Supplementary Material for "Learning Inter-Superpoint Affinity for Weakly Supervised 3D Instance Segmentation"

Linghua Tang, Le Hui, and Jin Xie(✉)

Nanjing University of Science and Technology, Nanjing, China
{tanglinghua, le.hui, csjxie}@njust.edu.cn

1 Overview

In this supplementary material, we provide more details on network architecture, visualization, and experiment results to validate the effectiveness of our method.

2 Network Architecture

Backbone network. The backbone network takes point cloud as input and outputs superpoint feature for subsequent superpoint-level prediction. Following [4,6,1,9], we frist convert raw point cloud into regular volumetric grids with the size of $2cm \times 2cm \times 2cm$. Then, voxel-level features are extracted by a 3D U-Net, which is constructed by stacking five submanifold convolution blocks [2]. The voxel-level features belonging to the same superpoint are averaged to produce superpoint features. After that, the edge-conditioned convolutions (ECC) [7] is employed to update superpoint features with superpoint graph structure. Finally, the 32-dimensional superpoint features are obtained for subsequent superpoint-level prediction.

Superpoint-level prediction. We use two multi-layer perception (MLP) upon superpoint features to predict semantic scores and offset for each superpoint, respectively. In inter-superpoint affinity mining module, we use a MLP branch to reduce the superpoint feature dimension from 32 to 7, and then the reduced superpoint features are used for computing loss $\mathcal{L}_{\text{aff}}$. In the volume-aware instance refinement module, two MLP branches are used to predict the number of voxels and the radius of the instance corresponding to the superpoint.

Network training. Our model is trained on a single TITAN RTX GPU. The AdamW optimizer with a base learning rate of 0.001 is adopted for the network training. We employ an unsupervised point cloud oversegmentation method [5] to generate superpoints for both ScanNet-v2 and S3DIS datasets. The process of network training is shown in Fig. 1. In the first stage, we use current pseudo labels to train the network for predicting superpoint-level semantic, offset and affinity. It is worth noting that the ground turth of the offset is obtained by calculating the offset between the centroid of superpoint and the centroid of current pseudo instance label after label propagation. After that, the predicted semantic and

affinity (in the blue dotted box of Fig. 1) are used for label propagation on the superpoint graph. The pseudo labels obtained by propagation are in turn used to train the network. In the second stage, based on the network trained in the first stage, The predicted semantic and offset (in the red dotted box of Fig. 1) is combined with weak labels to generate pseudo instances. Then, the object volume information is inferred from pseudo instances and retrains the network. Finally, the trained network is used for segmenting instances.

