Supplementary Material
One Thing One Click: A Self-Training Approach for Weakly Supervised 3D Semantic Segmentation

Zhengzhe Liu[1] Xiaojuan Qi[2] Chi-Wing Fu[1]
[1]The Chinese University of Hong Kong [2]The University of Hong Kong
{zzliu,cwfu}@cse.cuhk.edu.hk xjqi@eee.hku.edu.hk

In this supplementary document, we first show our results on ScanNet-v2 with fewer annotations. Then, we show more results of using our method with "One Thing One Click" on ScanNet-v2 and S3DIS (Section 2). Further, we discuss the relationship of our relation network and Prototypical Net [2] (Section 3). Finally, we show example super-voxels in Section 5. The code can be found in https://github.com/liuzhengzhe/One-Thing-One-Click.

1. Results with Fewer Annotations

To investigate the performance of our approach with even less annotated points, we further annotate ScanNet-v2 with a "Two Things One Click" scheme, where we annotate a single random point on half of the objects chosen randomly in the scene. In this way, only less than 0.01% points are annotated on ScanNet-v2. With the even sparse annotations, we still achieve 60.62% mIoU as shown in Table 1. This experiment also demonstrates that our method can still achieve decent performance even though the annotator ignores several objects by mistake in "One Thing One Click" scheme. We further investigate the performance drop with a more challenging "Four Things One Click" scheme. However, the model cannot converge well in the very first iteration due to the insufficient label and the self-training fails in this case.

Method	Annotation (%)	mIoU (%)
Two Things One Click*	0.01	54.71
Two Things One Click†	0.01	59.56
Two Things One Click	0.01	**60.62**

Table 1. Two Things One Click results and baselines on ScanNet-v2 val. set. * means the baseline model trained with the initial pseudo label shown in Figure 3 (d). † means disabling graph propagation and relation network during inference, but note that they are still used in training.

2. More Results on ScanNet-v2 and S3DIS

In this section, we show more results on ScanNet-v2 and S3DIS in Figures 1 and 2, respectively. Through these results, we demonstrate that even our approach is trained with only one annotated point per object in the scene, it can already produce segmentation results that are comparable to the fully supervised baseline [1]. See the error maps shown in (d) and (f) for better visualizations.

3. Relation to Prototypical Networks

In this section, we discuss the relationship between our relation network and Prototypical Networks [2]. First of all, [2] focuses on few-shot learning, and requires a strong generalization ability to classify the categories that are not seen in the training. In each training episode, [2] samples a subset of categories to simulate the unseen categories in testing. For better simulation, [2] does not require the prototype of each training category to be consistent in different training episodes. In this way, the network tends to regard the sampled training categories as unfamiliar, to better simulate the test case with new categories. Otherwise, the network will memorize the training categories themselves and lose the generalization ability to new categories in testing.

Very differently, in our method, the same categories are shared in both training and testing. To group the embedding of the same category and distinguish different categories, our categorical prototype should reveal the global mean representation of all samples in each category. To avoid the mean categorical embedding deviating from the actual categorical center, we design a memory bank in our model to update the prototypes with the moving average strategy, instead of relying on one single mini-batch. Hence, we can stabilize the prototypes in the training and ensure that they are still effective in the inference.

Secondly, [2] focuses on the classification task and assumes that there are plenty of samples of each category in the training set to support the set construction in each episode. However, in our 3D semantic segmentation task, we sample point clouds in each iteration, and there could be insufficient or even no samples for certain categories in a mini-batch.

