Self-Supervised Image Representation Learning with Geometric Set Consistency (Supplementary Material)

1. Details of Geometric Consistency Set

Here we provide more details about how to acquire the geometric consistency sets. For a 3D scene surface S we can have a sequence of corresponding RGB-D scanning images. We use the surface over-segmentation results produced by a normal-based graph cut method [2,5] as the geometric consistency sets $\{P_j\}$. In this way, the 3D scene will be divided into many small segments. Figure 2 shows some examples, and Figure 4 demonstrates the sets generated with different clustering edge weight thresholds. Then we project the 3D surface points together with the corresponding geometric consistency set id (here we use different ids to label the points in different geometric consistency sets) from 3D to 2D image views, and the pinhole camera model is used for 3D to 2D projection. Since the matching between the reconstructed 3D surface and the corresponding 2D image views may have a miss-match problem, we filter out invalid projections by comparing the depth difference between the projection points and the view depth maps. The projected points with the depth difference larger than a threshold, 0.05 in our experiments, will be regarded as invalid. Based on the valid projection points, we can have the corresponding projection $P_j^m = \{proj(s) \in I_m | s \in P_j\}$ of the geometric consistency set P_j from 3D onto 2D image view I_m.

2. Effect of the Initial Learning Rate

We study different choices of the initial learning rate in the pre-training stage to see how it will influence the fine-tuning results. Specifically, the networks are pre-trained with different initial learning rates, including 0.1 and 0.01, on the ScanNet [1] dataset and fine-tuned for the image semantic segmentation task on ScanNet and NYUv2 [8] datasets. Table 1 illustrates the performance of both our method and Pri3D [4]. Although the network performance varies with the initial learning rate, our method consistently outperforms Pri3D on all the settings.

Method	ScanNet	NYUv2
Pri3D (0.01)	59.7	54.8
Ours (0.01)	**60.3**	**55.4**
Pri3D (0.1)	61.7	51.4
Ours (0.1)	**63.1**	**54.1**

Table 1. **Effect of the initial learning rate.** Effect of different initial learning rates in the pre-training stage. ResNet50 is used as the backbone encoder, the network is pre-trained on ScanNet dataset and fine-tuned for the image semantic segmentation task on ScanNet and NYUv2 datasets. mIOU is used for evaluation.

3. Convergence

We study the convergence of the methods by fine-tuning the pre-trained networks on ScanNet [1] semantic segmentation dataset and reporting the performance on the validation set after different number of epochs. As shown in Figure 1, both Pri3D [4] and our proposed method converge within 10 epochs, and our method consistently outperforms Pri3D after that.

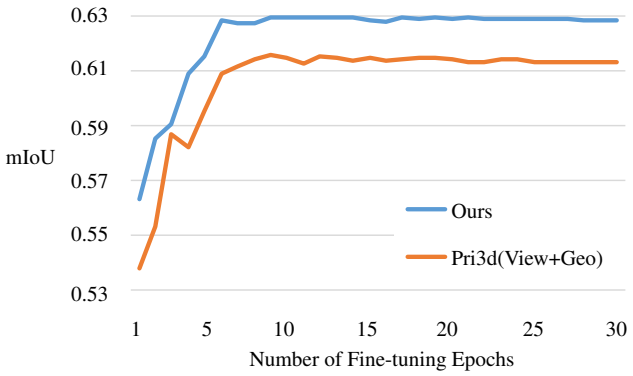

Figure 1. **Convergence of the methods.** We fine-tune the pre-trained networks on the ScanNet semantic segmentation task. The average performance of 3 runs for each method is reported.

4. Representation Space Analysis

We analyze the quality of the learned representation spaces by computing the coding rate [10], which measures the intra-category compactness, on ScanNet 2D semantic segmentation validation set.

