Scan2Part: Fine-grained and Hierarchical Part-level Understanding of Real-World 3D Scans

Alexandr Notchenko,Vladislav Ishimtsev, Alexey Artemov, Vadim Selyutin, Emil Bogomolov and Evgeny Burnaev

Skolkovo Institute of Science and Technology, Moscow, Russian Federation

Keywords: Semantic Segmentation, Scene Understanding, Volumetric Scenes, Part-level Segmentation.

Abstract: We propose Scan2Part, a method to segment individual parts of objects in real-world, noisy indoor RGB-D scans. To this end, we vary the part hierarchies of objects in indoor scenes and explore their effect on scene understanding models. Specifically, we use a sparse U-Net-based architecture that captures the fine-scale detail of the underlying 3D scan geometry by leveraging a multi-scale feature hierarchy. In order to train our method, we introduce the Scan2Part dataset, which is the first large-scale collection providing detailed semantic labels at the part level in the real-world setting. In total, we provide 242,081 correspondences between 53,618 PartNet parts of 2,477 ShapeNet objects and 1,506 ScanNet scenes, at two spatial resolutions of 2 cm^3 and 5 cm^3. As output, we are able to predict fine-grained per-object part labels, even when the geometry is coarse or partially missing. Overall, we believe that both our method as well as newly introduced dataset is a stepping stone forward towards structural understanding of real-world 3D environments.

1 INTRODUCTION

In the recent years, a wide variety of consumer-grade RGB-D cameras such as Intel Real Sense (Keselman et al., 2017), Microsoft Kinect (Zhang, 2012), or smartphones equipped with depth sensors, enabled inexpensive and rapid RGB-D data acquisition. Increasing availability of large, labeled datasets (e.g., (Chang et al., 2017; Dai et al., 2017)) made possible development of deep learning methods for 3D object classification and semantic segmentation. At the same time, acquired 3D data is often incomplete and noisy; while one can identify and segment the objects in the scene, reconstructing high-quality geometry of objects remains a challenging problem.

An important class of approaches, e.g., recent work (Avetisyan et al., 2019), uses a large dataset of clean, labeled geometric shapes (Chang et al., 2015), for classification/segmentation associating the input point or voxel data with object labels from the dataset, along with adapting geometry to 3D data. This approach ensures that the output geometry has high quality, and is robust with respect to noise and missing data in the input. At the same time, a "flat" classification/segmentation approach, with each object in the database corresponding to a separate label and matched to a subpart of the input data corresponding to the whole object, does not scale well as the number of classes grows and often runs into difficulties in the cases of extreme occlusion (only a relatively small part of an object is visible). Significant improvements can be achieved by considering object *parts*, or, more generally, part hierarchies. Part-based segmentation of 3D datasets promises to offer a significant improvement in a variety of tasks such as finding the best matching shape in the dataset or recognizing objects from highly incomplete data (e.g., from one or two parts).

Man-made environments (e.g., indoor scenes) commonly consist of objects that naturally form *part hierarchies* where objects and their parts can be divided into finer parts. In our work, we use scene and object representation based on such part hierarchies and focus on the key problem of semantic part segmentation of separate objects in the scenes, enabling further improvements in dataset-based reconstruction. To this end, we construct a volumetric part-labeled dataset of scanned 3D data suitable for machine learning applications. We use this new dataset to explore the limits of part-based semantic segmentation by training a variety of sparse 3D convolutional neural networks (CNNs) in multiple setups.

Our contributions include:

Notchenko, A., Ishimtsev, V., Artemov, A., Selyutin, V., Bogomolov, E. and Burnaev, E.
Scan2Part: Fine-grained and Hierarchical Part-level Understanding of Real-World 3D Scans.
DOI: 10.5220/0010848200003124
In *Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2022) - Volume 5: VISAPP*, pages 711-722
ISBN: 978-989-758-555-5; ISSN: 2184-4321

1. A new method Scan2Part, aiming to segment volumetric objects into semantic parts and their instances, leveraging a sparse volumetric 3D CNN model trained on a large-scale part-annotated collection of objects.
2. A new dataset Scan2Part, composed of 1,506 3D reconstructions of real-world scenes with 2,477 aligned 3D CAD models represented by real-world 3D geometry and annotated using hierarchical annotation, which links 3D scene reconstruction with part-annotation of indoor objects.

2 RELATED WORK

Deep Learning for 3D Scene Understanding. Deep learning has been applied for semantic 3D scene understanding in a variety of ways, and we only review the core work related to the semantic and instance segmentation of 3D scenes. Most relevant to our work, deep learning approaches have been used on 3D reconstructions of scenes represented by *3D volumetric grids* (Dai et al., 2017; Dai and Nießner, 2018; Hou et al., 2019; Liu and Furukawa, 2019). For volumetric grid, 3D convolutions may be defined analogously to 2D convolutions in image domains, giving rise to 3D convolutional neural nets (3D CNNs). Memory requirements have been addressed by adaptive data structures (Wang et al., 2017). Similar to this body of work, we operate on volumetric representations of 3D scenes, but perform segmentation of individual parts.

Methods operating on raw point clouds provide an alternative to volumetric 3D CNNs by constructing an appropriate operations directly on *unstructured point clouds* for a variety of applications including semantic labeling (e.g., (Qi et al., 2017a; Qi et al., 2017b; Klokov and Lempitsky, 2017; Wang et al., 2018; Wang et al., 2019b)). Most recently, instance (Elich et al., 2019; Liang et al., 2019; Elich et al., 2019; Yi et al., 2019; Yang et al., 2019; Zhang and Wonka, 2019; Engelmann et al., 2020) and joint semantic-instance (Wang et al., 2019a; Pham et al., 2019b) segmentation tasks on point clouds have been considered closely. While point-based methods require less computations, learning with irregular structures such as point clouds is challenging. To segment instances, recently proposed volumetric and point-based approaches use metric learning to extract per-point embeddings that are subsequently grouped into object instances (Elich et al., 2019; Yi et al., 2019; Lahoud et al., 2019; Liu and Furukawa, 2019). We take advantage of this mechanism in our instance segmentation methods.