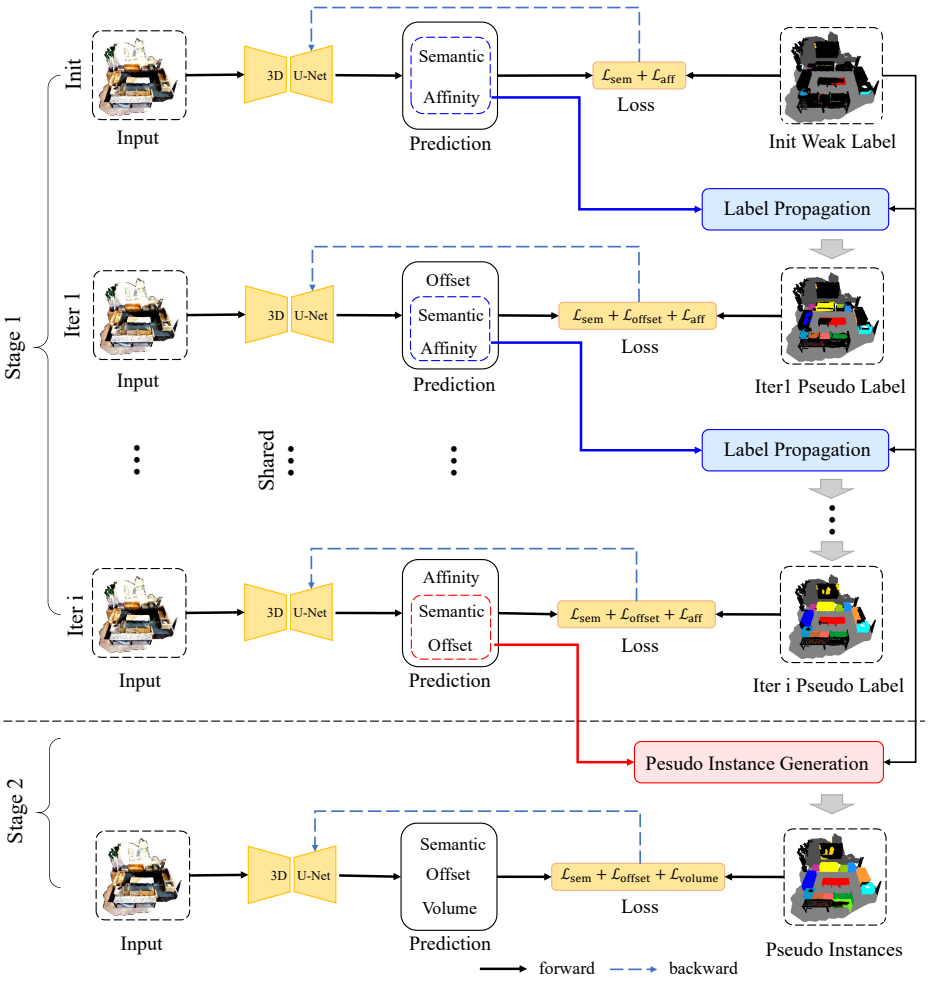

Fig. 1. The procedure of network training.

3 More Results

More visualizations of pseudo label generation. We show more visualizations of pseudo label generation on the ScanNet-v2 training set in Fig. 2. In different complex scenes, our method can effectively propagate label along superpoint graph and generate high-quality pseudo instances, thus providing a large number of effective supervision for network training.

More visualization results. Ins Fig. 3, we additionally show more visualization results of 3D semantic and instance segmentation on the ScanNet-v2 validation set and S3DIS. It can be observed that our method is able to obtain good 3D instance segmentation results in terms of extremely few labels.

ScaNet-v2 online test result. Tab. 2 reports the 3D instance segmentation results in terms of AP, AP_{25}, AP_{50} for all 18 categories on the ScanNet-v2 online test set. It can be observed that our method achieves state-of-the-art performance in the weakly supervised point cloud instance segmentation task, and even outperforms some fully supervised methods, such as GSPN [10] and 3D-SIS [3].

Different annotation rates. Tab. 1 reports the results on ScanNet-v2 validation set with different annotation rates. We gradually raise the rate of annotation from 0.02% (one annotated point per instance) to 0.10% (five annotated points per instance). The results show that the performance of our model can be improved as the annotation rate increases. The more annotation points, the more supervision in network training.

Table 1. 3D instance segmentation results with different annotations rates on the ScanNet-v2 validation set.

Annotations	0.02%	0.04%	0.06%	0.08%	0.10%	1%	5%	10%	100%
AP	28.1	31.8	33.4	34.4	34.9	35.1	35.5	35.6	**35.7**
AP_{50}	47.2	50.0	50.9	53.3	53.6	54.9	55.4	55.6	**55.9**
AP_{25}	67.5	68.3	69.3	70.0	71.6	72.0	72.2	72.4	**72.4**

