# BSNet: Box-Supervised Simulation-assisted Mean Teacher for 3D Instance Segmentation

## 1. Overview

In this supplementary material, we begin by presenting a more detailed comparison of quantitative metrics on ScanNetV2 [3] validation set (Section 4). We then provide additional details on gravity,collision constraints and more implementation details (Section 5). Consequently, we present statistical results regarding the real overlapping samples in the ScanNetV2 training set, such as the distribution of class pairs, the number of overlapping samples for each class pair, and etc (Section 6). To further validate the effectiveness of the proposed method, we conduct additional ablation studies (Section 7) and provide more visualizations (Section 8). Finally, we discuss the limitations of our approach and outline potential future research directions in this field (Section 9).

## 2. Results of S3DIS 6-fold cross-validation.

As shown in the following Table 1, our method outperforms existing SoTA approaches, demonstrating its effectiveness.

Table 1. **Results of S3DIS 6-fold cross-validation.**

| Method | mAP | AP@50 | Method | mAP | AP@50 |
|---|---|---|---|---|---|
| GaPro + SoftGroup | 51.4 | 65.8 | GaPro + ISBNet | 51.5 | 66.8 |
| Ours + SoftGroup | **54.1** (+2.7) | **67.0** (+1.2) | Ours + ISBNet | **57.9** (+6.4) | **68.5** (+1.7) |

## 3. Results of 3D object detection.

Due to the same level of annotations between box-supervised 3D instance segmentation and fully supervised 3D object detection, our approach can be effectively extended to 3D object detection. As illustrated in Table 2, our approach performs well in 3D object detection, surpassing the current state-of-the-art methods and the GaPro versions of 3DIS methods. It achieves a notable increase of 2.6 in Box AP@50.

Table 2. **3D object detection results on ScanNetV2 validation set.**

| Method | Venue | Box AP@50 | Box AP@25 |
|---|---|---|---|
| VoteNet [12] | ICCV 19 | 33.5 | 58.6 |
| 3DETR [9] | ICCV 21 | 47.0 | 65.0 |
| GroupFree [7] | ICCV 21 | 52.8 | 69.1 |
| HyperDet3D [20] | CVPR 22 | 57.2 | 70.9 |
| FCAF3D [13] | ECCV 22 | 57.3 | 71.5 |
| CAGroup3D [18] | NeurIPS 22 | 61.3 | 75.1 |
| GaPro [10] + SPFormer [15] | ICCV 23 | 65.9 | 78.9 |
| GaPro + ISBNet [11] | ICCV 23 | 67.0 | 77.1 |
| Ours + SPFormer | - | 67.0 | **80.0** |
| Ours + ISBNet | - | **69.6** | 79.3 |

## 4. Detailed Results on ScanNetV2 Validation Set

The detailed results for each category on ScanNetV2 validation set are reported in Table 3. As the table illustrates, ours + ISBNet achieves the best performance in 7 out of 18 categories, ours + SPFormer achieves the best performance in 8 out of

18 categories. The two of them work together to achieve 14 out of 18 categories. The superior performance demonstrates the effectiveness of our method.

Table 3. **Full quantitative results of mAP on ScanNetV2 validation set.** For reference purposes, we show the results of fully supervised methods in gray. Best performance of box supervised methods is in boldface.

