitioners make better-informed decisions and advancements in point cloud processing techniques for indoor scenes.
- One of the most significant challenges in DL-based monocular depth estimation is the scarcity of datasets with ground truth information, and obtaining such datasets can be a costly endeavor.

9 Conclusions

This comprehensive study aims to provide a detailed exploration of DL-based computer vision methods applied to point

clouds. The review encompasses various aspects, including point cloud acquisition technologies, and computer vision tasks involving registration, segmentation, classification, detection, completion, and compression. Additionally, it compares traditional methods with DL approaches, explores the differences between 3D point clouds and other modalities like 2D and depth images, and evaluates well-known benchmark datasets for different tasks.

Moreover, this work delves into the challenges arising from the advancement of 3D technologies, providing a comprehensive understanding of the obstacles faced in the field. Through an in-depth investigation of these challenges, it identifies key areas for future research, thereby highlighting trends, research gaps, and valuable insights to guide further studies effectively. Comparison of experimental results of DL methods for all tasks will be considered as future work.

Acknowledgment This work was supported by the MPhil program “Advanced Technologies in Informatics and Computers”, hosted by the Department of Computer Science, International Hellenic University, Kavala, Greece.

Author contributions All authors contributed to the study’s conception and design. Material preparation, data collection and analysis were performed by KAT and EV. The first draft of the manuscript was written by KAT, EV and GAP, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by HEAL-Link Greece. No funding was received for conducting this study.

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

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

References

- Wiley, V., Lucas, T.: Computer vision and image processing: a paper review. *Int. J. Artif. Intell. Res.* **2**, 22 (2018)
- Qian, R., Lai, X., Li, X.: 3D object detection for autonomous driving: a survey. *Pattern Recognit.* **130**, 108796 (2022). <https://doi.org/10.1016/j.patcog.2022.108796>
- Griffiths, D., Boehm, J.: A review on deep learning techniques for 3D sensed data classification. *Remote Sens.* **11**, 1499 (2019). <https://doi.org/10.3390/rs11121499>
- Cao, K., Xu, Y., Cosman, P.C. (2018) Patch-aware averaging filter for scaling in point cloud compression. In: 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP). pp. 390–394. IEEE
- Bi, S., Yuan, C., Liu, C., Cheng, J., Wang, W., Cai, Y.: A survey of low-cost 3D laser scanning technology. *Appl. Sci.* **11**, 3938 (2021). <https://doi.org/10.3390/app11093938>
- Tychola, K.A., Tsimperidis, I., Papakostas, G.A.: On 3D reconstruction using RGB-D cameras. *Digital.* **2**, 401–421 (2022). <https://doi.org/10.3390/digital2030022>
- Kingsland, K.: Comparative analysis of digital photogrammetry software for cultural heritage. *Digit. Appl. Archaeol. Cult. Herit.* **18**, e00157 (2020). <https://doi.org/10.1016/j.daach.2020.e00157>
- Kamnik, R., Nekrep Perc, M., Topolšek, D.: Using the scanners and drone for comparison of point cloud accuracy at traffic accident analysis. *Accid. Anal. Prev.* **135**, 105391 (2020). <https://doi.org/10.1016/j.aap.2019.105391>
- Tian, Y., Chen, L., Song, W., Sung, Y., Woo, S.: DGCB-Net: dynamic graph convolutional broad network for 3D object recognition in point cloud. *Remote Sens.* **13**, 66 (2020). <https://doi.org/10.3390/rs13010066>
- He, Y., Huang, H., Fan, H., Chen, Q., & Sun, J. (2021). FFB6D: A full flow bidirectional fusion network for 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3003-3013). <https://doi.org/10.48550/ARXIV.2103.02242>
- Debeunne, C., Vivet, D.: A review of visual-LiDAR fusion based simultaneous localization and mapping. *Sensors.* **20**, 2068 (2020). <https://doi.org/10.3390/s20072068>
- Alaba, S.Y., Ball, J.E.: A survey on deep-learning-based LiDAR 3D object detection for autonomous driving. *Sensors.* **22**, 9577 (2022). <https://doi.org/10.3390/s22249577>
- Stilla, U., Xu, Y.: Change detection of urban objects using 3D point clouds: a review. *ISPRS J. Photogramm. Remote Sens.* **197**, 228–255 (2023). <https://doi.org/10.1016/j.isprsjprs.2023.01.010>
- You, Y., Cao, J., Zhou, W.: A survey of change detection methods based on remote sensing images for multi-source and multi-objective scenarios. *Remote Sens.* **12**, 2460 (2020). <https://doi.org/10.3390/rs12152460>
- Hansen, L., Heinrich, M.P. (2021). Deep learning based geometric registration for medical images: How accurate can we get without visual features?. In: Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27 (pp. 18-30). Springer International Publishing. <https://doi.org/10.48550/ARXIV.2103.00885>
- Acar, H., Karsli, F., Ozturk, M., Dihkan, M.: Automatic detection of building roofs from point clouds produced by the dense image matching technique. *Int. J. Remote Sens.* **40**, 138–155 (2019). <https://doi.org/10.1080/01431161.2018.1508915>
- Bucksch, A., Lindenbergh, R., Menenti, M.: SkelTre. *Vis. Comput.* **26**, 1283–1300 (2010). <https://doi.org/10.1007/s00371-010-0520-4>
- Liu, L., He, J., Ren, K., Xiao, Z., Hou, Y.: A LiDAR-camera fusion 3D object detection algorithm. *Information* **13**, 169 (2022). <https://doi.org/10.3390/info13040169>
- Liu, S., Zhang, M., Kadam, P., Kuo, C.-C.J.: Introduction. In: 3D Point Cloud Analysis. pp. 1–13. Springer International Publishing, Cham (2021)
- Zhang, J., Zhao, X., Chen, Z., Lu, Z.: A review of deep learning-based semantic segmentation for the point cloud. *IEEE Access.* **7**, 179118–179133 (2019). <https://doi.org/10.1109/ACCESS.2019.2958671>
- Wu, Y., Wang, Y., Zhang, S., Ogai, H.: Deep 3D object detection networks using LiDAR data: a Review. *IEEE Sens. J.* **21**, 1152–1171 (2021). <https://doi.org/10.1109/JSEN.2020.3020626>
- Peng, C., Yang, M., Zheng, Q., Zhang, J., Wang, D., Yan, R., Wang, J., Li, B.: A triple-thresholds pavement crack detection method leveraging random structured forest. *Constr. Build. Mater.*

