Campus3D: A Photogrammetry Point Cloud Benchmark for Hierarchical Understanding of Outdoor Scene

Xinke Li[1], Chongshou Li[1,*], Zekun Tong[1], Andrew Lim[1], Junsong Yuan[2]
Yuwei Wu[1], Jing Tang[1], Raymond Huang[1]
[1]Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore
[2]Department of Computer Science and Engineering, State University of New York at Buffalo, Buffalo, NY, USA
{xinke.li,zekuntong}@u.nus.edu,{iselc,isealim}@nus.edu.sg,jsyuan@buffalo.edu
ywwu@u.nus.edu,{isejtang,raymond.huang}@nus.edu.sg

ABSTRACT

Learning on 3D scene-based point cloud has received extensive attention as its promising application in many fields, and well-annotated and multisource datasets can catalyze the development of those data-driven approaches. To facilitate the research of this area, we present a richly-annotated 3D point cloud dataset for multiple outdoor scene understanding tasks and also an effective learning framework for its hierarchical segmentation task. The dataset was generated via the photogrammetric processing on unmanned aerial vehicle (UAV) images of the National University of Singapore (NUS) campus, and has been point-wisely annotated with both hierarchical and instance-based labels. Based on it, we formulate a hierarchical learning problem for 3D point cloud segmentation and propose a measurement evaluating consistency across various hierarchies. To solve this problem, a two-stage method including multi-task (MT) learning and hierarchical ensemble (HE) with consistency consideration is proposed. Experimental results demonstrate the superiority of the proposed method and potential advantages of our hierarchical annotations. In addition, we benchmark results of semantic and instance segmentation, which is accessible online at https://3d.dataset.site with the dataset and all source codes.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; **Neural networks**; **3D imaging**.

KEYWORDS

Point cloud; scene understanding; hierarchical learning; semantic segmentation; instance segmentation

ACM Reference Format:
Xinke Li[1], Chongshou Li[1,*], Zekun Tong[1], Andrew Lim[1], Junsong Yuan[2] and Yuwei Wu[1], Jing Tang[1], Raymond Huang[1]. 2020. Campus3D: A Photogrammetry Point Cloud Benchmark for Hierarchical Understanding of Outdoor Scene. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413661

*Corresponding author: Chongshou Li (iselc@nus.edu.sg).

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413661

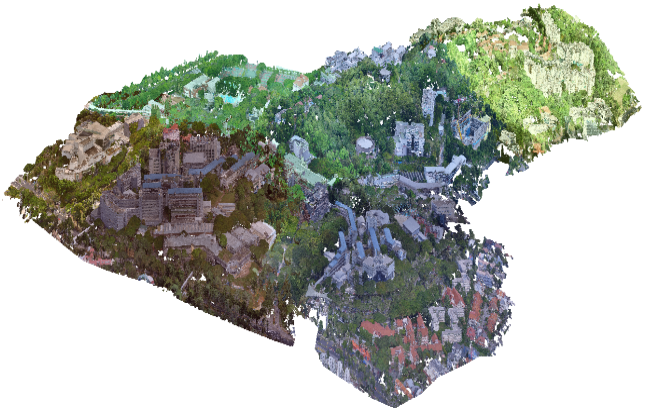

Figure 1: Overview of six regions of the Campus3D dataset.

1 INTRODUCTION

Due to the significant progress of 3D sensoring technologies in recent years, multiple sources of 3D point cloud become affordable and easily acquired. Reconstruction of outdoor scene from point cloud has also received an increasing interest, which is critical for various areas such as urban planning and management [5], vehicle navigation [4], virtual reality [7] as well as simulation [21]. As the fundamental step of reconstruction, scene understanding with point cloud data can be greatly facilitated by recent advances of machine learning techniques especially the deep learning. Large and well-annotated datasets play a leading role for the successful application of these techniques.

Although dozens of 3D scene-based point cloud datasets are proposed [1, 3, 9, 10, 14, 25, 26, 28, 31], majority of them are not perfectly fit for outdoor scene reconstruction. Firstly, the datasets may face various limitations from their sources which are are either RGB-D images [1, 9, 10, 28] or light detection and ranging (LiDAR) based mobile laser scanning (MLS) [3, 25, 26, 31] and terrestrial laser scanning (TLS) [14]. The RGB-D data can be easily obtained and processed via a mature pipeline [9], while it is likely prevented from capturing outdoor environment by the limited measurement range. The LiDAR scanner usually results in unavoidable severe occlusions and expensive equipment costs although it is good at capturing large-scale scenes [19]. Secondly, the annotations of extant datasets are not targeted for outdoor scene reconstruction. Following the

well-established data format CityGML [17], a standard urban model should contain fine structures of building and other artifacts. However, such fineness is not presented by current annotations which mainly consist of indoor objects or traffic elements [3]. Thus, it is necessary to build new datasets with the aim of supporting scene understanding based automatic reconstruction.

In this work, we construct a photogrametry point cloud dataset Campus3D from UAV imagery over the National University of Singapore (NUS) campus of 1.58 km^2 area. Due to the recent progress of Structure from Motion (SfM), Multi-View Stereo (MVS) and UAV techniques [11, 39], photogrammetry point cloud is easily accessible from unmanned aerial vehicle (UAV) imagery. This type of data source is able to fulfill the requirement of scene reconstruction because UAV imagery is robust to occlusion, and can effectively obtain the holistic view of the scene.

Inspired by the multiple levels of details (LoD) in CityGML [17] for the reconstruction, we point-wisely annotate this dataset with hierarchical multi-labels for both semantic and instance segmentation. For a data point, an example annotations is construction->building->wall/roof. The fine-grained label (e.g., wall/roof) can match the LoD2 for reconstruction [32], where the building model is detailed to roof and wall structures. For the further study on it, we organize the labels as a tree with five hierarchical (granularity) levels displayed by Figure 3. In the end, the whole dataset present a holistic view of scene, which contains 0.94 billion points with 2,530 modality-based instances, 24 semantic classes and 6 pattern-based regions as displayed by Figure 1.

