Point Cloud Semantic Scene Completion from RGB-D Images

Shoulong Zhang[1], **Shuai Li**[1,2,*], **Aimin Hao**[1,2], **Hong Qin**[3,*]
[1]Beihang University, Beijing, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Stony Brook University (SUNY), Stony Brook, USA
{shoulong.zhang, lishuai, ham}@buaa.edu.cn, qin@cs.stonybrook.edu

Abstract

In this paper, we devise a novel semantic completion network, called point cloud semantic scene completion network (PCSSC-Net), for indoor scenes solely based on point clouds. Existing point cloud completion networks still suffer from their inability of fully recovering complex structures and contents from global geometric descriptions neglecting semantic hints. To extract and infer comprehensive information from partial input, we design a patch-based contextual encoder to hierarchically learn point-level, patch-level, and scene-level geometric and contextual semantic information with a divide-and-conquer strategy. Consider that the scene semantics afford a high-level clue of constituting geometry for an indoor scene environment, we articulate a semantics-guided completion decoder where semantics could help cluster isolated points in the latent space and infer complicated scene geometry. Given the fact that real-world scans tend to be incomplete as ground truth, we choose to synthesize scene dataset with RGB-D images and annotate complete point clouds as ground truth for the supervised training purpose. Extensive experiments validate that our new method achieves the state-of-the-art performance, in contrast with the current methods applied to our dataset.

Introduction and Motivation

In our everyday life, we acquaint a new 3D indoor environment where the objects frequently occlude each other via our visual perception system routinely. Human beings could easily speculate the full geometry of invisible parts based on the semantic knowledge accumulated from our prior experiences. Motivated by this insight, semantic scene completion (SSC) has become a very active research area in 3D scene understanding in recent years. Thanks to the critical demand for high-quality 3D scene representation equipped with real-world scenario analysis, the SSC task promises its significance in robot navigation, auto-driving, and virtual & augmented reality (VR/AR) applications. However, due to the tremendous computational expenses brought by humongous 3D volumetric representation and the lack of favorable annotation datasets, the SSC research still suffers from its intrinsic challenges.

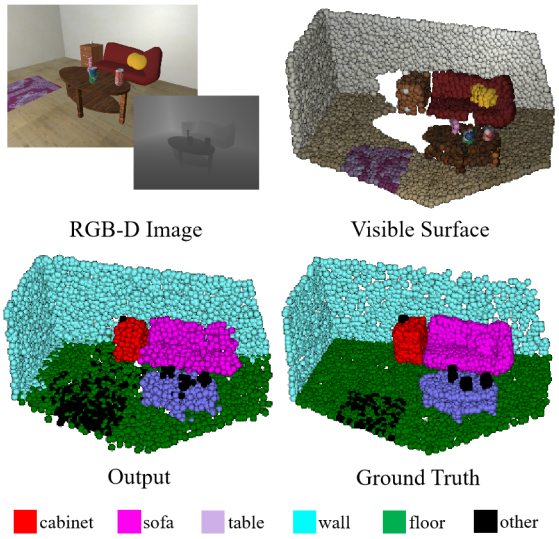

Figure 1: The input of our model (PCSSC-Net) is a partial scene point cloud generated by single view of RGB-D images. PCSSC-Net predicts a complete scene point cloud with semantic labels as output.

Previous SSC works adopt a volumetric representation (Song et al. 2017; Dai et al. 2018; Wang, Liu, and Tong 2020) and attempt to predict the occupancy and semantic category of each voxel grid. However, the volumetric map suffers from the low-resolution problem due to the high computational cost of the 3D convolution operation. In addition, the discretization causes the loss of geometric and semantic information. To ameliorate, the point cloud is a more favorable representation for complex structure and content (e.g., indoor scene environment involving many furnitures) thanks to its simplicity. Compared with the well-developed point cloud semantic segmentation techniques, existing point cloud deep generative models (Achlioptas et al. 2018; Lin, Kong, and Lucey 2018; Yang et al. 2018) are still less capable of recovering a complex and delicate structure from partial input in the indoor scene environment. Most of the available methods encode the incomplete point cloud into a global feature vector and generate the full geometry from a holistic decoder as a less meaningful point

*Corresponding authors.

set. Nevertheless, single high-dimensional global vector has its limitation in containing sufficient geometric and contextual information, and the entire geometry decoding task is not powerful enough to retrieve complex geometry neglecting crucial information like semantics. Although some works (Zhao et al. 2019b; Groueix et al. 2018; Liu et al. 2020) enhance the completion quality with a part-to-whole strategy in the latent space, it still has a large room to improve when faced with a complicated scene in the indoor environment.

To tackle the challenges mentioned above, this paper designs a novel network for semantic completion solely based on the point cloud in the indoor environment, and we name our new network design as point cloud semantic scene completion network (PCSSC-Net). In contrast with previous point cloud completion networks, our method first uses a divide-and-conquer strategy to hierarchically encode each sub-region of scene geometry and extract the contextual relationship among all involved sub-parts. We believe that the concatenation of multi-level (i.e., point-wise, patch-wise, and scene-wise) geometric and contextual features can contain sufficient amount of information and support a latent space where the decoder could infer the full geometry and their semantics. Second, given the fact that the semantics constitute an essential clue of geometry and natural partition of complicated indoor scene geometry, we devise a semantics-guided completion decoder that explicitly resorts to the semantic prediction that is of critical value to the network structure-inference ability at the semantic level in each category explained above.