| Method | mAP | bathtub | bed | bookshe. | cabinet | chair | counter | curtain | desk | door | other | picture | frige | s. curtain | sink | sofa | table | toilet | window |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PointGroup [5] | 34.8 | 59.7 | 37.6 | 26.7 | 25.3 | 71.2 | 6.9 | 26.6 | 14.0 | 22.9 | 33.9 | 20.8 | 24.6 | 41.6 | 29.8 | 43.4 | 38.5 | 75.8 | 27.5 |
| SSTNet [6] | 49.4 | 77.7 | 56.6 | 25.8 | 40.6 | 81.8 | 22.5 | 38.4 | 28.1 | 42.9 | 52.0 | 40.3 | 43.8 | 48.9 | 54.9 | 52.6 | 55.7 | 92.9 | 34.3 |
| SoftGroup [17] | 45.8 | 66.6 | 48.4 | 32.4 | 37.7 | 72.3 | 14.3 | 37.6 | 27.6 | 35.2 | 42.0 | 34.2 | 56.2 | 56.9 | 39.6 | 47.6 | 54.1 | 88.5 | 33.0 |
| DKNet [19] | 50.8 | 73.7 | 53.7 | 36.2 | 42.6 | 80.7 | 22.7 | 35.7 | 35.1 | 42.7 | 46.7 | 51.9 | 39.9 | 57.2 | 52.7 | 52.4 | 54.2 | 91.3 | 37.2 |
| Mask3D [14] | 55.2 | 78.3 | 54.3 | 43.5 | 47.1 | 82.9 | 35.9 | 48.7 | 37.0 | 54.3 | 59.7 | 53.3 | 47.7 | 47.4 | 55.6 | 48.7 | 63.8 | 94.6 | 39.9 |
| ISBNet [11] | 54.5 | 76.3 | 58.0 | 39.3 | 47.7 | 83.1 | 28.8 | 41.8 | 35.9 | 49.9 | 53.7 | 48.6 | 51.6 | 66.2 | 56.8 | 50.7 | 60.3 | 90.7 | 41.1 |
| SPFormer [15] | 56.3 | 83.7 | 53.6 | 31.9 | 45.0 | 80.7 | 38.4 | 49.7 | 41.8 | 52.7 | 55.6 | 55.0 | 57.5 | 56.4 | 59.7 | 51.1 | 62.8 | 95.5 | 41.1 |
| Box2Mask [2] | 39.5 | 70.6 | 41.7 | 23.1 | 27.4 | 73.8 | 8.8 | 31.0 | 14.4 | 27.1 | 45.1 | 31.5 | 34.3 | 44.3 | 46.0 | 51.1 | 31.4 | 83.6 | 25.9 |
| WISGP [4] + PointGroup | 31.3 | 40.2 | 34.7 | 26.2 | 27.2 | 69.1 | 5.9 | 19.9 | 8.7 | 18.2 | 30.9 | 26.2 | 30.7 | 33.1 | 23.8 | 33.9 | 39.1 | 73.7 | 22.4 |
| WISGP + SSTNet | 35.2 | 45.5 | 32.8 | 23.8 | 30.4 | 75.3 | 8.8 | 23.9 | 17.6 | 27.8 | 33.0 | 28.4 | 31.4 | 23.1 | 32.9 | 42.7 | 39.4 | 83.4 | 25.9 |
| GaPro [10] + ISBNet | 50.6 | 76.3 | 45.5 | 28.5 | 46.0 | 82.7 | 21.8 | **41.3** | 22.0 | 51.3 | 51.3 | 55.9 | 44.5 | 52.8 | 59.7 | 49.5 | **52.8** | 90.2 | 39.5 |
| GaPro + SPFormer | 51.1 | **78.3** | 47.2 | **41.2** | 47.0 | 80.0 | 21.3 | 39.5 | 19.2 | 50.2 | 54.5 | 54.7 | 44.8 | 52.1 | 54.7 | **57.2** | 52.0 | 86.3 | 39.7 |
| **Ours + ISBNet** | 52.8 | 65.3 | **57.1** | 36.7 | 45.3 | **83.0** | **25.5** | 40.0 | **28.2** | **53.5** | **57.5** | 57.3 | 46.2 | **62.7** | 56.4 | 52.9 | 52.2 | 90.7 | 39.2 |
| **Ours + SPFormer** | **53.3** | 73.6 | 52.1 | 40.3 | **49.6** | 82.4 | **25.5** | 38.3 | 22.0 | 53.1 | 56.6 | **59.6** | **49.5** | 55.9 | **62.0** | 55.0 | **52.8** | **90.9** | **40.3** |

## 5. More Model Details

**Gravity and Collision Constraints.** Regarding the gravity constraint, we only need to change the z coordinates of the two foreground instances. We directly align their bottoms with the ground (the XY plane). After that, we need to move the instance to maintain collision constraint. In this stage, only x and y coordinates will be changed. Specifically, we start by voxelizing the two instances seperately. Once these two instances occupy the same voxel space, we consider them to be in collision. Subsequently, we perform an offset along the x-axis or y-axis for one of the instances until the collision is resolved.

