SP^{2}T: Sparse Proxy Attention for Dual-stream Point Transformer

Supplementary Material

Indoor Sem. Seg.	ScanNet [7]		ScanNet200 [11]	
Methods with same TTA	Val	Test	Val	Test
MinkUnet [5]	72.2	73.4	25.0	25.3
OctFormer [15] (Rep.)	74.6	70.7	31.9	31.0
OctFormer [15] (Off.)	75.7	-	32.6	-
Swin3D [18] (Off.)	76.4	-	-	-
Swin3D [18] (Rep.)	76.6	71.4	-	-
PTv3 [17] (Off.)	77.5	73.6	35.2	34.0
SP^{2}T	**78.7**	**74.9**	**37.0**	**35.2**

Table 1. Indoor instance segmentation with same TTA between Val and Test set. Rep. means the model uses the code of Pointcept and reproduces it by ours. Off. means the model uses official weight and code.

Algorithm 1 Pytorch-Style Pesade-code of Spatial-wise Sampling

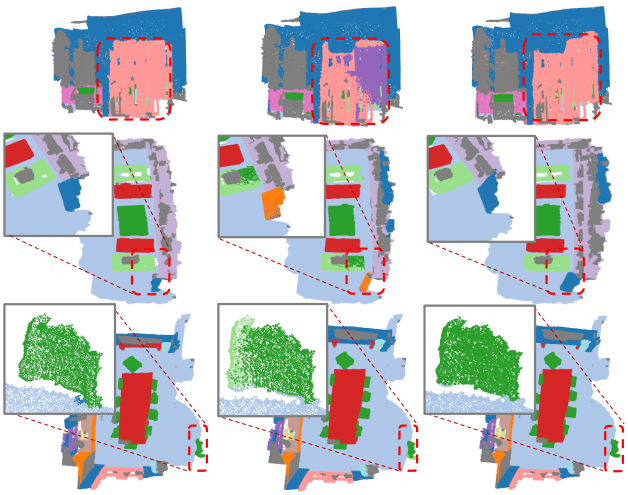
```
def Spatial_Wise_Sampling(
    s_min: float, s_max: float,     # min/max cell size
    cnt_low: int, cnt_high: int,    # target count range
    max_iter: int,                  # max iterations
    grid_range: float               # AABB size
) -> float:                         # optimal grid size
    l, r = s_min, s_max
    for _ in range(max_iter):
        grid_size = (l + r) / 2
        grid_shape = ceil(grid_range / grid_size)
        cell_count = prod(grid_shape)

        if cnt_low <= cell_count <= cnt_high:
            return grid_size
        elif cell_count < cnt_low:
            r = grid_size           # too sparse
        else:
            l = grid_size           # too dense
    return (l + r) / 2              # fallback
```

A. Experiment and Discussion

A.1. ScanNet Test Set

According to [12, 17, 18], there is a significant test time augmentation (TTA) difference between the val and test sets of Scannet [7] and Scannet200 [11]. The additional TTA includes incorporating data from the validation set for training, combining results from multiple training models [10], and applying over-segmentation [8].

For a fair comparison, we evaluated some SOTA models [5, 15, 17, 18] with an official or reproduction weights file using Val's TTA, as shown in Fig. 1. The experimental findings indicate that, without employing additional TTA, the result in the test set for the majority models [15, 17, 18] tends to be less than those on the validation set, rather than exceeding them. And it can be found that our model archive the best result both in the val and test set of Scannet and Scannet200.



(a) Ground Truth (b) PTv3 (c) Ours

Figure 1. Result Comparison in ScanNet.

(a) Ground Truth (b) PTv3 (c) Ours

Figure 2. Failed cases in ScanNet.

We are still working on over-segmentation and may update our model's test results employing over-segmentation in the final version of the paper. Furthermore, we recommend that future research ensure consistency in the TTA between the validation and test sets or at least make the TTA on the test set openly available. In 3D understanding, the focus should be on improving network design and training methodologies rather than using more testing tricks.

A.2. Pesade-code of Spatial-wise Sampling

The pesade-code of spatial-wise sampling is shown in Alg. 1. Spatial-wise sampling efficiently discerns the ideal proxy spacing by considering the AABB sizes of various points. The method is designed to maintain the proxies count within the bounds of N_{max} and N_{min}, using a bisection approach to determine the optimal proxy spacing L_p. If the number of proxies exceeds N_{max}, the proxy spacing L_p is reduced, and the total number of proxies is recalculated.