Part-aware segmentation methods commonly focus on meshes or complete, clean point clouds constructed from 3D CAD models. The most closely related work is semantic parts labeling (e.g., (Yi et al., 2016; Wang et al., 2017; Qi et al., 2017a; Mo et al., 2019b; Yi et al., 2019; Zhang and Wonka, 2019)) and part instance segmentation (Zhang and Wonka, 2019) for voxelized or point-sampled 3D shapes. Other works focus on leveraging parts structure of clean shapes for co-segmentation (Chen et al., 2019; Zhu et al., 2019), hierarchical mesh segmentation (Yi et al., 2017), shape assembly/generation (Mo et al., 2019a; Wu et al., 2019b; Wu et al., 2019a; Mo et al., 2020), geometry abstraction (Russell et al., 2009; Li et al., 2017; Sun et al., 2019), and other applications. In comparison, our focus is on learning part-based semantic and instance segmentation of noisy and fragmented real-world 3D scans. Very recently, initial approaches to semantic 3D segmentation have been proposed (Bokhovkin et al., 2021; Uy et al., 2019) but for a significantly less extensive part hierarchy. More specifically, (Bokhovkin et al., 2021) targets predicting part hierarchy at object and coarse parts levels, discarding smaller parts altogether; in contrast, we are able to predict parts at finer levels in the hierarchy.

Other methods have been studied in the context of 3D scene segmentation, such as complementing CNNs with conditional random fields (Pham et al., 2019b; Pham et al., 2019a; Wang et al., 2017), however, these are beyond the scope of this paper.

3D Scene Understanding Datasets. A body of work focuses on rendering-based methods, aiming to realistic 3D scenes procedurally (Fisher et al., 2012; Handa et al., 2016; Song et al., 2017; McCormac et al., 2017; Li et al., 2018; Garcia-Garcia et al., 2018). Such datasets can in principle provide arbitrarily fine semantic labels but commonly suffer from the reality gap caused by synthetic images; in contrast, our proposed dataset is built by transferring part annotations to real-world noisy scans. Recent advances in RGB-D sensor technology have resulted in the development of a variety of 3D datasets capturing real 3D scenes (Armeni et al., 2016; Hua et al., 2016; Dai et al., 2017; Chang et al., 2017; Armeni et al., 2017; Straub et al., 2019), however, none of these provide part-level object annotations. In contrast, our dataset provides semantic and instance part labels for a large-scale collection of indoor 3D reconstructions. ScanObjectNN (Uy et al., 2019) provides parts annotation for objects in real-world scenes, however, it does not include annotations in occluded regions and only specifies parts labeling at the coarsest levels in the parts hierarchy. Thus, this collection thus does not allow reasoning about the fine grained structure of the

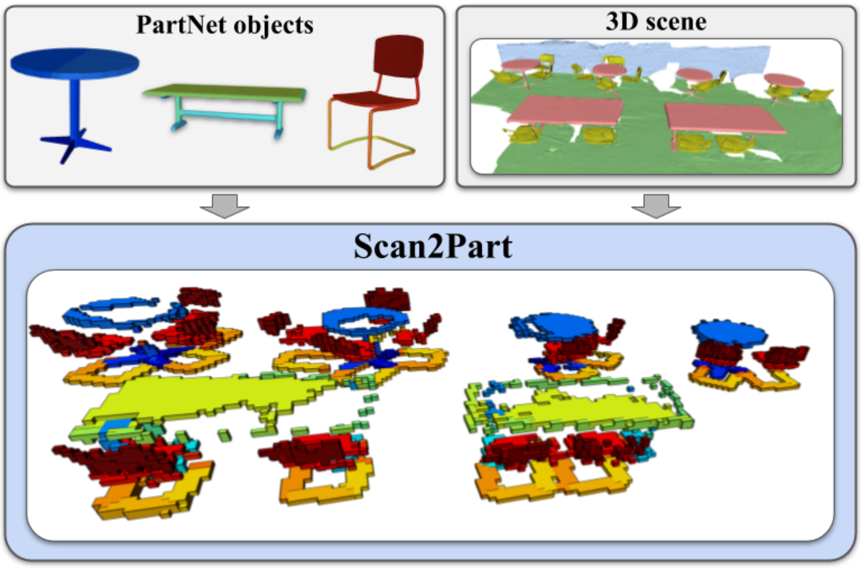


Figure 1: A pipeline for obtaining Scan2Part dataset. We project the PartNet (Mo et al., 2019b) labels to the ShapeNet (Chang et al., 2015) coordinate system (left), then use alignments in Scan2CAD (Avetisyan et al., 2019) (bottom) to map labels to real indoor scenes from ScanNet (Dai et al., 2017) (right).

objects; differently, our data construction approach allows to flexibly select part representation levels, and our experiments include results for 13, 36, and 79 part categories at first through third levels in the hierarchy.

Early collections of part-annotated meshes (Chen et al., 2009) are limited by their relatively smaller scale. With the introduction of a comprehensive ShapeNet benchmark (Chang et al., 2015), a coarse semantic part annotation has been created using active learning (Yi et al., 2016). More recently, a large-scale effort to systematically annotate 3D shapes within a coherent hierarchy was presented (Mo et al., 2019b). Still, none of these CAD-based collections include real-world 3D data, limiting their potential use. Our benchmark is designed to address this reality gap.

Large-scale 3D understanding datasets commonly require costly manual annotations by tens to hundreds of expert crowd workers (annotators), preceded by the development of custom labeling software (Armeni et al., 2016; Hua et al., 2016; Dai et al., 2017; Chang et al., 2017; Armeni et al., 2017; Straub et al., 2019; Yi et al., 2016; Mo et al., 2019b). Moreover, annotating parts in 3D objects from scratch is connected to inherent ambiguity in part definitions, as revealed by (Yi et al., 2016; Mo et al., 2019b). This challenge is even more pronounced for noisy, incomplete 3D scans produced by RGB-D fusion. We have chosen to instead build our Scan2Part dataset fully automatically by leveraging correspondences between four publicly available 3D collections: ScanNet (Dai et al., 2017), Scan2CAD (Avetisyan et al., 2019), ShapeNet (Chang et al., 2015), and PartNet (Mo et al., 2019b) datasets. As a result, we (1) become free from ambiguity in part definitions by re-using consistent, well-defined labels from (Mo et al., 2019b), and (2) are able to compute appropriate levels of semantic detail for our benchmark without manual re-labeling.

Assembly from Parts and Hierarchy. A lot of researchers over the last 20 years came to the idea that scenes and images are better represented as discrete structures with relational and hierarchical properties. New datasets that make explicit relations on visual objects were created recently (Krishna et al., 2017), spurring new research in scene graphs (Johnson et al., 2018; Xu et al., 2017) and reconstruction. In cognitive science, it have been conjectured for a long time that ability to compose complex objects and scenes from parts is a fundamental part of human perception (Hoffman and Richards, 1984; Biederman, 1987). The concept of ”Recognition-by-components” is closely related to ”analysis-by-synthesis” (Yuille and Kersten, 2006; Yildirim et al., 2015) approach in machine learning. Mumford and Zhu in (Zhu and Mumford, 2006) assume that spatial scenes can be defined by a ”grammar” of spatial objects and geometric primitives, similar to sentences in natural language, where a sequence of words can have multiple parsing trees providing multiple explanations to single sentence. Multiple solutions are expected because solving an under-constrained inverse problem (in this case inverse graphics problem) can have multiple solutions. The solution can be made unique either with stronger priors or can be resolved in downstream processing if uncertainty is propagated.

