Weakly-supervised 3D semantic scene reconstruction

Alexander Baumann
Technical University of Munich
alex.baumann@tum.de

Sophia Wagner
Technical University of Munich
sophiajohanna.wagner@tum.de

Abstract

In this work, we present a model for semantic reconstruction in a weakly-supervised fashion. It takes a point cloud of an indoor scene as input and outputs the mesh and the semantic label of the detected objects. The first part consists of the object detection pipeline. Unlike other models, we extract the features directly from the point cloud and do not convert the point cloud into a regular grid (e.g. voxels). Afterwards, we reconstruct the shape of the objects without any ground-truth labels. This procedure is employed by a shape prior which is pretrained on a synthetic dataset. Since the full 3D geometry is often unknown, our weakly-supervised approach is more appropriate since it does not require any shape annotations in the training data. Our code is available at http://github.com/alexanderbaumann99/Weakly-supervised-3D-semantic-scene-reconstruction.

1. Introduction

Semantic scene reconstruction aims to recover labels, poses and geometries of objects in 3D scenes and can be applied to various tasks such as robot navigation and interior design. Many previous works reconstruct 3D scenes by converting point clouds into regular grids [4–6, 10, 11, 13, 18, 19]. Due to the usage of 3D convolutions, these methods suffer from the resolution problem. RfD-Net [14] in contrast directly reconstructs objects from raw point clouds, which leads to more efficient learning thanks to their sparsity. The main drawback of the RfD-Net is that it requires full supervision. However, the full 3D geometry of the objects is unknown due to occlusions, view constraints and weak illumination. Therefore, we propose a weakly-supervised approach for 3D semantic scene reconstruction, which does not need any ground truth values for the shape completion.

Our model is based on the RfD-Net. It localizes the objects globally and predicts their corresponding instance classes. Then, shape reconstruction is implemented locally and conducted for every detected object. In contrast to [14], the shape reconstruction does not require ground-truth values. It is implemented by a shape prior, which is trained on a different, synthetic dataset in a fully-supervised manner.

In summary, our contributions are the following:

- We provide a novel approach for semantic instance reconstruction. To our knowledge, it is the first approach to reconstruct the shapes from a real-world point cloud scan without using any supervision on this dataset.
- We propose a shape prior which can predict the signed distance field directly from a given input point cloud, without voxelizing the scene.

2. Related Work

In this section, we give the reader an overview of some relevant literature on shape reconstruction. The objective is to reconstruct the shape of the object from a partial scan like a point cloud. There are various ways on how to approach this problem.

One common method is to learn an implicit function like the occupancy field [14] or the signed distance field [15]. Having such a function, one can apply Marching Cubes [12] to extract a mesh. A downside of this method is that one needs somehow ground-truth function values to learn the function. This is possible for synthetic datasets, e.g. [2]. However, such labels do not exist or are not reliable when dealing with real-world point cloud scans like [3].

That is why it would be very useful to apply this technique in a weakly-supervised manner. This was deployed in [13]. The authors trained a shape prior on a synthetic data set which predicted the signed distance field. Then they used pre-trained parts of this module for reconstructing the shapes on a real-world dataset. However, they still used some limited amount of ground-truth labels for their real-world dataset to finetune the shape prior.

A different technique is to retrieve the object shapes with existing CAD models. For this, one first needs to predict the instance class of the object and then find a well-fitting CAD model which models the shape of the object. In this way, one can reconstruct the shapes from RGB-D scans [1] or from 2D images [9]. Of course, to apply this method, it is necessary to have a CAD lookup table. Furthermore, one can hardly reconstruct characteristic details of the object since the object will be replaced by a standard CAD model.

3. Method

The architecture of our model is illustrated in figure 1. As in [14] we follow the principle of ”reconstruction from detection” with the difference of having only weak supervision. The problem consists of three subtasks. First of all, objects are detected from the input point cloud using a 3D proposal network backbone. In this step, the model benefits from full supervision. The predicted object box proposals are then aligned into a canonical coordinate system. Finally, we reconstruct the shape of the object using a shape prior trained on a synthetic dataset. In order to integrate the shape prior into the full pipeline, we propose two different approaches. For both approaches, we freeze the decoder of the shape prior and train an encoder such that the output mesh resembles the meshes from the corresponding output category. Details are explained in the following sections.