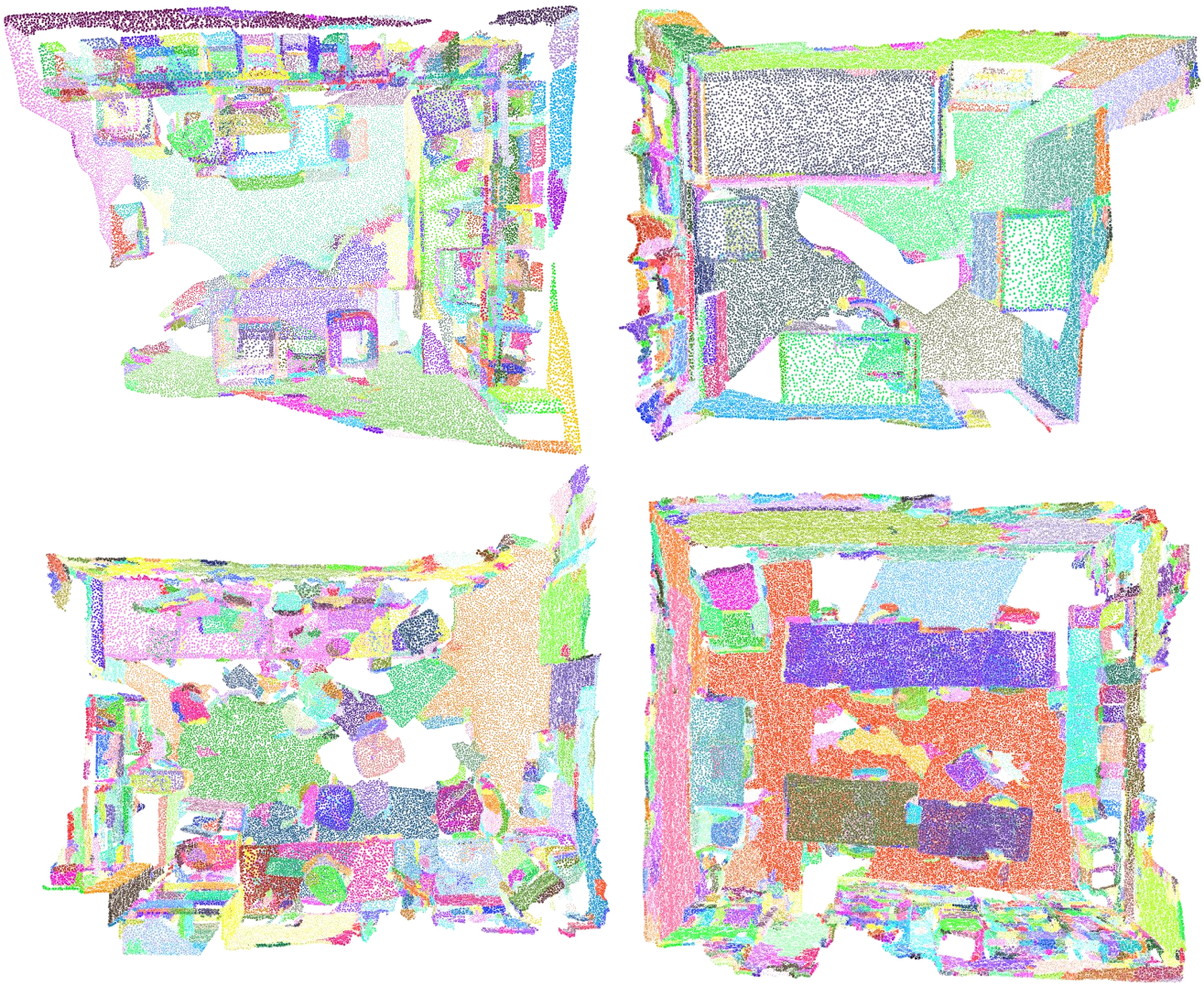

Figure 2. **Geometric consistency sets on 3D.** Different colors indicate different geometric consistency sets.

Specifically, let $\mathbf{F} \in \mathbb{R}^{d \times m}$ be the matrix containing m feature vectors with dimension d. The coding rate of $\mathbf{F}$ can be defined as:

$$R(\mathbf{F}, \epsilon) = \frac{1}{2} \log \det(\mathbf{I} + \frac{d}{m\epsilon^2} \mathbf{F}\mathbf{F}^\mathsf{T}) \qquad (1)$$

where I is the identity matrix, and ϵ is the distortion parameter. For each image, we extract pixel features with pre-trained networks. Since the pixel features extracted with different pre-trained networks may have different overall scales, we scale the features by dividing by average feature length. Then, we compute the coding rate for the features within each ground truth category. The coding rate of an image can be computed by averaging the coding rates of all the categories within this image. The average coding rates of Pri3D [4] and ours are 54.04 and 34.05 respectively. This means our pre-trained representations are more compact than Pri3D. Moreover, we also visualize the learned features by PCA. As shown in Figure 3, our features are cleaner and more separable.