- 263, 120080 (2020). <https://doi.org/10.1016/j.conbuildmat.2020.120080>
23. Fei, B., Yang, W., Chen, W.-M., Li, Z., Li, Y., Ma, T., Hu, X., Ma, L.: Comprehensive review of deep learning-based 3D point cloud completion processing and analysis. *IEEE Trans. Intell. Transp. Syst.* **23**, 22862–22883 (2022). <https://doi.org/10.1109/TITS.2022.3195555>
 24. Cao, C., Preda, M., Zaharia, T.: 3D point cloud compression. In: *The 24th International Conference on 3D Web Technology*. pp. 1–9. ACM, New York, NY, USA (2019)
 25. Golla, T., Klein, R.: Real-time point cloud compression. In: *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 5087–5092. IEEE (2015)
 26. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 652–660). (2016). <https://doi.org/10.48550/ARXIV.1612.00593>
 27. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: deep hierarchical feature learning on point sets in a metric space, <http://arxiv.org/abs/1706.02413>, (2017)
 28. Wang, F., Zhuang, Y., Gu, H., Hu, H.: Automatic generation of synthetic LiDAR point clouds for 3-D data analysis. *IEEE Trans. Instrum. Meas.* **68**, 2671–2673 (2019). <https://doi.org/10.1109/TIM.2019.2906416>
 29. Fang, J., Zhou, D., Yan, F., Zhao, T., Zhang, F., Ma, Y., Wang, L., Yang, R.: Augmented LiDAR simulator for autonomous driving. *IEEE Robot. Autom. Lett.* **5**, 1931–1938 (2020). <https://doi.org/10.1109/LRA.2020.2969927>
 30. Manivasagam, S., Wang, S., Wong, K., Zeng, W., Sazanovich, M., Tan, S., Yang, B., Ma, W.-C., Urtasun, R.: LiDARsim: Realistic LiDAR simulation by leveraging the real world. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11167–11176). (2020). <https://doi.org/10.48550/ARXIV.2006.09348>
 31. Wang, C., Ning, X., Li, W., Bai, X., Gao, X.: 3D Person re-identification based on global semantic guidance and local feature aggregation. *IEEE Trans. Circuits Syst. Video Technol.* (2023). <https://doi.org/10.1109/TCSVT.2023.3328712>
 32. Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J.: Argoverse: 3D tracking and forecasting with rich maps. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8748–8757). (2019). <https://doi.org/10.48550/ARXIV.1911.02620>
 33. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: learning from RGB-D data in indoor environments, <http://arxiv.org/abs/1709.06158>, (2017)
 34. Song, X., Wang, P., Zhou, D., Zhu, R., Guan, C., Dai, Y., Su, H., Li, H., Yang, R.: ApolloCar3D: a large 3D car instance understanding benchmark for autonomous driving. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5452–5462). (2018). <https://doi.org/10.48550/ARXIV.1811.12222>
 35. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3D point clouds: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 4338–4364 (2021). <https://doi.org/10.1109/TPAMI.2020.3005434>
 36. Wang, R., Peethambaran, J., Chen, D.: LiDAR point clouds to 3-D urban models: a review. *IEEE J Sel Top. Appl. Earth Obs. Remote Sens.* **11**(606), 627 (2018)
 37. Malleon, C., Guillemot, J.-Y., Hilton, A.: 3D reconstruction from RGB-D data. In: *Rosin, P.L., Lai, Y.-K., Shao, L., Liu, Y. (eds.) RGB-D Image Analysis and Processing*, pp. 87–115. Springer International Publishing, Cham (2019)
 38. Bamler, R., Eineder, M., Adam, N., Zhu, X., Gernhardt, S.: Interferometric potential of high resolution spaceborne SAR. *Photogramm. - Fernerkundung - Geoinf.* **2009**, 407–419 (2009). <https://doi.org/10.1127/1432-8364/2009/0029>
 39. Ahmed, E., Saint, A., Shabayek, A.E.R., Cherenkova, K., Das, R., Gusev, G., Aouada, D., Ottersten, B.: A survey on deep learning advances on different 3D data representations, <http://arxiv.org/abs/1808.01462>, (2019)
 40. Liu, W., Sun, J., Li, W., Hu, T., Wang, P.: Deep learning on point clouds and its application: A Survey. *Sensors.* **19**, 4188 (2019). <https://doi.org/10.3390/s19194188>
 41. Vinodkumar, P.K., Karabulut, D., Avots, E., Ozcinar, C., Anbarjafari, G.: A survey on deep learning based segmentation, detection and classification for 3D point clouds. *Entropy* **25**, 635 (2023). <https://doi.org/10.3390/e25040635>
 42. Ioannidou, A., Chatzilari, E., Nikolopoulos, S., Kompatsiaris, I.: Deep learning advances in computer vision with 3D data: a survey. *ACM Comput. Surv.* **50**, 1–38 (2018). <https://doi.org/10.1145/3042064>
 43. Camuffo, E., Mari, D., Milani, S.: Recent advancements in learning algorithms for point clouds: an updated overview. *Sensors.* **22**, 1357 (2022). <https://doi.org/10.3390/s22041357>
 44. Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J.: Review: deep learning on 3D point clouds. *Remote Sens.* **12**, 1729 (2020). <https://doi.org/10.3390/rs12111729>
 45. Xiao, A., Huang, J., Guan, D., Zhang, X., Lu, S., Shao, L.: Unsupervised point cloud representation learning with deep neural networks: A Survey. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5452–5462). (2023). <https://doi.org/10.1109/TPAMI.2023.3262786>
 46. Zhang, Z., Dai, Y., Sun, J.: Deep learning based point cloud registration: an overview. *Virtual Real. Intell. Hardw.* **2**, 222–246 (2020). <https://doi.org/10.1016/j.vrih.2020.05.002>
 47. Zhang, H., Wang, C., Tian, S., Lu, B., Zhang, L., Ning, X., Bai, X.: Deep learning-based 3D point cloud classification: a systematic survey and outlook. *Displays* **79**, 102456 (2023). <https://doi.org/10.1016/j.displa.2023.102456>
 48. Hooda, R., Pan, W.D., Syed, T.M.: A Survey on 3D point cloud compression using machine learning approaches. In: *Southeast Con 2022*. pp. 522–529. IEEE (2022)
 49. Xiao, A., Zhang, X., Shao, L., Lu, S.: A survey of label-efficient deep learning for 3D point clouds. *arXiv*. 2305.19812, (2023)
 50. Li, Z., Xiang, N., Chen, H., Zhang, J., Yang, X.: Deep learning for scene flow estimation on point clouds: a survey and prospective trends. *Comput. Graph. Forum.* (2023). <https://doi.org/10.1111/cgf.14795>
 51. Grilli, E., Menna, F., Remondino, F.: A review of point clouds segmentation and classification algorithms. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-2/W3*, 339–344 (2017). <https://doi.org/10.5194/isprs-archives-XLII-2-W3-339-2017>
 52. Kitchenham, B.: *Procedures for performing systematic reviews*. UK, Keele University, Keele (2004)
 53. Wang, X., Pan, H., Guo, K., Yang, X., Luo, S.: The evolution of LiDAR and its application in high precision measurement. *IOP Conf. Ser. Earth Environ. Sci.* **502**, 12008 (2020). <https://doi.org/10.1088/1755-1315/502/1/012008>
 54. Yue, X., Wu, B., Seshia, S.A., Keutzer, K., Sangiovanni-Vincentelli, A.L.: A LiDAR point cloud generator: from a virtual world to autonomous driving. In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval* (pp. 458–464). (2018). <https://doi.org/10.48550/ARXIV.1804.00103>
 55. Li, Y., Ibanez-Guzman, J.: Lidar for autonomous driving: the principles, challenges, and trends for automotive Lidar and perception systems. *IEEE Signal Process. Mag.* **37**, 50–61 (2020). <https://doi.org/10.1109/MSP.2020.2973615>