**More Implementation Details.** As to the training setting of the pseudo-labeler SAFormer, we use adamw [8] and a cosinelr scheduler with a maximal learning rate of 10-4. We voxelize the point clouds with the size of 0.02m. We set the number of layers of both local-structure attention and global-context attention to 4. The decay of the Mean Teacher paradigm [16] is set as 0.999. The above settings remain the same for ScanNetV2 and S3DIS [1].

## 6. Statistical Results Regarding the Real Overlapping Samples

As shown in Figure 1, we draw the distribution of class pairs that make up the true overlapping samples in the ScanNetV2 training set. The blue square represents the existence of the corresponding class pair, the gray square represents a class pair symmetric to the class pair represented by the blue square, the white square represents the absence of a corresponding class pair. It is worth noting that, we only consider the 18 instance categories on ScanNetV2 while the finer-grained categories on ScanNet200 are ignored. Take the first row as an example, there are 17 class pairs, including (A: cabinet, A: cabinet), (A: cabinet, B: bed), (A: cabinet, C: chair), ..., and (A: cabinet, R: otherfurniture). Consequently, we count the number of overlapping samples for each class pair. The statistical results are shown in Figure 2. The horizontal axis is in order: (A, A), (A, B), (A, C), ..., (B, B), (B, C), (B, D), ..., (R, R). From Figure 2, we can find that the class pair (chair, table) is the largest class pair among all class pairs. As shown in Figure 3 and Figure 4, we calculate the mean $\mu$ and variance $\sigma$ of the distances between the center points of each class pair.

### 6.1. About the results from the class-agnostic approach.

As shown in Table 4, we replace the original SD with a class-agnostic approach (random) and observe a slight decrease in mAcc, but it still surpasses the "Base". Additionally, we compare 3DIS results, as indicated in the columns for mAP and