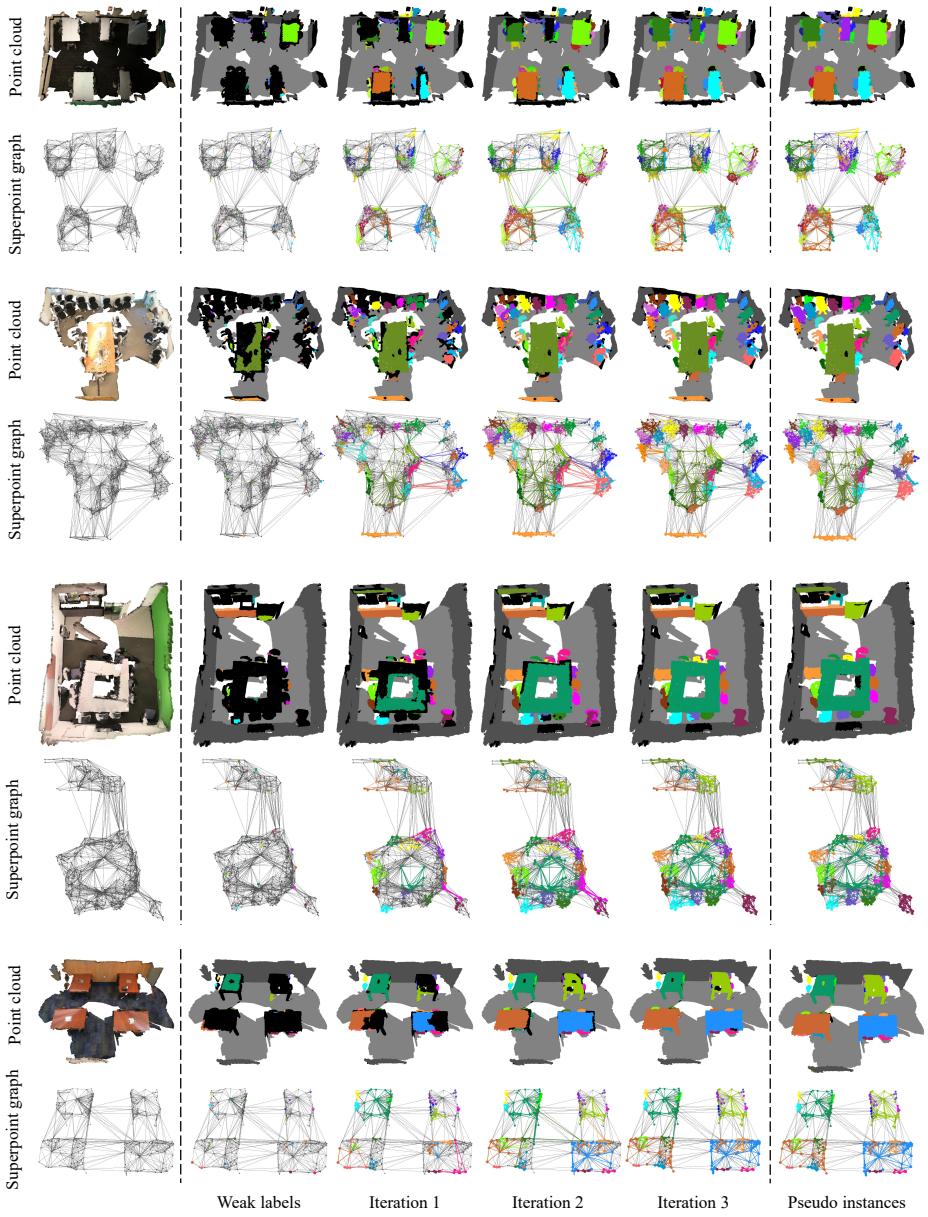

Fig. 2. Visualizations of pseudo label generation. **Left**: original point cloud and its superpoint graph. **Middle**: pseudo labels generated by random walks with predicted affinity and semantic at different iterations in the first stage. **Right**: predicted pseudo instances generated by applying clustering on the superpoint graph. Note that we remove the superpoints on the walls and floor for a better view.