3.1. Object Detection and Alignment

We follow [14] in order to detect and align objects into the canonical coordinate system. Given a point cloud $P \in \mathbb{R}^{N\times3}$ as input of the model, we apply the VoteNet [16] as a backbone network to propose N_p boxes, each having a D_p-dimensional feature vector $F_p \in \mathbb{R}^{N_p\times D_p}$. We use these features to predict semantics and box geometry information of each object. These include the center $c \in \mathbb{R}^3$, the scale $s \in \mathbb{R}^3$, the heading angle $\theta \in \mathbb{R}$, the semantic label l of the box and the so-called objectness score s_{obj} of the corresponding bounding box. The score indicates if the predicted box is close to (<0.3m, positive) or far away from (>0.6m, negative) any ground-truth object center. These proposal features are then regressed using a 2-layer MLP. This part is trained in a supervised fashion.

After the detection, only N_d boxes with the highest objectness scores are kept by the model. Then, we sample points from the point cloud and group them. For this, a group layer clusters the points that are within a radius r of the box centers $\{c_i\}$, resulting in N_d clusters consisting of M_p points each. As a last step, we normalize each point cluster by aligning their points into a canonical coordinate system. Grouping and aligning the points can be done in a fully unsupervised fashion.

3.2. Shape Prior

In order to generate meshes from the resulting within-box point clouds we train a separate shape prior on the synthetic ShapeNet dataset [2]. The Shape Prior is illustrated in figure 2. It consists of an encoder and decoder. The encoder extracts a shape embedding from the input point cloud via the ResPointNet [17]. The decoder takes query points as an input and outputs the corresponding signed distance values. The architecture can be described as follows. A convolutional 1D layer is followed by five blocks each consisting of two convolutional 1D layers intervened by Conditional Batch Norm layers [7] and ReLU activations. Another block of a Conditional Batch Norm layer and a convolutional 1D layer is added. Finally, a tanh layer outputs the SDF values. The Conditional Batch Norm layers are conditioned on the shape embeddings extracted by the encoder.

As proposed in [13] we use the following loss function. Given a batch of query points $Q = \{q_i\}_{i=1}^N \in \mathbb{R}^{3\times N}$, the corresponding ground truth signed distance values $S = \{s_i\}_{i=1}^N \in \mathbb{R}^N$ and the extracted shape embeddings $E = \{e_i\}_{i=1}^N \in \mathbb{R}^{D\times N}$, the loss function is given as:

$$\mathcal{L}(Q, S, E|f) = \frac{1}{N}\sum_{i=1}^{N} \|f(q_i|e_i) - sign(s_i)\|^2$$

where $f(.)$ is the conditional decoder function and $sign(.)$ is the sign function. We use the sign function as we are only interested in whether a query point is inside or outside the object and not in the distance to the object itself.

3.3. Shape Retrieval

We propose two different approaches to integrate the trained Shape Prior into the full pipeline such that we can reconstruct object meshes from a 3D indoor scene point cloud. We cannot directly use the trained encoder and decoder of the shape prior, as the input point cloud during the training of the shape prior looks significantly different from the input during testing. For the training of the shape prior the synthetic ShapeNet dataset [2] is used while during testing the ScanNet dataset [3] is used. As point clouds in the real-world ScanNet dataset are sparse and inconsistent, this leads to bad results.

Therefore, we propose a shape retrieval approach. We only freeze the pretrained decoder of the shape prior. The encoder can then be trained using two different approaches:

- **Pretrained Encoder.** We reuse the pretrained encoder of the shape prior and finetune it such that the resulting shape embedding resembles the mean shape embedding of the corresponding object class.
- **Skip Propagation.** We discard the encoder of the shape prior and create a new ResPointNet [17] encoder. As in [14], we use a skip propagation. Hereby, the encoder not only takes the within-box point cloud as input, but also the box proposals that were predicted during the detection. Therefore, the predicted semantic label as well as further geometry information of the object is included in the extraction of the shape embedding.

