Towards Part-Based Understanding of RGB-D Scans

Alexey Bokhovkin\textsuperscript{1,2}, Vladislav Ishimtsev\textsuperscript{2}, Emil Bogomolov\textsuperscript{2}, Denis Zorin\textsuperscript{3,2}, Alexey Artemov\textsuperscript{2}, Evgeny Burnaev\textsuperscript{2}, Angela Dai\textsuperscript{1}

\textsuperscript{1}Technical University of Munich
\textsuperscript{2}Skolkovo Institute of Science and Technology
\textsuperscript{3}New York University

Figure 1: From an input RGB-D scan (left), we propose to detect objects in the scan and predict their complete part decompositions as semantic part completion; that is, we predict the part masks for the complete object, inferring the part geometry of any missing or unobserved regions in the scan. To achieve this, we predict the part structure of each detected object to drive a geometric prior-driven prediction of the complete part masks.

Abstract

Recent advances in 3D semantic scene understanding have shown impressive progress in 3D instance segmentation, enabling object-level reasoning about 3D scenes; however, a finer-grained understanding is required to enable interactions with objects and their functional understanding. Thus, we propose the task of part-based scene understanding of real-world 3D environments: from an RGB-D scan of a scene, we detect objects, and for each object predict its decomposition into geometric part masks, which composed together form the complete geometry of the observed object. We leverage an intermediary part graph representation to enable robust completion as well as building of part priors, which we use to construct the final part mask predictions. Our experiments demonstrate that guiding part understanding through part graph to part prior-based predictions significantly outperforms alternative approaches to the task of semantic part completion.

1. Introduction

Recently, we have seen remarkable advances in 3D semantic scene understanding, driven by efforts in large-scale data collection and annotation of 3D reconstructions of RGB-D scanned environments [5, 2], coupled with exploration of 3D deep learning approaches across 3D representations such as sparse or dense volumetric grids [55, 39, 5, 15, 4], point clouds [38, 40], meshes [13, 23], and multi-view [7, 50]. This has led to significant progress in both 3D semantic segmentation as well as 3D semantic instance segmentation [16, 15, 4, 25]. These have enabled a basis for 3D perception at the level of objects, which is essential for semantic understanding, but lacks finer-grained understanding often critical for enabling interactions with objects and reasoning about functionality (e.g., the seat part of a chair is for sitting on, a knob or handle enables opening doors or drawers).

At the same time, notable progress has been made in part segmentation for shapes [33, 32, 18]. However, these methods have been developed on synthetic datasets such as ShapeNet [3], of objects in isolation; this scenario is
much less complex than the objects observed in real-world environments. Thus, we aim to bring these two directions together and propose the task of semantic part completion, predicting the part decomposition of objects in real-world 3D environments, where observations are often cluttered and geometrically incomplete (e.g., due to occlusions, sensor limitations, etc). That is, from an RGB-D scan of a scene, we detect objects characterized by 3D bounding boxes and class labels, and for each object, we predict its complete part decomposition into binary part masks, with each part mask reflecting the part geometry of the complete object, including unobserved missing regions, to achieve a holistic understanding of the objects in an observed scene.

To achieve this part-based understanding of a scene, we propose to predict the full part graph for each detected object, and based on the predicted part graph, the geometric masks for each complete part. Predicting the part graph structure enables capturing the complete semantic structure of the object in a low-dimensional representation, allowing reliable prediction of missing and unobserved parts (e.g., for a four-legged table with one leg unobserved, the missing leg is easy to predict based on commonly observed table part patterns). Furthermore, this enables us to build and exploit strong part geometry priors for each predicted part in the part graph. We can then predict the part masks by finding similar part priors and refining them to produce final part mask predictions. This enables a robust decomposition of an RGB-D scan of a scene into its component objects and their constituent parts, including regions of objects that have been unobserved. We believe that this takes an important step towards enabling local interactions with objects and functionality analysis in real-world 3D scenes.