Moreover, to better prepare for more feasible datasets and their possible training, we are faced with intrinsic difficulties associated with real-world scanned ground truth such as being incomplete, and frequently sparse, even for the indoor scene environment. Therefore, we construct a synthesized dataset which contains RGB-D images and complete annotated ground truth in the point cloud format. Based on this new dataset, our extensive experiments demonstrate that PCSSC-Net can generate a complete scene point cloud with the semantic prediction capability from partial input. Our method achieves better performance than other existing state-of-the-art methods. The primary contributions of this paper could be summarized as follows:

- We design a novel semantic scene completion network on point clouds, and the new network takes the point cloud generated by single RGB-D image as input.
- We propose a patch-based contextual (PBC) encoder to hierarchically extract both local geometric and contextual relationships among the sub-regions, with a goal of obtaining sufficient information to infer the full geometry. In addition, we articulate a semantics-guided completion (SGC) decoder to reconstruct the scene based on the semantic clusters.
- Extensive experiments confirm that our new method achieves the state-of-the-art performance, in comparison with existing methods on a geometrically complete and well-annotated synthetic indoor scene dataset being created in this paper.

Related Work

Semantic Scene Completion. The volumetric approach of semantic reconstruction from RGB-D images is known as semantic scene Completion (SSC). SSCNet (Song et al. 2017) first tackles the SSC problem with the development of 3D convolutional neural networks (3DCNN) and the dilated convolution. The subsequent SSC models (Garbade et al. 2019; Zhang et al. 2019; Li et al. 2019b) embrace the similar encoding strategy with the SSCNet. Currently, to avoid the costly Truncated Signed Distance Function (TSDF) encoding process, some methods (Li et al. 2019a, 2020b,a) attempt to extract features from 2D images and employ the 2D-3D projection. In addition, a newly-proposed method (Zhong and Zeng 2020) treats voxels as the regularly-arranged points to extract features via point convolution rather than 3D convolution. However, the intrinsic limitations of computational consumption and low-resolution representation still exist. Consider the simplicity of point clouds, our PCSSC-Net aims to improve the SSC task on point clouds.

Point Cloud Semantic Segmentation. The point cloud semantic segmentation techniques have been well developed starting from the seminal works, PointNet (Qi et al. 2017a) and PointNet++ (Qi et al. 2017b). As a basic application, various point cloud analysis models (Li et al. 2018; Li, Chen, and Lee 2018; Liu et al. 2019; Wu, Qi, and Li 2019; Zhao et al. 2019a; Wu et al. 2019) evaluate the segmentation performance as a crucial index of their methods. Moreover, there are specialized models for semantic segmentation. Similarity Group Proposal Network (SGPN) (Wang et al. 2018) predicts point-wised similarity matrix and confidence map for an accurate group proposal for semantic and instance clustering. Associatively Segmenting Instances and Semantics (ASIS) (Wang et al. 2019) method associates the semantic and instance segmentation as a highly intertwined multi-task learning problem. Besides, Superpoint Graph (SPG) (Landrieu and Simonovsky 2018) method divides point clouds into patches and executes a message passing for the contextual information. SPG introduces the possibility to deal with the semantic segmentation using a patch-based solution.

Point Cloud Completion. Guibas et al. introduces the first generative auto-encoder (Achlioptas et al. 2018) on point clouds. It retrieves the geometry from a global feature vector optimized by a generative adversarial network (GAN) in the latent space. Following the encoder-decoder architecture, FoldingNet (Yang et al. 2018) concatenates an abstract 2D grid to the global feature and integrally deforms the surface to fit the complete geometry. To recover more delicate details, one class of approaches adopts a multi-resolution technique. Point Completion Network (PCN) (Yuan et al. 2018) reshapes a coarse point set using the global feature and locally ”folds” tiny 2D grid to each point to generate a dense point cloud continuously. Point Fractal Network (PF-Net) (Huang et al. 2020) generates three-level resolution point sets to fit the complete ground truth by preserving all actual details. Another class of approaches employs an implicit part-to-whole strategy. 3D point capsule networks (3D-Capsule) (Zhao et al. 2019b) encode a point set with a