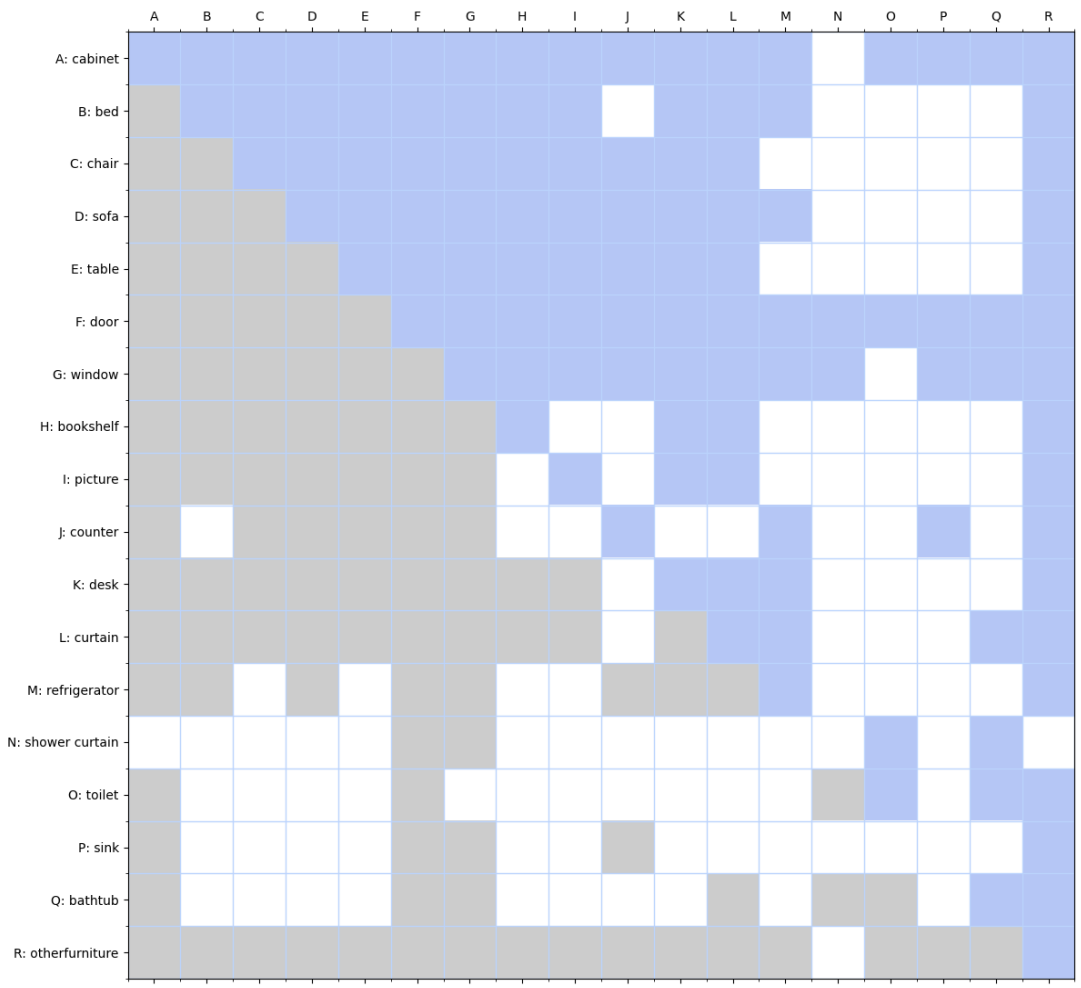


Figure 1. **Distribution of class pairs which make up the true overlapping samples.** Here, the blue square represents the existence of the corresponding class pair, the gray square represents a class pair symmetric to the class pair represented by the blue square, the white square represents the absence of a corresponding class pair.

AP@50. The results demonstrate that even with the adoption of a class-agnostic approach, high-quality pseudo-labels can still be generated, leading to accurate 3DIS outcomes. We also show results on ScanNet200 validation set in Table 5, which demonstrate the effectiveness of our method for datasets with a large number of classes.

Table 4. **Quality of pseudo-labels in overlapping areas.** Here, "Base" refers to C4 in Table 4 of the mainscript.

| Setting | mAcc | mAP | AP@50 | Setting | mAcc | mAP | AP@50 |
|---|---|---|---|---|---|---|---|
| Base | 55.3 | / | / | Base + random + GCC + ABP | 57.5 | 52.0 | 70.9 |
| Base + random | 56.7 | / | / | Base + SD + GCC + ABP | **59.6** | **52.8** | **71.7** |
| Base + SD | 58.5 | / | / | / | / | / | / |

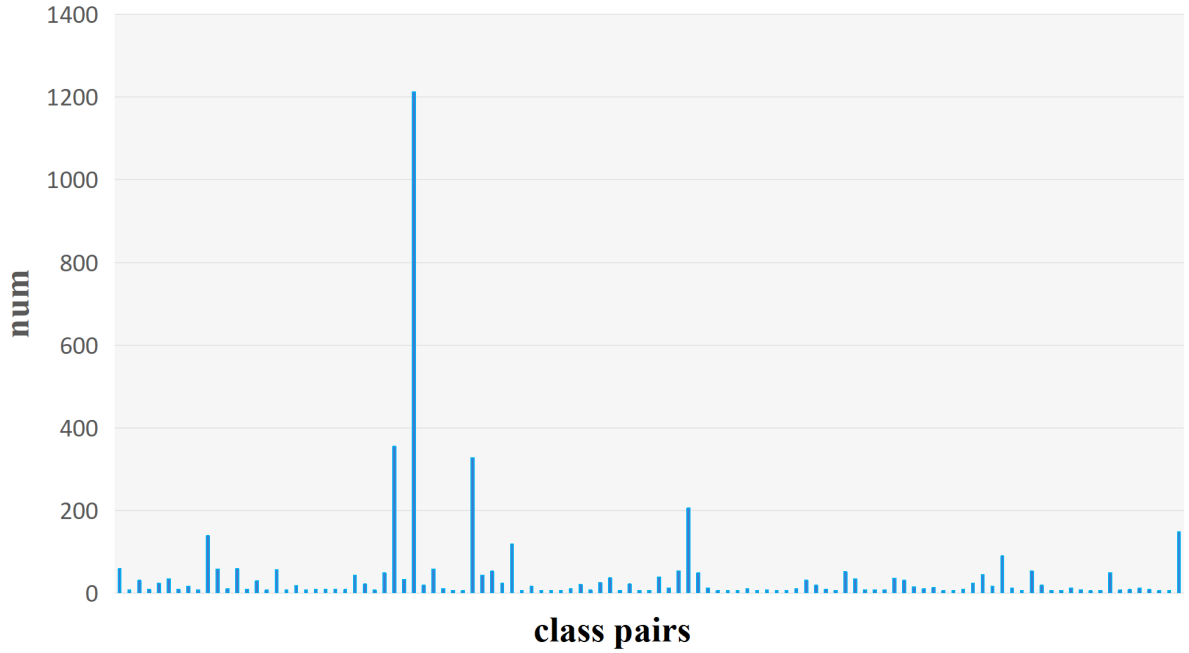


Figure 2. **Distribution of the number of class pairs.** The horizontal axis is in order: (A,A), (A,B), (A,C), ..., (B,B), (B,C), (B,D), ..., (R, R).


