LiDAR-Net: A Real-scanned 3D Point Cloud Dataset for Indoor Scenes

Yanwen Guo[1]*, Yuanqi Li[1], Dayong Ren[1], Xiaohong Zhang[1], Jiawei Li[1], Liang Pu[1], Changfeng Ma[1],
Xiaoyu Zhan[1], Jie Guo[1], Mingqiang Wei[2], Yan Zhang[1], Piaopiao Yu[2], Shuangyu Yang[1], Donghao Ji[1],
Huisheng Ye[1], Hao Sun[1], Yansong Liu[1], Yinuo Chen[1], Jiaqi Zhu[1], Hongyu Liu[1]
[1]Nanjing University, Nanjing, China
[2]Nanjing University of Aeronautics and Astronautics, Nanjing, China
{ywguo,guojie,zhangyannju}@nju.edu.cn, rdyedu@gmail.com
{yuanqili, xiaohongzhang, lijiawei, puliang, changfengma}@smail.nju.edu.cn
{zhanxy, shuangyuyang, donghaoji, huishengye, hao.sun, yansongliu}@smail.nju.edu.cn
{mqwei, yupiaopiao}@nuaa.edu.cn, {yinuochen, jiaqizhu, hongyu_liu}@smail.nju.edu.cn

Abstract

In this paper, we present LiDAR-Net, a new real-scanned indoor point cloud dataset, containing nearly 3.6 billion precisely point-level annotated points, covering an expansive area of 30,000m^2. It encompasses three prevalent daily environments, including learning scenes, working scenes, and living scenes. LiDAR-Net is characterized by its non-uniform point distribution, e.g., scanning holes and scanning lines. Additionally, it meticulously records and annotates scanning anomalies, including reflection noise and ghost. These anomalies stem from specular reflections on glass or metal, as well as distortions due to moving persons. LiDAR-Net's realistic representation of non-uniform distribution and anomalies significantly enhances the training of deep learning models, leading to improved generalization in practical applications. We thoroughly evaluate the performance of state-of-the-art algorithms on LiDAR-Net and provide a detailed analysis of the results. Crucially, our research identifies several fundamental challenges in understanding indoor point clouds, contributing essential insights to future explorations in this field. Our dataset can be found online: http://lidar-net.njumeta.com.

1. Introduction

Owing to its reliability and accuracy, Light Detection And Ranging (LiDAR) technology becomes increasingly popular, finding broad applications in autonomous vehicles for perception and localization [23, 26, 54] as well as expanding its applications in consumer electronics [3, 34, 37]. Introduced in 2020, the LiDAR scanner on the iPhone has enabled a multitude of innovative applications in virtual reality (VR) and 3D measurements. Deep learning has achieved remarkable success in computer vision, particularly in analyzing and understanding 3D scenes [35, 36, 49, 56]. Datasets undoubtedly have played a critical role in this process, since training a deep models usually requires extensive annotated datasets, especially for supervised learning [2, 8, 11, 19, 42, 53]. In the domain of 3D computer vision for indoor scenes, the most widely recognized datasets in the academic community are S3DIS [2], ScanNet [11], and the recent ScanNet++ [53], *etc.* S3DIS and ScanNet have been used to train models for various 3D scene understanding tasks, such as 3D object detection, along with semantic and instance segmentation. S3DIS and ScanNet collect RGB-D images, and then reconstruct and annotate textured meshes. ScanNet++ relies on a scanner to capture point clouds, however, the provided annotations are attached to the Poisson reconstruction instead of raw points.

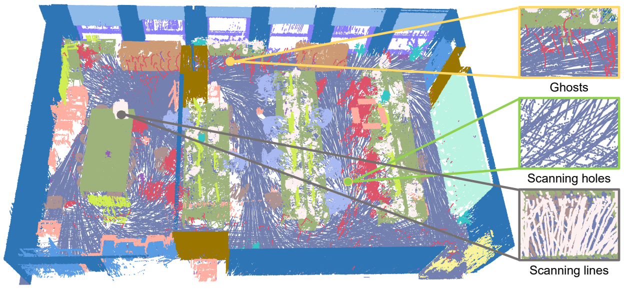

Figure 1. LiDAR-Net comprises 3.6 billion real-scanned points, each annotated with semantic and instance labels. Points of LiDAR-Net have real-scanned characteristics including scanning holes, scanning lines, and anomalies such as ghosts (indicated by red points) caused by moving persons. These characteristics significantly enhance the ability of trained deep models to generalize to real-world applications.

[*] Corresponding author.

#Name	#Year	#Spatial size	#Classes	#Points	# Raw points	# Anomalies	#Sensors
S3DIS [1]	2017	$6{\times}10^3m^2$	13	273M	✗	✗	Matterport
ScanNet [11]	2017	$3.4{\times}10^4m^2$	20	242M	✗	✗	RGB-D
ScanNet++ [53]	2023	$1.5{\times}10^4m^2$	1000+	446M	✗	✗	TLS
LiDAR-Net (Ours)	2023	$3.0{\times}10^4m^2$	24	3619M	✓	✓	MLS

Table 1. In comparison with well-known indoor scene datasets, LiDAR-Net exceeds both S3DIS and ScanNet++ regarding spatial size and resolution. Although ScanNet covers a larger spatial size than ours, we provide points with annotations at a higher resolution. TLS is Terrestrial Laser Scanning system. MLS is Mobile Laser Scanning system.

