Sketchy Bounding-box Supervision for 3D Instance Segmentation

Supplementary Material

This supplementary material provides more details about our method, in which we provide more ablation studies and analysis in Sec. A, more quantitative and qualitative results in Sec. B, and we discuss the limitations and potential applications in Sec. C.

A. More ablation studies and analysis

Number of the query vectors. Since the performances of the query-based methods are sensitive to the predefined number of the query vectors, following existing methods [12, 13], we conducted the ablation studies on the number of the query vectors. Table 1 presents the experimental results of various numbered query vectors on the ScanNetV2 validation set. It can be observed that the performance improved nearly as the increasing of the number of the query vectors, and our method obtains considerable performance at the number of the query vectors $Q = 400$, when the number is equal to 600, the performance degrades. Thus we set $Q = 400$ in our final experiments.

Number of the multi-level attention blocks. We implement an ablation experiment about the block number of the Multi-level Attention Block. As shown in Table 2, the best performance is achieved when the block number is set as 6.

Ablation study about the hyper-parameter λ. We perform the ablation experiment about the loss weights of coarse-to-fine instance segmentator in Table 3, and it can be observed that the best performance is achieved when the λ_1, λ_2, λ_3 are set as 0.5, 1.0, 0.5, respectively.

Efficiency of the model training. In Fig. 1, we make a qualitative comparison between GaPro [9] and our method, which both utilize the same training strategy as SPFormer [13]. On the one hand, it can be observed that our model converges quicker in the first 200 epochs, which illustrates that our method can converge at a higher speed. On the other hand, our method trained for 400 epochs while the GaPro is trained for 512 epochs in total, which verifies that our method requires less training cost.

Effectiveness of the multi-level attention blocks. In Fig. 2, we compare the visualization of instance segmentation results on the ScanNetV2 validation set. The coarse instances are the visualization of the initially predicted instances of the coarse-to-fine instance segmentator, while the fine instances are the corresponding visualization after deploying the multi-level attention blocks. It can be observed that without the multi-level attention blocks, the unreliable region of the table can not be segmented precisely (illustrated both in the 1st and 2nd rows), and the chairs closely can not be differed (illustrated in the 3rd row). With the multi-level attention blocks, the segmented objects are refined, and our model obtains well-detailed instances, which verifies that our proposed multi-level attention blocks can segment instances in a coarse-to-fine manner.

Table 1. **Ablation Study of the number of the query vectors** using accurate bounding boxes.

Number	AP	AP_{50}	AP_{25}
200	43.7	66.4	82.4
400	**46.0**	**68.8**	**83.6**
600	45.6	68.1	83.5

Table 2. **Ablation Study of the number of the multi-level attention blocks** using S_4 sketchy bounding boxes.

Number	AP_{50}	AP_{25}
2	60.6	77.4
4	61.9	76.7
6	**62.5**	**80.1**
8	60.2	76.4

Table 3. **Ablation study about the hyper-parameter λ** for the instance segmentation loss using S_4 sketchy bounding boxes.

λ_1	λ_2	λ_3	AP_{50}	AP_{25}
0.1	1.0	0.5	38.3	46.1
0.5	0.5	0.5	57.1	74.7
0.5	1.5	0.5	59.7	75.1
0.5	1.0	1.0	58.6	74.1
0.5	1.0	1.5	54.9	70.5
0.5	1.0	0.5	**62.5**	**80.1**

B. More quantitative and qualitative results

Results of 3D object detection. The instance predictions can be transformed into bounding box predictions by obtaining the two corner coordinates of the predicted binary masks. in Table 4, we compare the object detection result with the state-of-the-art 3D object detection methods [6, 8, 10, 11, 14–18] and 3D instance segmentation methods [5, 12]. And the leading performance of our method on the 3D object detection task has further verified that our designed Sketchy-3DIS can explore the underlying instance characteristics indicated by the bounding boxes.

Quantity comparisons of pseudo labels on ScanNetV2 training set. We compare the quality of the pseudo labels of

Table 4. **3D Object detection results on the ScanNetV2 validation set.**

Method	Task	Box_AP_{50}	Box_AP_{25}
VoteNet [10]		33.5	58.6
MLCVNet [16]		41.4	64.5
3DETR [8]		47.0	65.0
H3DNet [17]		48.1	67.2
Group-free [6]	Detection	52.8	69.1
RBGNet [15]		55.2	70.6
HyperDet3D [18]		57.2	70.9
FCAF3D [11]		57.3	71.5
CAGroup3D [14]		**61.3**	**75.1**
Mask3D [12]		56.2	70.2
MAFT [5]	Segmentation	63.9	73.5
Sketchy-3DIS (ours)		**64.0 (+2.7)**	**79.4 (+4.3)**

Table 5. **Comparison of parameters and training time.** The parameters and training time are the total statistics for pseudo labeling and instance segmentation.