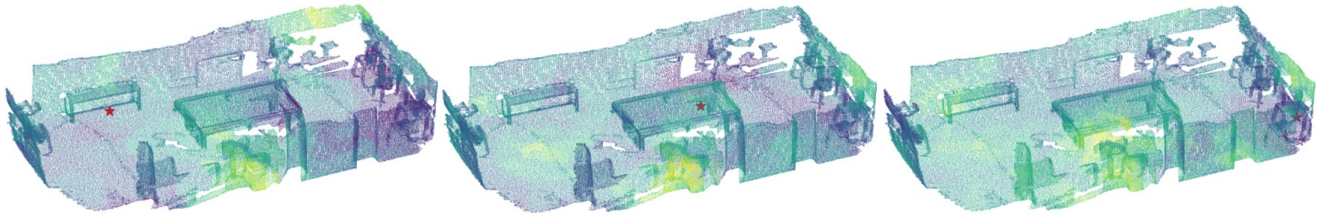
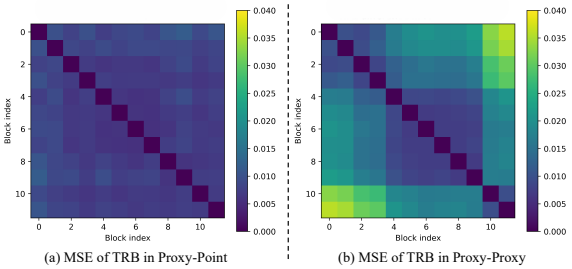


Figure 3. Visualization of the point-point attention map under FPS-based sampling. The red star represents the current point.

TRB	Share TRB	mIoU	mAcc	allAcc	Time
✔	✗	78.33	86.17	92.36	81ms
✔	✔	**78.71**	**86.23**	**92.51**	**74ms**

Table 2. Ablation study about sharing of TRB.



(a) MSE of TRB in Proxy-Point (b) MSE of TRB in Proxy-Proxy

Figure 4. (a) MSE of TRB in different proxy-point interaction layers. (b) MSE of TRB in different proxy-proxy interaction layers.

A.3. Visualization

Result Comparison. Fig. 1 compares the visualization of our method and PTv3 [17] on the Scannet dataset [7]. Specifically, Fig. 1 (a) presents the ground truth, Fig. 1 (b) depicts the result from PTv3, and Fig. 1 (c) illustrates our result. The visualizations indicate that, due to the global receptive field provided by the proxy, our method achieves more consistent and dependable segmentation results overall, thus enhancing segmentation performance.

Failed Cases. Fig. 2 illustrates the failed cases of our method on the Scannet dataset [7]. In Fig. 2, it is apparent that the proxy has failed to address the classification error attributed to location fusion. Consequently, achieving a proper balance between local and global information still requires further investigation.

Over-fitting due to FPS-based Sampling. Within the ablation study, the experiments show that FPS-based sampling performs poorly compared to alternative sampling methods. We visualized the point-proxy attention maps for FPS-based sampling methods to investigate this further, as illustrated in Fig. 3. The visualization demonstrates that the attention map resulting from FPS-based sampling exhibits static and repetitive patterns, with its attention not influenced by the proxy's location. Consequently, we contend that the sampling method based on FPS leads to significant overfitting of the model because of the scene leakage from FPS.

Efficiency	Indoor (ScanNet [7])		Outdoor (nuScenes [4])	
Methods	mIOU	Latency	mIOU	Latency
MinkUNet [5]	72.2	90ms	73.3	48ms
PTv2 [16]	75.4	191ms	80.2	146ms
PTv3 [17]	77.5	**61ms**	80.4	**44ms**
PTv3† [17]	77.6	73ms	80.5	53ms
SP^{2}T	**78.7**	74ms	**81.2**	54ms

Table 3. Ablation study about efficiency of SP^{2}T. PTv3† refers to PTv3 with an increased number of channels to maintain the same latency as SP^{2}T.

A.4. Shared Table Relative Bias

During the implementation phase, the point and proxy positions remain unchanged within the same layer. This means that for each instance of sparse attention in this layer, the relative position input provided to the relative position encoding module remains constant. Consequently, it is possible to compute the relative bias for a layer with a single invocation. Building on this optimization, we share all relative bias values across each layer, thus reducing model complexity and computational demand. Additionally, since the proxy positions are constant throughout the network, we test sharing the relative bias amongst proxies across the entire network and different layers.