The proposed dataset with hierarchical annotations is expected to promote better outdoor scene understanding. Based on the constructed label tree, we formulate a hierarchical learning (HL) problem for semantic segmentation, and propose a new metric for consistency across granularity levels named *Consistency Rate* (CR). Besides accuracy, prediction consistency is an important issue for the HL. For example, if one point is predicted as "*roof*" at fine-grained level, the results at the corresponding coarse level must be "*building*" and "*construction*"(see Figure 3), otherwise, it is a violation of the hierarchical relationship. Taking this into consideration, we introduce a two-stage method consist of multi-tasking (MT) learning and hierarchical ensemble (HE). The MT based on neural models jointly learns semantic labeling on different granularity levels. The post-processing HE rigidly ensures the results to fulfill the hierarchical consistency by choosing the most likely root-to-leaf path of the label tree. The results of CR and segmentation task suggest the goodness that the HL method utilizes the hierarchical relationship and the chance that hierarchical annotations assists segmentation tasks.

Furthermore, we establish the benchmarks on the dataset via applying deep models for two classic scene understanding tasks: (1) semantic segmentation and (2) instance segmentation. For the concern of computational efficiency and compatibility to point-based models, we investigate the data prepossess technique and two sampling methods: (1) random block sampling (RBS) and (2) random centered K nearest neighbor (RC-KNN) sampling. And the RBS is chosen as the unified sampling method for benchmarks in view of its better performance.

We summarize the contributions of this paper as follows:

- A photogrammetry point cloud dataset with hierarchical and instance-based annotations is present. Moreover, an accessible workflow of the acquisition and annotation is provided.
- An effective two-stage method for the formulated hierarchical semantic segmentation on point cloud is proposed. Experimental results demonstrate the superiority of our HL methods over the non-HL method in terms of both hierarchical consistency and segmentation performance.
- We propose new benchmarks for semantic segmentation and instance segmentation on 3D point cloud, and release the source codes[1] of the training/evaluation framework as well as the dataset. These benchmarks are standardized with consideration of the unified data prepossess techniques and sampling methods.

2 RELATED WORK

In this section, we firstly review the existing 3D scene-based point cloud datasets and compare our dataset with them in detail below. Based on the application area, we briefly divide the existing datasets into two categories: (1) indoor scene datasets and (2) outdoor scene datasets. A summary of comparison between Campus3D and the widely-used datasets is provided by Table 1, and additional comparisons in terms of annotation are provided in the supplementary document. Secondly, we briefly review the existing deep neural models for point cloud segmentation.

Indoor Dataset. Indoor scene understanding is an active research area, and many datasets have been reported in literature [1, 2, 6, 9, 16, 27–29, 40]. These datasets are usually generated by RGB-D images which can be easily got by cheap sensors (e.g., Microsoft Kinect). Early datasets NYUv2 [28], SUN3D [40] and SUN RGB-D [29] were annotated by either polygons in 2D [28, 29, 40] or bounding box in 3D [29], of which the information for 3D scene reconstruction (e.g., semantic segmentation, surface reconstruction, meshes, etc.) is limited. Recently released indoor scene datasets [1, 2, 6, 9, 16] contain more information. For instance, ScanNet [9] supplies estimated camera parameters, surface segmentation, textured meshes and semantic segmentations; however, comparing with the proposed photogrammetry dataset Campus3D, these datasets generated by RGB-D sensors have their limitations of short measurement range and sensitivity to the sunlight's infrared spectrum [14]. These natural limitations prevent the RGB-D datasets from applications of outdoor environment understanding.

Outdoor Dataset. Several outdoor scene 3D datasets [14, 25, 26, 31] are released in recent years. These datasets are generated via either MLS [3, 25, 26, 31] or TLS [14]. Points generated by the LiDAR are the raw output of the laser scanner, which are of high quality and large scales. The MLS point cloud datasets are always annotated with rich traffic elements to push the frontier of the autonomous driving field. One notable MLS point cloud dataset is a part of KITTI which was constructed by Geiger et al. [12, 13] and generated from 6 hours of traffic scenarios. Based on it, a point cloud dataset, semanticKITTI, has been proposed recently for outdoor semantic scene understanding [3]. However, different from our

[1] https://github.com/shinke-li/Campus3D

Table 1: Comparison between Campus3D and popular scene-based point cloud datasets.

Dataset	Data Source Type	Area/Length	Scene Type	Point #	Designed 3D Task
ScanNet [9]	RGB-D	floor: 34,453 m^2 surface: 78,595 m^2	Indoor	-	Object classification; Instance & semantic segmentation; CAD model retrieval
S3DIS[2]	RGB-D	6000 m^2	Indoor	695.9M	Object detection
Matterport3D[6]	RGB-D	floor: 46,561 m^2 surface: 219,398 m^2	Indoor	-	Instance & semantic segmentation
SemanticKITTI [3]	Velodyne HDL-64E (MLS)	39.2 km	Outdoor	4,549M	Semantic segmentation Semantic scene completion
Semantic3D[14]	Terrestrial Laser Scanner (TLS)	-	Outdoor	4,000M	Semantic segmentation
Paris-Lille-3D[25]	Velodyne HDL-32E (MLS)	1940 m	Outdoor	143.1M	Instance & semantic segmentation
Campus3D (Ours)	**UAV photogrammtry**	**1.58** $\times 10^6$ m^2	**Outdoor**	**937.1M**	**Hierarchical semantic segmentation** **Instance segmentation**

dataset collected by UAV imagery, LiDAR devices always suffer from occlusions thus lack for a holistic view of the scene.

Deep Segmentation Model. Semantic segmentation and instance segmentation are the major scene understanding tasks concerned by reconstruction. As the pioneering work PointNet and PointNet++ proposed by Qi et al. [23, 24], point-based deep neural models come to widely-studied in point cloud segmentation field since it can directly process the point cloud. Categories of point-based deep learning models mainly include feature pooling models [15, 23, 24, 43], convolution-based models [20, 30, 34], graph-based models [18, 33, 37] and attention-based models [41, 42]. Although most of these models are proposed for a single task, they can be involved in multitask learning. The examples are PointNet++ in ASIS [36] and PointNet in JSIS3D [22], which jointly learn instance embedding and semantic labeling in one structure. To jointly learn semantic labeling on multiple granularity levels, we propose our modification of point-based models fitting for multitask learning. The PointNet++ is applied as backbone since its general structure and high compatibility to multitask [35, 36].

3 CAMPUS3D DATASET

We note that the Campus3D is online accessible. Not only data can be downloaded there but also online interactive visualization and Github link for source codes are provided.

3.1 Data Acquisition

The point cloud of Campus3D dataset was constructed by the technique of Structure from Motion with Multi-View Stereovision (SfM-MVS) [38] on UAV images. Here we briefly describe our workflow for obtaining the dataset. Firstly, we flew drones over all areas and took images with exact GPS coordinates. The device to capture imagery was DJI Phaton 4 Pro drones equipping cameras with a 1-inch 2 MP CMOS sensors, and the drone flight planning mobile apps used in our application were DJI GS Pro and Pix4D Capture. Then the points would be generated by photogrammetry processing and registration from captured images and coordinates using Pix4D as SfM-MVS software.