Figure 3. **The mean $\mu$ of the distances between the center points of each class pair.** The horizontal axis is in order: (A,A), (A,B), (A,C), ..., (B,B), (B,C), (B,D), ..., (R, R).

### 6.2. About the statistical distance.

In the first row of Figure 6, the distance modification is indeed large due to the existence of individual specificity, but in the other rows, it is small. In fact, the distance modification induced by constraints is minimal in the majority of samples, as shown in Table 6. Assuming the distance mean for each class pair is denoted as $m_i$ and the distance variance is $v_i$, the number of class pairs is $N$. AvgMean = $\frac{\sum_{i=1}^{N} m_i}{N}$ , AvgStd = $\frac{\sum_{i=1}^{N} v_i}{N}$ , AvgErr = $\frac{\sum_{i=1}^{N} |s_i^{ori} - s_i^{sim}|}{N}$ $(s \in \{m, v\})$. To further validate the effectiveness of statistical distance, the corresponding experiment is conducted in Table 7.


Figure 4. **The variance $\sigma$ of the distances between the center points of each class pair.** The horizontal axis is in order: (A,A), (A,B), (A,C), ..., (B,B), (B,C), (B,D), ..., (R, R).

Table 5. **Results on ScanNet200 validation set.** Here, "Ours + ISBNet" refers to training the network using pseudo-labels generated through the class-agnostic approach.

| Method | mAP | %full | AP@50 | %full |
|---|---|---|---|---|
| ISBNet | 24.5 | / | 32.7 | / |
| Ours + ISBNet | 23.0 | 93.9% | 31.2 | 95.4% |

Table 6. **Difference between origin and simulated samples.**

| Sample | AvgMean | AvgStd |
|---|---|---|
| Origin samples | 1.085 | 0.330 |
| Simulated samples | 0.989 | 0.336 |
| AvgErr | 0.096 | 0.018 |

Table 7. **Influence of statistical distance.**

| Setting | mAcc |
|---|---|
| W statistical distance | **59.6** |
| W/o statistical distance | 58.7 |

# 7. More Ablation Studies

## 7.1. About the hyperparameters in Equation 11.

We perform the experiment in the Table 8.

## 7.2. Designs for the layer number of Local-Global Aware Attention

We implement an ablation experiment about the layer number of Local-Global Aware Attention. As shown in the Table 9, the best performance is achieved when both Num_LA and Num_GA are fixed at 8, which yields only a 0.2 improvement over

Table 8. **Ablation study about the hyperparameters in Equation 11.**

| $\widehat{\lambda_1}$ | $\widehat{\lambda_2}$ | $\widehat{\lambda_3}$ | mAP | AP@50 | $\widehat{\lambda_1}$ | $\widehat{\lambda_2}$ | $\widehat{\lambda_3}$ | mAP | AP@50 |
|---|---|---|---|---|---|---|---|---|---|
| 0.1 | 1 | 1 | 51.8 | 70.8 | 0.5 | 1.5 | 1 | 52.6 | **71.7** |
| 0.5 | 1 | 1 | **52.8** | 71.6 | 0.5 | 1 | 0.5 | 52.2 | 71.2 |
| 1 | 1 | 1 | 51.8 | 70.9 | 0.5 | 1 | 1.5 | 52.5 | 71.3 |
| 0.5 | 0.5 | 1 | 52.3 | 71.0 | / | / | / | / | / |

the performance when Num_LA and Num_GA are fixed at 4. Therefore, considering the balance between performance and parameter quantity, we set both Num_LA and Num_GA as 4.