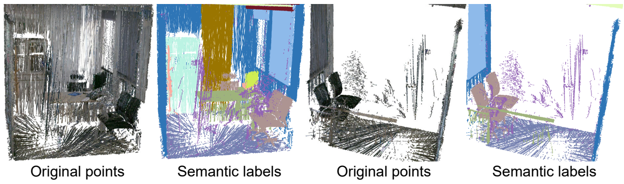

Figure 2. Illustration of reflection noise. These noise points usually suspends in the air. Purple highlights reflection noise in the semantic label figures

The 3D accuracy of these datasets is inferior to that of raw point clouds directly collected using laser scanners. Annotated data in existing datasets also lack characteristic distributions of real-scanned points, including scanning lines and scanning holes, as illustrated in Fig. 1. Consequently, deep models trained on these datasets tend to perform poorly when applied to actual, real-scanned point cloud data, significantly impeding their practical applicability.

In this paper, we contribute a new dataset, LiDAR-Net, to the academic community. Distinct from existing indoor scene datasets, LiDAR-Net is captured by real laser scanners and contains precise semantic and instance annotations for each raw point. LiDAR-Net encompasses three common daily scenes: learning, working, and living environments. It contains 3.6 billion 3D points, covering $30,000m^2$ of indoor space. Using professional LiDAR scanning systems, the 3D point clouds the dataset ensures the high authenticity of its 3D point clouds. Detailed information about the data collection process is provided in Section 3. Across all scenes, each point is carefully annotated and classified into 24 semantic categories, such as floor, ceiling, and window.

Compared to previous indoor scene datasets primarily collected using RGB-D cameras or annotated on reconstructed meshes, LiDAR-Net offers several advantages:

- Firstly, LiDAR-Net distinguishes itself by its authentically captured point cloud data, which leads to non-uniform point distributions, characterized by scanning holes and lines, as depicted in Fig. 1.
- Secondly, LiDAR-Net includes comprehensive annotations of scanning anomalies including reflection noises (purple points in Fig. 2) and ghosts caused by moving objects, particularly persons (red points in Fig. 1).
- Thirdly, LiDAR-Net covers corridors and stairs connecting each room, creating seamless and complete spaces crucial for integrated tasks across entire areas.

Table 1 presents a comparative analysis with other popular indoor scene datasets. While many outdoor point cloud datasets [43, 44] contain real-scanned data, the significant disparity in scene characteristics presents challenges for adapting outdoor datasets to indoor tasks. Moreover, outdoor point clouds generally show a higher level of sparsity and incompleteness compared to LiDAR-Net.

Using LiDAR-Net, we recognize several challenges and conduct studies on them, as outlined in Section 5. We first investigate the process of annotating a large indoor point cloud dataset. Then, we discuss the significance of real-scanned point cloud data for semantic learning, emphasizing the advantages of raw LiDAR-based point clouds over those from depth cameras or from reconstructed meshes. Additionally, we investigate the imbalance of semantic classes in indoor environments. While this paper does not provide exhaustive solutions, it introduces these challenges for further exploration by the research community. Our discussion on the complexities of multi-task learning in indoor 3D point clouds aims to spur developments in fields such as intelligent home systems, digital twins, indoor robotic navigation, asset management in extensive indoor infrastructures, and smart construction sites, *etc.* To summarize, our work makes the following contributions:

- We introduce LiDAR-Net, a comprehensive large-scale 3D point cloud dataset. This dataset is equipped with point-level semantic and instance annotations, making it ideal for multi-task training and testing.
- LiDAR-Net is characterized by its distinct features such as non-uniform point distributions, scanning anomalies, and seamless spatial coverage. These attributes notably enhance the generalization capabilities of trained deep models for real-world applications.
- We assess the efficacy of current algorithms within real-scanned point cloud environments and identify key challenges in this field.

2. Related Work

Booming deep learning-based 3D scene understanding methods desire various 3D scene datasets. Guo et al. [15] give a comprehensive survey on this topic. We briefly summary 3D scene datasets and 3D point cloud scene understanding methods below.

2.1. 3D Scene Datasets

According to the application scenario, existing 3D datasets can be divided into three categories: (i) 3D objects, (ii) indoor scenes, (iii) outdoor scenes.

3D objects. ModelNet [50] and ShapeNet [9] are both 3D synthetic mesh datasets for shape classification. Mo et al. [32] propose PartNet containing hierarchical labels for part segmentation. ABC dataset [22] contains 1M CAD models with structure information. Deitke et al. [12] propose Objaverse-XL, which is a universe of 10M+ 3D objects.

Indoor scenes. Early indoor 3D datasets are acquired by using commodity short-range depth scanners such as NYUv2 [41] and SUN RGB-D [42], comprising short RGB-D sequences with limited resolution. While SceneNN [19] and PiGraphs [39] provide reconstructed and labeled scenes, their scene counts are restricted. Armeni et al. [2] propose S3DIS, which is a widely used benchmark in many tasks, by collecting RGB-D images and reconstructing 3D textured mesh model for each scene. The colored 3D point clouds in S3DIS are densely and uniformly sampled from the textured meshes. ScanNet [11] also provide 3D reconstructions captured by an iPad equipped with a structure sensor, accompanied by annotations. Matterport3D [8] generates low-resolution meshes from panoramic RGB-D images. ARKitScenes [5] offers high-resolution ground truth geometry by using laser scans, but it only provides bounding box annotations. Novel view synthesis methods, such as neural radiance fields (NeRFs) [31], development rapidly in recent years [4, 28]. Yeshwanth et al. [53] propose ScanNet++, containing 460 scenes with DSLR images and iPhone RGB-D frames. They obtain each scene by using the Faro Focus Premium laser scanner and implementing Poisson reconstruction [20, 21] to obtain meshes. Then the meshes are annotated. Comparing to S3DIS [2], ScanNet [11], and ScanNet++ [53], LiDAR-Net provides annotated raw points, which can be directly employed for training deep models specializing in 3D point cloud scene understanding. LiDAR-Net meticulously preserves the characteristics of raw point clouds captured by sensors, encompassing non-uniform distributions like scanning holes and lines, as well as anomalies like reflection noises and ghosts.