In image collection process, we applied two types of flight routing strategies for UAV photography: (1) grid and (2) circular, which were accessible in the drone flight planning mobile apps. For relatively high buildings, we applied multiple circular flights at different height levels. During UAV image capturing, the drone were flown when the clear view was guaranteed by the weather. More detailed settings can be found in the supplementary document.

3.2 Data Annotation

To present more complicated geometric features, we annotated the point cloud with point-wise labels. In general, there are two approaches to perform 3D point-wise annotation: (1) label the pre-segmented clusters in 3D and (2) label the projected 2D image and assign labels to 3D points. Our strategy follows the second approach and performs a two-level of 2D projection segmentation, which avoids inherent error induced by pre-segmenting methods and lack of details in 2D projections of stationary angles. Initially, we divided the annotation tasks into hierarchical stages from coarse-grained label to fine-grained label. In each stage, annotation was firstly done by 2D polygon partitions in three orthogonal view-angles. To refine the details, the obtained 3D partitions were then pruned in user-defined rotation angles. All the tasks were completed by opensource software CloudCompare [8] and its add-on functions. Multiple annotators were hired to perform above labeling task after taking training courses for days. To ensure the accuracy and consistency of annotations, we divided annotators to several groups, and work on labeling and verifying, respectively, for each stage. We require that every point is labeled at least three times by different annotators and verified to an exact label.

According to CityGML [17], objects from urban scene are modeled in different granularity levels defined by the LoD, which can cope with applications in different scales. Motivated by this concept, the category labels used in the Campus3D are constructed as a hierarchical structure with various granularity levels and displayed by Figure 3. The hierarchies of the structure can work similarly to the LoDs. Each label is formed based on two criteria: (1) semantic attribute and (2) geometrical attribute. They may mutually assist each other to parse the points into refined parts. For example of both "*roof*" and "*driving_road*" with plane structure, they are difficult to be distinguished in geometric features but need to be separated due to the semantic difference and practical function. All labels are self-explanatory except for the following ones. And we provide explanations for them: (1) "*unclassified*" refers to unrecognized or over-sparse regions. Instead of removing these data-points, this category is set for reserving the completeness of dataset; (2) "*path&stair*"

is only for pedestrians while "*driving_road*" is only for vehicles; (3) "*artificial_landscape*" is referring to man-made landscape such as artificial pool while "*others*" represents some individual objects because there do not exist enough instances to group them as a new category. All the labels are defined in a rigid way for consistency of annotation.

3.3 Parsing and Statistics

To label the raw point clouds, we propose a hierarchical parsing method for decomposing the data into individually labeled points, which is naturally generated by the hierarchical annotation in previous section. The resulting Campus3D dataset can fulfill multiple tasks. We firstly divide the entire dataset into six identified regions: FASS, FOE, PGP, RA, UCC and YIH according to their architecture styles and functions. A descriptive summary of points of these six regions is given by Table 2.

Table 2: Area, mean height, points and points per area of data points of each region.

Region	Area (m^2)	Mean height (m)	# of points	# of points per area (m^2)
FASS	276,746	48.74	114,599,515	414.10
FOE	247,924	49.76	34,347,821	138.54
PGP	277,468	50.19	29,595,347	106.66
RA	365,537	61.62	54,446,114	148.95
UCC	127,572	30.09	333,404,689	2613.46
YIH	284,775	42.08	354,491,876	1244.81

Due to the hierarchical annotation strategy, class labels of the Campus3D can be defined by a tree-like structure. Based on the this structure, the coarse-grained level data can be simply obtained by merging their sub-class data including all leaf node, which is flexible for multi-level tasks. For example, class "*building*" data could be obtained by merging "*wall*" and "*roof*" data. After labeling each point by a hierarchical class tree, we performed instance labeling for each countable class, which may benefit 3D model reconstruction and scene understanding. For instance, to boost the LoD of the building model, it is necessary to distinguish various planar pieces from a roof. Figure 2 illustrates this parsing. We also note that more descriptive statistics of class and instances are provided in the supplementary document.

3.4 Data Preprocessing

To practically perform the machine learning algorithms on the data, we need data simplification on point cloud with consideration of imbalanced density and processing efficiency. We provide a reduced dataset from the original points. This reduced dataset is voxelly sampled from the original dataset with a sampling size of 0.15 meter. The sampling method thins the data points and also inhibits the imbalanced distributions of points among different regions and instances, which is caused by the varies of morphology. Moreover, the 0.15m sample size can keep the smallest object in the whole campus. We term this dataset as Campus3D-reduced. Note that all experimental studies, scene understanding tasks and benchmarks in this paper are run on the Campus3D-reduced. Table 3 shows the training, validation and test splits. This splitting makes sure that training set and test/validation set have all types of instances. And the performance of the class "*unclassified*" is not included in current study, which follows the convention in this arena [20, 23, 24].

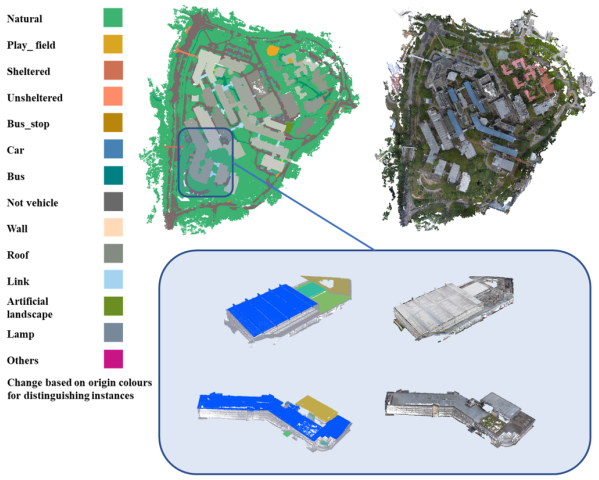

Figure 2: Instance segmentation of building and roof. Upper left: annotated ground truth instance; upper right: raw point cloud. The bottom is a zoomed-in example of instance annotation and different colors represent different roof pieces. And there are two building instances.