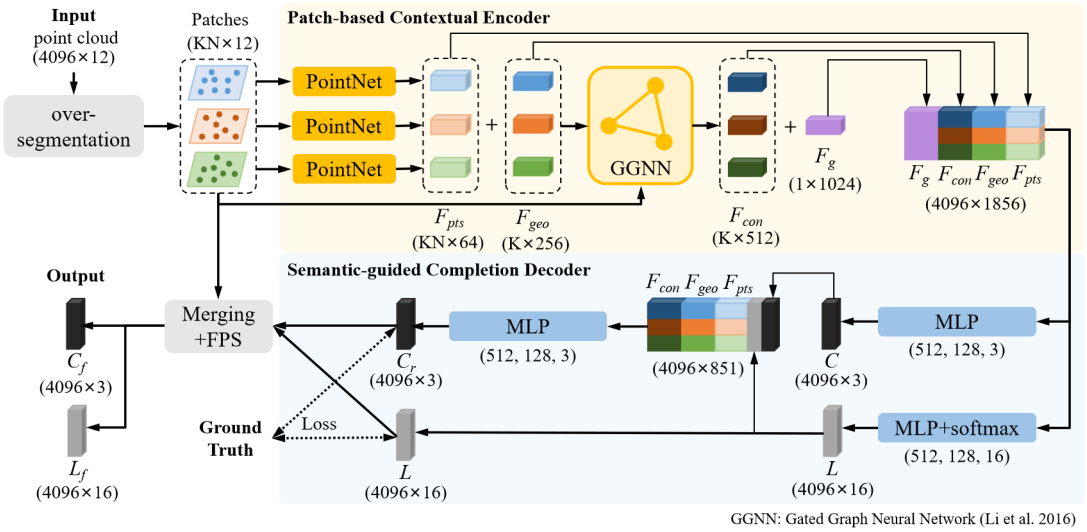

Figure 2: The pipeline of the PCSSC-Net. We first over-segment the input point could into patches. Given K patches (only 3 for illustration) with N points in each, our PBC encoder next learns point-wise, patch-wise, and scene-wise features for encoding both geometry and contextual information. The SGC decoder then retrieves the labeled complete geometry in two-stream where the semantic predictions are explicitly considered as a feature for geometric refinement. We finally merge the predicted results with the input to preserve the actual structures, and output the new point locations C_f with semantic assignments L_f.

series of encoders to obtain several codes to map different locations of the shape. Similarly, AtlasNet (Groueix et al. 2018) and Morphing and Sampling Network (MSN) (Liu et al. 2020) decode a global vector using a group of parametric surfaces in the latent space to recover disparate pieces of the point cloud. Unlike previous works, our PCSSC-Net explicitly extracts the geometric and contextual information in patches and produce points based on their semantic categories instead of less meaningful regions.

New Method

Overview. Given a partial scene point cloud, our approach is expected to predict a point cloud representing the complete scene geometry with the semantic assignment to each point. We illustrate the architecture of our PCSSC-Net in Figure 2. Our method semantically completes a scene from an RGB-D view in three steps. In the first step, we over-segment the incomplete point cloud into a collection of geometrically simple patches using an existing algorithm (Landrieu and Simonovsky 2018). Large patches may result in complex geometry due to multiple objects. The over-partitioned patches no longer carry valuable local information, and also lead to the increase of the model complexity (i.e., in $O(n^2)$). We utilize roughly 200 patches per scene on average. We also construct the adjacency graph of the sub-regions. In our experiment, a sub-cloud is connected to the five nearest local regions in the Euclidian space. The patches are sampled to contain an equal number of points for the batch training technique. Second, an autoencoder processes the sub-regions with the proposed PBC encoder and the SGC decoder. The PBC encoder hierarchically extracts point-wise, patch-wise, and scene-wise geometric and contextual features. With the concatenated multi-level features, the SGC decoder can take advantage of the encoded information to predict the semantic labels and the new point locations. Finally, we merge the partial input with the generated labeled points to preserve the original delicate geometric structures. The training algorithm is detailed in Algorithm 1.

Patch-based Contextual Encoder. Our PBC encoder is designed to hierarchically learn both geometric and contextual semantic information of the partial scene point cloud. Unlike previous point cloud completion networks, our PBC encoder takes two steps to encode the point cloud with an explicit divide-and-conquer strategy.

The first step is aimed to encode the low-level geometric features in each sub-region of the scene. Each sub-cloud includes N (N=25 in our experiments) points. The initial annotation of the point is a 12-dimension feature, including the coordinates, the RGB values, the normal vector, and the differential coordinates. The differential coordinate contains the curvature information of each location. Considering the simplicity and the availability of PointNet (Qi et al. 2017a), we use a shared-weights PointNet to extract point-wise features F_{pts} and a patch-wise feature F_{geo} individually for each sub-cloud. The point-wise 64-dimension features F_{pts} differ each point from the others for a satisfactory distinguishability in category assignment. Since PointNet deals with only one limited region at a time, the patch-wise 128-dimension feature F_{geo} can describe the patch location and the local geometric patterns.

For the second step, we adopt the propagation model of GGNN (Li et al. 2015) to extract the contextual relationship of patches from their adjacency graph. The initial node annotations are the combination of the patch-level geometric features: F_{geo}, the average coordinates, and the average normal vectors of patches. Using the node-level output and

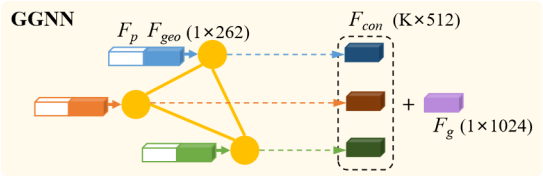

Figure 3: Gated graph neural network (GGNN) (Li et al. 2015). The initial node annotation is the concatenation of F_{geo}, the coordinates of center and the average normal vector (F_p, 6-dimension). The GGNN outputs the patch-wise contexutal features F_{con} and the global description F_g.

the graph-level output models of GGNN, we can achieve a patch-wise contextual feature F_{con}, and a scene-wise global description F_g in this step. The F_{con} aggregates the contextual semantic information from the nearest five units, which can be treated as expanding the preception field. The F_g encodes the scene context with the largest preception field.

Thus, our PBC encoder can provide both low-level geometric features and high-level semantic contextual features. It is our belief that the concatenation of those multi-level feature vectors can support a latent space in which our decoder can retrieve the complete scene structure from the partial input and predict the semantic class of each point.