We formulate the task of semantic part completion for 3D scene understanding, informing comprehensive part-based object understanding of real-world scans. To address this part understanding, we propose an approach to decompose a 3D scan of a scene into its complete object parts, outperforming state-of-the-art alternative approaches for the task:

- We propose to predict part graph information for objects in real-world scan scenes as an intermediary representation that enables robust, part-based completion of objects.
- We leverage the predicted part graphs to guide prior-based prediction for effective inference of geometric part mask decomposition for the objects of a scanned scene.

2. Related Work

3D Object Detection and Instance Segmentation. Following the success of convolutional neural networks for object detection and instance segmentation in 2D images [12, 42, 41, 19], we are now seeing notable advances in 3D object localization and segmentation. Earlier approaches leveraging 3D convolutional neural networks developed methods operating on dense voxel grids using 3D region proposal techniques for detection and segmentation [47, 20]. Sparse volumetric backbones have also been leveraged to enable effective feature extraction on high-resolution inputs for improved 3D detection and segmentation performance [10, 16]. Recently, VoteNet [37] introduced a Hough Voting-inspired scheme for 3D object detection on point clouds. This was extended by MLCVNet [56] to incorporate multi-scale contextual information for improved detection performance. These approaches have now shown impressive performance for instance-level scene understanding; we aim to build upon this and propose to infer finer-grained part decomposition for each object in a 3D scan.

3D Scan Completion. Repairing and completing holes or broken meshes has been well-studied for 3D shapes. Traditional methods have mainly focused on repairing small holes by fitting geometric primitives, continuous energy minimization, or leveraging surface reconstruction for interpolation of missing regions [34, 59, 49, 27, 28]. Structural or symmetry priors have also been leveraged for shape completion [52, 31, 36, 46, 49]. Recently, generative deep learning approaches have been developed, with significant progress in 3D shape reconstruction and completion [55, 9, 17, 35].

In addition to operating on the limited spatial context of shapes, generative deep learning approaches have also been developed for completion of 3D scenes. Song et al. [48] developed a voxel-based approach to predict geometric occupancy of a single depth frame, leveraging a large-scale synthetic 3D dataset of scenes. Dai et al. [8] proposed an autoregressive approach for scan completion, enabling very large scale completion. SG-NN [6] presented a self-supervised approach towards 3D scan completion, enabling training only on real scan data. These approaches operate on geometric completion but without knowledge of individual object instances, which is fundamental to many perception-based tasks. RevealNet [21] introduced an approach to detect objects in a 3D scan and infer each object’s complete geometry, joining together geometric reconstruction with object-based understanding. We similarly aim to infer each object’s complete geometry from a partial scan observation, but infer a part decomposition of the object structure, enabling both finer-grained understanding as well as more effective object completion through its part structure.

Part Segmentation of 3D Shapes. Understanding the structure of a 3D shape by identifying shape parts has been long-studied in shape analysis. Various approaches have been developed for finding a consistent segmentation across a set of shapes without supervision of part labels [14, 24, 45, 22]. Recently, deep learning based approaches
have been developed to find part segmentation of shapes in a data-driven fashion [26, 58, 18]. To better capture more complex structures in the part layout of shapes, several methods propose to parse object parts as hierarchies [54, 53, 57, 33, 32]. Such hierarchically structured representations have also been adopted for 3D scene synthesis, leveraging a scene graph [11, 60, 30], where object instances rather than parts form the node primitives. We also adopt a relational inference of parts, but aim to operate on noisy, incomplete real-world scans of scenes with multiple objects, and so propose to combine our hierarchical part decomposition with strong geometric part priors.

3. Method

3.1. Overview

We address the problem of simultaneous part segmentation and completion of objects of real-world RGB-D scans, which are often noisy and incomplete. An overview of our approach is illustrated on Fig. 2. Given an input 3D scan $S$, we aim to predict a set of parts for each object in the scan, with each part representing the complete geometry of the part, including any missing or unobserved regions. From $S$, we first detect a set of object instances $O = \{o_i\}$ in the scene, as 3D bounding box locations and class category predictions. For each detected object in $O$, we then convert it into a $32^3$ occupancy grid representation, to inform our part segmentation and completion.