Method	Dataset	T (h)	P (M)
GaPro [9] + SPFormer		80	17.6
BSNet [7] + SPFormer	ScanNetV2	37	20
Sketchy-3DIS (ours)		**23**	**14.2**
GaPro [9] + ISBNet		150	31.1
BSNet [7] + ISBNet	S3DIS	72	33.1
Sketchy-3DIS (ours)		**59**	**30.7**

the GaPro [9] and our method. The experimental results are shown in Table 6, the per-category results of AP_{50} indicate that our method can generate more accurate pseudo labels than GaPro.

Comparison of parameters and training time. In Table 5, we compare the training time and parameters of the pseudo labeling task and instance segmentation task on GaPro [9], BSNet [7] and ours Sketchy-3DIS. Compared with these two state-of-the-art approaches, our method requires less training time and parameters, especially on the ScanNetV2 dataset, which verifies that our method can achieve the box-supervised 3D instance segmentation efficiently and economically.

Visualization of predicted instances on the ScanNetV2 test set. In Fig. 3, we present some samples of predicted instances on the ScanNetV2 test set, in which we utilize various colors to represent different objects. For an input point cloud, our proposed Sketchy-3DIS can correctly segment each instance and produce well-detailed segmentation results.

Per-category instance segmentation results. We show per-category instance segmentation results on ScanNetV2 validation set and test set in Table 7 and Table 8, respectively. It can be observed that our method achieves considerable performance on both the ScanNetV2 validation set and the test set.

C. The limitations and potential applications

Limitations. The proposed Sketchy-3DIS is a bounding-box supervised method that tolerates the inaccurate annotated boxes, however, the performance would suffer degradation once the annotated boxes are with huge inaccuracy. Additionally, the bounding-box annotations obtained from the point-wise annotations of the input point cloud is limited, which are the same as the existing methods [2, 4, 9]. Last but not least, the related datasets [1, 3] are preprocessed and aim for academic research, there may be some variations in real situations.

Potential applications. The proposed Sketchy-3DIS can be applied as a pre-annotation strategy for tasks that require dense-level annotations. Moreover, this work can facilitate the development of some applications of robotics and autonomous driving, in which the navigation system can obtain the objective signals benefit by the high performance on AP_{25} of our proposed Sketchy-3DIS.

Future work. In the future, we will consider adapting our method to dynamic situations to make full use of the robustness characteristic of our proposed Sketchy-3DIS. We also consider expanding our method to a wide range of tasks such as the few-shot instance segmentation, object detection, and part segmentation.

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 *CVPR*, 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 *ECCV*, 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 *CVPR*, pages 5828–5839, 2017. 2

[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 *WACV*, pages 4271–4280, 2023. 2

[5] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. Mask-Attention-Free Transformer for 3D instance segmentation. In *ICCV*, pages 3693–3703, 2023. 1, 2

[6] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3D object detection via transformers. In *ICCV*, pages 2949–2958, 2021. 1, 2

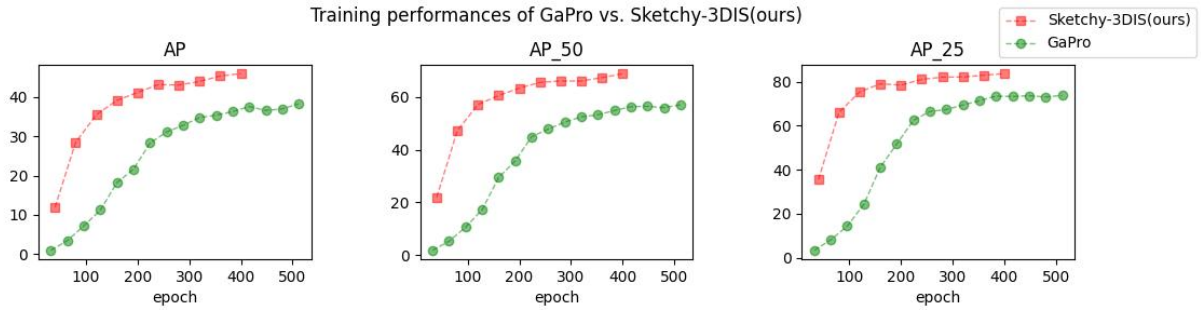

Figure 1. **Efficiency of the model training.** The figure compares the performance of GaPro [9] and ours Sketchy-3DIS during the training process on the ScanNetV2 validation set.