56. Kurdi, F.T., Gharineiat, Z., Campbell, G., Dey, E.K., Awrangjeb, M.: Full series algorithm of automatic building extraction and modelling from LiDAR data. In: 2021 Digital Image Computing: Techniques and Applications (DICTA). pp. 1–8. IEEE (2021)
57. Zollhöfer, M.: Commodity RGB-D sensors: data acquisition. In: Rosin, P.L., Lai, Y.-K., Shao, L., Liu, Y. (eds.) RGB-D image analysis and processing, pp. 3–13. Springer International Publishing, Cham (2019)
58. Alexandrov, S. V., Prankl, J., Zillich, M., Vincze, M.: Calibration and correction of vignetting effects with an application to 3D mapping. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4217–4223. IEEE (2016)
59. Niemiropo, T.T., Viitanen, M., Vanne, J.: Open3DGen: open-source software for reconstructing textured 3D models from RGB-D images. In: MMSys '21: 12th ACM Multimedia Systems Conference. pp. 12–22. ACM (2021)
60. Zollhöfer, M., Stotko, P., Görnitz, A., Theobalt, C., Nießner, M., Klein, R., Kolb, A.: State of the art on 3D reconstruction with RGB-D cameras. *Comput. Graph. Forum.* **37**, 625–652 (2018). <https://doi.org/10.1111/cgf.13386>
61. Zhao, C., Sun, Q., Zhang, C., Tang, Y., Qian, F.: Monocular depth estimation based on deep learning: an overview. *Sci. China Technol. Sci.* **63**, 1612–1627 (2020). <https://doi.org/10.1007/s11431-020-1582-8>
62. Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A.: End-to-end learning of geometry and context for deep stereo regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 66–75. IEEE (2017)
63. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5667–5675. IEEE (2018)
64. Kuznetsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2215–2223. IEEE (2017)
65. Bovenga, F.: Special Issue “Synthetic aperture radar (SAR) techniques and applications.” *Sensors.* **20**, 1851 (2020). <https://doi.org/10.3390/s20071851>
66. Zhang, G., Geng, X., Lin, Y.-J.: Comprehensive mPoint: a method for 3D point cloud generation of human bodies utilizing FMCW MIMO mm-wave radar. *Sensors.* **21**, 6455 (2021). <https://doi.org/10.3390/s21196455>
67. Stephan, M., Santra, A., Fischer, G.: Human target detection and localization with radars using deep learning. In: Wani, M.A., Khoshgoftaar, T.M., Palade, V. (eds.) Deep Learning Applications, 2: 173–197. Springer Singapore, Singapore (2021)
68. Cha, D., Jeong, S., Yoo, M., Oh, J., Han, D.: Multi-input deep learning based FMCW radar signal classification. *Electronics* **10**, 1144 (2021). <https://doi.org/10.3390/electronics10101144>
69. Atkinson, K.B.: Introduction to modern photogrammetry. *Photogramm. Rec.* **18**, 329–330 (2003). https://doi.org/10.1046/j.0031-868x.2003.024_01.x
70. González-Jorge, H., Martínez-Sánchez, J., Bueno, M., Arias, A.P.: Unmanned aerial systems for civil applications: a review. *Drones.* **1**, 2 (2017). <https://doi.org/10.3390/drones1010002>
71. Fan, J., Saadeghvaziri, M.A.: Applications of drones in infrastructures: challenges and opportunities. *Int. J. Mech. Mechatron. Eng.* **13**(10), 649–655 (2019). <https://doi.org/10.5281/ZENODO.3566281>
72. Kaimaris, D., Patias, P., Sifnaiou, M.: UAV and the comparison of image processing software. *Int. J. Intell. Unmanned Syst.* **5**, 18–27 (2017). <https://doi.org/10.1108/IJIUS-12-2016-0009>
73. Moon, D., Chung, S., Kwon, S., Seo, J., Shin, J.: Comparison and utilization of point cloud generated from photogrammetry and laser scanning: 3D world model for smart heavy equipment planning. *Autom. Constr.* **98**, 322–331 (2019). <https://doi.org/10.1016/j.autcon.2018.07.020>
74. Rahaman, H., Champion, E.: To 3D or not 3D: choosing a photogrammetry workflow for cultural heritage groups. *Heritage* **2**, 1835–1851 (2019). <https://doi.org/10.3390/heritage2030112>
75. Zhu, X.X., Bamler, R.: Super-resolution power and robustness of compressive sensing for spectral estimation with application to spaceborne tomographic SAR. *IEEE Trans. Geosci. Remote Sens.* **50**, 247–258 (2012). <https://doi.org/10.1109/TGRS.2011.2160183>
76. Shahzad, M., Zhu, X.X., Bamler, R.: Façade structure reconstruction using spaceborne TomoSAR point clouds. In: 2012 IEEE International Geoscience and Remote Sensing Symposium. pp. 467–470. IEEE (2012)
77. Shi, Y., Zhu, X.X., Bamler, R.: Nonlocal compressive sensing-based SAR tomography. *IEEE Trans. Geosci. Remote Sens.* **57**, 3015–3024 (2019). <https://doi.org/10.1109/TGRS.2018.2879382>
78. Zhou, Y., Tuzel, O.: VoxelNet: End-to-end learning for point cloud based 3D object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4490–4499. IEEE (2018)
79. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7675–7684. IEEE (2019)
80. Shi, C., Li, J., Gong, J., Yang, B., Zhang, G.: An improved lightweight deep neural network with knowledge distillation for local feature extraction and visual localization using images and LiDAR point clouds. *ISPRS J. Photogramm. Remote Sens.* **184**, 177–188 (2022). <https://doi.org/10.1016/j.isprsjprs.2021.12.011>
81. Vayghan, S.S., Salmani, M., Ghasemkhani, N., Pradhan, B., Alamri, A.: Artificial intelligence techniques in extracting building and tree footprints using aerial imagery and LiDAR data. *Geocarto Int.* **37**, 2967–2995 (2022). <https://doi.org/10.1080/10106049.2020.1844311>
82. Islam, M.M., Newaz, A.A.R., Karimodini, A.: A pedestrian detection and tracking framework for autonomous cars: efficient fusion of camera and LiDAR data. In: 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1287–1292). IEEE. (2021). <https://doi.org/10.48550/ARXIV.2108.12375>
83. Haala, N., Hastedt, H., Wolf, K., Ressel, C., Baltrusch, S.: Digital photogrammetric camera evaluation generation of digital elevation models. *Photogramm. - Fernerkundung - Geoinf.* **2010**, 99–115 (2010). <https://doi.org/10.1127/1432-8364/2010/0043>
84. Babatunde, O.H., Armstrong, L., Leng, J., Diepeveen, D.: A survey of computer-based vision systems for automatic identification of plant species. *J. Agric. Inf.* **6**(1), 61–71 (2015)
85. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015). <https://doi.org/10.1038/nature14539>
86. Fathi, H., Brilakis, I.: Automated sparse 3D point cloud generation of infrastructure using its distinctive visual features. *Adv. Eng. Informatics.* **25**, 760–770 (2011). <https://doi.org/10.1016/j.aei.2011.06.001>
87. Han, X.-F., Sun, S.-J., Song, X.-Y., Xiao, G.-Q.: 3D Point cloud descriptors in hand-crafted and deep learning age: state-of-the-art. *arXiv:1802*, (2018)
88. Nurunnabi, A., West, G., Belton, D.: Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data. *Pattern Recognit.* **48**, 1404–1419 (2015). <https://doi.org/10.1016/j.patcog.2014.10.014>
89. Li, X., Liu, J., Dai, S.: Point cloud super-resolution based on geometric constraints. *IET Comput. Vis.* **15**, 312–321 (2021). <https://doi.org/10.1049/cvi2.12045>
90. Liu, Y., Zou, B., Xu, J., Yang, S., Li, Y.: Denoising for 3D point cloud based on regularization of a statistical low-dimensional