Semantics-Guided Completion Decoder. Our SGC decoder uses a two-stream structure for semantic segmentation and completion tasks. The multi-level feature concatenation of F_{pts}, F_{geo}, F_{con}, and F_g is the input of our decoder. Since the concatenated feature size KN is probably not equal to 4096, we use the farthest point sampling (FPS) or duplication to obtain the expected size. For the segmentation stream, we use a three-layer multi-layer perceptron (MLP) block (512, 128, 16) and a softmax layer to shrink the feature channels to predict the category label L for each point. The semantic label matrix has a size of 4096×16. For the completion stream, the concatenated feature is also the input for the first MLP. The first MLP (512, 128, 3) is designed to reshape the entire point cloud globally and output the intermediate point coordinates C with a size of 4096×3. We observe that the global code F_g is not much helpful in retrieving the detailed structures. Since the second MLP (512, 128, 3) is designed to refine the predicted point cloud geometry, we replace the F_g with the semantic assignments and the intermediate point coordinates as extra information to the second MLP. Since the semantics can be a natural partitioning of the scene, the point-wise semantic prediction feature offers an explicit hint to cluster points in a latent space. Furthermore, with the cooperation of the proposed loss function, the semantic labels can also guide the second MLP to fit a complete geometry based on the semantic class in the training process. In the end, our SGC decoder outputs the semantic class label L and the refined point coordinates C_r in the size of 4096×16 and 4096×3 respectively.

Loss. We choose the chamfer distance for the reconstruction loss. The chamfer cistance (CD) and the earth mover's distance (EMD) are often considered in evaluating the similarity of two different point sets due to their irrelevant point order. The CD is used in our training phase because of its better efficiency and less computational consumption than the EMD. The CD value of the output point sets S and the ground truth point set S_{gt} is:

$$\mathcal{L}_{com} = \frac{1}{|S|}\sum_{x\in S}\min_{y\in S_{gt}}\|x-y\|_2 + \frac{1}{|S_{gt}|}\sum_{y\in S_{gt}}\min_{x\in S}\|y-x\|_2. \tag{1}$$

We choose the cross-entropy loss for the semantic segmentation task. The cross-entropy loss between predicted labels L and the ground truth L_{gt} is defined as:

$$\mathcal{L}_{seg} = \sum_{\hat{l}\in L, l\in L_{gt}} -l\log(\hat{l}). \tag{2}$$

Considering the completion task probably changes the location of points, we assume that the points maintain their original semantic classes after reshaping. Thus, we can construct the semantic label ground truth to calculate the order-dependent cross-entropy loss.

The CD is only the global constraint for the entire geometry and is badly influenced by the majority of the points. To encourage the SGC decoder to retrieve the object-level structure, we employ our semantic predictions as guidance for calculating the CD of each category. We add the weighted sum of these semantics-based CDs to the loss function in order to constrain the network to complete each class delicately rather than only the global geometry. In summary, our multi-task loss function is defined as the linear combination of the reconstruction loss, the cross-entropy loss, and the semantics-based CDs. The formulation is:

$$\mathcal{L}_{ssc} = \alpha\mathcal{L}_{com} + \beta\mathcal{L}_{seg} + \gamma\sum_{c=1}^{C}\omega_c\mathcal{L}_{com}^c, \tag{3}$$

Algorithm 1: Main Steps of PCSSC-Net Training.

Input: The incomplete point cloud $C_{init} \in \mathbf{R}^{4096\times 3}$ and hyperprameters α, β, γ, and $\{\omega_c\} \in \mathbf{R}^C$
Output: Point localtions $C_f \in \mathbf{R}^{4096\times 3}$ and semantic assignments $L_f \in \mathbf{R}^{4096\times 16}$.

```
Over-segment C_init into K patches P;
while not converged do
    foreach patch in P do
        Extract point-wise F_pts and patch-wise F_geo
          geometric feature using PointNet;
    Extract patch-wise contexutal feature F_con and
      global feature F_g using GGNN;
    Concatenate F_g, F_con, F_geo, and F_pts as F;
    Predict point-wise semantic labels L with F;
    Predict intermediate point locations C with F;
    Concatenate F_con, F_geo, F_pts, L, and C as G;
    Refine point locations C_r with G;
    Optimize total loss L_ssc;
Assign L to input points and merge C_init with C_r;
Down-sample merged points to the target number by
  FPS;
```

models	CD	bathtub	bed	shelf	cabinet	chair	desk	door	floor	sink	sofa	table	toilet	wall
DDRNet	1.87	64.80	16.02	45.71	133.64	28.45	12.04	33.56	**1.01**	172.23	29.76	20.28	148.25	**3.08**
Ours	**1.58**	**35.93**	**5.28**	**13.75**	**58.51**	**26.91**	**8.72**	**2.96**	2.18	**7.38**	**27.59**	**18.56**	**92.31**	4.35

Table 1: Semantic scene completion results. One metric is the chamfer distance (CD) between the output and the ground truth. The other metric is the semantic chamfer distance (SCD) of each class (all values $\times 1000$).