Table 3: Training, validation and test set splitting

	Training	Validation	Test
Region	FASS, YIH, RA, UCC	PGP	FOE

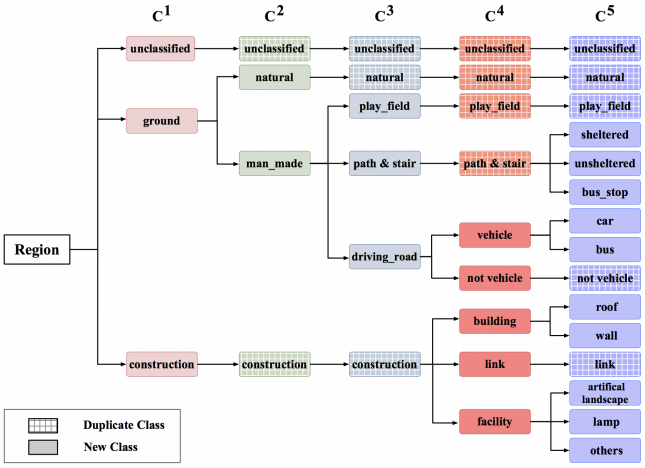

Figure 3: Label tree $\mathcal{T}$: each internal or leaf node with solid fill represents a class; the class name of each is inside each node. Each data point of Campus3D dataset is annotated by a path of the tree with solid nodes. e.g. construction -> building -> wall. For entirely partitioning the data in each level, some nodes are duplicates of its parent nodes which are filled by grids. The tree has five ($H = 5$) granularity levels: $C^1 = \{\textbf{\textit{unclassified}}, \textbf{\textit{ground}}, \textbf{\textit{construction}}\}, \cdots, C^5 = \{\textbf{\textit{unclassified}}, \textbf{\textit{natural}}, \cdots, \textbf{\textit{others}}\}$.

4 HIERARCHICAL LEARNING

In order to learn on hierarchical annotations of our dataset, we construct a five-level label tree displayed by Figure 3, where labels in each hierarchy can completely partition the entire dataset. In that case, each point possesses five parallel semantic labels, learning of which can be consider as a multi-label segmentation tasks. Compared with single label learning, the key problem towards hierarchical multi-label learning is how to leverage the relationship among hierarchies, while the hierarchical structure of labels should be kept. Therefore, we propose a simple yet effective framework, which includes a multi-task learning network and an ensemble process to maintain hierarchical structure. Before the methodology, we first proceed to the problem and performance metrics.

4.1 Problem and Metric Description

Let $(\mathbb{C}, \leq_\eta)$ represent the class hierarchy, where $\mathbb{C}$ is a set of classes and $\leq_\eta$ is a partial order representing the superclass relationship. For any $c_1, c_2 \in \mathbb{C}$, $c_2 \leq_\eta c_1$ if and only if c_1 is a superclass of c_2 or $c_1 = c_2$. Data point i with hierarchical annotation is denoted as (X^i, S^i) with $X^i \in \mathcal{X} = \mathbb{R}^D$ and S^i is a maximal chain of $\mathbb{C}$. The problem of such label is that length of the label set $|S_i|$ is not coherent from point to point. To construct a multi-label with coherent length, we further extend the definition of hierarchical learning by allowing duplication. We first notate the set of all maximal elements in $\mathbb{C}$ by C_{max} and the set of all minimal elements C by C_{min}. Note that C_{max} and C_{min} both belong to $\mathcal{A}_\mathbb{C}$, the set of all antichains in $\mathbb{C}$. We define a relationship to compare the any two antichains named parent antichain:

Definition 4.1 (Parent Antichain). For two distinct sets $C_c, C_p \in \mathcal{A}_\mathbb{C}$, if $\forall c_j \in C_c$, $\exists c'_j \in C_p$ let $c_j \leq_\eta c'_j$, then C_p is called a parent antichain of C_c with notation $C_c \prec_\eta C_p$.

Then we can obtain a sequence of antichains (sets) between $C_{\min}$ and $C_{\max}$ if $C_{\min} \neq C_{\max}$, namely, $\{C^h\}_{h=1}^H = (C^1 = C_{\max}, C^2, \ldots, C^H = C_{\min})$ with length $H = \max |S_i|$ that $C^H \prec_\eta C^{H-1}, \ldots, \prec_\eta C^1$ and $\cup_{h=1}^H C^h = \mathbb{C}$. Based on the sequence, a tree $\mathcal{T}$ can be constructed and displayed in Figure 3. The nodes in h^{th} layer of the tree can be associates with C^h, while the edge is defined as the partial order relationship between classes. Now we define the hierarchical learning problem. A dataset $\mathcal{D} = \{(X^i, Y^i) | i \in \mathbb{Z}, 1 \leq i \leq N\}$, where N is the number of points, $X^i \in \mathcal{X}$ and $Y^i \in \mathcal{Y} = C^1 \times C^2 \times \ldots \times C^H$. The hierarchical learning problem is to learn a function $f: \mathcal{X} \mapsto \mathcal{Y}$ from a hierarchically annotated dataset $\mathcal{D}$.

Given a HL method $f(\cdot)$, the performance can be evaluated by the conventional classification measurements such as accuracy, precision, recall, etc. However, they fail to take consistency into account which is critical for the HL. It is possible that a HL algorithm performs good in terms of conventional measurements, but generates highly inconsistent results violating the hierarchical relationship which are meaningless. Therefore, we propose a new measurement and quantitatively evaluate such consistency. Considering a solution (prediction) $Y \in \mathcal{Y}$, we first define the *fully consistent (FC)* for a solution at Definition 4.1. The set of all FC solutions is denoted as $\mathcal{Y}^{\text{FC}}$ ($\mathcal{Y}^{\text{FC}} \subset \mathcal{Y}$), and it includes all paths from root to leaf nodes in tree $\mathcal{T}$. Based on it, we propose, *consistency proportion (CP)*, to measure the consistency degree for solution Y^i. The CP value is between 0 and 1, and being one represents a FC solution. Then, for a set of solutions $\{Y^1, \ldots, Y^N\}$, the *consistency rate (CR)* is defined with parameter α being the desired consistency level for each solution.

Definition 4.2 (Fully Consistent). Solution $Y = (y_1, \ldots, y_H) \in \mathcal{Y}$ is defined as fully consistent (FC) if it satisfies $y_H \leq_\eta y_{H-1} \leq_\eta \ldots \leq_\eta y_1$.

Definition 4.3 (Consistency Proportion). The consistency proportion (CP) of $Y^i = (y_1^i, \ldots, y_H^i)$ is defined as:

$$\text{CP}(Y^i) = \frac{\max\limits_{(y_1,\ldots,y_H) \in \mathcal{Y}^{\text{FC}}} \sum_{h=1}^H \mathbb{1}(y_h^i = y_h)}{H}, \tag{1}$$

here $\mathbb{1}(x) = 1$ if x is True; 0 otherwise.