We then predict the part segmentation and completion for each detected object $o_i \in O$, resulting in a set of volumetric binary part masks. First, for a detected object $o_i$, we predict its semantic part structure $T_i$, with elements representing part class types, and the adjacency relations between the parts. This enables encoding the high-level, semantic part structure of the shape, which both facilitates completion of the shape structure, as missing parts are easy to identify in their semantic part structure, as well as guides the prediction of the geometry of each part. In particular, this allows us to leverage geometric part priors built for each part category. We construct the part priors based on clustering of train part masks for each part category, and learn to predict similar priors for each leaf in our predicted $T_i$, followed by a refinement of these priors to predict the final part mask geometry. This produces a semantic part decomposition of objects in a 3D scan while simultaneously inferring their complete part geometry.

3.2. Object Detection

From an input 3D scan, we first detect objects in the scene. We leverage a state-of-the-art 3D object detection approach, MLCVNet [56], as our object detection backbone. The input scan sampled to a point cloud, and object proposals are produced by voting [37], leveraging global contextual information at various scales. As output, we obtain 3D bounding box locations for each detected object. We then resample the input scan geometry within each detected box into $32^3$ occupancy grids $o_i \in O$ to inform our part decomposition.

For a detected object $o_i$ from the scan, represented as a $32^3$ occupancy grid of the scan geometry within its predicted bounding box, we encode the occupancy grid with four 3D convolutional blocks (consisting of convolution, group normalization and ReLU activation) and extract a fea-
tured encoding \( z_i \) of dimension 128, which is used to inform the part decomposition.

Object Orientation Prediction Since our object detection backbone predicts axis-aligned bounding boxes for each object, we additionally predict the orientation \( r_i \) of each object \( o_i \) from its feature \( z_i \) using an MLP. We assume that the up (gravity) vector is known in the scene, and thus predict the angle around the up vector by classifying the angle in \( n_\alpha = 8 \) bins of discretized angles \( \{0^\circ, 45^\circ, \ldots, 315^\circ\} \) with a cross entropy loss. The predicted object orientation helps to guide our prior-based part decomposition as described in Section 3.4.

3.3. Semantic Part Decomposition

For a detected object \( o_i \) from the scan, represented as a \( 32^3 \) occupancy grid of the scan geometry within its predicted bounding box, we aim to capture its high-level part structure from its cluttered and partial observation. We predict the semantic part structure \( T_i \) of the object; this facilitates completion of the object by predicting its high-level structure, as well as enables our prior-guided part geometry prediction.

We first encode the occupancy grid of \( o_i \) with a four 3D convolutional blocks (consisting of convolution, group normalization and ReLU activation), and extract a feature encoding \( z_i \) of dimension 128. We then decode \( z_i \) into a semantic part prediction, constructing a part set \( T_i \) with each element represented by its predicted part category and a 128-dimensional feature encoding. Inspired by StructureNet [32], we leverage a message-passing graph neural network for our semantic part prediction which enables relational inference between semantic parts. From \( z_i \), we predict part elements using an MLP to predict \( n_{\text{parts}} = 10 \) latent vectors \( \{z'_k\} \) that correspond to potential parts of \( o_i \). We additionally predict a tuple \( t_k = (c_k, s_k) \) for every part \( z'_k \), where \( c_k \) is the probability of part existence, \( s_k \) is the one-hot representation of the part category label. For each pair \( (z'_i, z'_j) \) of parts, we predict if they are adjacent or not, enforcing structural features to be learned by the message-passing network. We employ a cross entropy loss for the part category label, and binary cross entropy losses for part existence and adjacency relationships. This produces a high-level part summary of \( o_i \), where nodes \( \{z'_k\} \) represent part semantic information of the complete structure of \( o_i \), even if \( o_i \) has been partially observed. Note that this semantic part decomposition can be extended hierarchically to predict a full part tree, though we consider the first level children for our semantic part structure. We leverage this part semantic information to guide our final part decomposition as geometric part masks.