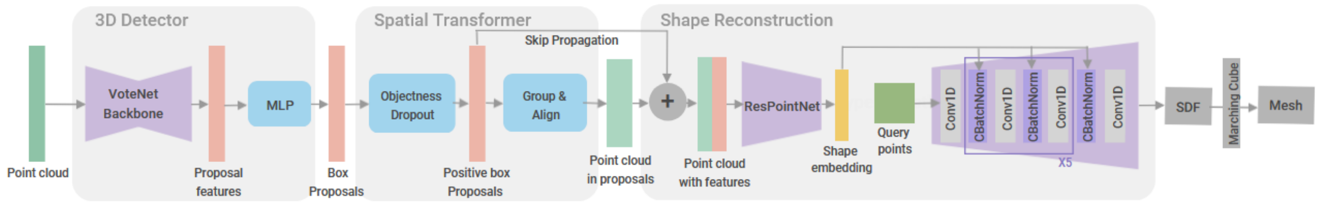


Figure 1. Overview of the network architecture. First, VoteNet [16] learns to predict N_p proposal features (each having dimension D_p) from an input point cloud which are regressed using a 2-layer MLP (to dimension D_b). Then, we sample points from the input cloud within the N_d boxes with the highest objectness scores and align these points into a canonical coordinate system. Optionally, we use a skip propagation where we concatenate the resulting points with box proposals. These are then fed into a ResPointNet [17]. This predicts a shape embedding which we condition on in the decoder. The decoder takes query points as an input and consists of 3 layers of fully-connected and Conditional Batch Norm [7] layers each. A final fully connected layer outputs a signed distance field which is converted to a mesh using the Marching Cubes algorithm [12].

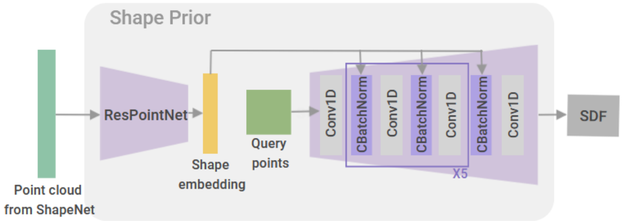


Figure 2. Shape prior training. The ResPointNet [17] consumes a point cloud from the ShapeNet [2] dataset as input and outputs a shape embedding. The decoder takes query points as an input and conditions on the shape embedding. It consists of multiple blocks of convolutional 1D layers and Conditional Batch Norm [7] layers in order to predict signed distance values.

Both approaches pursue the objective of creating meshes that are similar to meshes of the same semantic class category as the one predicted. For this, we compute the mean of all shape embeddings of each class from the ShapeNet data [2] using the trained encoder of the shape prior. For a predicted shape embedding z_i with ground-truth semantic class x and the mean $\tilde{z}_i^+$ of all shape embeddings of class x, we propose the following loss:

$$\mathcal{L}_{ret}(z_i, \tilde{z}_i^+) = \|z_i - \tilde{z}_i^+\|_2^2.$$

Instead of taking the mean, we also tried to use a random shape embedding of class x as $\tilde{z}_i^+$ and resample it in each iteration.
In the testing mode, one can reconstruct the meshes using the Marching Cubes algorithm [12] given the predicted signed distance values.

4. Datasets

As our approach suggests, we need two different datasets, a synthetic and a real-world dataset. As the real-world dataset, we use ScanNet [3] which consists 1,513 point cloud scans with instance labels. The shape prior is trained on parts of the synthetic ShapeNet dataset [2]. There, one only uses the object meshes of ShapeNet which also have a corresponding instance label in ScanNet. Hence, the resulting dataset consists of 22,702 objects.
We prepared the ShapeNet data by sampling 2,048 points from the meshes with the farthest point sampling algorithm [8]. As query points we used 1,024 points near the object surface and 1,024 points in a hypercube, all uniformly resampled in each iteration.

5. Results

5.1. Shape Prior

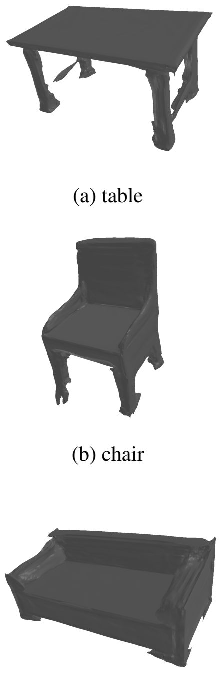
As the ShapeNet dataset has very unbalanced classes, we decided to use the whole dataset for training. This is possible since the evaluation is executed on the ScanNet dataset which is split into training and validation set. We trained the shape prior for 170 epochs and evaluated the shape prior by looking at chamfer distance from the point cloud towards the mesh. We also extracted the shapes from the mean shape embeddings. These must look reasonable since the corresponding shape embeddings are retrieved in the training of the shape reconstruction part. Some of those can be seen in 3.