3 METHOD

3.1 Dataset Construction

Data Preparation. Our first objective is to develop a large-scale 3D scene understanding dataset with part-level annotations. We use the 3D geometry in 1,506 3D scenes in Scannet dataset (Dai et al., 2017) reconstructed from RGB-D scans in the form of truncated Signed Distance Function (SDF) at voxel resolution = 2 cm and 5 cm. For labeling the scan geometry, we use the parts taxonomy in PartNet dataset (Mo et al., 2019b) represented as a tree structure where nodes encode parts at various detail levels and edges encode the “part of” relationship. We associate to each 3D scene a per-voxel mask storing leaf part IDs from the taxonomy (i.e., the most fine-grained categories). To label the 3D scene at a given level d of semantic detail ($d = 1$ meaning whole objects and $d = 8$ meaning finest parts), we start with the leaf labels in each voxel and traverse the taxonomy tree until hitting depth d.

We further describe the main steps taken to create our Scan2Part benchmark below, leaving the detailed discussion of the technicalities for the supplementary. Annotation schema for our dataset is shown in Figure 1. The final result of our procedure is a collection of scenes comprised of volumetric instances of separate objects, where background and non-object information is removed, allowing to focus on part based segmentation.

Transferring Labels to Volumetric 3D Grids. To obtain ground-truth semantic parts annotations for real-world 3D scenes, we establish correspondences between each volume in the 3D scene in Scannet and a set of part-annotated mesh vertices in a registered 3D CAD model from PartNet. To this end, we first find accurate 9 degrees of freedom (9 DoF) transformations between PartNet 3D models and their original versions from ShapeNet, and next use the manually annotated 9 DoF transformations and their respective object categories provided by Scan2CAD to obtain the final scan-to-part alignments. We further perform a simple majority voting, selecting only the most frequent (among the vertices) part label as the ground truth voxel label.

This procedure results in 242,081 correspondences represented as 9 DoF transformations between 1,506 reconstructions of real-wold Scannet scenes and 53,618 unique parts of 2,477 ShapeNet objects. Parts of each object have a tree structure, similar to (Mo et al., 2019b). Note that the majority-based voting implementation of annotation transfer results in some semantic parts labels not being represented in the 3D scan, ultimately affecting parts taxonomy, which we discuss below.

Parts Taxonomy Processing. The taxonomy of 3D shape parts is represented as a single tree structure based on (Mo et al., 2019b). However, as the voxel resolution of the real-world 3D scan is relatively coarse, particular part categories (e.g., small shape details such as keyboard buttons or door handles) cannot be represented by sufficient number of 3D data points, thus implying a reduction in the original taxonomy. We proceed with this reduction by first choosing an appropriate occurrence threshold (we pick 1800 voxels, but demonstrate the effect of different threshold values in the supplementary) and remove classes that have smaller number of representatives in the dataset. We finalize the taxonomy by pruning trivial paths in the tree (i.e., if a vertex has only one child, then we delete this vertex by connecting the child and the parent of this vertex), but keeping the leaf labels intact. We display the number of vertices at different granularities in the original and resulting part taxonomy levels in Table 1. Some parts (leafs in the part taxonomy) are not represented in ScanNet data, so we remove these from the tree. Figure 3 demonstrates example annotations produced by our automatic procedure.

Table 1: Number of parts on each tree level.

Level	**1** **(object)**	**2**	**3**	**4**	**5**	**6**	**7**	**8** **(part)**
Full Taxonomy	18	50	133	223	269	302	306	307
Pruned Taxonomy	13	36	79	—	—	—	—	—

To aid reproducibility, we will release all the necessary code to combine datasets with minimal efforts while respecting their licence terms.

3.2 Evaluation Protocol

Based on our dataset, we propose a novel benchmark for part-level scan 3D understanding, offering three core tasks, namely semantic labeling, hierarchical semantic segmentation, and semantic instance segmentation; we overview these tasks in the context of our datasets below. As inputs in all tasks we supply the voxelized SDF (with RGB information), representing 3D geometry and appearance of individual objects, already separated from the background voxels in the scene. Note that our tasks are similar to part-level object understanding but operate on real-world 3D shapes.

Evaluation Tasks. In *semantic labeling,* the goal is to associate a set of n semantic part labels $y_j = (y_j^{d_1}, \dots, y_j^{d_n})$ with each voxel v_j, at detail levels $d_1, \dots, d_n$. Compared to object-level segmentation, the part-level task becomes even more challeg-

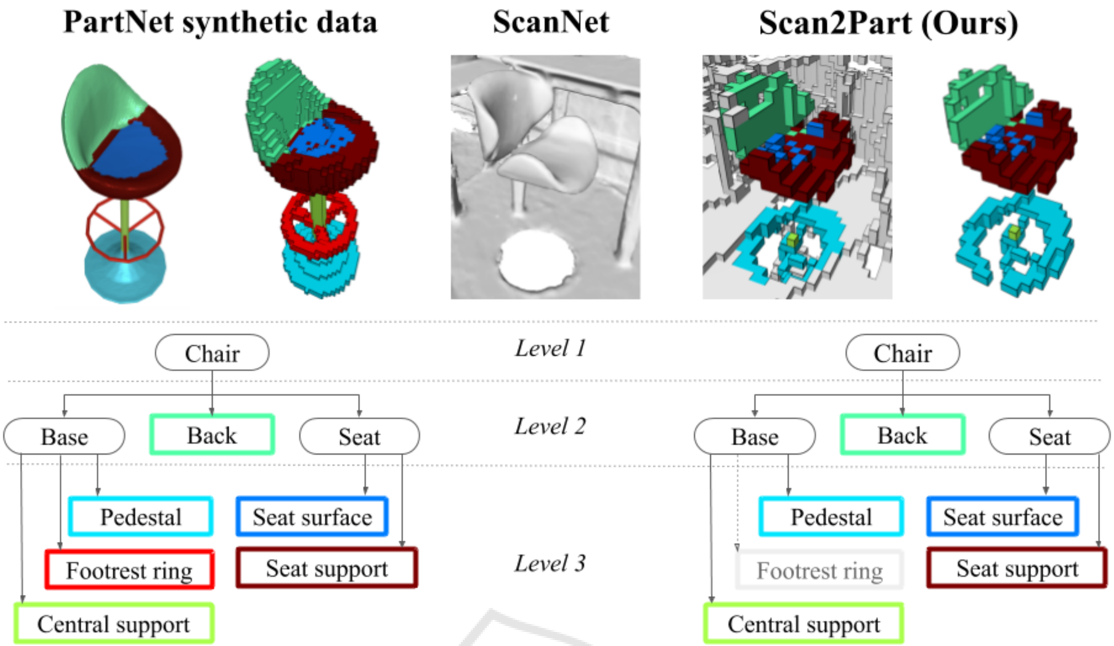


Figure 2: Top: our dataset is obtained by combining PartNet synthetic data with ScanNet sensor data. Bottom: the PartNet object hierarchy is compressed to include only parts sufficiently well represented in ScanNet data.