3.4. Prior-guided Geometric Part Decomposition

We then predict the final part decomposition by generating part masks for each element in the predicted semantic part arrangement \( T_i \), where each mask represents the complete geometry associated with the part, including regions that were unobserved in the initial scan observation. Rather than directly reconstructing the part geometry of each predicted part, we observe that object parts often maintain very similar geometry structures, which we leverage to obtain our final part decomposition. That is, we construct geometric part priors to aid in generating our complete part mask predictions, and learn to find similar geometric part priors which we then refine for a final prediction.

We construct our geometric part priors by \( k \)-means clustering of the binary part masks in the training set, inspired by the ShapeMask [29] construction of priors for novel 2D object segmentation. For each part type, we find \( K = 10 \) centroids of the part masks, and perform the clustering on the part masks in \( 32^3 \) grids of the canonical object space. This produces a set of part priors \( \{P_1, \ldots, P_M\} \) with \( M = n_{\text{classes}} K \). Various resulting part priors are visualized in Figure 3. Since objects in the real-world scan inputs may not be oriented in the canonical orientation of the object, we use the predicted orientation \( r_i \) to transform the priors to \( \{P_1^r, \ldots, P_M^r\} \).

Thus, to predict the part geometry associated for an element in the predicted semantic part set \( T_i \) with feature encoding \( z'_k \) and predicted part type \( t \), we use a one-layer MLP which takes as input \( z'_k \) and predicts a set of weights \( w_m \) used to construct an initial part reconstruction as:

\[
P^\text{coarse}_k = \sum_{m=1}^{M_i} w_m P^r_{m, i}
\]

where \( w = \text{softmax}(\phi(z'_k)) \), and \( \phi \) is a linear layer. We employ a proxy loss on this initial part reconstruction, using a mean squared error with a target part mask.

Such prior-guided part decomposition helps to reconstruct global structures in part masks such as symmetry and geometry in missing regions in the input observation. We
Table 1: Evaluation on semantic part completion on Scan2CAD [1]. We compare with state-of-the-art approaches for scan completion [6], followed by object detection [56], and then part segmentation [25, 32, 38]. By leveraging part structures to guide our prior-based approach, we obtain more accurate part decompositions.

<table>
<thead>
<tr>
<th>Method</th>
<th>Chamfer Distance (↓)</th>
<th>IoU (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chair</td>
<td>table</td>
</tr>
<tr>
<td>SG-NN + MLCVNet + PointNet++</td>
<td>0.078</td>
<td>0.111</td>
</tr>
<tr>
<td>SG-NN + MLCVNet + UNet</td>
<td>0.050</td>
<td>0.118</td>
</tr>
<tr>
<td>SG-NN + MLCVNet + PointGroup</td>
<td>0.074</td>
<td>0.102</td>
</tr>
<tr>
<td>MLCVNet + StructureNet</td>
<td>0.029</td>
<td>0.095</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>0.033</td>
<td>0.089</td>
</tr>
</tbody>
</table>

4. Results

We evaluate our proposed approach in comparison to alternative approaches for semantic part completion on real-world RGB-D scans. We use scans from the ScanNet dataset [5], containing 1513 reconstructed RGB-D scans, and evaluate with their train/val/test split of 1045/156/312 scenes, respectively. To train and evaluate the complete part decomposition for each object, we use the Scan2CAD [1] annotations of CAD model alignments from ShapeNet [3] to the ScanNet scans, coupled with the PartNet [33] annotations for the part decomposition of the ShapeNet CAD models. We train and evaluate on 6 object class categories representing the majority of parts (45 part types in total that we train and evaluate on) for these annotations. For a detailed specification of the part types used, we refer to the appendix.