Table 9. **Ablation study on Local-Global Aware Attention.** Here, Num_LA represents the number of layers of the local-structure attention, Num_GA represents the number of layers of the global-context attention.

| Num_LA | Num_GA | mAcc |
|---|---|---|
| 2 | 2 | 57.7 |
| 2 | 4 | 58.5 |
| 4 | 2 | 58.9 |
| 4 | 4 | 59.6 |
| 8 | 8 | **59.8** |

### 7.3. Designs for the threshold $\tau$ of Simulation-assisted Mean Teacher

We conduct an ablation experiment on the threshold $\tau$ of Simulation-assisted Mean Teacher. As shown in Table 10, the best result is achieved when $\tau$ is 0.9. A high threshold will filter out many significant pseudo-labels, while a low threshold will retain too many noisy ones. Therefore, we conduct such experiments to select hyperparameters to optimize this trade-off.

Table 10. **Ablation study on the threshold $\tau$ of Simulation-assisted Mean Teacher.**

| $\tau$ | mAcc |
|---|---|
| 0.6 | 57.3 |
| 0.7 | 58.9 |
| 0.8 | 58.7 |
| 0.9 | **59.6** |
| 0.95 | 58.2 |

## 8. More Visualization

In Figure 5, we visualize and compare the generated pseudo-labels of several methods. As shown in this figure's blue circles, our method can generate more precise pseudo-labels. Take the second row as an example, Box2Mask incorrectly predicts some table points as chair, and Gapro incorrectly predicts some chair points as table. At the same time,our approach successfully distinguishes between instances of the two categories, obtaining more accurate pseudo-labels.

To offer a more comprehensive illustration of the simulated sample generation process, we present qualitative results in Figure 6. The visualization highlights the success of our method in generating simulated samples, showcasing the meaningful combination of individual 3D shapes. This visual representation underscores the effectiveness and quality of our simulated samples, contributing to a better understanding of the generation process.

## 9. Discussions about the Limitations and Future Research

Our approach endeavors to leverage the neural network to learn local structure features and global relationship information from overlapping samples. In non-overlapping regions, where point clouds exist solely within a single bbox, the labels are definite. However, in overlapping areas, where point clouds belong to two different bboxes, the labels are undetermined. To address the lack of labels for points in overlapping areas, we employ the Mean Teacher approach to generate and continuously refine pseudo-labels. To enhance the quality of pseudo-labels and expedite the convergence of Mean Teacher, we introduce the Simulation-assisted branch. By utilizing simulated samples, this branch fundamentally equips the network with the ability to distinguish overlapping areas. However, a challenge arises: there are a few background points within non-overlapping areas. When assigning labels to non-overlapping regions, inevitably some of these background points are misclassified as foreground. Effectively filtering out background points in non-overlapping areas becomes crucial. This helps the network recognize the correct instance shapes and enhances the ability to filter out background points in overlapping areas. Therefore, how to filter out background points in non-overlapping areas may be a future research hotspot. One potential solution is to leverage the RGB-D images and a strongly robust segmentation model, such as SAM, to filter out background points in non-overlapping regions. We will also seek inspiration from some 2D amodal instance segmentation works.

## 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. 2

[2] Julian Chibane, Francis Engelmann, Tuan Anh Tran, and Gerard Pons-Moll. Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes. In *European Conference on Computer Vision*, pages 681–699. Springer, 2022. 2

[3] 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

[4] Heming Du, Xin Yu, Farookh Hussain, Mohammad Ali Armin, Lars Petersson, and Weihao Li. Weakly-supervised point cloud instance segmentation with geometric priors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4271–4280, 2023. 2

[5] 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. 2

[6] Zhihao Liang, Zhihao Li, Songcen Xu, Mingkui Tan, and Kui Jia. Instance segmentation in 3d scenes using semantic superpoint tree networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2783–2792, 2021. 2

[7] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 1

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 2

[9] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2906–2917, 2021. 1

[10] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Gapro: Box-supervised 3d point cloud instance segmentation using gaussian processes as pseudo labelers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17794–17803, 2023. 1, 2

[11] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. Isbnet: a 3d point cloud instance segmentation network with instance-aware sampling and box-aware dynamic convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 1, 2

[12] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 1

[13] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022. 1

[14] 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. 2

[15] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2393–2401, 2023. 1, 2

[16] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 2

[17] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 2

[18] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35:29975–29988, 2022. 1

[19] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 235–252. Springer, 2022. 2