Tab. 2 compares the accuracy of the model w/ and w/o the shared TRB. The experiment shows that the shared TRB improves the model's accuracy and reduces the inference time. Furthermore, Fig. 4 shows the MSE distance for TRB during point-proxy and proxy-proxy interaction in different layers. TRB demonstrates stage-level similarity in the proxy-proxy interaction, while all TRB is similar in the point-proxy interaction. Consequently, the sharing of TRB improves model accuracy and reduces inference time.

A.5. Model efficiency

Tab. 3 presents the model's performance based on accuracy and latency for both the indoor (ScanNet) and outdoor (nuScenes) datasets, tested on a single RTX 4090. It is important to note that scaling the PTv3 model to match the size of SP2T does not substantially improve the metrics for either dataset. Our model achieves an ideal compromise between accuracy and speed while maintaining consistent accuracy.

Indoor Semantic		Indoor Instance	
Config	Value	Config	Value
framework	/	frame	PointGroup
optimizer	AdamW	optimizer	AdamW
scheduler	Cosine	scheduler	Cosine
criteria	CrossEntropy (1) Lovasz [3] (1)	criteria	/
weight decay	5e-2	weight decay	5e-2
batch size	12	batch size	12
datasets	ScanNet / S3DIS	datasets	ScanNet
First Stage:			
learning rate	5e-3	learning rate	5e-3
block lr scaler	0.1	block lr scaler	0.1
warmup epochs	40	warmup iters	40
epochs	800	epochs	800
Second Stage:			
learning rate	2e-4	learning rate	2e-4
block lr scaler	1.0	block lr scaler	1.0
warmup epochs	20	warmup iters	20
epochs	400	epochs	400

Table 4. Indoor semantic / instance segmentation settings.

Outdoor Semantic			
Config	Value	Config	Value
optimizer	AdamW	batch size	12
scheduler	Cosine	weight decay	5e-3
criteria	CrossEntropy (1) Lovasz [3] (1)	datasets	NuScenes Sem.KITTI Waymo
First Stage:			
learning rate	2e-3	epochs	50
block lr scaler	1e-1	warmup epochs	2
Second Stage:			
learning rate	2e-4	epochs	30
block lr scaler	1.0	warmup epochs	1

Table 5. Outdoor semantic segmentation settings.

Outdoor Detection			
Config	Value	Config	Value
optimizer	AdamW	datasets	Waymo
scheduler	Cosine	weight decay	1e-2
framework	CenterPoint	batch size	12
First Stage:			
learning rate	3e-3	epochs	24
block lr scaler	1e-1	warmup epochs	0
Second Stage:			
learning rate	3e-4	epochs	12
block lr scaler	1.0	warmup epochs	0

Table 6. Outdoor object detection settings.

B. Implementation Details

Our implementation primarily utilizes Pointcept [6], a specialized codebase focusing on point cloud perception and representation learning. The details of our implementation are detailed in this section.

B.1. Training Settings

Datasets and metrics. The ScanNet dataset [7, 11], frequently utilized in indoor real-world down-stream tasks, contains 1,513 room scans derived from RGB-D frames, with 1,201 scenes designated for training and 312 reserved for validation. Each point was categorized into one of the 20 semantic labels in ScanNet [7] and 200 semantic labels in ScanNet200 [11]. In contrast, the S3DIS dataset [1] covers 271 rooms in six areas within three buildings with 13 categories.

nuScenes [4] consists of 40,157 annotated samples, each containing six monocular camera images that cover a 360-degree field of view and a 32-beam LiDAR. According to the specifications of nuScenes, the dataset comprises 1000 scenarios, 1.4M images, and 400K point clouds. The training set covers 700 scenarios, and the validation and test sets contain 150 scenarios each. SemanticKITTI [2] originates from the KITTI Vision Benchmark Suite and is comprised of 22 sequences, with 19 designated for training and the other 3 reserved for testing. Waymo [14] is a frequently utilized benchmark for outdoor 3D perception, comprising a total of 1,150 point cloud sequences (exceeding 200K frames). Each frame encompasses an extensive perception range of 150m × 150m.