Outdoor scenes. These datasets are usually captured by static Terrestrial Laser Scanners (TLS), Mobile Laser Scanners (MLS), and Airborne Laser Scanners (ALS). The representative datasets in this domain comprise roadway-level datasets, such as KITTI [14], Paris-rue-Madame [40], IQmulus [46], Semantic3D [16], Paris-Lille-3D [38], Argoverse [10], SemanticKITTI [6], SemanticPOSS [33], Toronto-3D [44], nuScenes [7], and Waymo dataset [43]. Additionally, urban-level aerial 3D point cloud datasets include DublinCity [57], DALES [47], LASDU [52], Campus3D [24], and SensatUrban [18].

2.2. Deep Learning on 3D Scene Understanding

Pioneered by PointNet [35], neural networks are used to process the unordered point cloud data [25, 36, 49, 56], leading to numerous innovative applications [13, 27, 30, 55]. We choose three representative tasks to evaluate LiDAR-Net dataset. **1) Semantic segmentation.** Thomas et al. [45] propose Kernel Point Convolution (KPConv) operators for 3D point clouds using a set of learnable kernel points. Hu et al. [17] propose RandLA-Net utilizing random point sampling to achieve remarkably high efficiency. **2) Instance segmentation.** ASIS [48] uses a feature fusion structure to enhance both semantic and instance branches by leveraging information from each other. Yang et al. [51] propose a single-stage and anchor-free network called 3D-BoNet for instance segmentation. **3) Object detection.** VoteNet [37] implements a unique voting scheme among points to infer object bounding box centers. GroupFree3D [29] leverages a sophisticated attention mechanism that obviates the need for point candidate grouping, enabling direct inference of object instances.

3. Dataset Acquisition and Annotation

3.1. Point Cloud Acquisition

Acknowledging the marked advantages of LiDAR technology in securing authentic three-dimensional data over traditional RGBD data acquisition methods, our research utilizes Leica BLK2GO, a device combining dual-axis LiDAR technology. The BLK2GO is equipped not only with a commercial dual-axis LiDAR and a 12-megapixel detail camera but also with a 4.8-megapixel, 3-shot panoramic camera. This integration enables stable collection of point cloud data with an accuracy of $\pm 3\, mm@10\, m$ and a capture speed of 420,000 points/second within a $0.5 \sim 25\, m$ range, while simultaneously obtaining high-resolution color images. Significantly, the BLK2GO implements GrandSLAM dual-positioning technology, seamlessly blending laser SLAM and visual SLAM. This approach involves using point cloud from LiDAR and inertial navigation from IMU for one aspect of positioning, and determining the current position and orientation by analyzing disparities between consecutive frames captured by the panoramic camera. For the LiDAR-Net, all data acquisition routes are meticulously pre-planned by professional surveyors, ensuring comprehensive and consistent coverage of the targeted area. To mirror the various interferences typical in real-world point cloud acquisition scenarios, all data are manually collected by human surveyors. Due to the device's battery capacity constraints, each data collection session is strategically limited to approximately one hour. To encompass the entire area, data from multiple individual sessions are manually integrated.