- manifold. *Sensors*. **22**, 2666 (2022). <https://doi.org/10.3390/s22072666>
91. Qi, C.R., Su, H., Niessner, M., Dai, A., Yan, M., Guibas, L.J.: Volumetric and multi-view CNNs for object classification on 3D data. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5648–5656). (2016). <https://doi.org/10.48550/ARXIV.1604.03265>
 92. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., Smola, A.: Deep sets. *Adv. Neural Inf. Process. Syst.* (2017). <https://doi.org/10.48550/ARXIV.1703.06114>
 93. Kalogerakis, E., Averkiou, M., Maji, S., Chaudhuri, S.: 3D Shape segmentation with projective convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6630–6639. IEEE (2017)
 94. Wu, W., Qi, Z., Fuxin, L.: PointConv: deep convolutional networks on 3D point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9613–9622. IEEE (2019)
 95. Li, R., Li, X., Fu, C.-W., Cohen-Or, D., Heng, P.-A.: PU-GAN: a point cloud upsampling adversarial network. (2019). <https://doi.org/10.48550/ARXIV.1907.10844>
 96. Sauder, J., Sievers, B.: Self-supervised deep learning on point clouds by reconstructing space. <http://arxiv.org/abs/1901.08396>, (2019)
 97. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. (2015). <https://doi.org/10.48550/ARXIV.1505.05192>
 98. Zamorski, M., Zięba, M., Klukowski, P., Nowak, R., Kurach, K., Stokowiec, W., Trzcziński, T.: Adversarial autoencoders for compact representations of 3D point clouds. *Comput. Vision Image Understand* **193**, 102921 (2018)
 99. Hua, B.-S., Tran, M.-K., Yeung, S.-K.: Pointwise convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 984–993). (2017). <https://doi.org/10.48550/ARXIV.1712.05245>
 100. Lin, C.-H., Kong, C., Lucey, S.: Learning efficient point cloud generation for dense 3D object reconstruction. *Proc. AAAI Conf. Artif. Intell.* **32**, (2018). <https://doi.org/10.1609/aaai.v32i1.12278>
 101. Djahel, R., Vallet, B., Monasse, P.: Towards efficient indoor/outdoor registration using planar polygons. *ISPRS Ann Photogramm. Remote Sens. Spat. Inf. Sci.* **2**, 51–58 (2021). <https://doi.org/10.5194/isprs-annals-V-2-2021-51-2021>
 102. Besl, P.J., McKay, N.D.: A method for registration of 3-D shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**, 239–256 (1992). <https://doi.org/10.1109/34.121791>
 103. Viroli, C., McLachlan, G.J.: Deep Gaussian mixture models. *Stat. Comput.* **29**, 43–51 (2019). <https://doi.org/10.1007/s11222-017-9793-z>
 104. Zhu, H., Guo, B., Zou, K., Li, Y., Yuen, K.-V., Mihaylova, L., Leung, H.: A review of point set registration: from pairwise registration to groupwise registration. *Sensors*. **19**, 1191 (2019). <https://doi.org/10.3390/s19051191>
 105. Deng, H., Birdal, T., Ilic, S.: PPFNet: global context aware local features for robust 3D point matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 195–205). (2018). <https://doi.org/10.48550/ARXIV.1802.02669>
 106. Zhou, J., Wang, M.J., Mao, W.D., Gong, M.L., Liu, X.P.: SiamesePointNet: a siamese point network architecture for learning 3D shape descriptor. *Comput. Graph. Forum.* **39**, 309–321 (2020). <https://doi.org/10.1111/cgf.13804>
 107. Wang, Y., Solomon, J.: Deep closest point: learning representations for point cloud registration. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3522–3531. IEEE (2019)
 108. Brightman, N., Fan, L., Zhao, Y.: Point cloud registration: a mini-review of current state, challenging issues and future directions. *AIMS Geosci.* **9**, 68–85 (2023). <https://doi.org/10.3934/geosci.2023005>
 109. Hu, S.-M., Cai, J.-X., Lai, Y.-K.: Semantic labeling and instance segmentation of 3D point clouds using patch context analysis and multiscale processing. *IEEE Trans. Vis. Comput. Graph.* **26**, 2485–2498 (2020). <https://doi.org/10.1109/TVCG.2018.2889944>
 110. Cheraghian, A., Rahman, S., Campbell, D., Petersson, L.: Transductive zero-shot learning for 3D point cloud classification. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 912–922. IEEE (2020)
 111. Yan, Z., Duckett, T., Bellotto, N.: Online learning for 3D LiDAR-based human detection: experimental analysis of point cloud clustering and classification methods. *Auton. Robots.* **44**, 147–164 (2020). <https://doi.org/10.1007/s10514-019-09883-y>
 112. Yuan, W., Khot, T., Held, D., Mertz, C., Hebert, M.: PCN: point completion network. In: 2018 International Conference on 3D Vision (3DV). pp. 728–737. IEEE (2018)
 113. Choy, C., Dong, W., Koltun, V.: Deep global registration. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2511–2520. IEEE (2020)
 114. Pais, G.D., Ramalingam, S., Govindu, V.M., Nascimento, J.C., Chellappa, R., Miraldo, P.: 3DRegNet: a deep neural network for 3D point registration. (2019). <https://doi.org/10.48550/ARXIV.1904.01701>
 115. Huang, X., Fan, L., Wu, Q., Zhang, J., Yuan, C.: Fast registration for cross-source point clouds by using weak regional affinity and pixel-wise refinement. (2019). <https://doi.org/10.48550/ARXIV.1903.04630>
 116. Lu, W., Wan, G., Zhou, Y., Fu, X., Yuan, P., Song, S.: DeepVCP: An end-to-end deep neural network for point cloud registration. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12–21. IEEE (2019)
 117. Choy, C., Park, J., Koltun, V.: Fully convolutional geometric features. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8957–8965. IEEE (2019)
 118. Poiesi, F., Boscaini, D.: Learning general and distinctive 3D local deep descriptors for point cloud registration. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(3), 3979–3985 (2022). <https://doi.org/10.1109/TPAMI.2022.3175371>
 119. Yew, Z.J., Lee, G.H.: RPM-Net: robust point matching using learned features. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11821–11830. IEEE (2020)
 120. Zhao, Y., Fan, L.: Review on deep learning algorithms and benchmark datasets for pairwise global point cloud registration. *Remote Sens.* **15**, 2060 (2023). <https://doi.org/10.3390/rs15082060>
 121. Yew, Z.J., Lee, G.H.: REGTR: End-to-end point cloud correspondences with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6677–6686). (2022). <https://doi.org/10.48550/ARXIV.2203.14517>
 122. Dong, K., Gao, S., Xin, S., Zhou, Y.: Probability driven approach for point cloud registration of indoor scene. *Vis. Comput.* **38**, 51–63 (2022). <https://doi.org/10.1007/s00371-020-01999-y>
 123. Sedghi, A., Luo, J., Mehrtash, A., Pieper, S., Tempny, C.M., Kapur, T., Mousavi, P., Wells, W.M.: Semi-supervised deep metrics for image registration. *arXiv preprint arXiv:1804.01565*. (2018). <https://doi.org/10.48550/ARXIV.1804.01565>
 124. McClelland, J.R., Modat, M., Arridge, S., Grimes, H., D’Souza, D., Thomas, D., Connell, D.O., Low, D.A., Kaza, E., Collins, D.J., Leach, M.O., Hawkes, D.J.: A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for partial image data or full images. *Phys. Med. Biol.* **62**, 4273–4292 (2017). <https://doi.org/10.1088/1361-6560/aa6070>

125. Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F.C., Miao, S., Maier, A.K., Ayache, N., Liao, R., Kamen, A.: Robust non-rigid registration through agent-based action learning. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*, pp. 344–352. Springer International Publishing, Cham (2017)
126. Hering, A., Kuckertz, S., Heldmann, S., Heinrich, M.P.: Enhancing label-driven deep deformable image registration with local distance metrics for state-of-the-art cardiac motion tracking. In: Handels, H., Deserno, T.M., Maier, A., Maier-Hein, K.H., Palm, C., Tolxdorff, T. (eds.) *Bildverarbeitung für die Medizin 2019*, pp. 309–314. Springer Fachmedien Wiesbaden, Wiesbaden (2019)
127. Ferrante, E., Oktay, O., Glocker, B., Milone, D.H.: On the adaptability of unsupervised CNN-based deformable image registration to unseen image domains. In: Shi, Y., Suk, H.-I., Liu, M. (eds.) *Machine Learning in Medical Imaging*, pp. 294–302. Springer International Publishing, Cham (2018)
128. Kim, B., Kim, J., Lee, J.-G., Kim, D.H., Park, S.H., Ye, J.C.: Unsupervised deformable image registration using cycle-consistent CNN. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.-T., Khan, A. (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 166–174. Springer International Publishing, Cham (2019)
129. Fan, Y., Wang, M., Geng, N., He, D., Chang, J., Zhang, J.J.: A self-adaptive segmentation method for a point cloud. *Vis. Comput.* **34**, 659–673 (2018). <https://doi.org/10.1007/s00371-017-1405-6>
130. Xie, Y., Tian, J., Zhu, X.X.: Linking points with labels in 3D: a review of point cloud semantic segmentation. *IEEE Geosci. Remote Sens. Mag.* **8**, 38–59 (2020). <https://doi.org/10.1109/MGRS.2019.2937630>
131. Akagic, A., Krivic, S., Dizdar, H., Velagic, J.: Computer vision with 3D point cloud data: methods, datasets and challenges. In: *2022 XXVIII International Conference on Information, Communication and Automation Technologies (ICAT)*. pp. 1–8. IEEE (2022)
132. Boulch, A., Saux, B. Le, Audebert, N.: Unstructured point cloud semantic labeling using deep segmentation networks. *Eurographics Work. 3D Object Retr.* 8-pages (2017). <https://doi.org/10.2312/3DOR.20171047>
133. Millioto, A., Vizzo, I., Behley, J., Stachniss, C.: RangeNet +: fast and accurate LiDAR semantic segmentation. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 4213–4220. IEEE (2019)
134. Zhang, Y., Zhou, Z., David, P., Yue, X., Xi, Z., Gong, B., Foroosh, H.: PolarNet: An improved grid representation for online LiDAR point clouds semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9601–9610). (2020). <https://doi.org/10.48550/ARXIV.2003.14032>
135. Honti, R., Erdélyi, J., Kopáčik, A.: Automation of cylinder segmentation from point cloud data. *Pollack Period.* **14**, 189–200 (2019). <https://doi.org/10.1556/606.2019.14.3.18>
136. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3431–3440. IEEE (2015)
137. Riegler, G., Ulusoy, A.O., Geiger, A.: OctNet: learning deep 3D representations at high resolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3577–3586). (2016). <https://doi.org/10.48550/ARXIV.1611.05009>
138. Graham, B., Engelcke, M., van der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9224–9232). (2017). <https://doi.org/10.48550/ARXIV.1711.10275>
139. Wang, S., Suo, S., Ma, W.-C., Pokrovsky, A., Urtasun, R.: Deep parametric continuous convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2589–2597). (2021). <https://doi.org/10.48550/ARXIV.2101.06742>
140. Engelmann, F., Kontogianni, T., Leibe, B.: Dilated point convolutions: on the receptive field size of point convolutions on 3D point clouds. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 9463–9469). IEEE (2019). <https://doi.org/10.48550/ARXIV.1907.12046>
141. Fan, H., Yang, Y.: PointRNN: point recurrent neural network for moving point cloud processing. *arXiv preprint arXiv:1910.08287*. (2019). <https://doi.org/10.48550/ARXIV.1910.08287>
142. Pirasteh, S., Rashidi, P., Rastveis, H., Huang, S., Zhu, Q., Liu, G., Li, Y., Li, J., Seydipour, E.: Developing an algorithm for buildings extraction and determining changes from airborne LiDAR, and comparing with R-CNN method from drone images. *Remote Sens.* **11**, 1272 (2019). <https://doi.org/10.3390/rs11111272>
143. Engelmann, F., Kontogianni, T., Hermans, A., Leibe, B.: Exploring spatial context for 3D semantic segmentation of point clouds. In: *Proceedings of the IEEE international conference on computer vision workshops* (pp. 716–724). (2018). <https://doi.org/10.48550/ARXIV.1802.01500>
144. Ye, X., Li, J., Huang, H., Du, L., Zhang, X.: 3D recurrent neural networks with context fusion for point cloud semantic segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision – ECCV 2018*, pp. 415–430. Springer International Publishing, Cham (2018)
145. Zhao, Z., Liu, M., Ramani, K.: DAR-Net: dynamic aggregation network for semantic scene segmentation. (2019). <https://doi.org/10.48550/ARXIV.1907.12022>
146. Landrieu, L., Boussaha, M.: Point cloud oversegmentation with graph-structured deep metric learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7440–7449). (2019). <https://doi.org/10.48550/ARXIV.1904.02113>
147. Guo, M.H., Cai, J.X., Liu, Z.N., Mu, T.J., Martin, R.R., Hu, S.M.: PCT: point cloud transformer. *Comput. Vis. Med.* **7**, 187–199 (2021)
148. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, (2017)
149. Dai, A., Nießner, M.: 3DMV: Joint 3D-multi-view prediction for 3D semantic scene segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 452–468). (2018). <https://doi.org/10.48550/ARXIV.1803.10409>
150. Chiang, H.-Y., Lin, Y.-L., Liu, Y.-C., Hsu, W.H.: A unified point-based framework for 3D segmentation. In: *2019 International Conference on 3D Vision (3DV)* (pp. 155–163). IEEE. (2019). <https://doi.org/10.48550/ARXIV.1908.00478>
151. Luo, C., Li, X., Cheng, N., Li, H., Lei, S., Li, P.: MVP-Net: multiple view pointwise semantic segmentation of large-scale point clouds. *arXiv preprint arXiv:2201.12769*. (2022). <https://doi.org/10.48550/ARXIV.2201.12769>
152. Taghizadeh, M., Chalechale, A.: A comprehensive and systematic review on classical and deep learning based region proposal algorithms. *Expert Syst. Appl.* **189**, 116105 (2022). <https://doi.org/10.1016/j.eswa.2021.116105>
153. Muhammad Yasir, S., Muhammad Sadiq, A., Ahn, H.: 3D instance segmentation using deep learning on RGB-D indoor data. *Comput. Mater. Contin.* **72**, 5777–5791 (2022)
154. Zhang, F., Guan, C., Fang, J., Bai, S., Yang, R., Torr, P.H.S., Prisacariu, V.: Instance segmentation of LiDAR point clouds. In:

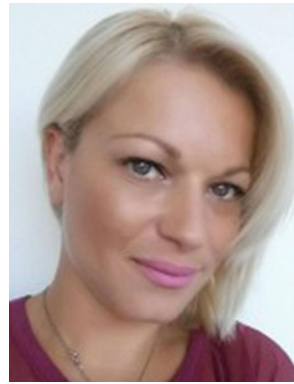
- 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9448–9455. IEEE (2020)
155. Pham, Q.-H., Nguyen, D.T., Hua, B.-S., Roig, G., Yeung, S.-K.: JSIS3D: joint semantic-instance segmentation of 3D point clouds with multi-task pointwise networks and multi-value conditional random fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8827–8836). (2019). <https://doi.org/10.48550/ARXIV.1904.00699>
 156. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.-W., Jia, J.: PointGroup: dual-set point grouping for 3D instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4867–4876). (2020). <https://doi.org/10.48550/ARXIV.2004.01658>
 157. Wang, W., Yu, R., Huang, Q., Neumann, U.: SGPN: similarity group proposal network for 3D point cloud instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2569–2578). (2017). <https://doi.org/10.48550/ARXIV.1711.08588>
 158. Wang, Z., Lu, F.: VoxSegNet: volumetric CNNs for semantic part segmentation of 3D shapes. *IEEE Trans. Vis. Comput. Graph.* **26**, 2919–2930 (2020). <https://doi.org/10.1109/TVCG.2019.2896310>
 159. Yi, L., Su, H., Guo, X., Guibas, L.: SyncSpecCNN: synchronized spectral CNN for 3D shape segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2282–2290). (2016). <https://doi.org/10.48550/ARXIV.1612.00606>
 160. Wang, P., Gan, Y., Shui, P., Yu, F., Zhang, Y., Chen, S., Sun, Z.: 3D shape segmentation via shape fully convolutional networks. *Comput. Graph.* **76**, 182–192 (2018)
 161. Yu, F., Liu, K., Zhang, Y., Zhu, C., Xu, K.: PartNet: a recursive part decomposition network for fine-grained and hierarchical shape segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9491–9500). (2019). <https://doi.org/10.48550/ARXIV.1903.00709>
 162. Wang, X., Liu, S., Shen, X., Shen, C., Jia, J.: Associatively segmenting instances and semantics in point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 4096–4105). (2019). <https://doi.org/10.48550/ARXIV.1902.09852>
 163. Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(7), 3523–3542 (2021)
 164. Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., Dou, D.: Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **64**(12), 3197–3234 (2022)
 165. Zhou, Z.-H.: A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **5**, 44–53 (2018). <https://doi.org/10.1093/nsr/nwx106>
 166. Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A.: Deep learning classifiers for hyperspectral imaging: A review. *ISPRS J. Photogramm. Remote Sens.* **158**, 279–317 (2019). <https://doi.org/10.1016/j.isprsjprs.2019.09.006>
 167. Wang, Y., Zhuo, W., Li, Y., Wang, Z., Ju, Q., Zhu, W.: Fully self-supervised learning for semantic segmentation. *arXiv preprint arXiv:2202.11981*. (2022). <https://doi.org/10.48550/ARXIV.2202.11981>
 168. Jiang, P., Ergu, D., Liu, F., Cai, Y., Ma, B.: A review of Yolo algorithm developments. *Procedia Comput. Sci.* **199**, 1066–1073 (2022). <https://doi.org/10.1016/j.procs.2022.01.135>
 169. Goel, V., Weng, J., Poupart, P.: Unsupervised video object segmentation for deep reinforcement learning. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (pp. 5688–5699) (2018). <https://doi.org/10.48550/ARXIV.1805.07780>
 170. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. pp. 234–241 (2015)
 171. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2018). <https://doi.org/10.1109/TPAMI.2017.2699184>
 172. Wang, C., Wang, C., Li, W., Wang, H.: A brief survey on RGB-D semantic segmentation using deep learning. *Displays* **70**, 102080 (2021). <https://doi.org/10.1016/j.displa.2021.102080>
 173. Zhang, J., Lin, X., Ning, X.: SVM-based classification of segmented airborne LiDAR point clouds in urban areas. *Remote Sens.* **5**, 3749–3775 (2013). <https://doi.org/10.3390/rs5083749>
 174. Atik, M.E., Duran, Z., Seker, D.Z.: Machine learning-based supervised classification of point clouds using multiscale geometric features. *ISPRS Int. J. Geo-Information.* **10**, 187 (2021). <https://doi.org/10.3390/ijgi10030187>
 175. Yan, Y., Mao, Y., Li, B.: SECOND: sparsely embedded convolutional detection. *Sensors*. **18**, 3337 (2018). <https://doi.org/10.3390/s18103337>
 176. Yang, Z., Wang, L.: Learning relationships for multi-view 3D object recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7504–7513. IEEE (2019)
 177. Wang, C., Pelillo, M., Siddiqi, K.: Dominant set clustering and pooling for multi-view 3D object recognition. *arXiv preprint arXiv:1906.01592* (2019). <https://doi.org/10.48550/ARXIV.1906.01592>
 178. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3D ShapeNets: a deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1912–1920). (2014). <https://doi.org/10.48550/ARXIV.1406.5670>
 179. Le, T., Duan, Y.: PointGrid: a deep network for 3D shape understanding. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9204–9214. IEEE (2018)
 180. Yang, J., Zhang, X., Huang, Y.: Graph attention feature fusion network for ALS point cloud classification. *Sensors*. **21**, 6193 (2021). <https://doi.org/10.3390/s21186193>
 181. Hermosilla, P., Ritschel, T., Vázquez, P.P., Vinacua, À., Ropinski, T.: Monte Carlo Convolution for learning on non-uniformly sampled point clouds. *ACM Trans. Graph. (TOG)* **37**(6), 1–12 (2018)
 182. Groh, F., Wieschollek, P., Lensch, H.P.A.: Flex-convolution (million-scale point-cloud learning beyond grid-worlds). In: *Asian Conference on Computer Vision* (pp. 105–122). Cham: Springer International Publishing. (2018). <https://doi.org/10.48550/ARXIV.1803.07289>
 183. Lei, H., Akhtar, N., Mian, A.: Octree guided CNN with spherical kernels for 3D point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9631–9640). (2019). <https://doi.org/10.48550/ARXIV.1903.00343>
 184. Zhang, K., Hao, M., Wang, J., de Silva, C.W., Fu, C.: Linked dynamic graph CNN: learning on point cloud via linking hierarchical features. *arXiv preprint arXiv:1904.10014*. (2019). <https://doi.org/10.48550/ARXIV.1904.10014>
 185. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: ULIP: Learning a unified representation of language, images, and point clouds for 3D

- understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1179–1189 (2023)
186. Wang, C.-S., Wang, H., Ning, X., Tian, S.-W., Li, W.-J.: 3D Point cloud classification method based on dynamic coverage of local area. *J. Softw.* **34**, 1962–1976 (2022)
 187. Wang, C., Ning, X., Sun, L., Zhang, L., Li, W., Bai, X.: Learning discriminative features by covering local geometric space for point cloud analysis. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–15 (2022). <https://doi.org/10.1109/TGRS.2022.3170493>
 188. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: a survey. *arXiv preprint arXiv:1904.10014* (2018). <https://doi.org/10.48550/ARXIV.1809.02165>
 189. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep Hough voting for 3D object detection in point clouds. In proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9277–9286). (2019). <https://doi.org/10.48550/ARXIV.1904.09664>
 190. Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., Qu, R.: A survey of deep learning-based object detection. *IEEE Access.* **7**, 128837–128868 (2019). <https://doi.org/10.1109/ACCESS.2019.2939201>
 191. Song, Y., Zhang, Y.-D., Yan, X., Liu, H., Zhou, M., Hu, B., Yang, G.: Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI: PCA classification using CNN From mp-MRI. *J. Magn. Reson. Imaging* **48**, 1570–1577 (2018). <https://doi.org/10.1002/jmri.26047>
 192. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P.: Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 318–327 (2020). <https://doi.org/10.1109/TPAMI.2018.2858826>
 193. Mehmood, K., Jalil, A., Ali, A., Khan, B., Murad, M., Khan, W.U., He, Y.: Context-aware and occlusion handling mechanism for online visual object tracking. *Electronics* **10**, 43 (2020). <https://doi.org/10.3390/electronics10010043>
 194. Yang, Z., Sun, Y., Liu, S., Shen, X., & Jia, J. (2019). Std: Sparse-to-dense 3d object detector for point cloud. In: Proceedings of the IEEE/CVF international conference on computer vision (pp. 1951–1960). <https://doi.org/10.48550/ARXIV.1907.10471>
 195. Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3dssd: Point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11040–11048). <https://doi.org/10.48550/ARXIV.2002.10187>
 196. Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1907–1915). <https://doi.org/10.48550/ARXIV.1611.07759>
 197. Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7345–7353). <https://doi.org/10.48550/ARXIV.2012.12397>
 198. Zeng, Y., Hu, Y., Liu, S., Ye, J., Han, Y., Li, X., Sun, N.: RT3D: real-time 3-D vehicle detection in LiDAR point cloud for autonomous driving. *IEEE Robot. Autom. Lett.* **3**, 3434–3440 (2018). <https://doi.org/10.1109/LRA.2018.2852843>
 199. Zarzar, J., Giancola, S., & Ghanem, B. (2019). PointRGCN: Graph convolution networks for 3D vehicles detection refinement. *arXiv preprint arXiv:1911.12236*. <https://doi.org/10.48550/ARXIV.1911.12236>
 200. Wang, Z., & Jia, K. (2019, November). Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1742–1749). IEEE. <https://doi.org/10.48550/ARXIV.1903.01864>
 201. Li, B. (2017, September). 3d fully convolutional network for vehicle detection in point cloud. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1513–1518). IEEE. <https://doi.org/10.48550/ARXIV.1611.08069>
 202. Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019, May). Mvx-net: Multimodal voxelnet for 3d object detection. In: 2019 International Conference on Robotics and Automation (ICRA) (pp. 7276–7282). IEEE. <https://doi.org/10.48550/ARXIV.1904.01649>
 203. Guo, Y., Wang, F., Xin, J.: Point-wise saliency detection on 3D point clouds via covariance descriptors. *Vis. Comput.* **34**, 1325–1338 (2018). <https://doi.org/10.1007/s00371-017-1416-3>
 204. Liu, H., Hu, Q., Li, B., Guo, Y.: Robust long-term tracking via instance-specific proposals. *IEEE Trans. Instrum. Meas.* **69**, 950–962 (2020). <https://doi.org/10.1109/TIM.2019.2908715>
 205. Giancola, S., Zarzar, J., & Ghanem, B. (2019). Leveraging shape completion for 3d siamese tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 1359–1368). <https://doi.org/10.48550/ARXIV.1903.01784>
 206. Zarzar, J., Giancola, S., & Ghanem, B. (2019). Efficient bird eye view proposals for 3D Siamese tracking. *arXiv preprint arXiv:1903.10168*. <https://doi.org/10.48550/ARXIV.1903.10168>
 207. Qi, H., Feng, C., Cao, Z., Zhao, F., & Xiao, Y. (2020). P2b: Point-to-box network for 3d object tracking in point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6329–6338). <https://doi.org/10.48550/ARXIV.2005.13888>
 208. Wang, Z., Li, S., Howard-Jenkins, H., Prisacariu, V., & Chen, M. (2020). Flownet3d++: Geometric losses for deep scene flow estimation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 91–98). <https://doi.org/10.48550/ARXIV.1912.01438>
 209. Hemalatha, C., Muruganand, S., Maheswaran, R.: A survey on real time object detection, tracking and recognition in image processing. *Int. J. Comput. Appl.* **91**, 38–42 (2014). <https://doi.org/10.5120/15969-5407>
 210. Pal, S.K., Pramanik, A., Maiti, J., Mitra, P.: Deep learning in multi-object detection and tracking: state of the art. *Appl. Intell.* **51**, 6400–6429 (2021). <https://doi.org/10.1007/s10489-021-02293-7>
 211. Wang, J., Ding, D., Li, Z., Ma, Z.: Multiscale point cloud geometry compression. In: 2021 Data Compression Conference (DCC). pp. 73–82. IEEE (2021)
 212. Wen, X., Wang, X., Hou, J., Ma, L., Zhou, Y., Jiang, J.: Lossy geometry compression of 3d point cloud data via an adaptive octree-guided network. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
 213. Quach, M., Chetouani, A., Valenzise, G., & Dufaux, F. (2021). A deep perceptual metric for 3D point clouds. *arXiv preprint arXiv:2102.12839*. <https://doi.org/10.48550/ARXIV.2102.12839>
 214. Quach, M., Valenzise, G., Dufaux, F.: Learning Convolutional Transforms for Lossy Point Cloud Geometry Compression. In: 2019 IEEE international conference on image processing (ICIP). pp. 4320–4324. (2019). IEEE. <https://doi.org/10.48550/ARXIV.1903.08548>
 215. Wang, J., Zhu, H., Ma, Z., Chen, T., Liu, H., Shen, Q.: Learned point cloud geometry compression. *arXiv preprint arXiv:1909.12037* (2019). <https://doi.org/10.48550/ARXIV.1909.12037>
 216. Huang, L., Wang, S., Wong, K., Liu, J., Urtasun, R.: OctSqueeze: octree-structured entropy model for LiDAR compression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1313–1323. (2020). <https://doi.org/10.48550/ARXIV.2005.07178>
 217. Nguyen, D.T., Quach, M., Valenzise, G., Duhamel, P.: Learning-based lossless compression of 3D point cloud geometry. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4220–4224. IEEE. (2020). <https://doi.org/10.48550/ARXIV.2011.14700>