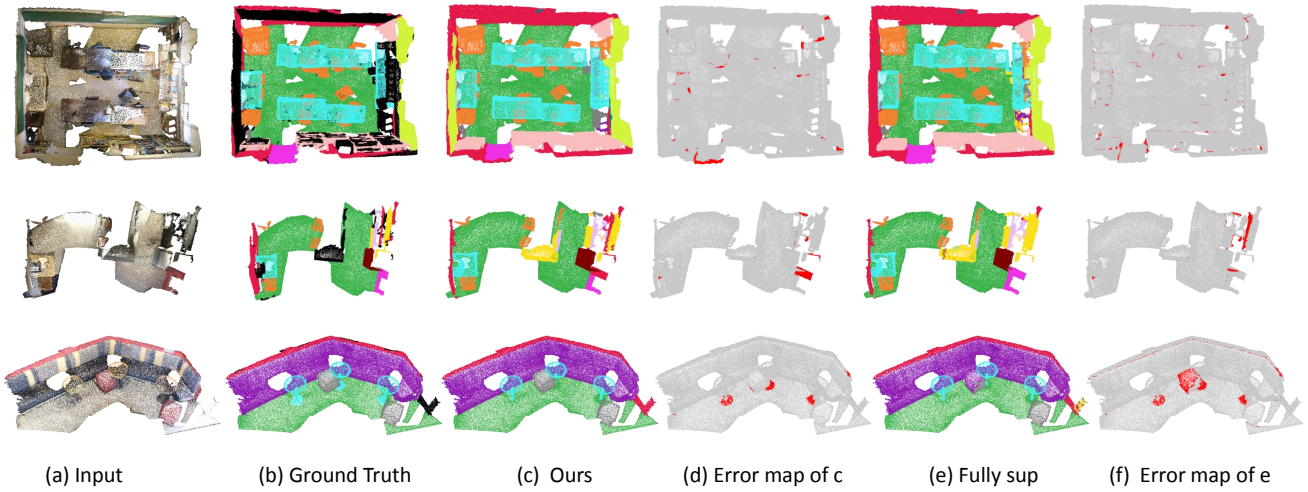

Figure 1. More results on ScanNet-v2. (c) is produced by our model trained only with "One Thing One Click" annotations. (e) is the fully supervised results of [1]. Red regions in (d) and (f) indicate the wrong predictions.

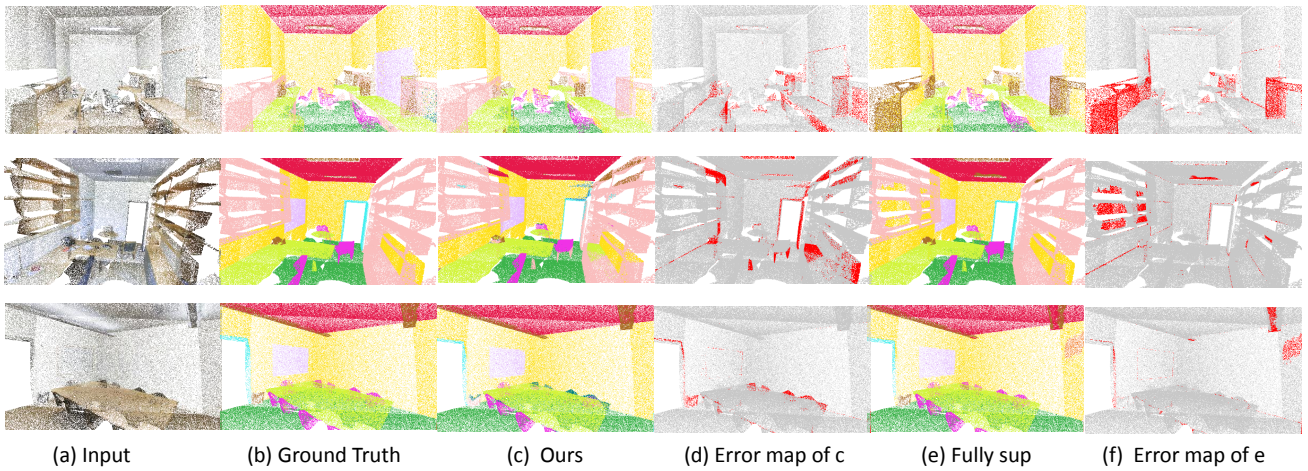

Figure 2. More results on S3DIS. (c) is produced by our model trained only with "One Thing One Click" annotations. (e) is the fully supervised results of [1]. Red regions in (d) and (f) indicate the wrong predictions.