To evaluate our part decompositions of the objects in a scan, we use a Chamfer Distance metric to capture structural consistency as well as an intersection over union (IoU) metric to capture more local consistency. For IoU, we evaluate 32^3 voxelizations of each predicted part in object space, compared to the Scan2CAD ground truth part. For Chamfer Distance, we use the predicted voxel centers as points, normalized to the unit box of the object. For both Chamfer Distance and IoU, we compute the metrics for each part type and average over all part types corresponding to an object class category. The class average is computed by averaging all resulting category numbers, and instance average computed by averaging the metrics of all part instances regardless of their object category. Note that to evaluate part segmentation without completion, we consider only predictions which overlap with the original scan geometry.

Comparison to alternative approaches. In Table 1, we compare to several state-of-the-art approaches for part segmentation and scan completion, coupled together to provide a complete part decomposition of the objects in a scan. As an alternative approach for this task, we consider scan completion followed by object detection and part instance segmentation. We employ the state-of-the-art scan completion approach SG-NN [6] to generate a prediction for the complete geometry of a partial scan observation, and then apply the object detections of with MLCVNet [56] (the same...
Table 2: Evaluation of part segmentation on Scan2CAD [1]. We evaluate part segmentation of visible geometry only, in comparison with state-of-the-art part segmentation [25, 32, 38].

<table>
<thead>
<tr>
<th>Method</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>class avg</th>
<th>inst avg</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>class avg</th>
<th>inst avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLCVNet + PointNet++</td>
<td>0.101</td>
<td>0.066</td>
<td>0.087</td>
<td>0.053</td>
<td>0.090</td>
<td>0.099</td>
<td>0.083</td>
<td>0.091</td>
<td>14.0</td>
<td>17.0</td>
<td>5.8</td>
<td>18.2</td>
<td>24.8</td>
<td>5.3</td>
<td>14.2</td>
<td>13.2</td>
</tr>
<tr>
<td>MLCVNet + UNet</td>
<td>0.052</td>
<td>0.082</td>
<td>0.062</td>
<td>0.034</td>
<td>0.093</td>
<td>0.068</td>
<td>0.065</td>
<td>0.060</td>
<td>24.1</td>
<td>13.4</td>
<td>9.3</td>
<td>14.3</td>
<td>16.1</td>
<td>9.4</td>
<td>18.5</td>
<td>17.9</td>
</tr>
<tr>
<td>MLCVNet + PointGroup</td>
<td>0.054</td>
<td>0.057</td>
<td>0.077</td>
<td>0.045</td>
<td>0.072</td>
<td>0.086</td>
<td>0.065</td>
<td>0.061</td>
<td>28.4</td>
<td>14.9</td>
<td>9.6</td>
<td>27.5</td>
<td>18.8</td>
<td>11.9</td>
<td>18.5</td>
<td>19.6</td>
</tr>
<tr>
<td>MLCVNet + StructureNet</td>
<td>0.039</td>
<td>0.084</td>
<td>0.062</td>
<td>0.034</td>
<td>0.075</td>
<td>0.083</td>
<td>0.063</td>
<td>0.056</td>
<td>32.6</td>
<td>2.1</td>
<td>9.4</td>
<td>23.1</td>
<td>16.1</td>
<td>15.4</td>
<td>16.5</td>
<td>15.4</td>
</tr>
<tr>
<td>Ours</td>
<td>0.037</td>
<td>0.071</td>
<td>0.060</td>
<td>0.031</td>
<td>0.069</td>
<td>0.058</td>
<td>0.054</td>
<td>0.048</td>
<td>36.9</td>
<td>15.3</td>
<td>11.1</td>
<td>29.3</td>
<td>27.4</td>
<td>21.5</td>
<td>23.6</td>
<td>27.8</td>
</tr>
</tbody>
</table>

Figure 4: Qualitative evaluation on semantic part completion in comparison with state of the art for part decomposition, including scan completion followed by part segmentation. Our approach produces more consistent, accurate part decompositions.

as those from the original scan, to eliminate any effect of possibly varying detections), obtain a final complete part decomposition by the state-of-the-art instance segmentation of PointGroup [25]. We also compare to StructureNet [32].
Table 3: Instance completion on Scan2CAD [1]. We evaluate object completion as a union of predicted part decompositions, and compare with state-of-the-art instance completion [21] and the union of StructureNet [32] parts as instances.