models	CD	bathtub	bed	shelf	cabinet	chair	desk	door	floor	sink	sofa	table	toilet	wall
Input	4.31	5.13	5.62	2.28	6.57	2.19	2.41	1.35	1.31	3.01	4.40	2.17	7.13	5.59
FCAE	4.68	9.24	6.00	8.39	6.73	9.51	5.55	7.11	1.61	62.66	5.72	10.44	12.91	8.39
FoldingNet	4.81	8.17	5.72	8.98	7.48	9.08	5.33	4.85	1.48	55.21	5.52	11.13	12.06	8.42
PCN	4.72	9.19	5.15	6.89	6.26	9.17	4.83	6.64	1.58	44.26	4.88	9.52	11.03	8.59
3D-Capsule	3.91	8.59	5.00	7.00	6.06	7.84	4.47	4.84	1.37	47.66	4.91	7.69	10.35	5.51
MSN	2.26	4.77	2.53	2.98	3.25	3.83	2.58	2.01	1.58	2.97	2.69	3.51	4.87	2.65
Ours	**1.58**	**2.64**	**1.88**	**1.65**	**2.50**	**1.81**	**1.24**	**1.29**	**1.10**	**2.57**	**1.94**	**1.64**	**3.55**	**1.80**

Table 2: Completion results. One metric is the CD between the predicted scene and the ground truth. We assign the semantic label to each output point as the same class of the nearest point in the ground truth to calculate the SCDs (all values $\times 1000$).

where α, β and γ are the weights for balancing the influence of each tasks, $\mathcal{L}_{com}^{c}$ is the CD of semantic class c between the output and the ground truth, and ω_c is the weight of $\mathcal{L}_{com}^{c}$, which is used to control the degree of isolation of class c against other semantic classes.

Merging and Sampling. Since the limited ability of the auto-encoder, not all geometric details can be learned in the model. As in MSN and PF-Net (Liu et al. 2020; Huang et al. 2020), the output is the combination of the predicted points and the geometry signal directly from the input. In the MSN, it uses the minimum density sampling (MDS) to sample the output and the input simultaneously and equally, and merge them with a residual block. For the PF-Net, it only predicts the missing parts and reserves all points of the input. Similarly, we merge the input points with the predicted scene point cloud preserving the original structures. As the features obey the original point order, our semantic prediction is also the segmentation result of the input partial point cloud. Therefore, we can uniformly and simultaneously sample 4096 points (C_f) and 4096 semantic labels (L_f) using the FPS algorithm.

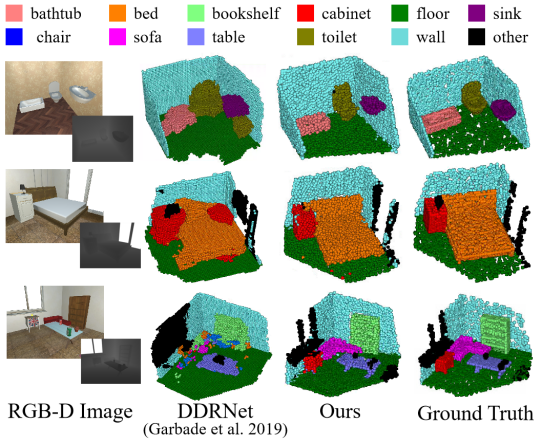

Figure 4: Semantic scene completion results. From left to right: input RGB-D image, DDRNet, our PCSSC-Net, and voxelized ground truth. The colors of the points represent the semantic classes. All voxels are illustrated as points.

Experiments and Evaluations

Dataset Preparation and Training Details. Existing real-world 3D scene datasets tend to be incomplete with occlusions and noises while being reconstructed from partial scans, so they are far from ideal if serving as the ground truth. At the same time, the existing volumetric synthesized datasets only provide partial observations and are not suitable either for our data format choice. Instead, to focus our key research effort on incompleteness due to objects' self-occlusion and inter-occlusion, we synthesized a new dataset with 12 layouts of the SceneNet RGB-D dataset (McCormac et al. 2017) and 263 typical indoor object models of the ShapeNet dataset (Chang et al. 2015) with realistic textures. Our dataset contains 1912 different viewports of 500 scene meshes generated through the Unity3D[1] engine, including bathroom, bedroom, living room, and office. We manually arranged camera viewports to guarantee the object meshes utterly present in the frustum so that the ground truth contains intact structures. The dataset has a similar scale of the NYUv2 dataset (Silberman et al. 2012), which includes 1449 scene data. Our synthesized dataset obeys typical room configurations (i.e., the chair is around the table, the TV is face to the sofa). Each viewport generates a group data of RGB image, depth image, and semantically annotated complete scene point cloud covering 16 semantic classes. We divide our dataset into 1520 and 392 scenes for the training and testing purpose. As a data-driven approach, our current model has potential to reconstruct real scenes with real-world training datasets in the future. We anticipate only slight modifications on the number of patches and perhaps denser adjacency graphs.

Our model is implemented in PyTorch. We trained our model on Nvidia RTX 2080Ti GPU for roughly 70 hours with a batch size of 8. We choose ADAM for the optimizer.