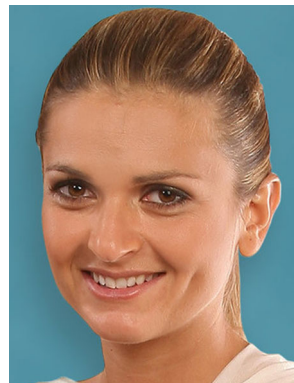
218. Nguyen, D.T., Quach, M., Valenzise, G., Duhamel, P.: Lossless coding of point cloud geometry using a deep generative model. *IEEE Trans. Circ. Syst. Video Technol.* **31**(12), 4617–4629 (2021)
219. Wiesmann, L., Milioto, A., Chen, X., Stachniss, C., Behley, J.: Deep compression for dense point cloud maps. *IEEE Robot. Autom. Lett.* **6**, 2060–2067 (2021). <https://doi.org/10.1109/LRA.2021.3059633>
220. Ochotta, T., Saube, D.: Image-based surface compression. *Comput. Graph Forum* **27**, 1647–1663 (2008). <https://doi.org/10.1111/j.1467-8659.2008.01178.x>
221. Hornung, A., Wurm, K.M., Bennewitz, M., Stachniss, C., Burgard, W.: OctoMap: an efficient probabilistic 3D mapping framework based on octrees. *Auton. Robots.* **34**, 189–206 (2013). <https://doi.org/10.1007/s10514-012-9321-0>
222. Abd-Alzhra, A.S., Al-Tamimi, M.S.: Image compression using deep learning: methods and techniques. *Iraqi J. Sci.* **63**(3), 1299–1312 (2022)
223. Wang, L., Wang, S.: A survey of image compression algorithms based on deep learning. In *Review* (2023)
224. Santurkar, S., Budden, D., Shavit, N.: Generative compression. In: 2018 Picture Coding Symposium (PCS). pp. 258–262. IEEE. (2017). <https://doi.org/10.48550/ARXIV.1703.01467>
225. Baig, M.H., Koltun, V., Torresani, L.: Learning to inpaint for image compression. arXiv e-prints. arXiv: 1709.08855. (2017). <https://doi.org/10.48550/ARXIV.1709.08855>
226. Gupta, S., Arbelaez, P., Girshick, R., Malik, J.: Aligning 3D models to RGB-D images of cluttered scenes. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4731–4740. IEEE (2015)
227. Li, Y., Dai, A., Guibas, L., Nießner, M.: Database-assisted object retrieval for real-time 3D reconstruction. *Comput. Graph. Forum.* **34**, 435–446 (2015). <https://doi.org/10.1111/cgf.12573>
228. Wang, W., Huang, Q., You, S., Yang, C., Neumann, U.: Shape inpainting using 3D generative adversarial network and recurrent convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2298–2306. (2017). <https://doi.org/10.48550/ARXIV.1711.06375>
229. Sarkar, K., Varanasi, K., Stricker, D.: Learning quadrangulated patches for 3D shape parameterization and completion. In: 2017 International Conference on 3D Vision (3DV). pp. 383–392. IEEE. (2017). <https://doi.org/10.48550/ARXIV.1709.06868>
230. Fu, Z., Hu, W., Guo, Z.: Local frequency interpretation and non-local self-similarity on graph for point cloud inpainting. *IEEE Trans. Image Process.* **28**(8), 4087–4100 (2018)
231. Nealen, A., Igarashi, T., Sorkine, O., Alexa, M.: Laplacian mesh optimization. In: GRAPHITE06: International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia 2006. pp. 381–389. ACM (2006)
232. Kazhdan, M., Hoppe, H.: Screened poisson surface reconstruction. *ACM Trans. Graph.* **32**, 1–13 (2013). <https://doi.org/10.1145/2487228.2487237>
233. Mitra, N.J., Guibas, L.J., Pauly, M.: Partial and approximate symmetry detection for 3D geometry. In: ACM SIGGRAPH 2006 Papers. p. 560. ACM Press (2006)
234. Rock, J., Gupta, T., Thorsen, J., Gwak, J., Shin, D., Hoiem, D.: Completing 3D object shape from one depth image. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2484–2493. IEEE (2015)
235. Yin, K., Huang, H., Zhang, H., Gong, M., Cohen-Or, D., Chen, B.: Morfit: interactive surface reconstruction from incomplete point clouds with curve-driven topology and geometry control. *ACM Trans. Graph.* **33**, 1–12 (2014). <https://doi.org/10.1145/2661229.2661241>
236. Dai, A., Qi, C.R., Nießner, M.: Shape completion using 3D-encoder-predictor CNNs and shape synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5868–5877. (2016). <https://doi.org/10.48550/ARXIV.1612.00101>
237. Zhang, W., Yan, Q., Xiao, C.: Detail preserved point cloud completion via separated feature aggregation. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 512–528. Springer International Publishing. (2020). <https://doi.org/10.48550/ARXIV.2007.02374>
238. Wen, X., Xiang, P., Han, Z., Cao, Y.-P., Wan, P., Zheng, W., Liu, Y.-S.: PMP-Net: point cloud completion by learning multi-step point moving paths. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7443–7452. (2020). <https://doi.org/10.48550/ARXIV.2012.03408>
239. Wang, X., Ang, M.H., Lee, G.H.: Voxel-based network for shape completion by leveraging edge generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13189–13198. (2021). <https://doi.org/10.48550/ARXIV.2108.09936>
240. Wen, X., Han, Z., Cao, Y. P., Wan, P., Zheng, W., & Liu, Y. S. (2021). Cycle4completion: Unpaired point cloud completion using cycle transformation with missing region coding. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13080-13089). <https://doi.org/10.48550/ARXIV.2103.07838>
241. Chen, Y., Hu, H.: An improved method for semantic image inpainting with GANs: progressive inpainting. *Neural. Process. Lett.* **49**, 1355–1367 (2019). <https://doi.org/10.1007/s11063-018-9877-6>
242. Zhao, G., Liu, J., Jiang, J., Wang, W.: A deep cascade of neural networks for image inpainting, deblurring and denoising. *Multimed. Tools Appl.* **77**, 29589–29604 (2018). <https://doi.org/10.1007/s11042-017-5320-7>
243. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2536–2544. (2016). <https://doi.org/10.48550/ARXIV.1604.07379>
244. Mahdaoui, A.E., Ouahabi, A., Moulay, M.S.: Image denoising using a compressive sensing approach based on regularization constraints. *Sensors.* **22**, 2199 (2022). <https://doi.org/10.3390/s22062199>
245. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6721–6729. (2016). <https://doi.org/10.48550/ARXIV.1611.09969>
246. Liu, G., Reda, F.A., Shih, K.J., Wang, T.-C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European conference on computer vision (ECCV). pp. 85–100. (2018). <https://doi.org/10.48550/ARXIV.1804.07723>
247. Xiang, H., Zou, Q., Nawaz, M.A., Huang, X., Zhang, F., Yu, H.: Deep learning for image inpainting: A survey. *Pattern Recognit.* **134**, 109046 (2023). <https://doi.org/10.1016/j.patcog.2022.109046>
248. Davis, J., Marschner, S.R., Garr, M., Levoy, M.: Filling holes in complex surfaces using volumetric diffusion. In: First International Symposium on 3D Data Processing Visualization and Transmission. pp. 428–861. IEEE Comput. Soc (2002)
249. Darabi, S., Shechtman, E., Barnes, C., Goldman, D.B., Sen, P.: Image melding: combining inconsistent images using patch-based synthesis. *ACM Trans. Graph.* **31**, 1–10 (2012). <https://doi.org/10.1145/2185520.2185578>
250. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution. In: Proceedings