For categories with insufficient samples, we sample them with replacement to match the number of samples of other categories. For categories with no samples in the batch, our memory bank helps to accumulate the embedding learned in the previous iterations and stabilize the prototypes of these categories relative to the actual categorical center.

Further, we conduct an ablation study to manifest the effectiveness of the memory bank in our relation network. In this ablation study (to be presented in the next section in this supplementary document), we adopt the same strategy as [2] to update the prototype of each category. Specifically, we directly use the average embedding of the sampled super-voxels in the current mini-batch as the prototype, instead of updating the prototype using the memory bank. From the ablation study results presented in Table 2, "3D U-Net+Rel (w/o MB)+GP" shows that without memory bank, the performance of "3D U-Net+Rel (w/o MB)+GP" degrades to be a value similar to "3D U-Net+GP," where the relation network is not used. Please see Section 4 for more details.

4. Summary of All the Ablation Studies

Table 2 summarizes all the ablation studies presented in the main paper. Their settings are listed below:

- 3D U-Net (w/o ST): Train [1] with the "Initial pseudo label" as Figure 3(d) illustrated in the main paper, and without using our self training (ST) mechanism.
- 3D U-Net: Self-training (ST) is adopted on "3D U-Net (w/o ST)." The pseudo labels for the second to fifth iterations are generated according to our self-training approach, *i.e.*, we use network predictions of high confidence to iteratively improve the results.
- 3D U-Net+GP: Based on the previous model "3D U-

Figure 3. Some example super-voxels in ScanNet-v2. They have large variations in shape, geometrical structure, density, and number of points.

Baselines	3D U-Net (w/o ST)	3D U-Net
mIoU	60.14	65.91
Baselines	3D U-Net+GP	3D U-Net+Rel (w/o MB)+GP
mIoU	67.92	67.98
Ours	3D U-Net+Rel+GP†	3D U-Net+Rel+GP
mIoU	69.12	**70.45**

Table 2. Summary of our ablation studies. "w/o" is the short of "without." "3D U-Net" means the U-Net architecture in [1]. "ST" means self-training, "GP" means graph propagation, "Rel" means relation network, "MB" means memory bank, and "†" indicates graph propagation and relation network are only utilized in training but not in inference. From the results, we can see that adding components (ST, GP, etc.) in our method gradually improves the performance, and our full model achieves the best performance.

Net," we add back the graph propagation (GP) in the pseudo label generation, and utilize the hand-crafted features, including the colors and coordinates as the pairwise term for the similarity measurement.

- 3D U-Net+Rel (w/o MB)+GP: Based on "3D U-Net+GP," the pairwise term further includes the embedding generated from the relation network (Rel). However, the categorical prototypes are derived as the mean of the sampled data in the current mini-batch without using the memory bank (MB). This setting is discussed earlier in Section 3 of this supplementary document but not included in the main paper.
- 3D U-Net+Rel+GP†: Based on "3D U-Net+Rel (w/o MB)+GP," we further utilize the memory bank (MB) to update the categorical prototypes. However, graph propagation and relation network are used only in the training but not in the inference. This is to evaluate the performance of our approach with the same computation complexity as "3D U-Net" in the inference.
- 3D U-Net+Rel+GP: This is our full model, for which its only difference compared to "3D U-Net+Rel+GP†" is that it utilizes graph propagation and relation network in both training and inference.

From Table 2, we can see that each of the key modules of our approach, i.e., self-training (ST), graph propagation (GP), relation network (Rel), and memory bank (MB), has its own contribution to the overall performance.

5. Examples of the Super-Voxel

Lastly, we show example super-voxels in ScanNet-v2 in Figure 3. Due to the irregular geometrical structures and complex shapes, hand-crafted features like colors and coordinates cannot fully describe their properties. To this end, we propose a relation network to learn the high-level similarities among them.

References

[1] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems (NeurIPS)*, pages 4077–4087, 2017.