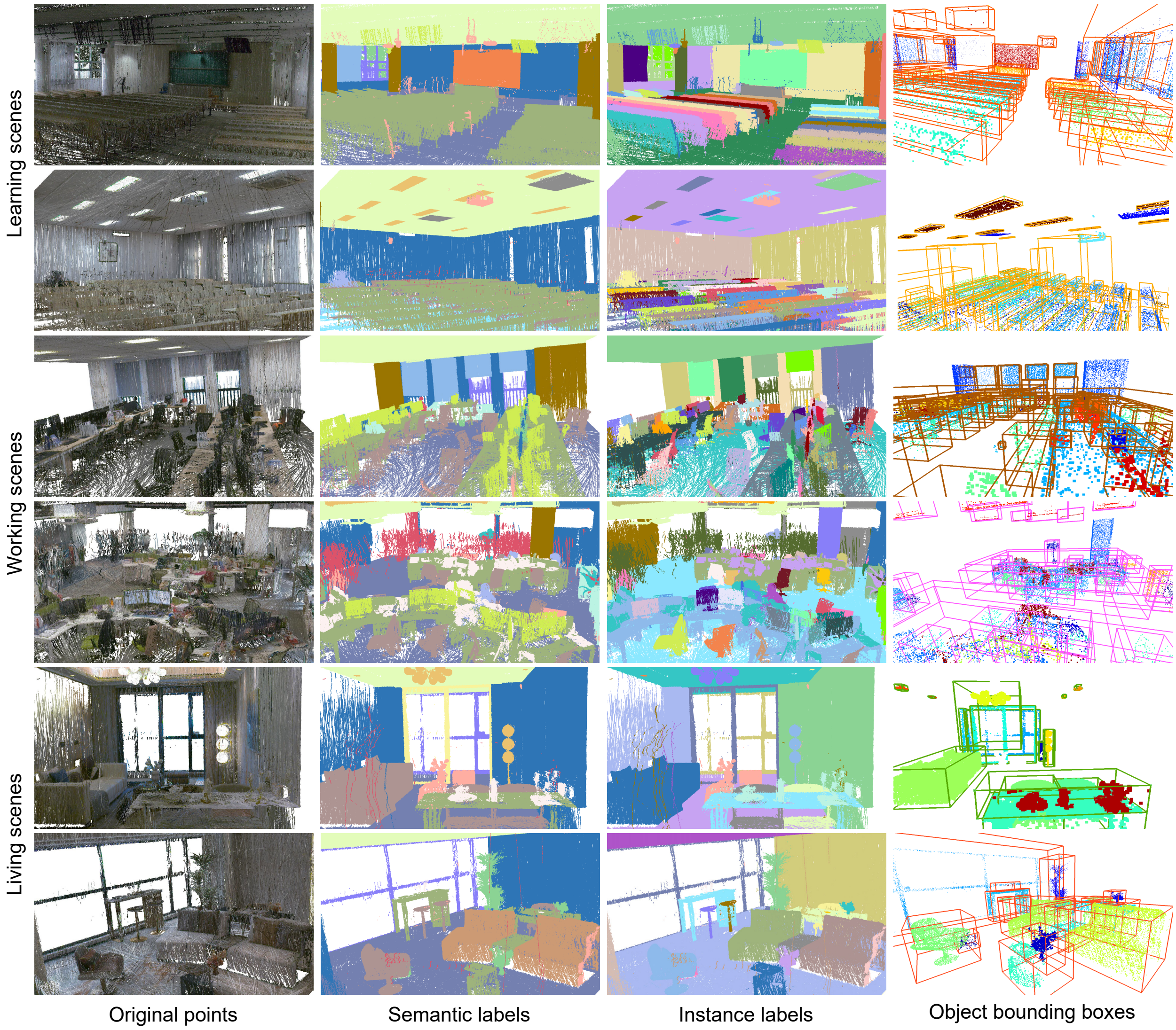

Figure 3. Semantic and instance labels. LiDAR-Net provides point-level annotations directly associated with each raw data point.

3.2. Point Cloud Annotation

In the dataset, we address three sub-tasks: semantic segmentation, instance segmentation, and object detection. The categorization criteria are based on two key principles: 1) Each category should possess a distinct and unambiguous semantic meaning, aligning with both academic research and practical business applications. This alignment facilitates advanced visual tasks such as 3D reconstruction, point cloud completion, and applications in virtual reality, building upon baseline tasks. 2) The categories should exhibit significant geometric or appearance differences, ensuring clear distinction across different groups. We provide a Point Cloud AnnoTation system (PCAT system) to annotate all point clouds in the LiDAR-Net with both semantic and instance labels. Subsequently, object detection labels, in the form of bounding boxes, are derived from these instance labels. The PCAT system is set to be released, allowing users to easily modify or subdivide labels according to their specific requirements. Each label in LiDAR-Net is rigorously manually annotated and verified, ensuring the consistency and high quality of the semantic categories. The entire dataset labeling consumes approximately 550 working hours. Illustrative examples of our annotations are displayed in Fig. 3.

We annotate various indoor components, including floors, stairs, and ceilings, as well as a wide range of furniture and fixtures such as tables, chairs, and sofas. Additionally, we carefully annotate anomalies captured by LiDAR, encompassing reflection noises and ghosts, as they introduce inaccuracies into the data and present challenges in data analysis. Specular reflections from smooth surfaces like glass or metal can disrupt LiDAR systems, resulting in reflection noise points. LiDAR systems would capture incomplete and repetitive ghosts when persons or other objects move in front of the scanner. These anomalies, commonly encountered in real-world applications like robotics and consumer electronics, are accurately recorded in our LiDAR-Net dataset to aid in training deep learning models capable of handling these challenges.

3.3. Benchmark

LiDAR-Net contains classrooms, study lounges, seminar rooms, auditoriums for learning scenes, offices, meeting rooms, lounges for working scenes, and living rooms, bedrooms, dining rooms, washrooms for living scenes. Fig. 4 shows the diversity of scenes. Furthermore, We record world coordinates of rooms to form connecting areas, including 2, 7, and 5 seamless areas of learning, working and living scenes, which facilitate navigation and understanding large-scale indoor scenes.

For the facilitation of GPU-accelerated batch training and testing, we divide the point clouds from various scenes into individual rooms based on their primary structural features, particularly the wall orientations. It is important to note that we provide coordinate correspondences for each room to facilitate tasks working in a whole area. Point clouds from learning scenes are divided into 216 rooms, including 173 rooms for training and 43 for testing. Similarly, we split the working scenes into 206 rooms, including 169 and 37 for training and testing, each room covering approximately $100m^2$. In living scenes, of the 43 rooms analyzed, 33 are used for training and 10 for testing. The distribution of room areas is illustrated in Fig. 5.

LiDAR-Net contains 24 semantic categories, including houseplant (low potted plants), tree (tall trees typically positioned on the floor), person, floor, stair, ceiling, pipe, wall, pillar, window, curtain, door, table, chair, sofa, blackboard (common in learning or working scenes), monitor, bookshelf, wardrobe, bed, light, tabletop others (miscellaneous small objects on tables), reflection noise, ghost (resulting from moving entities, such as people). Additionally, an 'unknown' class exists outside these 24 categories, encompassing points not classified into any specific category. Each point within the 24 semantic categories is assigned instance labels. 17 categories are used for object detection. Floors, ceilings, pipes, walls, pillars, reflection noises and ghosts, which lack definite boundaries, are excluded from object detection task, following [29, 37]. It is important to note that instance labels are provided for these excluded categories, facilitating use cases like smart construction sites and vectorized indoor modeling. The LiDAR-Net dataset will be publicly released, with the goal of establishing benchmarks in indoor point cloud semantic segmentation, instance segmentation, and object detection, which will be evaluated through an online public platform. Following ScanNet [11] and ScanNet++ [53], labels of the test set will remain hidden.

4. Experiments

4.1. Representative Baselines

In Section 2.2, we explore the three main tasks in learning-based 3D point cloud scene understanding. To comprehensively evaluate our LiDAR-Net dataset, we carefully select eight emblematic methods, applying them as benchmarks across different tasks to conduct an extensive benchmark test on our dataset. Precisely, we utilize PointNet [35], PointNet++ [36], KPConv [45], and RandLA-Net [17], for point cloud semantic segmentation, ASIS [48] and 3D-BoNet [51] for instance segmentation, VoteNet [37] and GroupFree3D [29] for object detection.

