TWIST: Two-Way Inter-label Self-Training for Semi-supervised 3D Instance Segmentation *supplementary material*

Ruihang Chu[1] Xiaoqing Ye[2] Zhengzhe Liu[1] Xiao Tan[2]
Xiaojuan Qi[3*] Chi-Wing Fu[1,4] Jiaya Jia[1,5]
[1]CUHK [2]Baidu Inc. [3]HKU [4]SHIAE [5]SmartMore

In this supplementary document, we present the detailed SparseConv-based network architecture of TWIST in Sec. A. We describe more details of self-training initialization in Sec. B. We show several extended experimental results in Sec. C and more visualizations of our model on the instance segmentation task with scarce labels for training in Sec. D.

A. Network Architecture

Fig. 6(a) shows the detailed architecture of the Sparse U-Net backbone in the 3D feature extraction network Φ, whereas Fig. 6(b) shows the architecture of the SparseConv encoder in the proposal re-correction module Ψ. Specifically, they are constructed by repeated Sparse (De)Conv ResBlocks, each with an architecture shown in Fig. 6(c).

B. Self-training Initialization Details

At the first round of the self-training (round 1), the learned model $\Phi(., \theta_0^r)$, which is trained only on the labeled point clouds, generates poor pseudo labels on unlabeled data. As the re-correction module has not been trained yet, these pseudo labels cannot be denoised in proposal level. For stable training, we only generate initial pseudo semantic labels $\tilde{S}^u$. At the *pseudo-label generation stage*, we select points whose largest semantic class probability fall above 0.9 to generate the corresponding semantic labels, and spread each label to all the points locally in a super-voxel for label propagation, following [2]. Pseudo offset vectors $\tilde{O}^u$ are not produced or involved in the *training stage* due to their unknown quality. This initialization strategy facilitates the self-training convergence and is removed at later rounds.

C. Additional Experiments

Ablations on instance certainty threshold. During the *pseudo-label generation stage* in each self-training round, only the instance proposal with an instance certainty score higher than a threshold will be involved later in the pseudo-label update procedure; otherwise, the proposal will be discarded. The final performance for adopting different thresholds is shown in Table 9. Selecting extremely low (*e.g.*, 0.1) or high (*e.g.*, 0.9) threshold values will lead to inferior instance segmentation accuracy, as they either tend to produce inaccurate pseudo labels from the poor-quality instance proposals or prevent many actually reliable instance proposals from generating pseudo labels. Dynamically setting the threshold cannot bring significant performance improvement, so we directly adopt 0.5 as the default threshold value.

Threshold	mAP	AP_{50}	AP_{25}
0.1	25.9	43.0	55.6
0.3	26.8	44.0	**56.4**
0.5	27.0	**44.1**	56.2
0.7	26.1	43.5	55.7
0.9	24.3	41.2	53.2
dynamic↑	26.4	43.8	55.9
dynamic↓	**27.1**	44.0	**56.4**

Table 9. Comparisons of using different instance certainty thresholds for instance proposal filtering during the *pseudo-label generation stage*. The dynamic↑ means that we vary the threshold as 0.1→0.3→0.5→0.7→0.9 in the five self-training rounds, whereas the dynamic↓ adopts an inverse threshold sequence of dynamic↑. These experiments were conducted on the ScanNet v2 validation set with 5% labeled data.

α	mAP	AP_{50}	AP_{25}
0.1	26.7	43.6	55.8
0.5	**27.0**	43.9	**56.2**
1	**27.0**	**44.1**	**56.2**
2	26.7	43.5	55.7
10	26.9	43.7	55.7

Table 10. Comparisons of using different loss ratios α. These experiments were conducted on the ScanNet v2 validation set with 5% labeled data.

*Corresponding Author

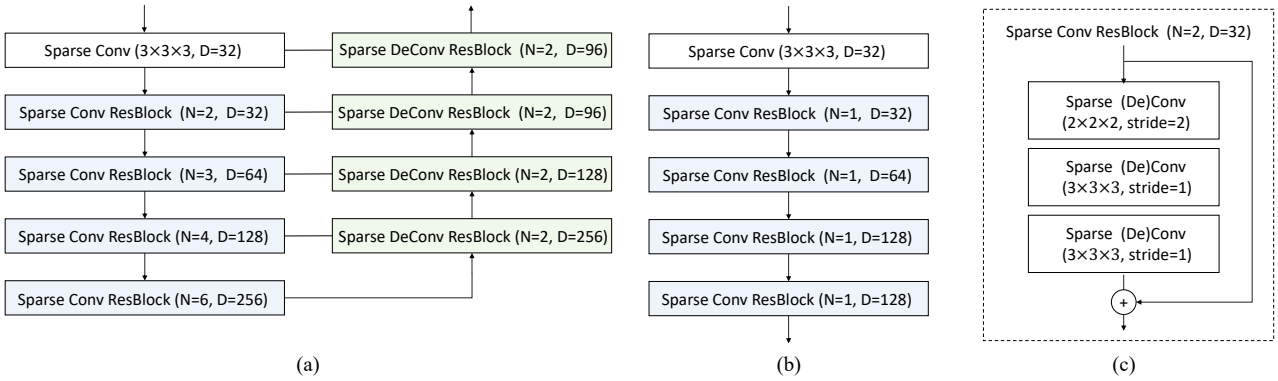

Figure 6. (a) The detailed architecture of the Sparse U-Net backbone in the 3D feature extraction network Φ, mainly constructed by stacked Sparse (De)Conv ResBlocks. (b) The architecture of the SparseConv encoder in the proposal re-correction module Ψ. (c) The architecture of the base module, *i.e.*, Sparse (De)Conv ResBlock. N denotes the number of repeated layers and D denotes the output channel dimension. Note that every (De)Conv layer is followed by a batch normalization (BN) layer and a ReLU operation.