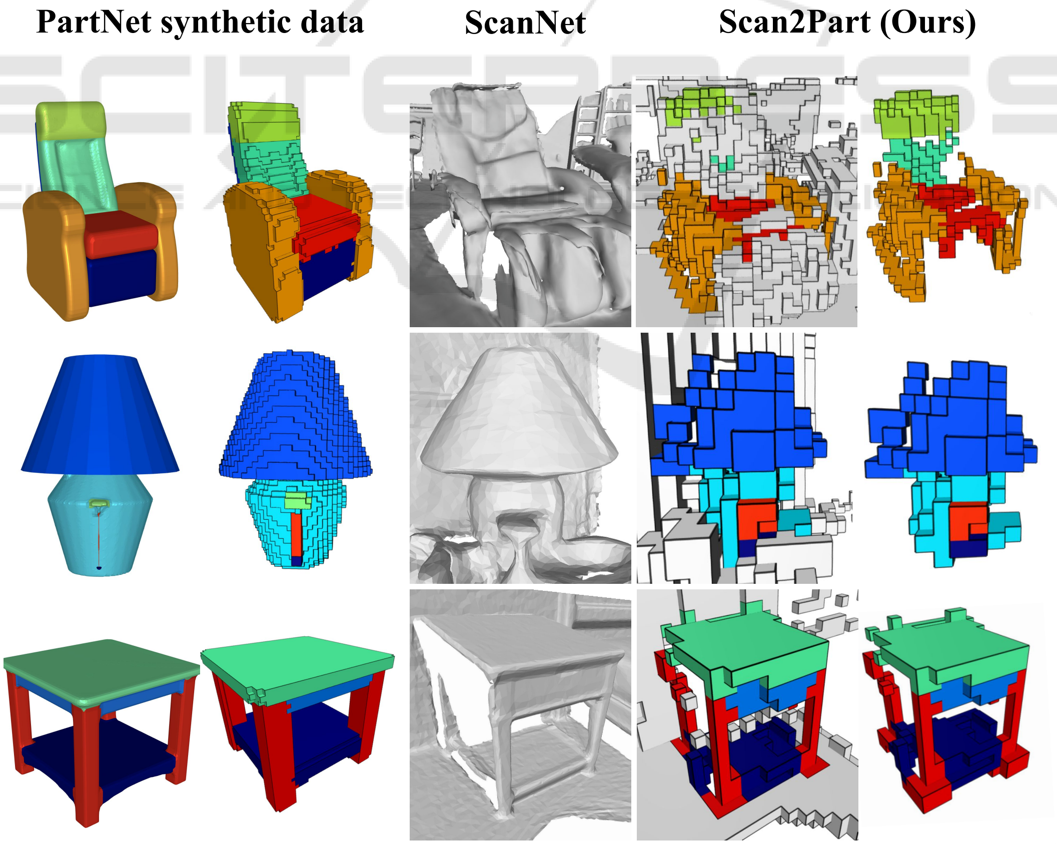


Figure 3: Example annotations produced by our automatic procedure at the spatial resolution of 5 cm^3. From left to right, part-annotated meshes and voxelized shapes in PartNet; fragments of their respective reconstructions in ScanNet; voxelized part-annotated shapes with and without background (non-object) voxels in Scan2Part.

ing, particularly at deeper levels in the taxonomy that require predicting at increasingly more fine-grained categories. We provide $1,506$ scenes with $53,618$ unique parts of $2,477$ objects.

For *hierarchical semantic segmentation,* one must perform segmentation at all levels in the hierarchy, inferring labels in coarse- and fine-grained detail levels simultaneously.

For *instance segmentation,* the goal is to simultaneously perform part-level semantic labeling and assign each voxel v_j a unique part instance ID (e.g., to differentiate separate legs of a table).

For both semantic and instance segmentations we follow original train/val division of Scannet dataset (Dai et al., 2017).

The Choice of Scene Understanding Levels. We evaluate the algorithms at three granularity levels for each object category: coarse-, middle- and fine-grained, roughly corresponding to evaluation in (Mo et al., 2019b).

Quality Measures. We evaluate semantic labeling and hierarchical segmentation models by inferring the semantic labels for entire input scenes and computing quality measures at each scene understanding level d_k separately. More specifically, for each class c present in the set of classes $\mathbb{C}_k$ at granularity d_k, we compute the standard Intersection over Union score IoU_c and the balanced accuracy score Acc_c. We report these per-class numbers along with mean IoU and mean balanced accuracy averaged over $\mathbb{C}_k$: $\text{mIoU}_k = 1/n_k \sum_{c \in \mathbb{C}_k} \text{IoU}_c$, $\text{mAcc}_k = 1/n_k \sum_{c \in \mathbb{C}_k} \text{Acc}_c$.

We additionally evaluate hierarchical semantic segmentation by averaging mIoU over all hierarchy levels $k \in \{1, \dots, K\}$.

Instance segmentation is assessed as object detection and thus evaluate this task using average precision (AP) with IoU threshold at 0.5. To generate object hypotheses, each instance is checked against a threshold of confidence equal to 0.25, to filter out noisy voxels.

3.3 Proposed Approach

The input to all our models is the voxelized object SDF, representing 3D geometry either with or without associated RGB values.

Part-level 3D Understanding. To predict parts in our multi-label formulation, we produce a set of softmax scores $p_j = (p_j^{d_1}, \dots, p_j^{d_n})$. Our models for semantic and instance segmentation tasks are 3D CNNs identical in architecture up to the last layer. We note that training a 3D CNN model for segmentation is a computationally challenging problem; thus, we opted to make it more tractable by using frameworks for sparse differentiable computations. Specifically, we use Minkowski Networks (Choy et al., 2019a; Choy et al., 2019b), a popular sparse CNN backbone, to implement large sparse fully-convolutional neural networks for geometric feature learning. We use the Res16UNet34C architecture that showed state-of-the-art performance on multiple scene understanding tasks. For each input voxel, the network produces a 32-dimensional feature vector that is further appropriately transformed by the last layer, accommodating a required number of predicted classes. All our networks are fully convolutional, enabling inference for scenes with arbitrary spatial extents.