4.2. Evaluation Metrics

Consistent with prevailing benchmarks, our evaluation framework is designed for a thorough assessment across various segmentation and detection tasks. For semantic segmentation, we focus primarily on mean Intersection-over-Union (mIoU) as the key evaluation metric, following [17, 45]. In instance segmentation, our analysis emphasizes mean precision (mPre) and mean recall rate (mRec) to measure algorithmic performance, following [48, 51]. Moreover, within the domain of object detection, we concentrate on the mean of per-category object detection average precision with 3D IoU threshold 0.25 and 0.5, denoted as mAP@0.25 and mAP@0.5, following [29, 37].

4.3. Benchmark Results

To ensure a fair comparison, we rigorously follow the experimental settings outlined in the original publications of each benchmark algorithm. Table 2 presents the quantitative results across three tasks. The overall performances in segmentation and detection has not reached a satisfactory level. For instance, examining RandLA-Net [17] and KPConv [45], which show the best performance in semantic segmentation, Table 3 details the IoU for each semantic category of the two methods. Table 3 reveals that certain key categories, such as window and bookshelf, were poorly segmented. RandLA-Net achieves 42.02% for the IoU of tree, which is higher than 7.94% of KPConv, however for the IoU of sofa, performance of KPConv is much better than that of RandLA-Net. There is a significant variation in performance across different algorithms within these challenging categories, and no single algorithm consistently outperforms the other. Comprehensive quantitative results of other semantic segmentation, instance segmentation and object detection methods for each category are given in the supplementary materials.

A key advantage of our LiDAR-Net is that its annotated raw points closely mirror the distribution of real-scanned

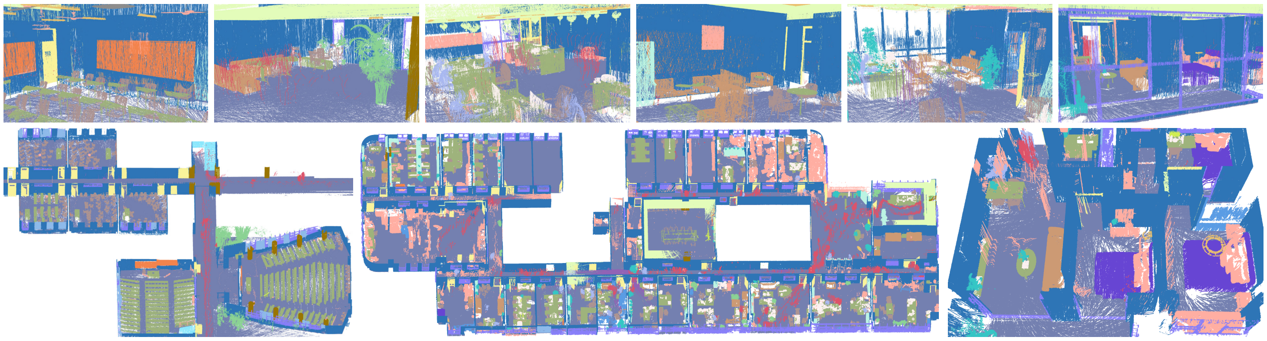

Figure 4. The diversity of scenes. Each point is color-coded by its semantic label. The bottom row displays three seamless areas.

Semantic segmentation	mIoU	Instance segmentation	mPre	mRec	Object detection	mAP@0.25	mAP@0.5
PointNet [35]	16.27	ASIS [48]	37.32	**25.08**	VoteNet [37]	44.83	16.71
PointNet++ [36]	28.81						
KPConv [45]	**44.29**	3D-BoNet [51]	**39.82**	24.63	GroupFree3D [29]	**46.80**	**18.86**
RandLA-Net [17]	32.60						

Table 2. Statistics of selected emblematic methods on LiDAR-Net. The percent sign (%) is omitted.

Methods	Houseplant	Tree	Person	Floor	Stair	Ceiling	Pipe	Wall	Pillar
KPConv [45]	26.98	7.94	45.22	85.43	25.70	86.51	38.49	74.92	5.68
RandLA-Net [17]	3.56	42.02	40.75	93.17	0.42	86.12	19.43	73.81	0.91
	Window	Curtain	Door	Table	Chair	Sofa	Blackboard	Monitor	Bookshelf
KPConv [45]	9.48	22.21	34.44	79.88	75.45	58.02	74.79	55.23	26.76
RandLA-Net [17]	18.00	41.33	51.22	61.69	42.28	0.00	74.77	6.06	7.73
	Wardrobe	Bed	Light	Tabletop others	Reflection noise	Ghost	Unknown	mIoU	
KPConv [45]	34.05	39.51	65.21	40.58	19.22	56.82	18.80	44.29	
RandLA-Net [17]	0.00	0.00	54.13	14.91	0.00	67.53	15.09	32.60	

Table 3. Comparison of IoU for each semantic category. The percent sign (%) is omitted.

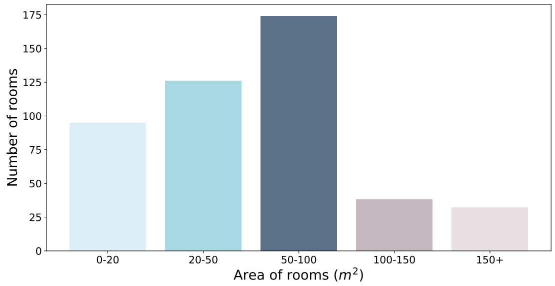

Figure 5. Distribution of room areas. LiDAR-Net encompasses a diverse array of rooms, reflecting a broad spectrum of spatial configurations.