Definition 4.4 (Consistency Rate). The consistency rate (CR) with CP level α for $\{Y^1, \ldots, Y^N\}$ is:

$$\text{CR}_\alpha = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[\text{CP}(Y^i) \geq \alpha]. \tag{2}$$

Here α is a threshold parameter and $0 \leq \alpha \leq 1$.

4.2 Methodology

We propose a two-stage framework to the HL (see Figure 4): (1) multi-task learning (MT) and (2) hierarchical ensemble (HE).

Multi-task Learning (MT). The main structure of MT networks contains a shared encoder and multiple parallel decoders with classification heads. To practically perform the MT, we utilized the feed forward architecture of PointNet++ [24]. Specifically, an feature map $N \times D$ of point cloud with size N and feature dimension D is fed as input. Then the shared encoder encodes them as embedding. Such embedding is then decoded parallelly into $N \times D^F$ by H decoders for H granularity levels. Decoder h computes the likelihood distributions of classes (C^h) at granularity level h for each data point. The loss of MT method is the sum of the losses of its branches,

$$\mathcal{L}_{\text{MT}} = \mathcal{L}_{\text{prediction}} + \mathcal{L}_{\text{consistency}} \tag{3}$$

where the prediction loss $\mathcal{L}_{\text{prediction}}$ is the weighted average of the cross entropy losses of H levels. And it is formulated as,

$$\mathcal{L}_{\text{prediction}} = \sum_{h=1}^H \beta_h \cdot L_{\text{prediction}}^h. \tag{4}$$

Here, for granularity level h, $L_{\text{prediction}}^h$ and β_h are the cross-entropy loss and weight respectively. The consistency loss is served as a regularization term to maintain the consistence structure of predicted distributions

$$\mathcal{L}_{\text{consistency}} = \sum_{h=1}^{H-1} \gamma_h \sum_{(y_h, y_{h+1}) \in \text{PC}^h} \left[P^{h+1}(y_{h+1}) - P^h(y_h) \right]_+^2 \tag{5}$$

here $\text{PC}^h = \{(y_h, y_{h+1}) | (y_1, \ldots, y_H) \in \mathcal{Y}^{\text{FC}}\}$, where γ^h is the loss weight of h^{th} level and $P^h(\cdot)$ is the predicted likelihood distribution over class set (antichain) C^h. By Definition 4.2, given FC solution $(y_1, \ldots, y_H)$, y_h is the superclass of or same as y_{h+1}. This loss is

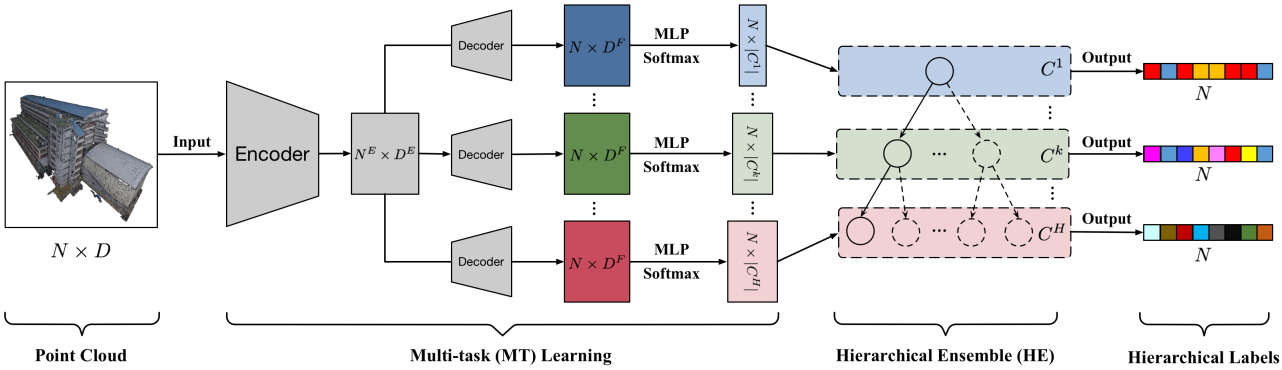

Figure 4: The framework of our method is divided into two stages: Multi-task Learning (MT) and Hierarchical Ensemble (HE). Point cloud data is fed as input. After feature extraction through the shared encoder, the features are decoded into multiple heads, and the predicted distributions at different granularity levels are obtained via MLP layers. Then, the HE stage utilizes hierarchical relationship to gain the final hierarchical labels. The width of the model depends on the hierarchical levels, where the k^{th}-level represents the middle parts. $N^E \times D^E$ is the size of embedding from the encoder. The MLP output dimension $|C^k|$ depends on the number of labels on the k^{th} granularity level.

the sum of the losses of all the parent-child pair in tree $\mathcal{T}$, which is to keep a smaller prediction score $P^{h+1}(y_{h+1})$ than its parent score $P^h(y_h)$ such that consistency is reserved. To investigate the effectiveness of the consistency loss, a loss without consistency loss branch named $\mathbf{MT}_{\text{nc}}$ is tested to perform hierarchical semantic segmentation as ablation study,

$$\mathcal{L}_{\text{MT}_{\text{nc}}} = \mathcal{L}_{\text{prediction}}. \tag{6}$$

Hierarchical Ensemble (HE). The HE is a post-processing method for initial predicted results. It computes the weighted sum of likelihood scores over all the root-to-leaf paths in tree $\mathcal{T}$. The path associated with largest score is the final predicted solution. It is formulated as equation (7). Note that solutions generated by HE are FC and the CR (and/or CP) value is 1.

$$Y_{\text{HE}} = \arg \max_{(y_1,\ldots,y_H)\in\mathcal{Y}^{\text{FC}}} \sum_{h=1}^{H} P^h(y_h) \tag{7}$$

In order to perform comparison analysis, we also apply a multi-classifier (**MC**) method which does not leverage the mutual relationship across levels, and only trains an independent segmentation classifier for each granularity level. And H classifiers are trained and evaluated separately. It performs conventional segmentation H times for the dataset based on PointNet++. Futhermore, a variant of the proposed two-stage method, MC+HE is also investigated, which uses the HE to post-process outputs of the MC.