<table>
<thead>
<tr>
<th>Method</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>avg</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLCVNet + StructureNet</td>
<td>0.0033</td>
<td>0.0016</td>
<td>0.0074</td>
<td>0.0046</td>
<td>0.0194</td>
<td>0.0025</td>
<td>0.0079</td>
<td>29.5</td>
<td>21.9</td>
<td>22.4</td>
<td>24.1</td>
<td>23.5</td>
<td>32.4</td>
<td>25.6</td>
</tr>
<tr>
<td>RevealNet</td>
<td>0.0035</td>
<td>0.0070</td>
<td>0.0043</td>
<td>0.0020</td>
<td>0.0076</td>
<td>0.0078</td>
<td>0.0053</td>
<td>35.1</td>
<td>26.2</td>
<td>46.1</td>
<td>38.3</td>
<td>19.6</td>
<td>24.7</td>
<td>31.7</td>
</tr>
<tr>
<td>MLCVNet + UNet</td>
<td>0.0038</td>
<td>0.0103</td>
<td>0.0050</td>
<td>0.0119</td>
<td>0.0028</td>
<td>0.0059</td>
<td>39.7</td>
<td>30.0</td>
<td>62.6</td>
<td>28.2</td>
<td>17.4</td>
<td>37.7</td>
<td>35.9</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>0.0038</td>
<td>0.0075</td>
<td>0.0022</td>
<td>0.0053</td>
<td>0.0061</td>
<td>0.0045</td>
<td>0.0049</td>
<td>38.6</td>
<td>30.7</td>
<td>57.4</td>
<td>33.6</td>
<td>37.6</td>
<td>37.8</td>
<td>39.3</td>
</tr>
</tbody>
</table>

Part segmentation on 3D scans. In addition to our task of semantic part completion, we evaluate our approach in comparison to state of the art on part segmentation in Table 2. To evaluate part segmentation, we consider only the part predictions that intersect with the original scan geometry, and compare to PointGroup [25], StructureNet [32], and a UNet baseline, using the object detection of by MLCVNet [56]. For part segmentation, we see that our part structure reasoning coupled with geometric priors also produces more consistent part segmentations of the objects in a scan, particularly in IoU as our approach results in more locally consistent part structures.

<table>
<thead>
<tr>
<th>Method</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>Chamfer Distance (l)</th>
<th>Chamfer Distance (l)</th>
<th>IoU (t)</th>
<th>IoU (t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o Part Msg Pass</td>
<td>0.037</td>
<td>0.094</td>
<td>0.069</td>
<td>0.039</td>
<td>0.077</td>
<td>0.096</td>
<td>0.069</td>
<td>0.057</td>
<td>20.0</td>
<td>6.6</td>
</tr>
<tr>
<td>w/o Priors</td>
<td>0.036</td>
<td>0.093</td>
<td>0.067</td>
<td>0.044</td>
<td>0.058</td>
<td>0.101</td>
<td>0.067</td>
<td>0.056</td>
<td>21.8</td>
<td>7.3</td>
</tr>
<tr>
<td>No Prior Refine</td>
<td>0.034</td>
<td>0.093</td>
<td>0.069</td>
<td>0.034</td>
<td>0.057</td>
<td>0.096</td>
<td>0.064</td>
<td>0.055</td>
<td>22.5</td>
<td>7.6</td>
</tr>
<tr>
<td>Prior Refine (Abs)</td>
<td>0.036</td>
<td>0.089</td>
<td>0.065</td>
<td>0.034</td>
<td>0.067</td>
<td>0.105</td>
<td>0.066</td>
<td>0.055</td>
<td>21.4</td>
<td>7.5</td>
</tr>
<tr>
<td>Ours</td>
<td>0.033</td>
<td>0.089</td>
<td>0.069</td>
<td>0.033</td>
<td>0.054</td>
<td>0.096</td>
<td>0.062</td>
<td>0.053</td>
<td>22.1</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Table 4: Ablation study for our design decisions, evaluated for semantic part completion on Scan2CAD [1].
Figure 5: Qualitative results on real-world ScanNet [5] scenes using Scan2CAD [1] and PartNet [33] targets. Our approach effectively predicts each object’s complete geometry as a decomposition into semantic parts.