5. Ablation of Two-stage Training

We test our method with different pre-training configurations, i.e. set-InfoNCE loss only, set-InfoNCE plus pixel-InfoNCE and two-stage training. As shown in Table 2, the best performance is achieved with two-stage training that learns from low-level to high-level. Similar strategies can also be found in research topics like curriculum learning.

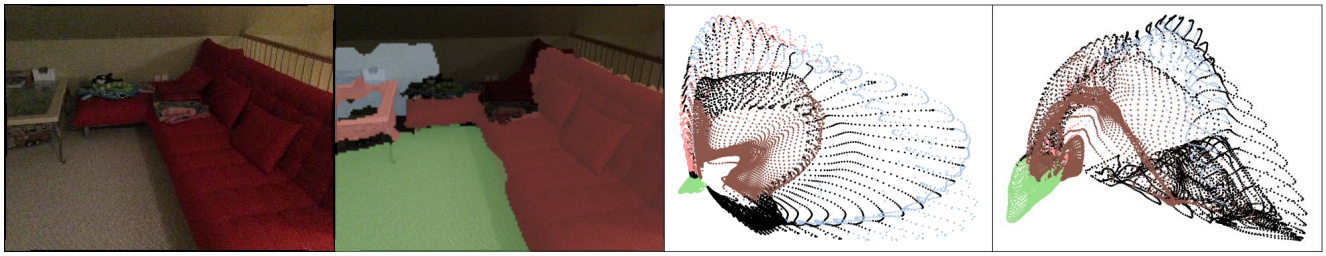

Figure 3. **PCA embedding of learned pixel features.** Different colors indicate different ground truth segmentation categories, and black color indicates the unlabeled regions.