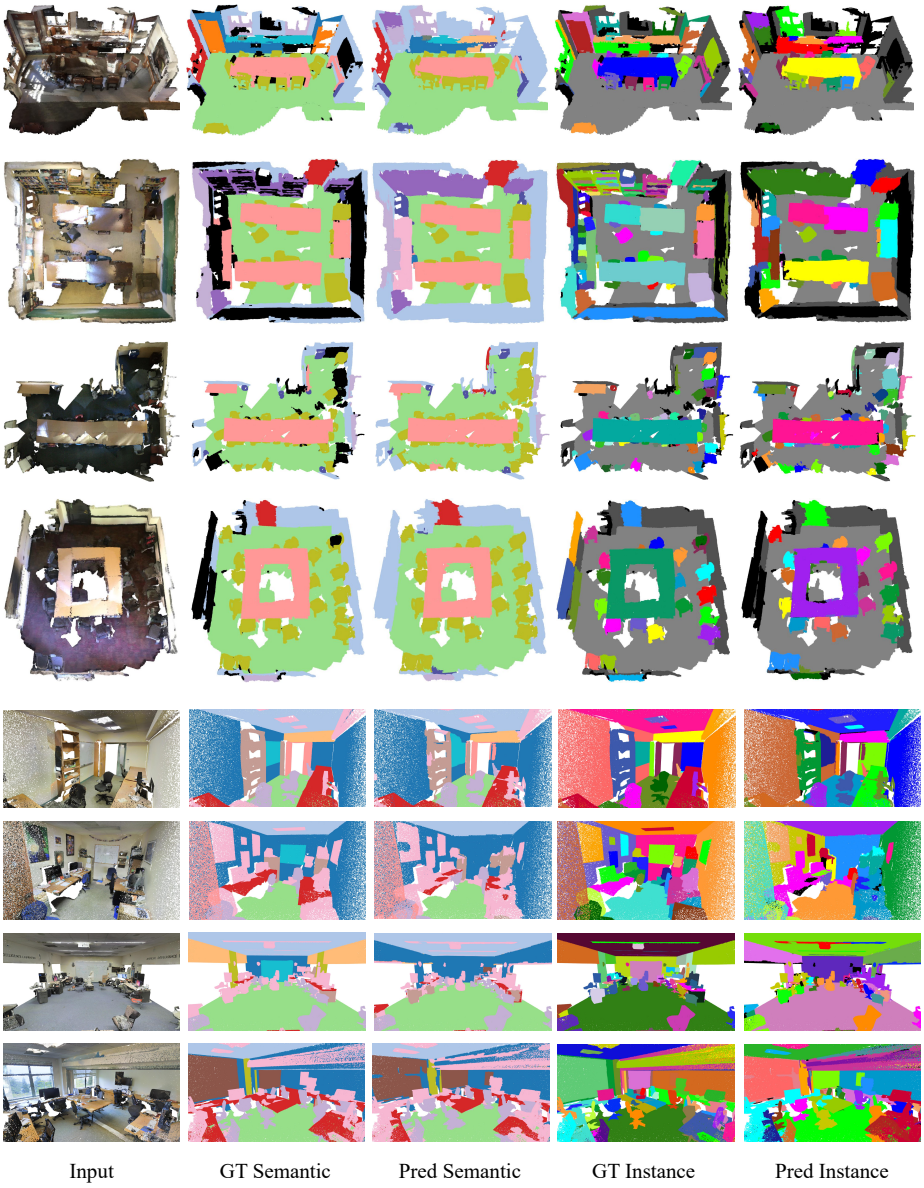

Fig. 3. Visualization of the 3D semantic and instance segmentation results on the validation of ScanNet-v2 (top) and S3DIS (bottom).

Table 2. Instance segmentation result on ScanNet v2 online test set in terms of AP, AP_{25}, AP_{50}.

Method	AP	bathtub	bed	bookshe	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	s.curtain	sink	sofa	table	toilet	window
Fully Sup.																			
GSPN [10]	15.8	35.6	17.3	11.3	14.0	35.9	1.2	2.3	3.9	13.4	12.3	0.8	8.9	14.9	11.7	22.1	12.8	56.3	9.4
3D-SIS [3]	16.1	40.7	15.5	6.8	4.3	34.6	0.1	13.4	0.5	8.8	10.6	3.7	13.5	32.1	2.8	33.9	11.6	46.6	9.3
PointGroup [4]	40.7	63.9	49.6	41.5	24.3	64.5	2.1	57.0	11.4	21.1	35.9	21.7	42.8	66.0	25.6	56.2	34.1	86.0	29.1
SSTNet [6]	**50.6**	73.8	54.9	49.7	31.6	69.3	17.8	37.7	19.8	33.0	46.3	57.6	51.5	85.7	49.4	63.7	45.7	94.3	29.0
HAIS [1]	45.7	70.4	56.1	45.7	36.4	67.3	4.6	54.7	19.4	30.8	42.6	28.8	45.4	71.1	26.2	56.3	43.4	88.9	34.4
SoftGroup [9]	50.4	66.7	57.9	37.2	38.1	69.4	7.2	67.7	30.3	38.7	53.1	31.9	58.2	75.4	31.8	64.3	49.2	90.7	38.8
Weakly Sup.																			
SegGroup [8]	24.6	55.6	33.5	6.2	11.5	49.0	0	29.7	1.8	18.6	14.2	8.3	23.3	21.6	15.3	46.9	25.1	74.4	8.3
3D-WSIS (**ours**)	**25.1**	38.0	27.4	28.9	14.4	41.3	0	31.1	6.5	11.3	13.0	2.9	20.4	38.8	10.8	45.9	31.1	76.9	12.7