4.3 Experimental Results

Based on the class label tree given by Figure 3, we build five granularity levels ($H = 5$). They are given in the first and second columns of Table 5. PointNet++ [24] is used as backbone. We set β_i=1 and $\gamma_i = 0.05$ with $i \in \{1, \cdots, H\}$, and more detailed settings are present in the supplementary material. We apply CR_1 ($\alpha = 1$), intersection-over-union (IoU) as well as overall accuracy (OA) for performance analysis.

Table 4: Test results (OA%) for different HL methods.

Method	Granularity Level				
	C^1	C^2	C^3	C^4	C^5
MC	85.3	79.5	78.3	76.3	74.0
MC+HE	89.1	81.4	79.9	77.9	73.5
MT_{nc}	90.2	82.2	80.9	78.8	74.6
MT	**90.7**	**83.1**	**81.7**	79.8	75.2
MT+HE	**90.7**	**83.1**	**81.7**	**80.0**	**75.4**

Comparisons between different HL methods. After removing points with ground-truth label of "*unclassified*" (unlabeled), for each class, the intersection and union sets of predicted point set and ground-truth are generated; then the IoUs are computed as the ratio of intersection set cardinality to that of the union. And the OA is computed as proportion of correct predictions to total points. Results for test set results are presented in Table 4 and Table 5. There are several observations: (1) in terms of average IoU and OA of the granularity level, it decreases significantly with granularity level changing from C^1 to C^5. It indicates that the difficulty of the problem increases as the label instances become small and distributed sparsely; (2) the performance of MC + HE is better than that of MC only for most cases; (3) overall, the HL methods (i.e., MT + HE, MT, MT_{nc}, MC + HE) taking hierarchical labels into account perform better than the MC without considering them. These observations demonstrate that hierarchical labels help and enhance the performance.

One possible reason of the better performance by the HL method is that the inherent relation among label layers provide additional geometrical information for semantic segmentation. A visual illustration is given by Figure 5. The MC semantic segmentation on the level C^5 and C^3 without other level information results in that "*roof*" is wrongly recognized as driving road (i.e."*road*"or "*not vehicle*") or natural ground "*natural*" (see C^5 (b) and C^3 (b) of Figure 5). We

found that they are geometrically similar but semantically different. Here we first define this phenomenon as *geometric ambiguity*: points with similar geometric features but significantly different semantic labels are wrongly classified to the same semantic class. As indicated by the result of MT ((c) column of Figure 5), hierarchical and multiple annotation can ameliorate this phenomenon. For the instance of C^1 level in Figure 5, points on roof belonging to "*construction*" are easily recognized as "*ground*" by the MC, while the MT framework is able to segment them correctly by leveraging information from finer levels.

Table 5: Test results (class IoU%) for HL methods

Granularity Level	Class	Method				
		MC	MC+HE	MT_{nc}	MT	MT+HE
C^1	ground	78.9	83.5	84.9	**85.5**	**85.5**
	construction	67.4	75.5	78.2	**79.1**	79.0
C^2	natural	66.8	69.4	71.5	**71.9**	**71.9**
	man_made	52.6	54.7	53.7	**54.8**	54.7
	construction	72.9	75.5	77.0	**79.1**	79.0
C^3	natural	67.1	69.4	71.8	**71.9**	**71.9**
	play_field	0.3	0.3	0.7	**3.0**	2.2
	path&stair	7.2	8.2	**7.8**	0.0	0.0
	driving_road	49.9	52.0	51.2	**52.5**	52.3
	construction	74.1	75.5	77.7	**79.1**	79.0
C^4	natural	67.8	69.4	**71.9**	71.8	**71.9**
	play_field	0.8	0.3	0.5	1.9	**2.2**
	path&stair	7.8	**8.2**	7.9	0.0	0.0
	vehicle	34.5	**38.7**	36.7	36.4	36.6
	not vehicle	48.6	51.1	50.1	51.1	**51.3**
	building	70.1	72.1	74.0	75.7	**76.0**
	link	2.1	2.2	**3.7**	0.5	0.5
	facility	0.0	0.0	**0.1**	0.0	0.0
C^5	natural	71.0	69.4	**72.2**	71.7	71.9
	play_field	1.6	0.3	2.0	1.8	**2.2**
	sheltered	**10.7**	10.4	**10.7**	1.4	0.0
	unsheltered	4.4	4.4	**4.7**	0.1	0.0
	bus_stop	0.4	**0.5**	0.1	0.0	0.0
	car	**40.7**	39.9	38.5	37.0	36.6
	bus	0.0	0.0	0.0	0.0	0.0
	not vehicle	50.6	51.1	50.1	51.0	**51.3**
	wall	57.3	56.0	57.2	57.1	**57.5**
	roof	58.6	57.2	60.4	61.1	**61.3**
	link	3.5	2.2	**3.6**	0.5	0.5
	artificial_landscape	0.0	0.0	0.0	0.0	0.0
	lamp	0.0	**0.1**	0.0	0.0	0.0
	others	**0.2**	0.0	**0.2**	0.0	0.0

Insights of Consistency Rate. As a metric evaluating consistency of hierarchical relationship, the result of CR_1 reveals that our framework leverages the mutual assistance among hierarchies. From Figure 6, the CR_1 of MT is over around 15% than that of MC which ignores the hierarchical annotation. It suggests that MT learning may correct the results in certain level according to features from other granularity levels. On another hand, with comparison of MC and MC+HE results from Table 5 and Table 4, HE also boosts the performance by maintaining hierarchical relationship forcibly, while this boosting is not significant from MT to MT+HE. The results of CR_1 quantitatively explain why HL methods can effectively address the geometric ambiguity discussed above.

Effectiveness of Consistency Loss. Performance differences between MT and MT_{nc} in Figure 6 and Table 4 demonstrate the effectiveness of the proposed consistency loss. With it, MT can significantly restrain the hierarchical violations in segmentation results, while MT_{nc}, ignoring it, results in around 10% decrease in terms of CR_1. Moreover, MT also performs better than MT_{nc} in terms of OA (see Table 4).