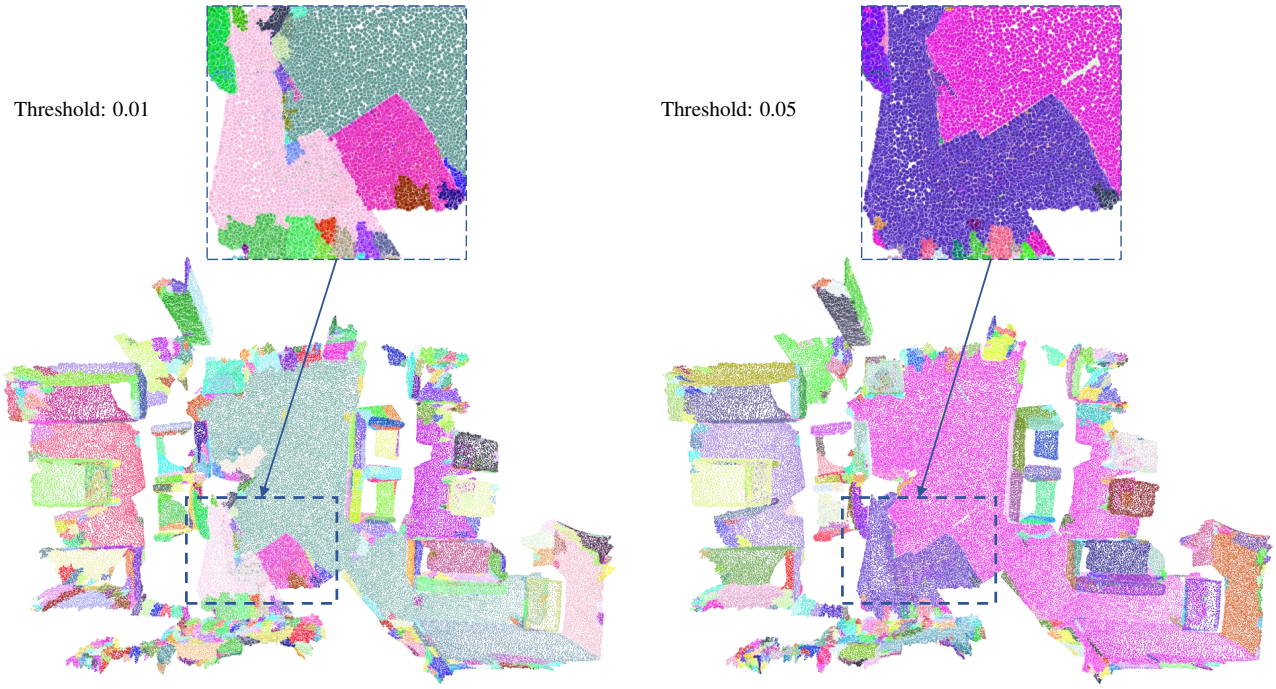

Figure 4. **3D geometric consistency sets generated by different parameters.** Different colors indicate different geometric consistency sets, and larger clustering edge weight threshold will lead to larger sets. Black color denotes the unlabeled regions.

	set-InfoNCE	set + pixel-InfoNCE	two stage
mIoU	60.6	61.0	63.1

Table 2. **Two-stage training ablation.** Performance of 2D semantic segmentation task on ScanNet with different pre-training configurations. ResNet50 is used as the backbone encoder.

6. Performance on SUN RGB-D Dataset

To further validate the transferability of our method, we fine-tune the ScanNet pre-trained representations on SUN RGB-D Dataset [9] for the 2D semantic segmentation task. Specifically, the dataset contains 5k images for training and 5k images for testing, and the networks are pre-trained with initial learning rates of 0.1 and 0.01 respectively. As shown in Table 3, our method achieves better performance compared with Pri3D [4].

7. Performance on KITTI Dataset

We further test our method on KITTI [3] dataset, to see the effectiveness of our method in the outdoor autonomous driving scenario. KITTI is a dataset captured by driving around a city with cars equipped with different kinds of sensors, including stereo camera, GPS, laser-scanner, etc. In our experiment, we use the unlabeled RGB-D sequences for pre-training and fine-tune the pre-trained network on the 2D image semantic segmentation task. To compute the geometric consistency sets, we use the Voxel Cloud Connectivity