[20] Yu Zheng, Yueqi Duan, Jiwen Lu, Jie Zhou, and Qi Tian. Hyperdet3d: Learning a scene-conditioned 3d object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5585–5594, 2022. 1

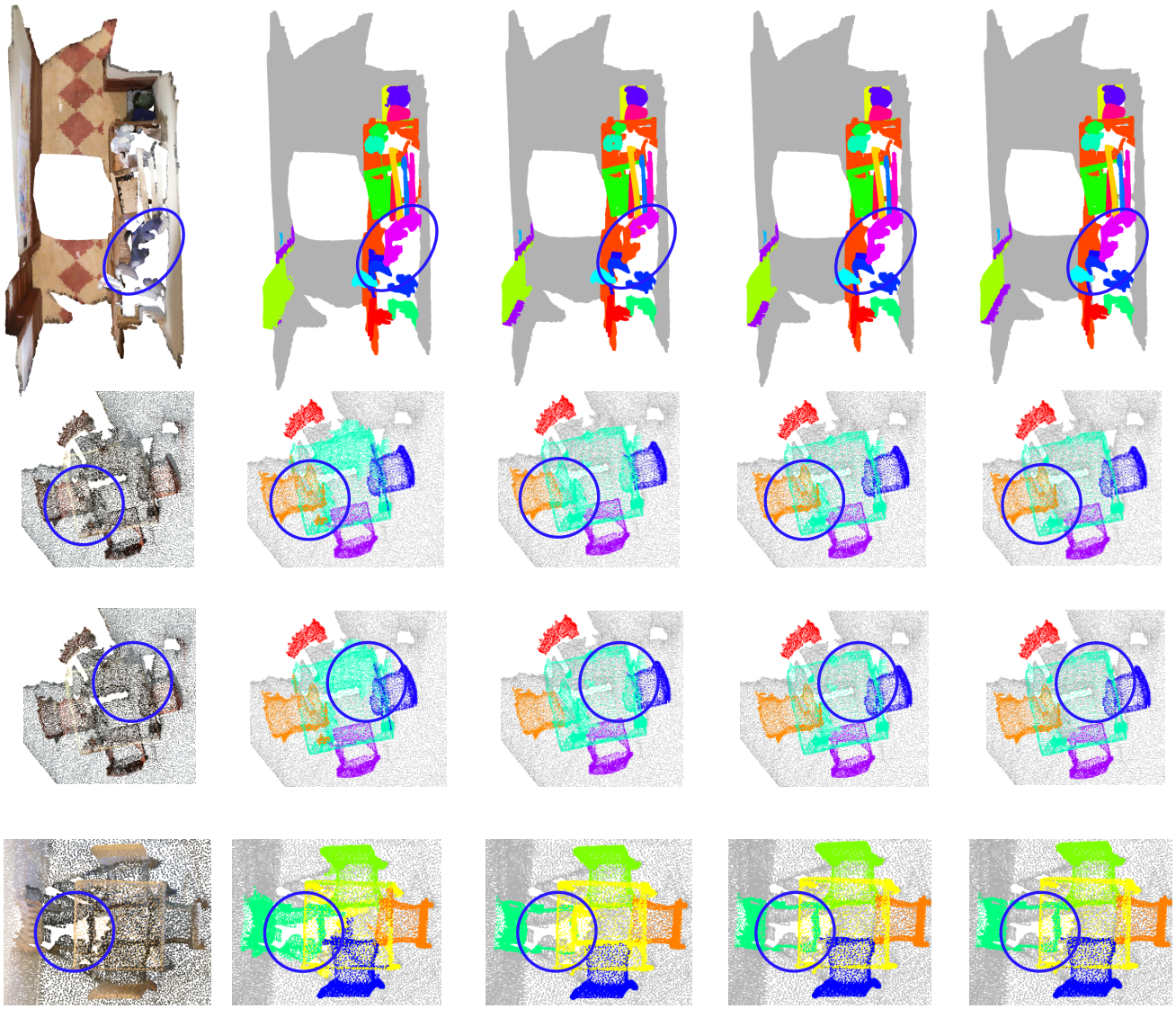


Figure 5. **More qualitative results on ScanNetV2 training set.** Our approach produces highly accurate pseudo instance masks, particularly in overlapping areas (blue circles).


Figure 6. **More qualitative visualization results of our Simulated Sample Generation.**