For segmentation metrics, we utilize the mean class-wise intersection over union (mIoU) as the principal metric in ScanNet, ScanNet200, and S3DIS. Furthermore, following previous work, area 5 in S3DIS is designated for testing with a 6-fold cross-validation. For dectection metircs, all results are assessed by the conventional protocol employing 3D mean Average Precision (mAP) and its weighted version based on heading accuracy (mAPH).

Indoor semantic segmentation. The setting for indoor semantic segmentation is displayed in Tab. 4. The SP^{2}T model was trained in two phases. In the first stage, emphasis is placed on local fusion, using the local fusion network for separate training on Scannet [7] or S3DIS [1]. Hence, the model incorporates the weights of local fusion into the second training phase.

Indoor instance segmentation. Followed by PTv3 [17], we use PointGroup [9] as our foundational framework. Specifically, our configuration mainly follows PTv3. In addition, as with semantic segmentation, the model was trained in two stages.

Outdoor semantic segmentation. Similarly to indoor segmentation, Tab. 5 outlines the training parameters for SP^{2}T when applied to outdoor segmentation. Similarly to our approach for indoor segmentation, the model undergoes a two-stage training process. In the first stage, a distinct local fusion network is explicitly trained for dataset [2, 4]. Then, for the second stage, the model is initialized with the

Config	Indoor	Outdoor
Proxy embedding depth	2	
Proxy embedding temperature	10	1
Proxy init method	Spatial-wise	
Proxy number	160	400
Proxy search range	[0.0, 1.0]	[0.0, 20.0]
Proxy search iter	10	16
Association method	Vertex-based	
Association dim	3	2
Attention channels per head	16	
Attention dropout	0.0	
TRB table size	16	
TRB table strength	1.0	
TRB table temperature	[0.5, 2.5]	
Point-Proxy TRB input scale	2.5	0.2
Proxy-wise TRB input scale	0.4	0.04
Drop path	0.3	

Table 7. Model settings.

Augmentations	Parameters	Indoor	Outdoor
random dropout	dropout ratio: 0.2, p: 0.2	✓	-
random rotate	axis: z, angle: [-1, 1], p: 0.5	✓	✓
	axis: x, angle: [-1 / 64, 1 / 64], p: 0.5	✓	-
	axis: y, angle: [-1 / 64, 1 / 64], p: 0.5	✓	-
random scale	scale: [0.9, 1.1]	✓	✓
random flip	p: 0.5	✓	✓
random jitter	sigma: 0.005, clip: 0.02	✓	✓
elastic distort	params: [[0.2, 0.4], [0.8, 1.6]]	✓	-
auto contrast	p: 0.2	✓	-
color jitter	std: 0.05; p: 0.95	✓	-
grid sampling	grid size: 0.02 (indoor), 0.05 (outdoor)	✓	✓
sphere crop	ratio: 0.8, max points: 128000	✓	-
normalize color	p: 1	✓	-

Table 8. Data augmentations.

parameters of this local fusion network and continues to train.

Outdoor object detection. Tab. 6 outlines the training parameters for SP^{2}T when applied to outdoor object detection. The model is also trained through a two-stage approach. Initially, a specific local fusion network is exclusively trained for Waymo [14]. Subsequently, in the second stage, the model starts with the parameters from this local fusion network, and training is resumed.

B.2. Model Settings

Tab. 7 presents a comprehensive overview of our model's configuration, focusing primarily on the proxy's initialization and association method, table-based relative bias and dropout [13]. Furthermore, the parameters for local fusion mirror those of the specific methods [5, 17], and the proxy channel is the same as the channel of local fusion.

B.3. Data Augmentations

As illustrated in Tab. 8, we adopted the PTv3 [17] data augmentation approach to maintain fairness during both training and evaluation. In addition, we applied the same data enhancement to the test set and evaluated other models using this augmentation.

References

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 3

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 3

[3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 3

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 2, 3

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1, 2, 4

[6] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. `https://github.com/Pointcept/Pointcept`, 2023. 3

[7] 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

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

[9] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 3

[10] Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8500–8509, 2022. 1

[11] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 1, 3

[12] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 1

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 4

[14] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 3, 4

[15] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 1

[16] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022. 2

[17] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4840–4851, 2024. 1, 2, 3, 4

[18] Yu-Qi Yang, Yu-Xiao Guo, Jian-Yu Xiong, Yang Liu, Hao Pan, Peng-Shuai Wang, Xin Tong, and Baining Guo. Swin3d: A pretrained transformer backbone for 3d indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 1