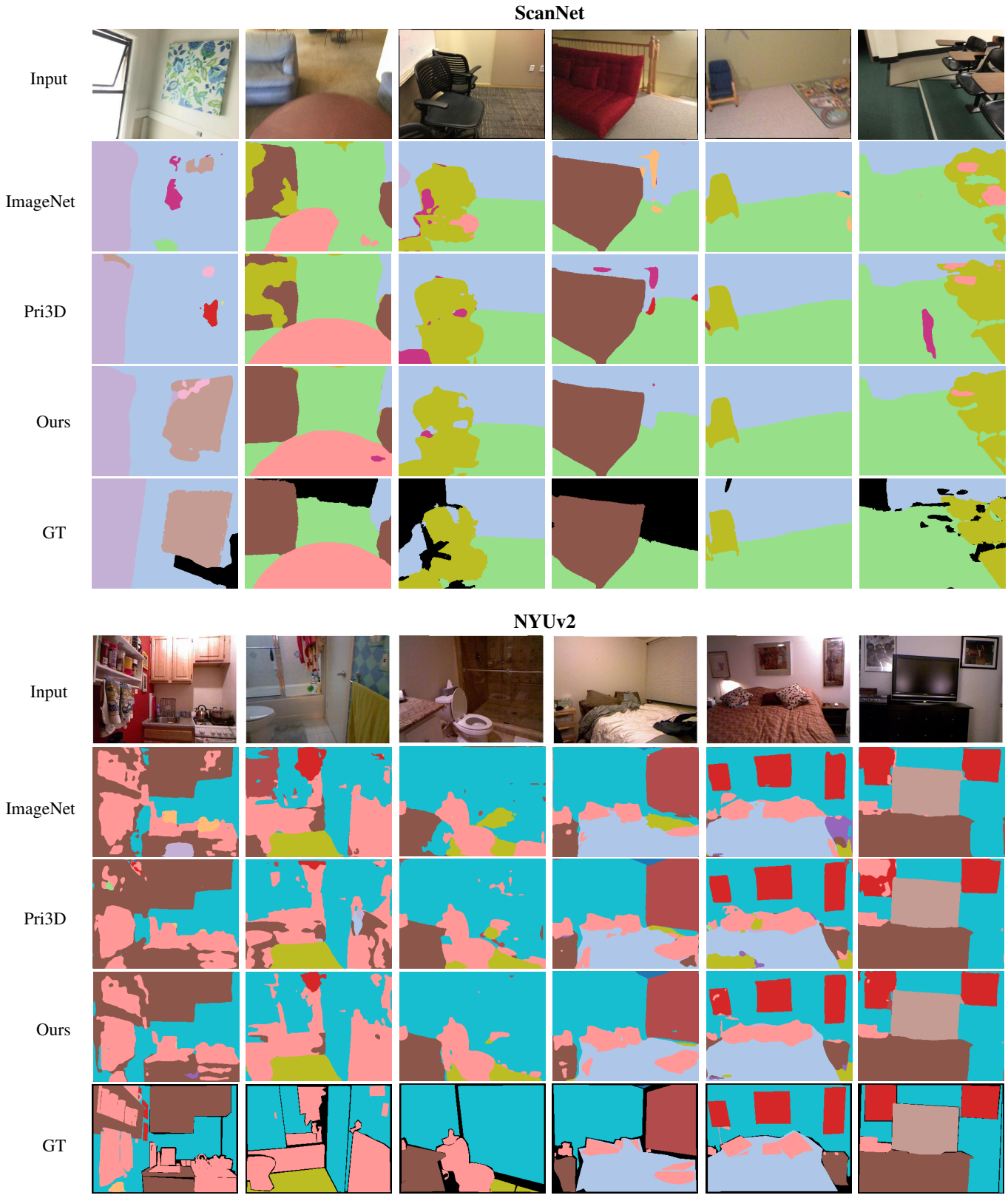

Figure 5. **More qualitative results of semantic segmentation task on ScanNet and NYUv2 datasets.** All the methods are pre-trained on ScanNet with ResNet50 as the backbone encoder.

Method	ResNet50
ImageNet Pre-training	34.8
Pri3D (0.1)	37.3
Ours (0.1)	38.1
Pri3D (0.01)	38.6
Ours (0.01)	**39.2**

Table 3. **Performance on SUN RGB-D dataset.** Pre-train on ScanNet with different learning rates and fine-tune on SUN RGB-D for 2D semantic segmentation. ResNet50 is used as the backbone encoder.

Method	ResNet50
ImageNet Pre-training	28.5
Pri3D	33.2
Ours	**33.7**

Table 4. **Performance on KITTI dataset.** ResNet50 is used as the backbone encoder, the network is pre-trained with unlabeled RGB-D sequence on KITTI and fine-tuned for the image semantic segmentation task. mIOU is used for evaluation.

Segmentation (VCCS) [6] method implemented by PCL [7] library to extract clusters on the per-view point clouds. During training, for each view pair, we use the geometric consistency sets from one view and project them onto the other, while the geometric consistency sets from the other view are ignored. In this way, we can have the same geometric consistency sets for the corresponding views. Table 4 illustrates the results. We note that the point clouds in KITTI are partial and noisy, and the moving objects in the scenes may lead to incorrect correspondences between views, which makes the results sensitive to different clustering parameters. The experiment here is to demonstrate the possibility of adapting our method to outdoor scenes; we hope our work can motivate future research in this direction.

8. More Qualitative Results

In Figure 5, we show more qualitative results of 2D semantic segmentation task on ScanNet [1] and NYUv2 [8] datasets. As is shown, the segmentation results produced with our method have less noise compared with those from other methods.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 1, 5

[2] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 1

[3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[4] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021. 1, 2, 3

[5] Andrej Karpathy, Stephen Miller, and Li Fei-Fei. Object discovery in 3d scenes via shape analysis. In *2013 IEEE International Conference on Robotics and Automation*, pages 2088–2095. IEEE, 2013. 1

[6] Jeremie Papon, Alexey Abramov, Markus Schoeler, and Florentin Worgotter. Voxel cloud connectivity segmentation-supervoxels for point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2027–2034, 2013. 5

[7] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011. 5

[8] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1, 5

[9] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 3

[10] Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *Advances in Neural Information Processing Systems*, 33:9422–9434, 2020. 1