[1]Unity3D engine. URL: https://unity.com/

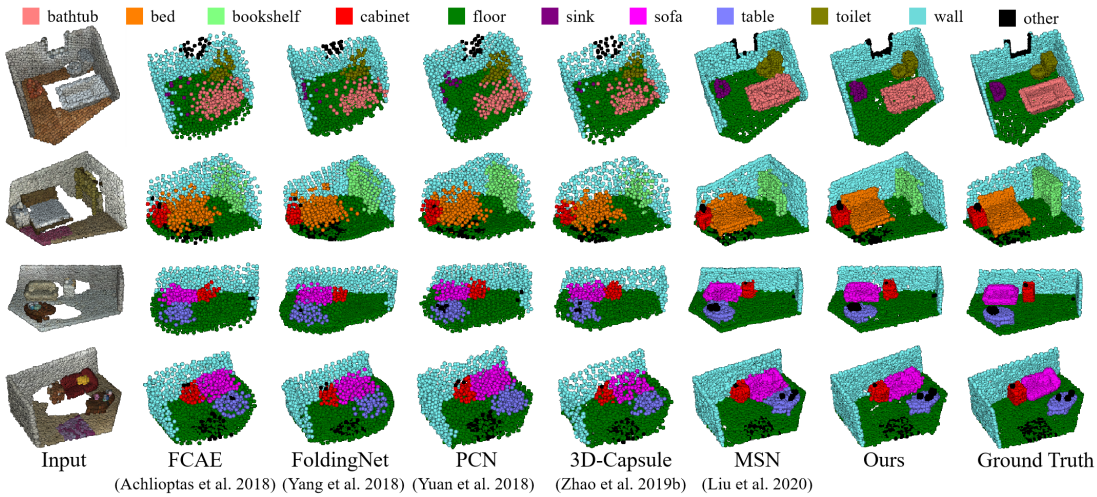

Figure 5: Visualization of the completion results. From left to right: input, FCAE, FoldingNet, PCN, 3D-Capsule, MSN, our PCSSC-Net and groud truth. The color of the input is the original value from the RGB image. Besides, the color of the output points does **NOT** indicate the point-level semantic prediction. It is assigned to the same class of the nearest point in the ground truth for better visualization. Each output contains 4096 points.

The initial learning rate is set to 0.001, and the decay rate is 0.7 for every ten epochs. We employed RELU as the activation function except for the GGNN with respect to the original implementation of the GGNN. For the parameters in the loss, we set $\alpha = 0.005$, $\beta = 1.5$ and $\gamma = 0$ for the first 10 training hours. Then, we set $\alpha = 0.015$, $\beta = 1.5$, and $\gamma = 1.5$ for geometry refinement based on their semantic classes. The CD weight for the chair class is set to 0.01 and 1 for other classes.

Evaluation Metrics. To compare with the volumetric representations, we treat each voxel as an abstract point so that the voxelized map can be seen as a scene point cloud. We adopt the CD as the metric to evaluate the global completion performance. The outputs of all evaluated methods contain or are sampled up to 4096 points with the same scale in a box of [-1, 1], which guarantees the validation of CD values' comparison. By taking the object-level structures and semantic segmentation results into account, we propose a semantic chamfer distance (SCD) metric, which is calculated between the same semantic class of the output point cloud and the ground truth. The SCDs depend on both semantic segmentation and completion performance.

To compare with the point cloud completion methods, we calculate SCDs for all models (including ours) differently. Since the completion methods do not involve the semantic prediction, we assign a label to each point according to the semantic class of the nearest point in the ground truth because it is most likely to complete that semantic class based on its location. We use the CD and the SCD to measure the performance of recovering the global geometry and object-level structure for each method.

Semantic Scene Completion. We compare our PCSSC-Net with a current semantic scene completion network **DDRNet** (Li et al. 2019a). DDRNet is a promising SSC method using the multi-level feature fusion encoding with residual 3DCNN modules. We train DDRNet on our synthesized dataset from scratch. With a different format of 3D representations, each voxel is regarded as a point in the center. The illustration of the results is shown in Figure 4. As we can see in Table 1, our PCSSC-Net performs better than DDRNet in the global recovery and the completion of the majority of semantic classes. In the experiment, the DDRNet can retrieve the floor and wall geometry more precisely than PCSSC-Net due to the structural arrangement of the voxel. The PCSSC-Net is competitive in reconstructing complicated geometry than the volumetric DDRNet.

Point Cloud Completion. We compare our PCSSC-Net with the state-of-the-art point cloud completion models. The **Input** is the original partial input point cloud. **FCAE** is proposed in (Achlioptas et al. 2018). We borrow the encoder of PointNet++ (Qi et al. 2017b) and three fully-connected layers as decoder. **FoldingNet** (Yang et al. 2018) folds an abstract 2D grid to fit the shape with a global code. **PCN** (Yuan et al. 2018) reshapes the geometry using the global description and refines the details with the local folding operation. FCAE, FoldingNet, and PCN all recover the complete point cloud based on a global feature vector. Besides, **3D-Capsule** (Zhao et al. 2019b) uses the multi-MLP blocks to implicitly encode the different sub-regions in the latent space. **MSN** (Liu et al. 2020) employs the multi-MLP modules to implicitly retrieve the parts of the complete structure and uses the expansion penalty to isolate them. We retrain those networks from scratch on our scene dataset. As we can see in Table 2, the FACE, FoldingNet, and PCN are not capable of retrieving the scene structure comparing with the results of the input. However, our PCSSC-Net performs the best with the smallest CD and SCD values. Our CD value is only 87.2% and 50.4% to that of the MSN and the 3D-Capsule. Different from MSN and 3D-Capsule manipulating patches implicitly in the latent space, we could notice from

models	Acc	mIoU	bathtub	bed	shelf	cabinet	chair	desk	door	floor	sink	sofa	table	toilet	wall
PN	94.5	73.3	83.5	81.8	59.2	55.0	52.1	77.3	87.9	99.1	61.2	67.1	63.2	63.0	96.6
PN++	96.9	87.4	**92.2**	95.5	82.2	77.8	82.1	86.8	83.3	96.7	**89.4**	87.1	86.0	82.9	97.1
Ours	**97.3**	**88.2**	90.5	**96.9**	**84.9**	**79.0**	**83.7**	**87.4**	**89.1**	**99.1**	89.2	**87.7**	**86.0**	**86.5**	**97.5**