Dataset	Method	1%			5%			10%			20%		
		mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}	mAP	AP_{50}	AP_{25}
ScanNet v2	Sup-only	5.1	9.8	17.6	18.2	32.0	47.0	26.7	42.8	58.9	29.3	47.9	63.0
	PC [3]	7.2	12.5	20.3	19.4	35.4	48.5	27.0	43.9	59.5	30.2	49.5	63.6
	TWIST	**9.6** (+4.5)	**17.1** (+7.3)	**26.2** (+8.6)	**27.0** (+8.8)	**44.1** (+12.1)	**56.2** (+9.2)	**30.6** (+3.9)	**49.7** (+6.9)	**63.0** (+4.1)	**32.8** (+3.5)	**52.9** (+5.0)	**66.8** (+3.8)
	TWIST + PC [3]	11.2 (+6.1)	19.5 (+9.7)	30.4 (+12.8)	28.0 (+9.8)	45.5 (+13.5)	57.3 (+10.3)	32.7 (+6.0)	51.1 (+8.3)	64.4 (+5.5)	33.9 (+4.6)	53.3 (+5.4)	67.5 (+4.5)
S3DIS	Sup-only	9.0	12.7	20.7	21.5	30.4	42.8	25.2	36.8	48.3	29.9	41.2	54.5
	PC [3]	13.4	15.9	23.1	22.9	33.6	44.5	27.1	38.7	50.2	31.2	43.1	56.6
	TWIST	**17.9** (+8.9)	**22.5** (+9.8)	**27.1** (+6.4)	**27.1** (+5.6)	**37.1** (+6.7)	**48.6** (+5.8)	**33.6** (+8.4)	**45.6** (+8.8)	**55.8** (+7.5)	**36.7** (+6.8)	**48.4** (+7.2)	**59.7** (+5.2)
	TWIST + PC [3]	18.6 (+9.6)	24.1 (+11.4)	28.7 (+8.0)	29.0 (+7.5)	38.7 (+8.3)	49.8 (+7.0)	34.2 (+9.0)	46.4 (+9.6)	57.1 (+8.8)	37.4 (+7.5)	49.3 (+8.1)	61.2 (+6.7)

Table 11. Results on the ScanNet v2 validation set (top) and S3DIS Area-5 set (bottom) for different label ratios: 1%, 5%, 10%, and 20%. "Sup-only" is the fully-supervised baseline model trained with only labeled data but no unlabeled data. From the results shown above, we can see that combining TWIST and PointContrast [3] (the rows for TWIST + PC) brings consistent performance improvement on all metrics.

Ablations on loss ratio. During the *training stage* of each self-training round, the 3D feature extraction network Φ and the proposal re-correction module Ψ are jointly optimized with the objective $\mathcal{L} = \mathcal{L}^{\Phi} + \alpha\mathcal{L}^{\Psi}$. The experimental results of using different α are presented in Table 10. Since Φ and Ψ focus on individual learning tasks, different loss ratios between them would impose minor impact on the overall performance. Hence, we choose $\alpha = 1$ as the default.

Combination with PointContrast [3]. To further validate the complementary strength of TWIST with the existing 3D pre-training approaches, we experiment with the model pre-trained on PointContrast [3], which is one of the 3D pre-training frameworks. In this way, we can initialize the self-training of TWIST in the first round. The comparison results on two datasets are exhibited in Table 11, from which we can see that combining TWIST and PointContrast (PC) leads to consistent performance improvement on all metrics, *e.g.*, 1.0-2.1% mAP increase on the ScanNet v2 validation set for all different label ratios.

Per-category results. We show the detailed per-category performance on the ScanNet data-efficient benchmark for comparison. The results trained with {1%, 5%, 10%, 20%} training set as the labeled data are exhibited in Table 12.

D. More Visualizations

Last, we show quantitative visualization results produced by TWIST when trained with only 10% labeled scenes on the ScanNet v2 validation set in Fig. 7. Note that distinct instances have different randomly-generated colors in the visualizations, so ground truths and predicted instances may not have the same color.

References

[1] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3D scene understanding with contrastive scene contexts. In *CVPR*, pages 15587–15597, 2021.

[2] Zhengzhe Liu, Xiaojuan Qi, and Chi-Wing Fu. One thing one click: A self-training approach for weakly supervised 3d semantic segmentation. In *CVPR*, pages 1726–1736, 2021.

[3] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *ECCV*, 2020.