Datasets	Semantic segmentation	Instance segmentation		Object detection	
	mIoU	mPre	mRec	mAP@0.25	mAP@0.5
S3DIS [2]	35.58	15.72	7.84	7.93	0.28
ScanNet [11]	33.18	15.88	14.44	20.91	1.21
ScanNet++ [53]	23.60	26.48	4.68	13.31	1.67
LiDAR-Net	**39.75**	**32.92**	**20.39**	**32.96**	**9.22**

Table 4. Quantitative comparisons of deep models trained on different datasets and tested on real-scanned data. The percent sign (%) is omitted.

points. To evaluate it, we train the benchmark methods on existing popular datasets, including S3DIS [2], ScanNet [11], and ScanNet++ [53]. Then, the deep models are tested on 5 real-scanned rooms, captured using a Leica BLK2GO scanner in locations not included in LiDAR-Net. Due to the categories of the four datasets are different, we test the models on the communal categories. Fig. 6 and Table 4 show the qualitative and quantitative comparisons, selecting RandLA-Net [17] for semantic segmentation, ASIS [48] for instance segmentation, and GroupFree3D [29] for object detection. Detailed statistics of other methods and each category are listed in the supplementary materials. Benefited from the realistic distribution of raw points in LiDAR-Net, deep models trained on it demonstrate superior performances across all three tasks.

5. Challenges

This section focuses on the key challenges presented by our dataset, which inherently stem from the authentic LiDAR collection process. Then, we provide an in-depth analysis of these challenges with a potential to drive advancements in the current field. Concurrently, we delve into tentative discussions of viable solutions. It is crucial to clarify that the aim of this paper is not to introduce a new method for addressing these challenges. Rather, our goal is to identify data biases potentially caused by existing datasets, to shed light on unresolved issues, and to provide analysis and in-

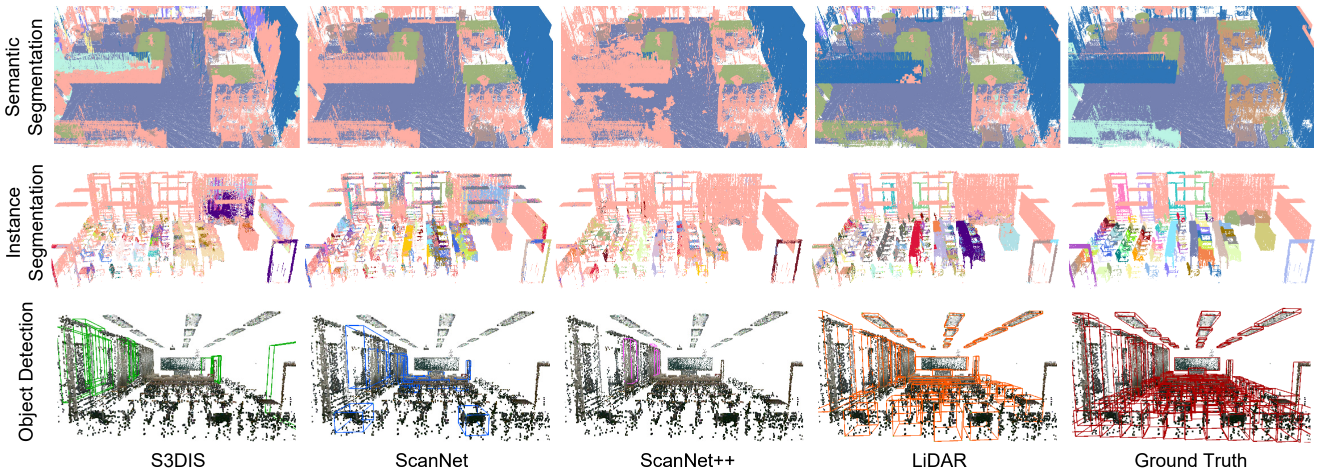

Figure 6. Qualitative comparisons of deep models trained on different datasets and tested on real-scanned data. Each row displays one task. Deep models are tested on the communal categories. It's important to note that ScanNet++ does not include instance labels for floors, walls, and ceilings, and therefore, these elements are excluded from the instance segmentation experiment.

sights. Ultimately, we seek to catalyze advancements in the understanding of indoor scene point clouds.

5.1. Differences in Data Distribution

LiDAR scanning. In previous discussions, we introduced a point cloud dataset derived from authentic LiDAR scanning, which significantly differs from existing indoor point cloud datasets. While LiDAR and RGB-D scanning produce point clouds with some similar attributes, notable differences exist in their characteristics. The non-uniform distribution of real-scanned points, characterized by scanning holes and lines, is a result of the laser scanning process. LiDAR measures distance by emitting laser pulses and measuring the time they take to return. The laser beam moves directionally, creating linear point captures. Gaps in these lines lead to scanning holes. Furthermore, the distance from the scanner affects point distribution. Points closer to the scanner are denser, while those at greater distances tend to be sparser, a phenomenon owing to the dispersion and weakening of laser pulse energy. LiDAR's capability to operate over long distances inherently leads to a noticeable decrease in point density at farther ranges.

RGB-D cameras. Conversely, RGB-D cameras, which utilize structured light or Time of Flight (ToF) technology, determine distance using a distinct methodology. Points closer to the sensor tend to be denser due to variations in observational angles, while farther points show relative sparsity owing to resolution limits. RGB-D cameras are capable of providing relatively dense data within a specific range, but this density noticeably decreases or even disappears beyond their operational limit. Therefore, although both authentic LiDAR and RGB-D cameras exhibit the challenge of variable point density, there is a fundamental difference in their respective point cloud distributions. In the indoor conditions, authentic LiDAR typically encompasses a point distribution radius spanning $2 \sim 7$ meters, in sharp contrast to the $2 \sim 5$ meters typically covered by RGB-D cameras (as illustrated in Fig. 7).