We train the network in multiple setups, differing by the structure of supervision available to the network, each time evaluating labeling performance at each detail level d_i. Specifically, we define our loss function L to be a weighted sum of cross entropy losses for each level of detail

$$L(p,y) = \sum_{k=1}^{K} \alpha_k L_{\text{CE}}(p^k, y^k) \qquad (1)$$

and specify a set of weighting schemes for $\alpha_1, \dots, \alpha_K$. This loss structure allows expressing both "flat" segmentation formulations (e.g., choosing $\alpha_k = \delta_{ki}$ to segment at level d_i only), that we view as baselines, and multi-task formulations that integrate training signal across multiple levels of semantic detail.

Similarily to (Mo et al., 2019b), we approach this task using *bottom-up,* and *top-down* methods. Bottom-up method performs segmentation at the most fine-grained level and propagates the labels to object level, leveraging the taxonomy structure. Conversely, the top-down approach infers labels first at coarse level (starting with objects) and subsequently at finer levels (parts), recursively descending along predicted taxonomy branches. We note that *multi-task training* using (1) also results in a hierarchical segmentation method, and include it in the comparison.

Part Instance Segmentation. We employ a discriminative loss function which has demonstrated its effectiveness in previous works (De Brabandere et al., 2017; Pham et al., 2019b) and integrates intra-instance clustering and inter-instance separation terms along with a small regularization component. Compared to architectures that use region proposal modules (Yi et al., 2019; Pham et al., 2019b; Engelmann et al., 2020), this results in a more computationally lightweight architecture, while reducing instance segmentation problem to a metric learning task. At inference time, we cluster voxel feature vectors to produce part instances in a scene using mean-shift algorithm (Comaniciu and Meer, 2002).

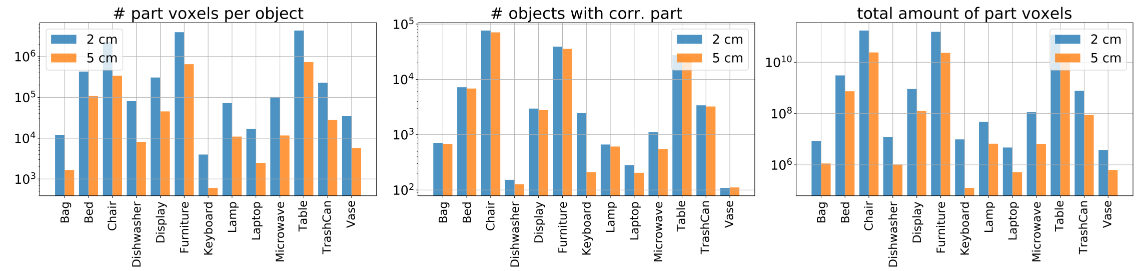


Figure 4: Histograms of object statistics for the coarsest ($d = 1$) level of details: LEFT: number of voxels per object, CENTER: the number of corresponding objects in Scan2Part, RIGHT: the total number of object voxels in Scan2Part.

4 EXPERIMENTS

4.1 Experimental Setup

Data and Training. We select 80% of the scenes in ScanNet for training, keeping 5% as a mini-validation to tune hyperparameters, and put aside 20% scenes for testing. Due to the high imbalance some classes can be under-represented in a testset. Testset was selected through iterative discrete optimization, using desired 80/20 proportion through all classes (especially smaller ones).

We optimize our models using Adam (Kingma and Ba, 2014), initializing learning rate proportionally to the batch size with base learning rate = 0.003 and $\beta_1 = 0.9$, $\beta_2 = 0.999$. Learning rate during training follows Multi-step exponential decay schedule with $\gamma = 0.2$, and a weight decay of 10^{-4}. We train until early stopping with patience parameter of 10, or until we reach a maximum epoch limit of 200. Depending on the specific model training can take several hours on our dedicated server equiped with 4 Nvidia GTX 1080.

Baselines. SparseConvNet (Graham et al., 2018) is an implementation of sparse 3D CNNs; though it has a smaller number of sparse layer types, it has enough features to implement a U-Net style fully-convolutional architecture we need to compute geometric features of voxels. As a baseline, we also train a Dense 3D CNN (Lee et al., 2017) model based on U-Net architecture with ResNet blocks.

To train the baselines, we extract 16 volumetric crops of size 64^3 from each voxelized scene and randomly shuffle them across all scenes in the training split. The batch size was selected from 4 to 48 in different experiments, subject to fitting the model into the available GPU memory. Because this model's implementation requires slicing of the scene, it cannot be evaluated for per-instance metrics.

Semantic Part Segmentation. Table 5 summarizes our model specifications, differing by the choice of weights in (1). The first three rows correspond to training a single level-of-detail semantic segmentation model. The last three rows define a Multi-Task-Training (MTT) objective where learned features of the occupied voxels are projected using linear layers to different level labels, and their loss functions are combined for training.

Part Instance Segmentation. This task test the ability of our models to separate parts solely based on their shape and mutual position within object, disregarding semantic class information.

Hierarchical Segmentation. Performing hierarchical segmentation of voxels in a scene according to a taxonomy of objects and their parts described in 3.1 is a challenging task that can be approached in different ways:

- **Top-Down.** approach assigns labels to a voxel by solving a sequence of smaller classification tasks. This approach predicts the distribution of semantic object labels first, followed by prediction of first-level parts, and continuing until a voxel can be assigned a leaf label from the taxonomy. Every prediction produces SoftMax distribution over possible sub-parts; voxels are masked based on the ground truth of the ”parent” part during evaluation.
- **Bottom-Up.** A model predicts a leaf label to a voxel on the finest level-of-detail, and semantic part segmentation on any level-of-detail can be computed by ”projecting” up. Probabilities of part labels having the same ”parent” part are added together.