L.R.	Method	$\mathbf{AP}_{50}$	bath	bed	bkshf	cab	chair	cntr	curt	desk	door	ofurn	pic	rfrig	showr	sink	sofa	table	toil	wind
1%	Sup-only	10.1	0.0	48.8	0.0	3.5	49.1	0.0	16.0	0.9	11.1	2.5	0.2	0.9	0.0	0.1	26.5	21.9	0.0	0.1
	PC [3]	11.9	0.0	53.4	4.1	4.7	56.0	0.0	30.7	0.0	8.3	7.1	6.8	6.4	0.0	0.0	18.6	17.6	0.0	0.0
	CSC [1]	11.7	0.0	49.2	4.0	3.2	52.4	0.0	17.4	2.1	13.9	9.6	**11.2**	3.9	0.0	1.8	25.8	15.0	0.0	0.1
	TWIST	14.2	0.0	52.6	0.0	10.5	60.5	**5.0**	27.9	**3.6**	14.8	10.9	1.3	14.8	0.0	**7.2**	25.6	21.1	0.0	0.0
	TWIST+CSC [1]	**18.6**	0.0	**54.5**	**15.3**	**14.8**	**66.7**	**5.0**	**42.0**	0.7	**21.0**	**17.5**	1.6	**21.8**	0.0	0.0	**51.1**	**22.7**	0.0	0.0
5%	Sup-only	27.3	66.7	56.7	10.6	20.3	68.5	1.3	0.2	13.0	26.9	23.4	12.9	10.3	0.0	6.3	55.7	38.5	75.3	5.7
	PC [3]	29.8	66.7	75.2	0.5	18.6	64.4	0.0	35.9	11.8	22.3	26.6	13.1	1.2	0.0	25.6	55.0	33.3	79.1	7.3
	CSC [1]	32.5	66.7	69.8	10.6	19.8	70.8	0.0	24.4	19.4	27.9	29.2	17.9	10.7	0.0	**44.6**	60.0	32.8	69.3	10.8
	TWIST	40.1	66.7	73.0	**34.9**	**38.7**	**79.1**	**2.9**	**47.1**	**26.1**	29.2	34.6	**24.8**	23.5	0.0	36.0	**66.6**	35.8	**88.9**	14.2
	TWIST+CSC [1]	**42.1**	66.7	**75.7**	33.3	35.8	77.0	0.8	43.6	25.4	**36.1**	**37.2**	22.4	**37.8**	**14.3**	30.3	64.3	**44.6**	**88.9**	**24.2**
10%	Sup-only	41.3	66.7	72.0	44.2	28.8	73.5	0.5	32.6	13.8	30.2	32.9	20.4	44.5	49.8	22.9	65.7	45.2	88.9	11.5
	PC [3]	43.2	66.7	75.7	56.0	27.8	74.0	0.3	43.5	12.3	30.9	34.7	10.9	52.2	42.9	22.3	73.9	43.4	94.4	14.9
	CSC [1]	44.0	66.7	73.7	41.8	21.8	79.1	**9.4**	32.8	**18.5**	25.1	38.2	27.3	**56.5**	53.9	37.7	58.8	37.1	**100**	12.8
	TWIST	46.6	66.7	72.8	**56.7**	**34.3**	78.5	2.9	51.6	12.2	**36.5**	**44.0**	34.3	35.6	42.9	45.5	**74.8**	43.6	88.2	18.5
	TWIST+CSC [1]	**48.1**	66.7	**76.0**	46.8	31.3	**80.2**	0.8	**52.9**	9.8	36.4	41.1	**34.8**	50.0	**57.1**	**50.4**	64.6	**53.0**	94.4	**20.1**
20%	Sup-only	47.3	52.8	77.3	63.2	39.1	78.4	5.0	51.5	27.1	39.2	40.0	12.3	48.6	38.7	32.2	62.7	**57.3**	100	26.8
	PC [3]	48.8	47.2	77.3	59.4	37.4	77.4	1.3	35.3	25.2	32.7	41.6	8.7	43.5	**85.7**	44.4	77.9	56.0	100	28.2
	CSC [1]	52.9	**100**	**77.3**	**70.4**	41.4	78.6	5.0	41.2	39.4	37.6	44.2	17.9	**54.2**	53.9	39.4	**79.3**	56.4	94.4	21.7
	TWIST	53.5	66.7	76.7	58.5	41.3	**81.4**	5.0	**54.2**	28.9	**43.9**	**49.9**	18.5	40.0	**85.7**	44.8	75.3	55.4	**100**	**35.9**
	TWIST+CSC [1]	**55.0**	**100**	75.8	57.0	**46.2**	79.7	5.0	38.9	**43.6**	43.3	48.4	**25.3**	49.1	57.1	**53.8**	76.0	56.1	**100**	35.4

Table 12. Instance segmentation performance on the ScanNet data-efficient benchmark with {1%, 5%, 10%, 20%} training data as the labeled data. We adopt the AP_{50} as the metric and show per-category performance over 18 classes. L.R. means the label ratio.

Figure 7. Visualization results of TWIST (trained with 10% labeled scenes). Distinct instances have different colors.