Table 3: Semantic segmentation results. The metric contains overall accuracy, mean IoU, and IoU of each semantic class (%). The result of our model is the segmentation prediction of the completed point cloud.

models	CD	bathtub	bed	shelf	cabinet	chair	desk	door	floor	sink	sofa	table	toilet	wall
A	1.68	2.66	2.27	1.67	3.03	**1.79**	1.38	1.21	1.22	2.98	2.09	1.72	3.66	1.84
B	1.62	2.64	2.07	1.73	2.67	1.86	1.30	**1.20**	1.17	3.02	2.02	1.71	**3.53**	1.85
Ours	**1.58**	**2.64**	**1.88**	**1.65**	**2.50**	1.81	**1.24**	1.29	**1.10**	**2.57**	**1.94**	**1.64**	3.55	**1.80**

Table 4: Completion results on ablated versions. Version A removes the semantic segmentation stream in the SGC decoder. Version B eliminates the semantics-based CD in the loss function during the training process (all values $\times 1000$).

our experiments that our explicit processing of the partitioning of the scene is the key to such significant improvement. The values of SCD vary in different categories. The classes with simple structures (i.e., floor, door, and wall) have relatively small SCD values, and other semantic classes are more challenging to be completed with larger SCD values. With the smallest SCD values, our PCSSC-Net can achieve a higher degree of semantic class isolation and completeness than the other point cloud completion models.

In addition, as an auxiliary task, our semantic segmentation results are also satisfactory comparing with the baseline methods, which are **PN** (Qi et al. 2017a) and **PN++** (Qi et al. 2017b). As shown in Table 3, our PCSSC-Net can predict semantic labels with an improvement of 0.8% to PN++ and 14.9% to PN in mIoU. The accuracy of 97.3% proves that the segmentation stream can provide relatively correct semantic information for the completion task.

Ablation Study. We also compare our model with its ablated versions. Version **A** eliminates the semantic stream in the SGC decoder. The input of the second MLP in the completion stream has no semantic hint for the refinement. Version **B** only uses the combination of the global CD and the cross-entropy loss as the final loss function. The semantics-based CD terms are excluded. The quantitative results are documented in Table 4. The results of Version A demonstrate that the semantic guidance can help the SGC decoder to refine the global geometry and the object-level structures with smaller SCDs. It is also noteworthy that our model already achieves a precise completion result without semantics comparing with the other point completion networks, which once again validates the efficacy of our PBC encoder design. Version B results are also presenting a shred of evidence to highlight the importance of semantics in the scene completion task. Besides, the semantics-based CD loss can constrain the network to re-arrange points inside the semantic cluster, which also validate our assumption in constructing the ground truth with semantic labels. For instance, in Figure 6, the leaky floor could be filled with wall points, which causes the misjudgment in segmentation results. The semantics-based CD loss can effectively isolate each class and enforce the network to complete the partial points inside their semantic clusters.

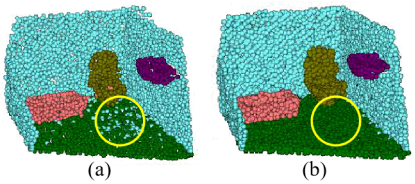

Figure 6: Ablation studies on the semantics-based CD loss: (a) The output without the semantics-based CD could have a mixture of semantic classes while completing the point clouds; (b) The semantics-based CD loss can isolate each semantic class with a satisfatory segmentation performance.

Discussion and Conclusion

We proposed a novel semantics scene completion network on the point cloud, PCSSC-Net, in this paper. It is shown that our PBC encoder can learn comprehensive multi-level features from partial inputs, and our SGC decoder can produce the semantically reconstructed scene geometry, benefiting from the semantic information. Comprehensive experiments validate that our method can produce more precise and better semantically completed results than existing alternative approaches.

Our method still has several limitations yet to be overcome. First, the PCSSC-Net fails to reconstruct large missing parts where our encoder could not capture enough information. Second, our model is highly dependent on the semantic segmentation's accuracy, where the wrong assignments decay the performance due to the SGC decoder architecture. Our near-term efforts are geared towards possible improvements. Other aspects of 3D scene understanding and applications could be investigated in the long term. For example, we could expand our general framework towards the understanding of dynamic scenes where the multi-views and motions provide extra contextual information along time. Moreover, it would be interesting to create virtual objects and enhance their interaction with the real environment with full semantic understanding. These pursuits would broaden the scope of 3D vision applications in 3D scene understanding and its integration with real world.

Acknowledgements

This research is supported in part by National Key R&D Program of China (No.2018YFB1700603), National Natural Science Foundation of China (No.61672077 and 61532002), Beijing Natural Science Foundation-Haidian Primitive Innovation Join Fund (L182016), and National Science Foundation of USA: IIS-1812606 and IIS-1715985.

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. J. 2018. Learning Representations and Generative Models for 3D Point Clouds. In *ICML*, volume 80, 40–49.

Chang, A. X.; Funkhouser, T. A.; Guibas, L. J.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; Xiao, J.; Yi, L.; and Yu, F. 2015. ShapeNet: An Information-Rich 3D Model Repository. *CoRR* abs/1512.03012.