4.2 Comparative Studies

Can We Recognize Parts in Real-world Scans? Table 4 demonstrates part-level semantic 3D understanding performance of our approach vs. the baseline methods. We show that adding semantic information from multiple levels of detail via part objectives

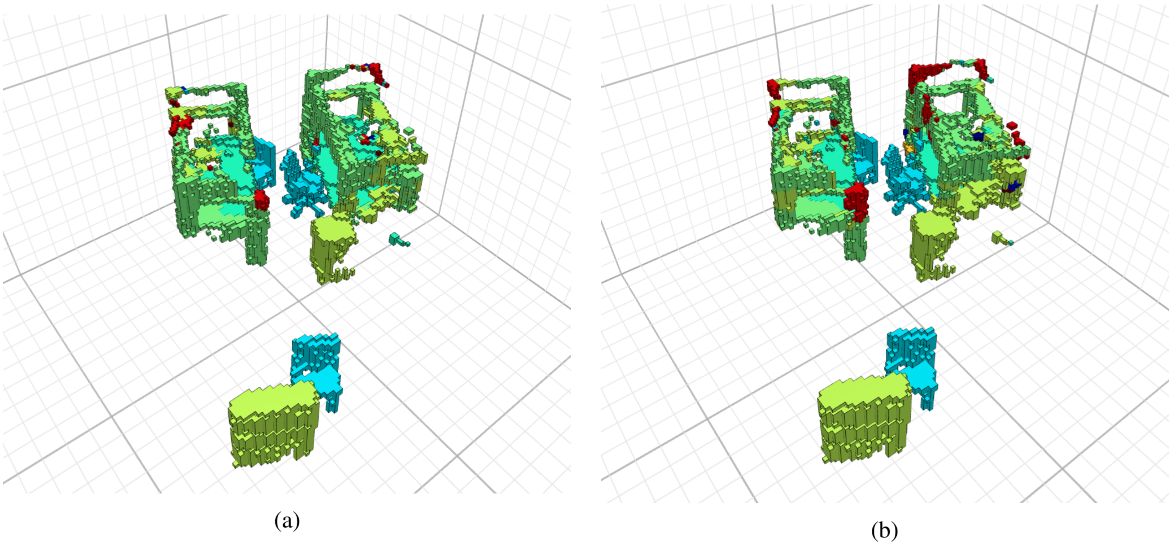


Figure 5: Semantic segmentation prediction from Minkowski Engine (a) and Submanifold (b) models.

Table 2: Instance segmentation results on levels d_1, d_2, and d_3.

Model	Lvl	Micr.	Disp.	Lamp	Lapt.	Bag	Stor.	Bed	Table	Chair	Dishw.	Trash.	Vase	Keyb.	**Avg**
Ours (w/ color)	d_1	0.790	0.676	0.794	0.852	0.808	0.798	0.795	0.852	0.900	0.884	0.840	1.000	0.874	0.839
	d_2	0.856	0.356	0.667	0.710	0.275	0.414	0.505	0.467	0.562	0.423	0.640	0.488	0.440	0.757
	d_3	0.228	0.146	0.200	0.234	0.484	0.279	0.743	0.667	0.490	0.176	0.333	0.244	0.807	0.711
num. instances		142	26	11	33	496	69	543	754	6	199	9	18	21	179

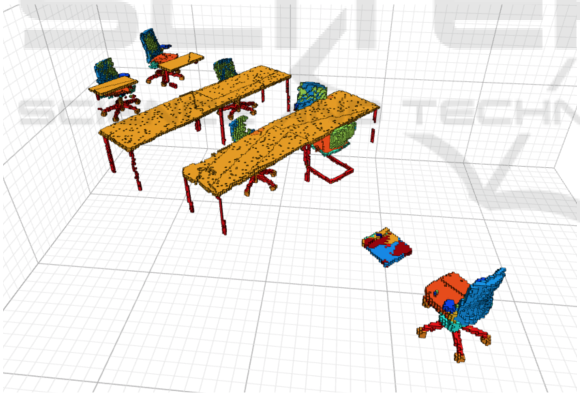


Figure 6: Qualitative hierarchical semantic segmentation results using our models, trained in each respective part-category level and our proposed method. Note the segmentation performance, particularly at finer levels in the parts taxonomy.

in in (1) significantly improves models segmentation performance across hierarchy levels.

Does Adding Part-level Annotation Improve Object-level 3D Understanding? Table 3 demonstrates results across different object classes and in different inference setups obtained at the object-level and lower level-of-detail.

Does Pre-training for Part Segmentation Help Achieve Hierarchical / Instance Tasks? We present performance of hierarchical semantic segmentation in Table 3. Note that despite the baselines are focusing solely on a single level of semantic detail, our method is able to leverage a multi-task objective in (1) to perform more efficient segmentation.

Part Instance Segmentation Results. We present instance segmentation results in Table 2, results indicate that segmentation accuracy decreases when the level of detail is reduced, presumably because object-part-masks become more generic, closet to shape primitives. The metric performs on par to the object level on particular objects with less shape variability (e.g., Bag, Bed, Table, Keyboard).

4.3 Ablative Studies

What is required for efficient part-level understanding? The following ablations can inform to what is more important in our problem domain.

Effect of Backbone Network. In Table 4 one can see that having a sparse backbone is crucial to the semantic segmentation ability of the model. There exists a moderate effect between model size/complexity and performance. We are confident that we got close to the limits of segmentation performance given the quality of data in the benchmark.

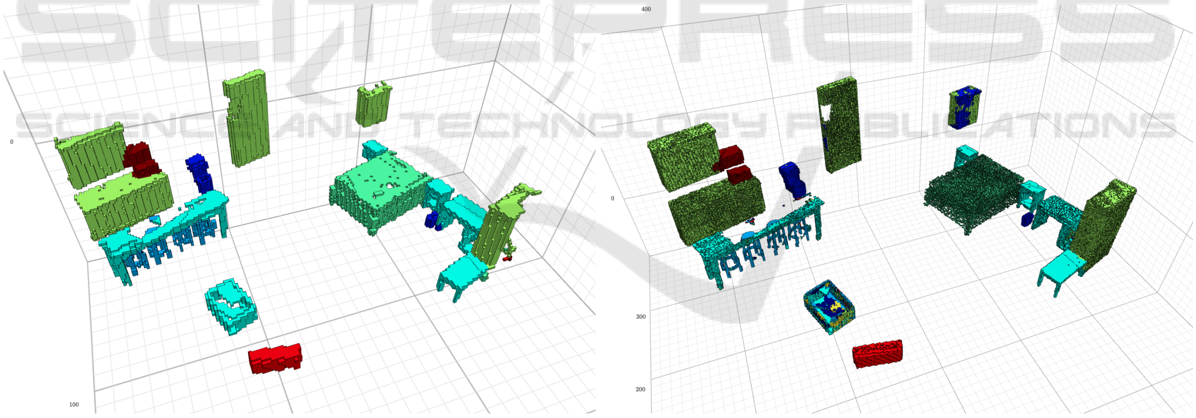
Effect of Loss. Most of our models in Table 4 are trained with weighted cross-entropy Loss to combat

Table 3: Hierarchical semantic segmentation results in terms of mean IoU and mean balanced accuracy for different semantic granularities. We include mIoU averaged over all hierarchy levels as an integral measure.