5.2. Integration into pipeline

As explained before, there are several ways to integrate the shape prior into the full pipeline for reconstructing the shape. We tried them all and evaluated the results by using the chamfer distance form the point cloud within the corresponding bounding box of the object to the resulting mesh. The results can be seen in 1. One can deduce that it is better to retrieve random embeddings than the mean embeddings. Furthermore, the table 1 shows that the approach using the pretrained encoder outperforms the one with the

	table	chair	bookshelf	sofa	trash bin	cabinet	display	bathtub	overall
RfD-Net [14]	**0.729**	**0.126**	**0.507**	1.325	0.020	**0.345**	**0.014**	0.118	**0.411**
SP + mean embeddings	1.570	0.308	0.972	**1.134**	**0.012**	0.586	0.018	0.104	0.747
SP + random embeddings	1.587	0.296	0.696	1.380	0.024	0.650	0.019	**0.088**	0.731
PE + mean embeddings	1.494	0.266	0.764	1.492	0.021	0.453	0.024	0.126	0.690
PE + random embeddings	1.310	0.270	0.706	1.757	0.026	0.526	0.025	0.135	0.676

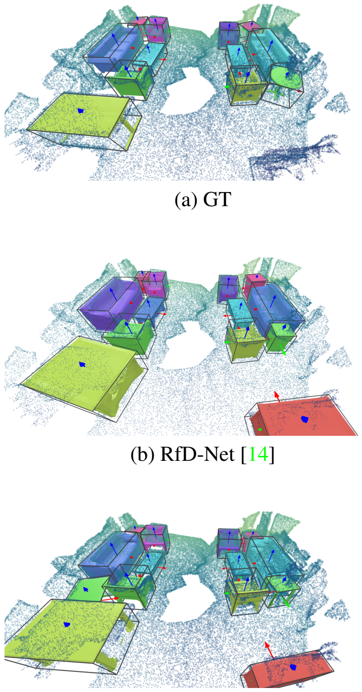
Table 1. average Chamfer distances from the input point cloud to the corresponding extracted mesh, multiplied with 1000.
SP = skip propagation, PE = pretrained encoder


(a) table

(b) chair

(c) sofa

Figure 3. Extracted meshes from the average shape embeddings of the instance classes


(a) GT

(b) RfD-Net [14]

(c) Ours

Figure 4. Ground-truth and predictions of a scene

skip propagation. Overall, our model can outperform the baseline in some of the semantic classes even though we do not use ground-truth occupancy values on the ScanNet data [3] in contrast to RfD-Net [14]. In 4, we compare the ground-truth of a scene, created by Scan2CAD [1] and the associating prediction of [14] and our model.

6. Conclusion

We present a novel approach for reconstructing the shapes without using any supervision on the given real-world dataset. Our method is based on a shape prior, which is trained on a synthetic dataset in a supervised manner, and on a retrieval of the shape embeddings. Despite the advantage of ground-truth labels of the RfD-Net [14], we can outperform this method in some of the semantic classes with respect to the chamfer distance. For having even better reconstruction results, one could add more object classes to our modified ShapeNet dataset. So, the shape prior would learn the shape of more instance classes which results to more appropriate meshes on ScanNet.

References

[1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 1, 4

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet:

An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2, 3

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 2, 3, 4

[4] Angela Dai, Christian Diller, and Matthias Niessner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[5] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott E. Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 1

[6] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[7] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 3

[8] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. 3

[9] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 1

[10] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 1

[11] Siqi Li, Changqing Zou, Yipeng Li, Xibin Zhao, and Yue Gao. Attention-based multi-modal fusion network for semantic scene completion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11402–11409, Apr. 2020. 1

[12] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 1, 3

[13] Mahyar Najibi, Guangda Lai, Abhijit Kundu, Zhichao Lu, Vivek Rathod, Thomas Funkhouser, Caroline Pantofaru, David Ross, Larry S Davis, and Alireza Fathi. Dops: Learning to detect 3d objects and predict their 3d shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11913–11922, 2020. 1, 2

[14] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4608–4618, 2021. 1, 2, 4

[15] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[16] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2, 3

[17] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 3

[18] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pages 190–198, United States, Nov. 2017. Institute of Electrical and Electronics Engineers Inc. 1

[19] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1