<table>
<thead>
<tr>
<th>Method</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>class avg</th>
<th>inst avg</th>
<th>chair</th>
<th>table</th>
<th>cab.</th>
<th>bkshlf</th>
<th>bed</th>
<th>bin</th>
<th>class avg</th>
<th>inst avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res. 16</td>
<td>0.034</td>
<td>0.088</td>
<td>0.072</td>
<td>0.054</td>
<td>0.061</td>
<td>0.109</td>
<td>0.070</td>
<td>0.055</td>
<td>28.4</td>
<td>10.5</td>
<td>13.5</td>
<td>20.9</td>
<td>18.5</td>
<td>21.2</td>
<td>18.8</td>
<td>22.8</td>
</tr>
<tr>
<td>Res. 32</td>
<td>0.033</td>
<td>0.089</td>
<td>0.069</td>
<td>0.033</td>
<td>0.054</td>
<td>0.096</td>
<td>0.062</td>
<td>0.053</td>
<td>22.1</td>
<td>7.7</td>
<td>13.0</td>
<td>18.1</td>
<td>17.3</td>
<td>22.0</td>
<td>16.7</td>
<td>18.3</td>
</tr>
<tr>
<td>Res. 64</td>
<td>0.045</td>
<td>0.098</td>
<td>0.058</td>
<td>0.044</td>
<td>0.067</td>
<td>0.100</td>
<td>0.069</td>
<td>0.060</td>
<td>18.8</td>
<td>5.6</td>
<td>9.9</td>
<td>10.5</td>
<td>14.7</td>
<td>19.3</td>
<td>13.1</td>
<td>15.4</td>
</tr>
</tbody>
</table>

Table 5: Evaluation of various object resolutions during training for semantic part completion on Scan2CAD [1].

there are many avenues for further development. For instance, a dense volumetric representation of parts may suffice for functionality analysis of furniture-type objects, but can struggle to generate very high resolution parts for small objects; we believe sparse [15, 4] or hierarchical [43, 51] approaches would complement our prior-based approach. Furthermore, objects are currently considered independently for each part decomposition, where relational inference between objects in a scene would help to explain noisy or unobserved part regions (e.g., multiple chairs or tables in a scene are often repeated instances of the same geometry).

5. Conclusion

In this paper, we have presented a new approach for the semantic part completion task of predicting a geometrically complete part decomposition for each object in a 3D scan. For each detected object in a scene, we exploit relational part structure prediction to guide a geometric part prior prediction, which is then refined to a final part decomposition, where each part is represented by its semantic part type and the geometry corresponding to the part, including any missing or unobserved regions in the scan. We show that our structural and prior-guided reasoning about object parts notably outperforms alternative approaches on this task. We believe that our approach makes an important step towards part-based understanding of 3D environments, and opens up new possibilities for part-level functionality analysis, autonomous agent interactions with an environment, and more.

6. Acknowledgements

We would like to thank the support of the Zentrum Digitalisierung.Bayern (ZD.B) and the Russian Science Foundation (Grant 19-41-04109). We additionally acknowledge the use of Skoltech CDISE HPC cluster Zhores.
References