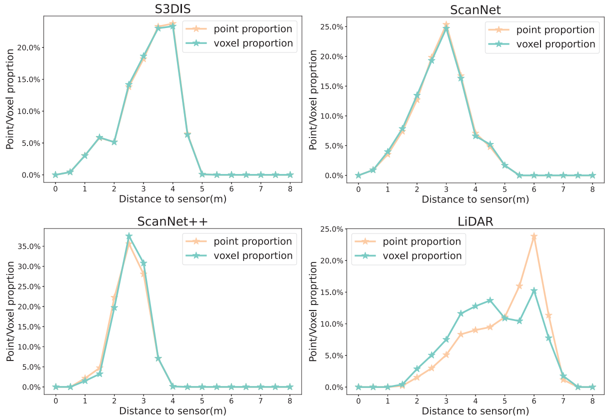

Figure 7. Point and voxel distributions in S3DIS, ScanNet, ScanNet++ and LiDAR-Net.

Comparison of the datasets. Due to the distinctions in acquisition technologies, LiDAR and RGB-D systems exhibit notable discrepancies in factors such as scanning precision, noise, color information, and performance in different lighting conditions. This type of data variance becomes particularly conspicuous in deep learning methodologies heavily reliant upon data. Fig. 7 illustrates the point and voxel distributions in existing indoor datasets compared to LiDAR-Net. Firstly, the three existing datasets predominantly contain points within a $2 \sim 5$ meters range from the scannner, while our LiDAR-Net provides a distinct distribution. It is worth to be explored that would the deep models trained on previous datasets over-fitted to the distribution of $2 \sim 5$ meters? Our LiDAR-Net, with its broader range of $2 \sim 7$ meters, provides an expanded scope for future research in this area. Secondly, since the points in the previous three datasets are generated by sampling on annotated 3D reconstructions, they mirror the voxel distribution. In contrast, the primary strength of LiDAR-Net lies in its annotated raw points, which are non-uniform, leading to distinct point and voxel distributions. This divergence introduces new challenges to the field. Our dataset adeptly encapsulates these genuine characteristics of real-world data, thereby establishing a robust foundation for various forthcoming applications in indoor perception technology.

5.2. Impact of Imbalance Class Distribution

Upon further investigation, we discern a more critical issue within indoor scene datasets: a conspicuous performance discrepancy resulting from imbalanced class distribution. We showcase the cumulative count of 3D points per semantic category in the top of Fig. 8, and the instance distribution is shown in the bottom. It is observable that the primary semantic categories, including ceilings, floors, walls, *etc.*, make up over 80% of the total points. In contrast, smaller yet crucial categories (*e.g.*, monitors) constitute a mere 0.16% of the total points, highlighting a stark imbalance in the distribution of semantic classes. This disproportion is inherent to indoor scenes, where walls, ceilings, and floors naturally cover large surface areas, while items like monitors are considerably smaller.

Data imbalance poses a significant challenge in 3D scene understanding, an issue that is also evident in LiDAR-Net. As shown in Table 3, the skew in segment headers is especially marked in the semantic segmentation task, owing to the disparity in point cloud quantities at the point level. For instance, looking at the results of RandLA-Net in specific categories, the semantic segmentation accuracy for floors is around 93%, while for stairs, it nearly drops to zero. Essentially, indoor scenes are dominated by a few categories such as ceilings and walls, whereas categories like stairs, though minor but critical, have an extremely sparse data distribution. When considering instance segmentation, the combination of instance-level and point-level imbalances exacerbates the problem, leading to a more severe dual-imbalance situation. Detailed instance segmentation results of each category are listed and discussed in the supplementary materials. This profound imbalance in distribution poses another significant challenge introduced by our dataset.

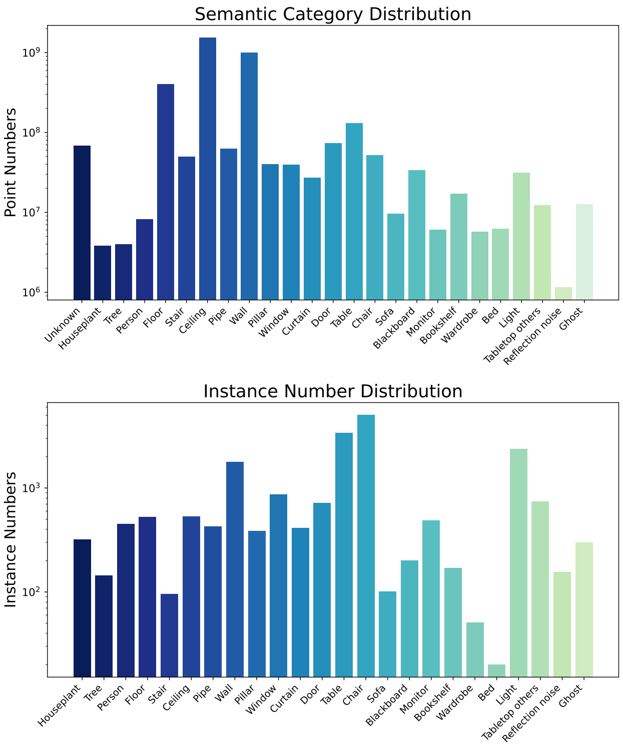

Figure 8. Top: Point number distribution of semantic categories. Bottom: Instance number distribution. Unknown points do not have instance labels.