Model	Lvl	Micr.	Disp.	Lamp	Lapt.	Bag	Stor.	Bed	Table	Chair	Dishw.	Trash.	Vase	Keyb.	**Avg**
Flat	d1	**0.408**	0.704	**0.558**	0.202	0.581	**0.811**	**0.777**	**0.783**	**0.896**	0.003	**0.639**	**0.403**	0.112	**0.529**
	d2	0.199	0.471	0.526	0.298	0.346	0.035	0.423	0.351	0.301	0.183	0.199	0.302	0.033	0.282
	d3	0.070	0.262	0.380	0.170	0.214	0.213	0.231	0.344	0.193	0.000	0.102	0.208	0.029	0.186
Top-Down	d1	0.099	0.266	0.079	0.091	0.338	0.431	0.281	0.527	0.610	0.009	0.276	0.054	0.019	0.237
	d2	-	**0.769**	0.442	**0.856**	**0.661**	0.499	0.339	0.391	0.171	**0.697**	0.313	0.199	**0.585**	0.494
	d3	-	0.522	0.631	0.439	0.529	0.288	-	0.741	0.355	0.682	0.402	-	-	0.510
Bottom-Up	d1	0.377	0.626	0.489	0.170	0.515	0.792	0.673	0.748	0.872	0.001	0.582	0.281	0.062	0.476
	d2	0.188	0.469	0.000	0.041	0.155	0.053	0.006	0.126	0.147	0.000	0.064	0.063	0.031	0.102
	d3	0.279	0.542	0.198	0.385	0.260	0.232	0.110	0.404	0.335	0.145	0.267	0.128	0.034	0.210

Table 4: Fine-grained semantic part segmentation performance in terms of mean IoU and mean per-Instance IoU for different semantic granularities, using our model in various setups and baselines.

	Categ. mIoU, % ↑			Inst. mIoU, % ↑		
Method	d_1	d_2	d_3	d_1	d_2	d_3
Dense 3D CNN (Lee et al., 2017)	0.225	0.131	0.058	-	-	-
SparseConvNet (Graham et al., 2018)	0.4212	0.2561	0.1789	0.7751	0.7151	0.4923
Ours (w/o color)	0.5179	0.2913	0.2185	0.8387	0.7219	**0.5140**
Ours (w/ color)	**0.5290**	0.2848	**0.2231**	**0.8422**	0.7217	0.5137
Ours (MTT balanced)	0.5209	**0.3104**	-	0.8364	**0.7363**	-
Ours (MTT fine-grained)	0.4953	0.2929	0.2123	0.8065	0.7217	0.5093
Ours (MTT coarse)	0.5091	0.2963	0.1886	0.841	0.7253	0.4414
Num. classes	13	36	79	13	36	79



(a) (b)

Figure 7: Semantic segmentation results for the same scene voxelized at 5 cm^3 (a) and 2 cm^3 (b) voxel resolution. Despite coarser geometry yields somewhat more robust segmentation, it does not allow representing finer parts.

high unbalance for part labels in scenes. Training on part segmentation in a multi-task setup with different choices of objective weights has demonstrated a trade-off in the effectiveness of the geometric features and their ability to predict labels on different level-of-details.

Effect of Color Information. Surprisingly, introducing color information only contributes slightly to the performance of semantic part segmentation (see Table 4). This result disagrees with the expectations that voxel color would be a highly predictive factor for part correspondence. We hypothesize that this could be due to the high variability in lightning and acquisition conditions in the original ScanNet dataset.

Effect of Voxel Resolution. We conducted experiments at a 2.5× finer voxel resolution of 2 cm^3 and present their results in Table 6. Despite higher complexity geometry, we have observed an improvement in performance metrics across levels of detail. However, due to a significant increase in computational

Table 5: Our objective configurations in (1).

Configuration	α_1	α_2	α_3
Base coarse	1	0	0
Base middle	0	1	0
Base fine	0	0	1
MTT-12	.5	.5	0
MTT-123-coarse	.7	.2	.1
MTT-123-fine	.1	.2	.7

Table 6: Our method is able to effectively work on various voxel resolution leves. On 5cm resolution tipical training time for equal number of epochs increased from several hours to a day.

	Categ. mIoU, % ↑			Inst. mIoU, % ↑		
Method	d_1	d_2	d_3	d_1	d_2	d_3
Voxel size 5 cm^3	0.5179	0.2913	0.2185	0.8387	0.7219	0.5140
Voxel size 2 cm^3	0.6465	0.4138	0.3415	0.8958	0.8263	0.6295
Num. classes	13	36	79	13	36	79

requirements, we performed the majority of experiments on the 5 cm^3 version of our dataset.

5 CONCLUSION

We introduced Scan2Part, a novel method and a challenging benchmark for part-level understanding of real-world 3D objects. The core of our method is to leverage structural knowledge of objects composition to perform a variety of segmentation tasks in setting with complex geometry, high levels of uncertainty due to noise. To achieve that, we explore the part taxonomies of common objects in indoor scenes, on multiple scales and methods of compressing them for more effective use in machine learning applications. We demonstrated that specific ways of training deep segmentation models like ours sparse Residual-U-Net architecture on these novel tasks we introduced are better at capturing inductive biases in structured labels on some parts of a taxonomy but not the others. Further research on relationships between structure of real-world scenes and perception models is required and we hope our benchmark and dataset will accelerate it. In the near future we plan on releasing a second version of the dataset with background of scenes is not removed, so that and we can study performance of our models in scenarios closer to raw sensor data.

ACKNOWLEDGEMENTS

This work has been supported by the Russian Science Foundation under Grant 19-41-04109. The authors acknowledge the use of Zhores (Zacharov et al., 2019) for obtaining the results presented in this paper.

REFERENCES

Armeni, I., Sax, A., Zamir, A. R., and Savarese, S. (2017). Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., and Savarese, S. (2016). 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1534–1543.

Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A. X., and Nießner, M. (2019). Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2614–2623.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.

Bokhovkin, A., Ishimtsev, V., Bogomolov, E., Zorin, D., Artemov, A., Burnaev, E., and Dai, A. (2021). Towards part-based understanding of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7484–7494.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., and Zhang, Y. (2017). Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Chen, X., Golovinskiy, A., and Funkhouser, T. (2009). A benchmark for 3D mesh segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 28(3).

Chen, Z., Yin, K., Fisher, M., Chaudhuri, S., and Zhang, H. (2019). Bae-net: branched autoencoder for shape co-segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8490–8499.

Choy, C., Gwak, J., and Savarese, S. (2019a). 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084.

Choy, C., Park, J., and Koltun, V. (2019b). Fully convolutional geometric features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8958–8966.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. *arXiv preprint arXiv:1702.04405*.

Dai, A. and Nießner, M. (2018). 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468.

De Brabandere, B., Neven, D., and Van Gool, L. (2017). Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*.

Elich, C., Engelmann, F., Schult, J., Kontogianni, T., and Leibe, B. (2019). 3d-bevis: birds-eye-view instance segmentation. *arXiv preprint arXiv:1904.02199*.

Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., and Nießner, M. (2020). 3d-mpa: Multi proposal aggregation for 3d semantic instance segmentation. *arXiv preprint arXiv:2003.13867*.

Fisher, M., Ritchie, D., Savva, M., Funkhouser, T., and Hanrahan, P. (2012). Example-based synthesis of 3d object arrangements. In *ACM SIGGRAPH Asia 2012 papers*, SIGGRAPH Asia '12.

Garcia-Garcia, A., Martinez-Gonzalez, P., Oprea, S., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Jover-Alvarez, A. (2018). The robotrix: An extremely photorealistic and very-large-scale indoor dataset of sequences with robot trajectories and interactions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6790–6797. IEEE.

Graham, B., Engelcke, M., and van der Maaten, L. (2018). 3d semantic segmentation with submanifold sparse convolutional networks. *CVPR*.

Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., and Cipolla, R. (2016). Understanding real world indoor scenes with synthetic data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4077–4085.

Hoffman, D. D. and Richards, W. A. (1984). Parts of recognition. *Cognition*, 18(1-3):65–96.

Hou, J., Dai, A., and Nießner, M. (2019). 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4421–4430.

Hua, B.-S., Pham, Q.-H., Nguyen, D. T., Tran, M.-K., Yu, L.-F., and Yeung, S.-K. (2016). Scenenn: A scene meshes dataset with annotations. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 92–101. IEEE.

Johnson, J., Gupta, A., and Fei-Fei, L. (2018). Image generation from scene graphs. *CoRR*, abs/1804.01622.

Keselman, L., Iselin Woodfill, J., Grunnet-Jepsen, A., and Bhowmik, A. (2017). Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klokov, R. and Lempitsky, V. (2017). Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 863–872.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Lahoud, J., Ghanem, B., Pollefeys, M., and Oswald, M. R. (2019). 3d instance segmentation via multi-task metric learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9256–9266.

Lee, K., Zung, J., Li, P., Jain, V., and Seung, H. S. (2017). Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*.

Li, J., Xu, K., Chaudhuri, S., Yumer, E., Zhang, H., and Guibas, L. (2017). Grass: Generative recursive autoencoders for shape structures. *ACM Transactions on Graphics (TOG)*, 36(4):52.

Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., and Leutenegger, S. (2018). Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference (BMVC)*.

Liang, Z., Yang, M., and Wang, C. (2019). 3d graph embedding learning with a structure-aware loss function for point cloud semantic instance segmentation. *arXiv preprint arXiv:1902.05247*.

Liu, C. and Furukawa, Y. (2019). Masc: multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*.

McCormac, J., Handa, A., Leutenegger, S., and J.Davison, A. (2017). Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation?

Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N., and Guibas, L. J. (2019a). Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*.

Mo, K., Wang, H., Yan, X., and Guibas, L. J. (2020). Pt2pc: Learning to generate 3d point cloud shapes from part tree conditions. *arXiv preprint arXiv:2003.08624*.

Mo, K., Zhu, S., Chang, A. X., Yi, L., Tripathi, S., Guibas, L. J., and Su, H. (2019b). Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 909–918.

Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. (2019a). Real-time progressive 3d semantic segmentation for indoor scenes. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1089–1098. IEEE.

Pham, Q.-H., Nguyen, T., Hua, B.-S., Roig, G., and Yeung, S.-K. (2019b). Jsis3d: joint semantic-instance segmentation of 3d point clouds with multi-task point-wise networks and multi-value conditional random fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8827–8836.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*.

Russell, C., Kohli, P., Torr, P. H., et al. (2009). Associative hierarchical crfs for object class image segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 739–746. IEEE.

Song, S., Yu, F., Zeng, A., Chang, A. X., Savva, M., and Funkhouser, T. (2017). Semantic scene completion from a single depth image. *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., Nardi, R. D., Goesele, M., Lovegrove, S., and Newcombe, R. (2019). The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Sun, C.-Y., Zou, Q.-F., Tong, X., and Liu, Y. (2019). Learning adaptive hierarchical cuboid abstractions of 3d shape collections. *ACM Transactions on Graphics (TOG)*, 38(6):1–13.

Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T., and Yeung, S.-K. (2019). Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1588–1597.

Wang, P.-S., Liu, Y., Guo, Y.-X., Sun, C.-Y., and Tong, X. (2017). O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics (TOG)*, 36(4):1–11.

Wang, W., Yu, R., Huang, Q., and Neumann, U. (2018). Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2569–2578.

Wang, X., Liu, S., Shen, X., Shen, C., and Jia, J. (2019a). Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4096–4105.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., and Solomon, J. M. (2019b). Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12.

Wu, R., Zhuang, Y., Xu, K., Zhang, H., and Chen, B. (2019a). Pq-net: A generative part seq2seq network for 3d shapes. *arXiv preprint arXiv:1911.10949*.

Wu, Z., Wang, X., Lin, D., Lischinski, D., Cohen-Or, D., and Huang, H. (2019b). Sagnet: Structure-aware generative network for 3d-shape modeling. *ACM Transactions on Graphics (TOG)*, 38(4):1–14.

Xu, D., Zhu, Y., Choy, C. B., and Fei-Fei, L. (2017). Scene graph generation by iterative message passing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, B., Wang, J., Clark, R., Hu, Q., Wang, S., Markham, A., and Trigoni, N. (2019). Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746.

Yi, L., Guibas, L., Hertzmann, A., Kim, V. G., Su, H., and Yumer, E. (2017). Learning hierarchical shape segmentation and labeling from online repositories. *arXiv preprint arXiv:1705.01661*.

Yi, L., Kim, V. G., Ceylan, D., Shen, I.-C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., and Guibas, L. (2016). A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35(6):1–12.

Yi, L., Zhao, W., Wang, H., Sung, M., and Guibas, L. J. (2019). Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3947–3956.

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., and Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual conference of the cognitive science society*, volume 1.

Yuille, A. and Kersten, D. (2006). Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308.

Zacharov, I., Arslanov, R., Gunin, M., Stefonishin, D., Bykov, A., Pavlov, S., Panarin, O., Maliutin, A., Rykovanov, S., and Fedorov, M. (2019). “zhores”—petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in skolkovo institute of science and technology. *Open Engineering*, 9(1):512–520.

Zhang, B. and Wonka, P. (2019). Point cloud instance segmentation using probabilistic embeddings. *arXiv preprint arXiv:1912.00145*.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10.

Zhu, C., Xu, K., Chaudhuri, S., Yi, L., Guibas, L., and Zhang, H. (2019). Cosegnet: Deep co-segmentation of 3d shapes with group consistency loss. *arXiv preprint arXiv:1903.10297*.

Zhu, S.-C. and Mumford, D. (2006). A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362.