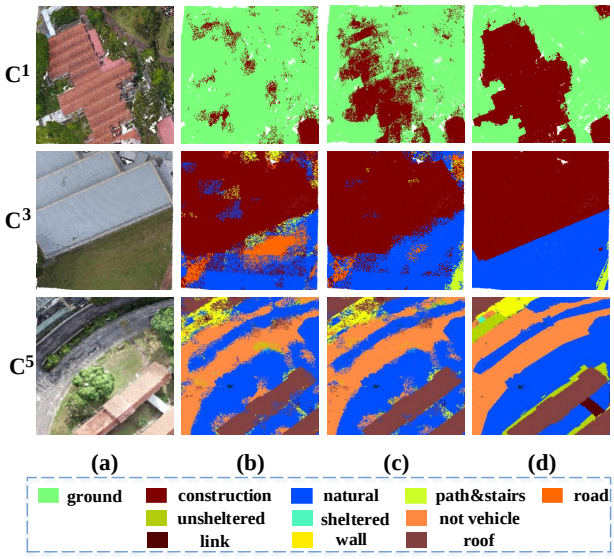

Figure 5: Visualization of hierarchical segmentation results. (a): raw point cloud; (b) and (c) are MC and MT results, respectively, in C^1, C^3 and C^5 levels; (d) : ground truth label.

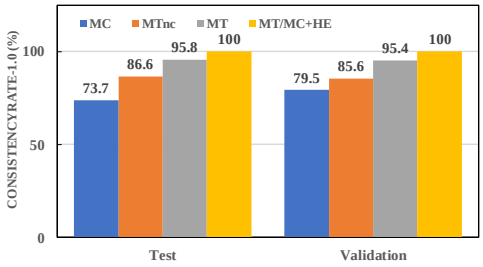

Figure 6: Results of CR_1 (%) using different HL methods.

5 HIERARCHICAL SCENE UNDERSTANDING TASKS AND BENCHMARKS

In this section, we apply the HL framework for scene understanding and build benchmarks on two tasks of hierarchical semantic segmentation and instance segmentation. We first investigate the effectiveness of sampling methods for feature learning on large-scale point cloud datasets.

5.1 Sampling Methods

For scene based point cloud datasets, sampling is necessary for feature learning because of requirements of efficiency and fixed size of model input. Typical sampling strategies are uniform sampling and farthest point sampling (FPS) [23]. Here we do not apply farthest point sampling (FPS) due to its computational inefficiency. The simplest sampling method is to randomly pick a fixed size of points with uniform distribution. However, randomly sampling a small size of points from large point cloud data would induce

considerable randomness into the samples, which may lead to failure of training process. Therefore, to conduct less bias and learnable sampling, we experimented two variations of uniform sampling, the details of which are presented in the supplementary document: (1) l-w random block sampling (l-w RBS), and (2) random centered K nearest neighor (RC-KNN). Given a set of points $\mathcal{D}_x = \{(x_1^1, x_2^1, x_3^1), (x_1^2, x_2^2, x_3^2), \ldots, (x_1^n, x_2^n, x_3^n)\} \in \mathbb{R}^3$ ($D = 3$ and three coordinates: latitude, longitude and height), we define them and illustrate how to select N ($N < n$) points from $\mathcal{D}_x$ as follows.

l-w **RBS** randomly chooses a point P_c (x_1^c, x_2^c, x_3^c) from $\mathcal{D}_x$ with an uniform distribution and then uniformly samples $\mathcal{D}'_x = \{(x_1^c, x_2^c, x_3^c),\ (x_1^{i_1}, x_2^{i_1}, x_3^{i_1}),\ \ldots, (x_1^{i_{(N-1)}}, x_2^{i_{(N-1)}}, x_3^{i_{(N-1)}})\}$ in a l-w block centered at P_c, namely, for any $1 \le j \le N-1$, $x_1^c - \frac{l}{2} \le x_1^{i_j} \le x_1^c + \frac{l}{2}$ and $x_2^c - \frac{w}{2} \le x_2^{i_j} \le x_2^c + \frac{w}{2}$.

RC-KNN randomly chooses a point P_c (x_1^c, x_2^c, x_3^c) from $\mathcal{D}_x$ with an uniform distribution, and then K ($K = N$) nearest neighors to point P_c in terms of Euclidean distance are chosen as the sampled points.

In order to investigate the effectiveness of the above two sampling methods on the HL methods, we apply them with PointNet++ [24] as the feature leaning network on our dataset. Specifically, in each training iteration, we use either RBS or RC-KNN to select N points from a randomly picked region in training set as a sample in a batch. The sample size N is set as 2048, and block size of both l and w in RBS are set as 12m. We compute the mean IoU (mIoU) across all classes for each granularity level. The test results are given in Table 6. From the results, we can see that the RBS method dominates the RC-KNN method by our setting. Note that the setting of RBS sampling is also utilized in Section 4.

Table 6: Semantic segmentation results (mIoU%) for RBS and RC-KNN sampling methods.

Model	Sampling Method	Granularity level C^1	C^2	C^3	C^4	C^5
MT	RBS	**82.3**	**68.6**	**41.3**	**29.7**	**20.1**
	RC-KNN	77.4	61.7	37.2	25.5	9.9
MT + HE	RBS	**82.2**	**68.5**	**41.1**	**29.8**	**20.1**
	RC-KNN	77.0	62.0	37.1	25.6	9.8

5.2 Semantic Segmentation

In this section, we benchmark the performance on semantic segmentation task. Three established models are applied: PointNet++ [24], PointCNN [20] and DGCNN [37]. The sampling method used here is the RBS sampling with block size of 12m ($l = w = 12$m). We take the hierarchical annotation into account and apply our the proposed MT+HE method. The mIoU across all classes for each granularity is used as the performance metric. The results of both test and validation are given by Table 7.

5.3 Instance Segmentation

In this section, we build the instance segmentation benchmark of current dataset. The training, validation and test splitting still follows Table 3. We perform this task for the granularity level four (C^4) only, where there exists the largest number of available classes and instances among all granularity levels for training, validation and test. The ASIS [36] and SGPN [35] method were used here for the baseline evaluation. For each class, the weight coverage (WCov) as introduced by Wang et al. in [36] is computed as the performance measurement. Results for both validation and test sets are shown in Table 8, which shows that ASIS [36] performs better than SGPN [35]. Note that classes "*natural*", "*path&stair*", "*not vehicle*" and "*facility*" are not countable, thus no instance segmentation results are for them.

Table 7: Semantic segmentation results (mIoU%) for three feature learning models with HL methods.

Dataset	Learning Model	Granularity level C^1	C^2	C^3	C^4	C^5
Validation	PointNet++	79.7	67.0	43.4	33.4	21.9
	PointCNN	86.9	77.4	51.1	38.0	27.2
	DGCNN	**88.1**	**79.8**	**53.0**	**39.5**	**29.5**
Test	PointNet++	79.7	67.0	43.4	33.4	21.9
	PointCNN	88.6	78.2	58.4	41.1	27.2
	DGCNN	**90.9**	**81.3**	**61.5**	**43.6**	**29.1**

Table 8: Instance segmentation results (WCov) for selected classes at granularity level four (C^4).