- of the IEEE/CVF international conference on computer vision. pp. 4471–4480. (2018). <https://doi.org/10.48550/ARXIV.1806.03589>
251. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839. (2017). <https://doi.org/10.48550/ARXIV.1702.04405>
 252. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: a RGB-D scene understanding benchmark suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 567–576. IEEE (2015)
 253. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631. (2019). <https://doi.org/10.48550/ARXIV.1903.11027>
 254. Huang, J., Guan, D., Xiao, A., Lu, S.: Cross-view regularization for domain adaptive panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10133–10144. (2021). <https://doi.org/10.48550/ARXIV.2103.02584>
 255. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schubert, P.: A2D2: Audi autonomous driving dataset. arXiv preprint. [arXiv:2004.06320](https://arxiv.org/abs/2004.06320). (2020). <https://doi.org/10.48550/ARXIV.2004.06320>
 256. Chen, Y., Wang, J., Li, J., Lu, C., Luo, Z., Xue, H., Wang, C.: DBNet: A large-scale dataset for driving behavior learning. Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 9, (2019)
 257. Vallet, B., Brédif, M., Serna, A., Marcotegui, B., Paparoditis, N.: TerraMobilita/iQmulus urban point cloud analysis benchmark. Comput. Graph. 49, 126–133 (2015). <https://doi.org/10.1016/j.cag.2015.03.004>
 258. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the KITTI dataset. Int. J. Rob. Res. 32, 1231–1237 (2013). <https://doi.org/10.1177/0278364913491297>
 259. Roynard, X., Deschaud, J.-E., Goulette, F.: Paris-Lille-3D: a large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. Int. J. Rob. Res. 37, 545–557 (2018). <https://doi.org/10.1177/0278364918767506>
 260. Munoz, D., Bagnell, J.A., Vandapel, N., Hebert, M.: Contextual classification with functional Max-Margin Markov Networks. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). pp. 975–982. IEEE (2009)
 261. Paris-rue-Madame Database - A 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods: In: Special Session on Urban Scene Analysis: interpretation, mapping and modeling. pp. 819–824. SCITEPRESS - Science and Technology Publications (2014)
 262. Uy, M. A., Pham, Q. H., Hua, B. S., Nguyen, T., & Yeung, S. K. (2019). Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 1588–1597). <https://doi.org/10.48550/ARXIV.1908.04616>
 263. Hua, B.-S., Pham, Q.-H., Nguyen, D.T., Tran, M.-K., Yu, L.-F., Yeung, S.-K.: SceneNN: a scene meshes dataset with annotations. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 92–101. IEEE (2016)
 264. Hackel, T., Savinov, N., Ladicky, L., Wegner, J.D., Schindler, K., Pollefeys, M.: Semantic3D.Net: a new large-scale point cloud classification benchmark. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. IV-1/W1, 91–98 (2017). <https://doi.org/10.5194/isprs-annals-IV-1-W1-91-2017>
 265. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Gall, J., Stachniss, C.: Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset. Int. J. Rob. Res. 40, 959–967 (2021). <https://doi.org/10.1177/02783649211006735>
 266. Xiao, J., Owens, A., Torralba, A.: SUN3D: a database of big spaces reconstructed using SfM and object labels. In: 2013 IEEE International Conference on Computer Vision (ICCV). pp. 1625–1632. IEEE (2013)
 267. De Deuge, M., Quadros, A., Hung, C., Douillard, B.: Unsupervised feature learning for classification of outdoor 3D scans. In: Australasian Conference on Robotics and Automation, ACRA (2013)
 268. Zeng, A., Song, S., Nießner, M., Fisher, M., Xiao, J., Funkhouser, T.: 3DMatch: learning local geometric descriptors from RGB-D reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1802–1811. (2016). <https://doi.org/10.48550/ARXIV.1603.08182>
 269. Tchapmi, L.P., Kosaraju, V., Rezaatofghi, H., Reid, I., Savarese, S.: TopNet: structural point cloud decoder. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 383–392. IEEE (2019)
 270. Wang, C., Dai, Y., Elsheimy, N., Wen, C., Retscher, G., Kang, Z., Lingua, A.: ISPRS benchmark on multisensory indoor mapping and positioning. ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. 5, 117–123 (2020). <https://doi.org/10.5194/isprs-annals-V-5-2020-117-2020>
 271. Pan, L., Chen, X., Cai, Z., Zhang, J., Zhao, H., Yi, S., Liu, Z.: Variational relational point completion network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8524–8533. (2021). <https://doi.org/10.48550/ARXIV.2104.10154>
 272. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 746–760 (2012)
 273. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: a large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 909–918. (2018). <https://doi.org/10.48550/ARXIV.1812.02713>
 274. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: an information-rich 3D model repository. arXiv preprint [arXiv:1512.03012](https://arxiv.org/abs/1512.03012). (2015). <http://arxiv.org/abs/1512.03012>,
 275. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2D-3D-semantic data for indoor scene understanding. arXiv preprint. [arXiv:1702.01105](https://arxiv.org/abs/1702.01105). (2017). <https://doi.org/10.48550/ARXIV.1702.01105>
 276. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3D semantic parsing of large-scale indoor spaces. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1534–1543. IEEE (2016)
 277. Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S.: Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. 36 (3). pp. 2795–2803. (2021). <https://doi.org/10.48550/ARXIV.2107.05399>

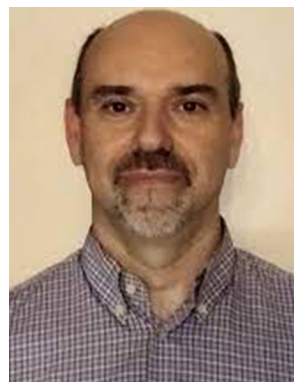
278. Griffiths, D., Boehm, J.: SynthCity: a large scale synthetic point cloud. arXiv preprint [arXiv:1907.04758](https://arxiv.org/abs/1907.04758). (2019). <https://doi.org/10.48550/ARXIV.1907.04758>
279. Dong, Z., Liang, F., Yang, B., Xu, Y., Zang, Y., Li, J., Wang, Y., Dai, W., Fan, H., Hyyppä, J., Stilla, U.: Registration of large-scale terrestrial laser scanner point clouds: a review and benchmark. *ISPRS J. Photogramm. Remote Sens.* **163**, 327–342 (2020). <https://doi.org/10.1016/j.isprsjprs.2020.03.013>
280. Siddiqi, K., Zhang, J., Macrini, D., Shokoufandeh, A., Bouix, S., Dickinson, S.: Retrieving articulated 3-D models using medial surfaces. *Mach. Vis. Appl.* **19**, 261–275 (2008). <https://doi.org/10.1007/s00138-007-0097-8>
281. Patil, A., Malla, S., Gang, H., Chen, Y.-T.: The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 9552–9557. IEEE (2019)
282. Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., Omari, S., Shah, S., Kulkarni, A., Kazakova, A., Tao, C., Platinsky, L., Jiang, W., Shet, V.: Lyft Level5 AV Dataset 2019. <https://level5.lyft.com/dataset/> (2023). Accessed 9 December 2023
283. Pham, Q.-H., Sevestre, P., Pahwa, R.S., Zhan, H., Pang, C.H., Chen, Y., Mustafa, A., Chandrasekhar, V., Lin, J.: A*3D dataset: towards autonomous driving in challenging environments. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 2267–2273. IEEE (2020)
284. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2019)
285. Rottensteiner, F., Sohn, G., Jung, J., Gerke, M., Baillard, C., Benitez, S., Breitkopf, U.: The ISPRS benchmark on urban object classification and 3D building reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **I-3**, 293–298 (2012). <https://doi.org/10.5194/isprsannals-I-3-293-2012>
286. Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Li, J.: Toronto-3D: a large-scale mobile LiDAR dataset for semantic segmentation of urban roadways. In: IEEE/CVF conference on computer vision and pattern recognition. pp. 202–203 (2020)
287. Varney, N., Asari, V.K., Graehling, Q.: DALES: A Large-scale aerial LiDAR data set for semantic segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 717–726. IEEE (2020)
288. Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., Yu, J., Xu, H., Xu, C.: One million scenes for autonomous driving: ONCE Dataset. arXiv: 2106.11037. (2021)
289. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., Chen, G., Cui, S., Han, X.: Mvimnet: A large-scale dataset of multi-view images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9150–9161 (2023)



Kyriaki A. Tychola received the Diploma Degrees from the Department of Geoinformation and Surveying from the Technological Educational Institute, Serres, Greece and in Landscape Architecture from the Technological Educational Institute of Kavala, Greece in 2005 and 2013 respectively and M.Sc Degree from the Department of Geography and applied Geoinformatics in 2022 specializing in GIS and Photogrammetry. She has publications in international scientific journals in computer vision field. As a researcher, she is a member of the Machine Learning and Vision (MLV) Research Group, where she participates in research projects.



Eleni Vrochidou received the Diploma, the M.Sc and Ph.D. Degrees from the Department of Electrical & Computer Engineering, Democritus University of Thrace (DUTH), Greece, in 2004, 2007, and 2016, respectively. She is currently a part-time lecturer in the Department of Computer Science (IHU) at the International Hellenic University. Her research interests are intelligent systems, signal processing, pattern recognition, and embedded systems. She has several publications in international scientific conferences, journals, and book chapters in these areas. As a researcher, she is a member of the Machine Learning and Vision (MLV) Research Group, where she participates in research projects.



George A. Papakostas received the diploma in Electrical and Computer Engineering in 1999 and M.Sc. and Ph.D. in Electrical and Computer Engineering in 2002 and 2007, respectively, from the Democritus University of Thrace (DUTH), Greece. He is a Tenured Full Professor in the Department of Computer Science International Hellenic University, Greece. He is the Head of the Machine Learning and Vision (MLV) Research Group. Prof. Papakostas has (co)authored more than 200 publications, his publications have over 3600 citations with an h-index 34 (Google Scholar). His research interests include machine learning, computer/machine vision, pattern recognition, and computational intelligence.