Method	AP_{50}	bathtub	bed	bookshe	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	s.curtain	sink	sofa	table	toilet	window
Fully Sup.																			
GSPN [10]	30.6	50.0	40.5	31.1	34.8	58.9	5.4	6.8	12.6	28.3	29.0	2.8	21.9	21.4	33.1	39.6	27.5	82.1	24.5
3D-SIS [3]	38.2	100	43.2	24.5	19.0	57.7	1.3	26.3	3.3	32.0	24.0	7.5	42.2	85.7	11.7	69.9	27.1	88.3	23.5
PointGroup [4]	63.6	100	0.765	0.624	0.505	79.7	11.6	69.6	38.4	44.1	55.9	47.6	59.6	100	66.6	75.6	55.6	99.7	51.3
SSTNet [6]	69.8	100	69.7	88.8	55.6	80.3	38.7	62.6	41.7	55.6	58.5	70.2	60.0	100	82.4	72.0	69.2	100	50.9
HAIS [1]	69.9	100	84.9	82.0	67.5	80.8	27.9	75.7	46.5	51.7	59.6	55.9	60.0	100	65.4	76.7	67.6	99.4	56.0
SoftGroup [9]	**76.1**	100	80.8	84.5	71.6	86.2	24.3	82.4	65.5	62.0	73.4	69.9	79.1	98.1	71.6	84.4	76.9	100	59.4
Weakly Sup.																			
SegGroup [8]	44.5	66.7	77.3	18.5	31.7	65.6	0	40.7	13.4	38.1	26.7	21.7	47.6	71.4	45.2	62.9	51.4	100	22.2
3D-WSIS (**ours**)	**47.0**	66.7	68.5	67.7	37.2	56.2	0	48.2	24.4	31.6	29.8	5.2	44.2	85.7	26.7	70.2	55.9	100	28.7

Method	AP_{25}	bathtub	bed	bookshe	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	s.curtain	sink	sofa	table	toilet	window
Fully Sup.																			
GSPN [10]	54.4	50.0	65.5	66.1	66.3	76.5	43.2	21.4	61.2	58.4	49.9	20.4	28.6	42.9	65.5	65.0	53.9	95.0	49.9
3D-SIS [3]	55.8	100	77.3	61.4	50.3	69.1	20.0	41.2	49.8	54.6	31.1	10.3	60.0	85.7	38.2	79.9	44.5	93.8	37.1
PointGroup [4]	77.8	100	90.0	79.8	71.5	86.3	49.3	70.6	89.5	56.9	70.1	57.6	63.9	100	88.0	85.1	71.9	99.7	70.9
SSTNet [6]	78.9	100	84.0	88.8	71.7	83.5	71.7	68.4	62.7	72.4	65.2	72.7	60.0	100	91.2	82.2	75.7	100	69.1
HAIS [1]	80.3	100	99.4	82.0	75.9	85.5	55.4	88.2	82.7	61.5	67.6	63.8	64.6	100	91.2	79.7	76.7	99.4	72.6
SoftGroup [9]	**86.5**	100	96.9	86.0	86.0	91.3	55.8	89.9	91.1	76.0	82.8	73.6	80.2	98.1	91.9	87.5	87.7	100	82.0
Weakly Sup.																			
SegGroup [8]	63.7	100	92.3	59.3	56.1	74.6	14.3	50.4	76.6	48.5	44.2	37.2	53.0	71.4	81.5	77.5	67.3	100	43.1
3D-WSIS (**ours**)	**67.8**	100	88.0	83.6	70.1	72.7	27.3	60.7	70.6	54.1	51.5	17.4	60.0	85.7	71.6	84.6	71.1	100	50.6

References

1. Chen, S., Fang, J., Zhang, Q., Liu, W., Wang, X.: Hierarchical aggregation for 3D instance segmentation. In: ICCV (2021)
2. Graham, B., Engelcke, M., Van Der Maaten, L.: 3D semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018)
3. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: CVPR (2019)
4. Jiang, L., Zhao, H., Shi, S., Liu, S., Fu, C.W., Jia, J.: PointGroup: Dual-set point grouping for 3D instance segmentation. In: CVPR (2020)
5. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: CVPR (2018)
6. Liang, Z., Li, Z., Xu, S., Tan, M., Jia, K.: Instance segmentation in 3D scenes using semantic superpoint tree networks. In: ICCV (2021)
7. Simonovsky, M., Komodakis, N.: Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: CVPR (2017)
8. Tao, A., Duan, Y., Wei, Y., Lu, J., Zhou, J.: SegGroup: Seg-level supervision for 3D instance and semantic segmentation. TIP (2022)
9. Vu, T., Kim, K., Luu, T.M., Nguyen, X.T., Yoo, C.D.: SoftGroup for 3D instance segmentation on 3D point clouds. In: CVPR (2022)
10. Yi, L., Zhao, W., Wang, H., Sung, M., Guibas, L.J.: GSPN: Generative shape proposal network for 3D instance segmentation in point cloud. In: CVPR (2019)