6. Conclusion

In this study, we present a comprehensive indoor point cloud dataset derived from authentic LiDAR scanning. This dataset encompasses three meticulously labeled scenes with a total of 3.6 billion points, covering 30,000m^2. Designed for a variety of applications including semantic segmentation, instance segmentation, and object detection, our dataset serves as a versatile resource for related research tasks. Our comprehensive benchmark testing not only highlights persistent challenges but also emphasizes the distinct advantages of our dataset, particularly regarding the differences between annotated raw points from genuine LiDAR and points sampled from annotated 3D reconstructions. We have thoroughly examined issues such as the pronounced effects of imbalanced class distribution and the adaptability of models to unseen real-scanned scenes. Looking towards a future where autonomous indoor robots navigate diverse human-centric environments, we underscore the importance of real-time perception using authentic LiDAR-scanned data. The precision and high resolution of real-world LiDAR point cloud data are crucial for emerging cyber-physical systems like smart homes and digital twins. We believe our dataset and benchmarks will spur further progress in related areas of study.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62032011), the Natural Science Foundation of Jiangsu Province (No. BK20211147), and the program B for Outstanding PhD candidate of Nanjing University.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1543, 2016. 2

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1, 3, 6

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6359–6367, 2020. 1

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 3

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 3

[6] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019. 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 3

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 3

[9] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[10] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8748–8757, 2019. 3

[11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 1, 2, 3, 5, 6

[12] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 3

[13] Gil Elbaz, Tamar Avraham, and Anath Fischer. 3d point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2017. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361. IEEE, 2012. 3

[15] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(12):4338–4364, 2020. 2

[16] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d. net: A new large-scale point cloud classification benchmark. *arXiv preprint arXiv:1704.03847*, 2017. 3

[17] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11108–11117, 2020. 3, 5, 6

[18] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343, 2022. 3

[19] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, pages 92–101. IEEE, 2016. 1, 3

[20] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):1–13, 2013. 3

[21] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, page 0, 2006. 3

[22] Sebastian Koch, Albert Matveev, Zhongshi Jiang, Francis Williams, Alexey Artemov, Evgeny Burnaev, Marc Alexa, Denis Zorin, and Daniele Panozzo. Abc: A big cad model dataset for geometric deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9601–9611, 2019. 3

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders

for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12697–12705, 2019. 1

[24] Xinke Li, Chongshou Li, Zekun Tong, Andrew Lim, Junsong Yuan, Yuwei Wu, Jing Tang, and Raymond Huang. Campus3d: A photogrammetry point cloud benchmark for hierarchical understanding of outdoor scene. In *Proceedings of the ACM International Conference on Multimedia*, pages 238–246, 2020. 3

[25] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[26] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (8):3412–3432, 2020. 1

[27] Yuanqi Li, Shun Liu, Xinran Yang, Jianwei Guo, Jie Guo, and Yanwen Guo. Surface and edge detection for primitive fitting of point clouds. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 3

[28] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 15651–15663, 2020. 3

[29] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2949–2958, 2021. 3, 5, 6

[30] Changfeng Ma, Yinuo Chen, Pengxiao Guo, Jie Guo, Chongjun Wang, and Yanwen Guo. Symmetric shape-preserving autoencoder for unsupervised real scene point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13560–13569, 2023. 3

[31] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[32] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. 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 (CVPR)*, pages 909–918, 2019. 3

[33] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 687–693. IEEE, 2020. 3

[34] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 523–540. Springer, 2020. 1

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1, 3, 5, 6

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 3, 5, 6

[37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9277–9286, 2019. 1, 3, 5, 6

[38] Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. Paris-lille-3d: A large and high-quality ground-truth urban point cloud dataset for automatic segmentation and classification. *The International Journal of Robotics Research*, 37(6):545–557, 2018. 3

[39] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 3

[40] Andrés Serna, Beatriz Marcotegui, François Goulette, and Jean-Emmanuel Deschaud. Paris-rue-madame database: a 3d mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. In *International Conference on Pattern Recognition, Applications and Methods (ICPRAM)*, 2014. 3

[41] Nathan Silberman and Rob Fergus. Indoor scene segmentation using a structured light sensor. In *IEEE/CVF International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 601–608. IEEE, 2011. 3

[42] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576, 2015. 1, 3

[43] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020. 2, 3

[44] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 202–203, 2020. 2, 3

[45] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 3, 5, 6

[46] Bruno Vallet, Mathieu Brédif, Andrés Serna, Beatriz Marcotegui, and Nicolas Paparoditis. Terramobilita/iqmulus urban point cloud analysis benchmark. *Computers & Graphics*, 49:126–133, 2015. 3

[47] Nina Varney, Vijayan K Asari, and Quinn Graehling. Dales: A large-scale aerial lidar data set for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 186–187, 2020. 3

[48] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4096–4105, 2019. 3, 5, 6

[49] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019. 1, 3

[50] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1912–1920, 2015. 3

[51] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 3, 5, 6

[52] Zhen Ye, Yusheng Xu, Rong Huang, Xiaohua Tong, Xin Li, Xiangfeng Liu, Kuifeng Luan, Ludwig Hoegner, and Uwe Stilla. Lasdu: A large-scale aerial lidar dataset for semantic labeling in dense urban areas. *ISPRS International Journal of Geo-Information*, 9(7):450, 2020. 3

[53] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–22, 2023. 1, 2, 3, 5, 6

[54] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021. 1

[55] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2790–2799, 2018. 3

[56] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 1, 3

[57] SM Zolanvari, Susana Ruano, Aakanksha Rana, Alan Cummins, Rogerio Eduardo da Silva, Morteza Rahbar, and Aljosa Smolic. Dublincity: Annotated lidar point cloud and its applications. *arXiv preprint arXiv:1909.03613*, 2019. 3