Table 6. **Quality of pseudo labels on ScanNet v2 training set.** The comparisons in AP_{50} of each object category is shown in this table.

Method	AP_{50}	bath	bed	bkshf	cabinet	chair	counter	curtain	desk	door	other fur.	picture	fridge	s. cur.	sink	sofa	table	toilet	window
GaPro [9]	81.9	99.1	85.9	82.4	88.2	74.8	48.9	74.5	45.6	81.0	84.4	95.6	88.4	94.9	87.0	91.3	72.2	92.7	88.0
Sketchy-3DIS (ours)	**86.8**	97.1	85.5	84.0	87.6	96.0	48.6	89.8	59.3	81.1	94.2	85.7	96.5	98.4	90.2	97.3	85.1	99.5	85.7

[7] Jiahao Lu, Jiacheng Deng, and Tianzhu Zhang. BSNet: Box-supervised simulation-assisted mean teacher for 3D instance segmentation. In *CVPR*, pages 20374–20384, 2024. 2

[8] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3D object detection. In *ICCV*, pages 2906–2917, 2021. 1, 2

[9] Tuan Duc Ngo, Binh-Son Hua, and Khoi Nguyen. GaPro: Box-supervised 3D point cloud instance segmentation using gaussian processes as pseudo labelers. In *ICCV*, pages 17794–17803, 2023. 1, 2, 3

[10] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3D object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 1, 2

[11] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. FCAF3D: Fully convolutional anchor-free 3D object detection. In *ECCV*, pages 477–493. Springer, 2022. 1, 2

[12] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *ICRA*, pages 8216–8223. IEEE, 2023. 1, 2

[13] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3D scene instance segmentation. In *AAAI*, pages 2393–2401, 2023. 1

[14] 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. *NIPS*, 35:29975–29988, 2022. 1, 2

[15] Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng Li, Bernt Schiele, and Liwei Wang. RBGNet: Ray-based grouping for 3D object detection. In *CVPR*, pages 1110–1119, 2022. 2

[16] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. MLCVNet: Multi-level context votenet for 3D object detection. In *CVPR*, pages 10447–10456, 2020. 2

[17] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3DNet: 3D object detection using hybrid geometric primitives. In *ECCV*, pages 311–329. Springer, 2020. 2

[18] Yu Zheng, Yueqi Duan, Jiwen Lu, Jie Zhou, and Qi Tian. HyperDet3D: Learning a scene-conditioned 3D object detector. In *CVPR*, pages 5585–5594, 2022. 1, 2

Table 7. **Per-category instance segmentation results of Sketchy-3DIS (ours) on ScanNetV2 validation set.** For reference purposes, we show the results of instance segmentation results on ScanNetV2 validation set.

	bath	bed	bookshelf	cabinet	chair	counter	curtain	desk	door
AP	75.1	23.4	29.5	39.3	76.5	15.5	41.0	9.4	46.3
AP_{50}	87.3	60.3	61.4	64.7	94.0	48.4	68.6	36.8	69.8
AP_{25}	87.3	60.3	61.4	64.7	94.0	48.4	68.6	36.8	69.8
	other fur.	picture	fridge	s. cur.	sink	sofa	table	toilet	window
AP	53.9	51.9	50.0	51.2	51.9	50.9	40.7	87.4	34.8
AP_{50}	69.2	67.7	68.7	70.2	73.7	78.1	66.7	94.7	57.7
AP_{25}	77.4	77.0	73.0	83.7	91.9	90.5	85.0	99.4	75.6

Table 8. **Per-category instance segmentation results of Sketchy-3DIS (ours) on ScanNetV2 test set.** For reference purposes, we show the results of instance segmentation results on ScanNetV2 test set.

	bath	bed	bookshelf	cabinet	chair	counter	curtain	desk	door
AP	74.1	32.0	31.0	37.1	70.4	5.7	39.0	11.6	41.0
AP_{50}	100.0	72.0	66.7	67.0	87.7	12.3	69.1	43.9	69.7
AP_{25}	100.0	95.5	76.1	80.5	95.0	78.5	75.7	91.3	90.8
	other fur.	picture	fridge	s. cur.	sink	sofa	table	toilet	window
AP	47.9	60.2	52.0	67.1	48.9	58.0	37.3	85.0	38.4
AP_{50}	62.7	67.0	64.5	100.0	80.7	77.6	60.3	100.0	60.4
AP_{25}	73.7	80.0	73.5	100.0	91.1	94.9	81.8	100.0	80.7