Instance class name	Validation ASIS	SGPN	Test ASIS	SGPN
play_field	0.0	0.0	**3.4**	0.0
vehicle	**34.0**	17.3	**44.1**	32.7
building	**53.4**	35.7	**40.3**	32.1
link	**13.3**	10.1	**11.2**	8.5
Average	**26.4**	15.8	**24.8**	18.3

6 CONCLUSION

A well-annotated point cloud dataset with two benchmarks, Campus3D, is proposed in this paper. It is annotated with multiple and hierarchical label for the better scene understanding and potential usage in reconstruction. We define the HL problem and propose a new measure to evaluate the consistency across granularity levels. A two-stage method MT+HE is presented to the HL. Experimental results demonstrate its effectiveness comparing with MC without taking multiple and hierarchical information into account. Moreover, we investigate two sampling methods for point cloud learning with HL methods and identify RBS as the useful one. Future users will benefit from these initial and basic explorations. In the end, we apply established models and benchmark performance for semantic and instance segmentation for future comparisons. Other potential tasks can be built based on the Campus3D such as hierarchical instance segmentation and 3D model reconstruction.

ACKNOWLEDGMENTS

This research is supported by National University of Singapore (NUS) Institute of Operations Research and Analytics (IORA) grant R-726-000-002-646 and National Research Foundation of Singapore grant NRF-RSS2016-004. The authors gratefully acknowledge the data collection support of Virtual NUS team.

REFERENCES

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

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

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

[4] Cindy Cappelle, Maan E El Najjar, François Charpillet, and Denis Pomorski. 2012. Virtual 3D city model for navigation in urban areas. *Journal of Intelligent & Robotic Systems* 66, 3 (2012), 377–399.

[5] Ludovico Carozza, David Tingdahl, Frédéric Bosché, and Luc Van Gool. 2014. Markerless vision-based augmented reality for urban planning. *Computer-Aided Civil and Infrastructure Engineering* 29, 1 (2014), 2–17.

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

[7] Arnis Cirulis and Kristaps Brigis Brigmanis. 2013. 3D outdoor augmented reality for architecture and urban planning. *Procedia Computer Science* 25 (2013), 71–79.

[8] CloudCompare. [n.d.]. CloudCompare: 3D point cloud and mesh processing software, Open Source Project. https://www.danielgm.net/cc/

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5828–5839.

[10] Michael Firman. 2016. RGBD datasets: Past, present and future. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 19–31.

[11] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. 2010. Building rome on a cloudless day. In *European Conference on Computer Vision*. Springer, 368–381.

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3354–3361.

[14] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan D Wegner, Konrad Schindler, and Marc Pollefeys. 2017. SEMANTIC3D. NET: A NEW LARGE-SCALE POINT CLOUD CLASSIFICATION BENCHMARK. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4 (2017).

[15] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. 2019. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. *arXiv preprint arXiv:1911.11236* (2019).

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

[17] Thomas H Kolbe, Gerhard Gröger, and Lutz Plümer. 2005. CityGML: Interoperable access to 3D city models. In *Geo-information for disaster management*. Springer, 883–899.

[18] Loic Landrieu and Martin Simonovsky. 2018. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4558–4567.

[19] Minglei Li, Liangliang Nan, Neil Smith, and Peter Wonka. 2016. Reconstructing building mass models from UAV images. *Computers & Graphics* 54 (2016), 84–93.

[20] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. 2018. PointCNN: Convolution On X-Transformed Points. In *Advances in Neural Information Processing Systems*. 828–838.

[21] Madeleine Manyoky, Ulrike Wissen Hayek, Kurt Heutschi, Reto Pieren, and Adrienne Grêt-Regamey. 2014. Developing a GIS-based visual-acoustic 3D simulation for wind farm assessment. *ISPRS International Journal of Geo-Information* 3, 1 (2014), 29–48.

[22] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. 2019. 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 Conference on Computer Vision and Pattern Recognition*. 8827–8836.

[23] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 652–660.

[24] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*. 5099–5108.

[25] Xavier Roynard, Jean-Emmanuel Deschaud, and François Goulette. 2018. 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 (2018), 545–557.

[26] Andrés Serna, Beatriz Marcotegui, François Goulette, and Jean-Emmanuel Deschaud. 2014. Paris-rue-Madame database: a 3D mobile laser scanner dataset for benchmarking urban detection, segmentation and classification methods. In *4th International Conference on Pattern Recognition, Applications and Methods ICPRAM 2014*.

[27] Nathan Silberman and Rob Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*. IEEE, 601–608.

[28] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*. Springer, 746–760.

[29] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.

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

[31] Bruno Vallet, Mathieu Brédif, Andrés Serna, Beatriz Marcotegui, and Nicolas Paparoditis. 2015. TerraMobilita/iQmulus urban point cloud analysis benchmark. *Computers & Graphics* 49 (2015), 126–133.

[32] Yannick Verdie, Florent Lafarge, and Pierre Alliez. 2015. Lod generation for urban scenes. *ACM Transactions on Graphics* 34, ARTICLE (2015), 30.

[33] Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan. 2019. Graph attention convolution for point cloud semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10296–10305.

[34] Shenlong Wang, Simon Suo, Wei-Chiu Ma, Andrei Pokrovsky, and Raquel Urtasun. 2018. Deep parametric continuous convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2589–2597.

[35] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. 2018. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2569–2578.

[36] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. 2019. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4096–4105.

[37] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)* 38, 5 (2019), 146.

[38] Matthew J Westoby, James Brasington, Niel F Glasser, Michael J Hambrey, and Jennifer M Reynolds. 2012. 'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* 179 (2012), 300–314.

[39] Changchang Wu, Sameer Agarwal, Brian Curless, and Steven M Seitz. 2011. Multicore bundle adjustment. In *CVPR 2011*. IEEE, 3057–3064.

[40] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE International Conference on Computer Vision*. 1625–1632.

[41] Saining Xie, Sainan Liu, Zeyu Chen, and Zhuowen Tu. 2018. Attentional shapecontextnet for point cloud recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4606–4615.

[42] Jiancheng Yang, Qiang Zhang, Bingbing Ni, Linguo Li, Jinxian Liu, Mengdie Zhou, and Qi Tian. 2019. Modeling point clouds with self-attention and gumbel subset sampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3323–3332.

[43] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. 2019. PointWeb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5565–5573.