Dai, A.; Ritchie, D.; Bokeloh, M.; Reed, S.; Sturm, J.; and Nießner, M. 2018. ScanComplete: Large-Scale Scene Completion and Semantic Segmentation for 3D Scans. In *CVPR*, 4578–4587.

Garbade, M.; Chen, Y.; Sawatzky, J.; and Gall, J. 2019. Two Stream 3D Semantic Scene Completion. In *IEEE CVPR Workshops*, 416–425.

Groueix, T.; Fisher, M.; Kim, V. G.; Russell, B. C.; and Aubry, M. 2018. A Papier-Mâché Approach to Learning 3D Surface Generation. In *IEEE CVPR*, 216–224.

Huang, Z.; Yu, Y.; Xu, J.; Ni, F.; and Le, X. 2020. PF-Net: Point Fractal Network for 3D Point Cloud Completion. In *IEEE/CVF CVPR*, 7659–7667.

Landrieu, L.; and Simonovsky, M. 2018. Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs. In *IEEE CVPR*, 4558–4567.

Li, J.; Chen, B. M.; and Lee, G. H. 2018. SO-Net: Self-Organizing Network for Point Cloud Analysis. In *IEEE CVPR*, 9397–9406.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020a. Anisotropic Convolutional Networks for 3D Semantic Scene Completion. In *IEEE/CVF CVPR*, 3348–3356.

Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; Zhao, C.; and Reid, I. D. 2019a. RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion. In *IEEE CVPR*, 7693–7702.

Li, J.; Liu, Y.; Yuan, X.; Zhao, C.; Siegwart, R.; Reid, I.; and Cadena, C. 2019b. Depth Based Semantic Scene Completion With Position Importance Aware Loss. *IEEE RAL* 5(1): 219–226.

Li, S.; Zou, C.; Li, Y.; Zhao, X.; and Gao, Y. 2020b. Attention-Based Multi-Modal Fusion Network for Semantic Scene Completion. In *AAAI*, 11402–11409.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; and Chen, B. 2018. PointCNN: Convolution On X-Transformed Points. In *NeuraIPS*, 828–838.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. 2015. Gated Graph Sequence Neural Networks. *arXiv: 1511.05493* .

Lin, C.; Kong, C.; and Lucey, S. 2018. Learning Efficient Point Cloud Generation for Dense 3D Object Reconstruction. In *AAAI*, 7114–7121.

Liu, M.; Sheng, L.; Yang, S.; Shao, J.; and Hu, S. 2020. Morphing and Sampling Network for Dense Point Cloud Completion. In *AAAI*, 11596–11603.

Liu, Y.; Fan, B.; Meng, G.; Lu, J.; Xiang, S.; and Pan, C. 2019. DensePoint: Learning Densely Contextual Representation for Efficient Point Cloud Processing. In *IEEE/CVF ICCV*, 5238–5247.

McCormac, J.; Handa, A.; Leutenegger, S.; and Davison, A. J. 2017. SceneNet RGB-D: Can 5M Synthetic Images Beat Generic ImageNet Pre-training on Indoor Segmentation? In *IEEE ICCV*, 2697–2706.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *IEEE CVPR*, 77–85.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *NeuraIPS*, 5099–5108.

Silberman, N.; Hoiem, D.; Kohli, P.; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*, 746–760.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. A. 2017. Semantic Scene Completion from a Single Depth Image. In *IEEE CVPR*, 190–198.

Wang, P.; Liu, Y.; and Tong, X. 2020. Deep Octree-based CNNs with Output-Guided Skip Connections for 3D Shape and Scene Completion. In *IEEE/CVF CVPR*, 1074–1081.

Wang, W.; Yu, R.; Huang, Q.; and Neumann, U. 2018. SGPN: Similarity Group Proposal Network for 3D Point Cloud Instance Segmentation. In *IEEE CVPR*, 2569–2578.

Wang, X.; Liu, S.; Shen, X.; Shen, C.; and Jia, J. 2019. Associatively Segmenting Instances and Semantics in Point Clouds. In *IEEE CVPR*, 4096–4105.

Wu, P.; Chen, C.; Yi, J.; and Metaxas, D. N. 2019. Point Cloud Processing via Recurrent Set Encoding. In *AAAI*, 5441–5449.

Wu, W.; Qi, Z.; and Li, F. 2019. PointConv: Deep Convolutional Networks on 3D Point Clouds. In *IEEE CVPR*, 9621–9630.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2018. FoldingNet: Point Cloud Auto-Encoder via Deep Grid Deformation. In *IEEE CVPR*, 206–215.

Yuan, W.; Khot, T.; Held, D.; Mertz, C.; and Hebert, M. 2018. PCN: Point Completion Network. In *3DV*, 728–737.

Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; and Yang, X. 2019. Cascaded Context Pyramid for Full-Resolution 3D Semantic Scene Completion. In *IEEE/CVF ICCV*, 7800–7809.

Zhao, H.; Jiang, L.; Fu, C.; and Jia, J. 2019a. PointWeb: Enhancing Local Neighborhood Features for Point Cloud Processing. In *IEEE CVPR*, 5565–5573.

Zhao, Y.; Birdal, T.; Deng, H.; and Tombari, F. 2019b. 3D Point Capsule Networks. In *IEEE CVPR*, 1009–1018.

Zhong, M.; and Zeng, G. 2020. Semantic Point Completion Network for 3D Semantic Scene Completion. In *ECAI*, 2824–2831.