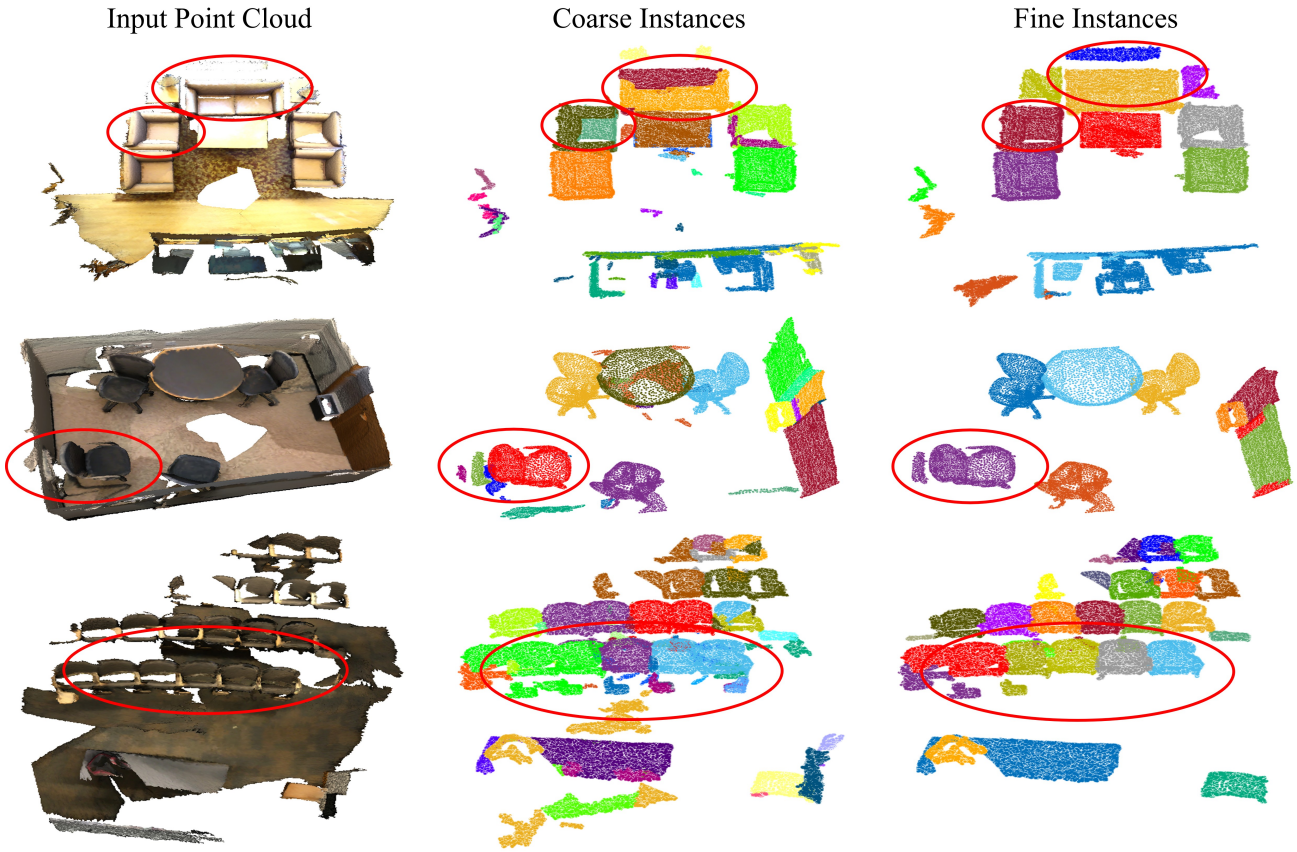

Figure 2. **Effectiveness of the Multi-level Attention Block.** "Coarse Instances" illustrate the predicted instances before the multi-level attention blocks in the coarse-to-fine instance segmentator, while "Fine Instances" illustrate the fine instances obtained after deploying the multi-level attention blocks. The red cycles highlight the key objects.

Figure 3. **Qualitative results on ScanNetV2 test set.** The top row denotes the input point cloud and the bottom row denotes the corresponding segmented instances.