GaPro: Box-Supervised 3D Point Cloud Instance Segmentation Using Gaussian Processes as Pseudo Labelers – Supplementary Material –

Tuan Duc Ngo Binh-Son Hua Khoi Nguyen
VinAI Research, Hanoi, Vietnam
{v.tuannd42, v.sonhb, v.khoindm}@vinai.io

1. Supplementary Material

In this supplementary material, we provide:

- Analysis on the impact of using superpoints in Gaussian Processes (Sec. 1.1).
- Results on S3DIS with 6-fold cross validation (Sec. 1.2).
- Runtime statistics including model parameters and training time (Sec. 1.3).
- Per-class AP on the ScanNetV2 validation set and hidden test set (Sec. 1.4).
- More qualitative results of our approach on all test datasets (Sec. 1.5).

1.1. Impact of Superpoints in Gaussian Processes

Due to the typically large number of points per instance in 3D point clouds, running Gaussian Process (GP) directly on a point level is often impractical. For example, the ScanNetV2 [3] dataset has around 1K-10K points per instance. To address this issue, we developed a point-level version of GP (row 3 in Tab. 3 in the main paper), which subsamples the top 800 nearest points from the determined points for each undetermined region. This version requires approximately 15 hours to generate pseudo labels for the entire ScanNetV2 training set on a single V100 GPU.

On the contrary, for the superpoint level version, the number of superpoints typically ranges from 10 to 1K, which allows us to run GP directly on all the superpoints. This version generates the pseudo labels in just 5 hours, as reported in the main paper. We found that the accuracy of our method when using superpoints is improved compared to using points because we can consider all superpoints and the superpoints are well aligned to the instance boundaries. These results are shown in Tab. 1.

1.2. Quantitative Results on S3DIS 6-fold Cross validation

Tab. 2 summarizes the results on 6-fold cross-validation of the S3DIS [1] dataset. We observe the same trend as the results on Area 5 in Tab. 2 of the main paper.

1.3. Run-time Statistics

Tab. 3 shows the parameters of the models and the training time of multiple methods on the ScanNetV2 dataset. For the **training time**, all the models are trained with batch size=8 on a single V100 GPU without mixed-precision training (FP16=False), and the other training details are kept the same as the original models. Our method generates pseudo-labels and can be plugged and played with any instance segmentation method. Therefore, there is no run-time overhead when applying our pseudo-labels for training compared to the full supervision case. The training time difference in Tab. 3 is mainly due to running time variations in data loaders and network optimizations.

	AP	**AP$_{50}$**	**Gen. Time**
GP with points	45.7	67.2	15 hours
GP with superpoints	**48.9**	**68.4**	**5 hours**

Table 1: Handling the undetermined regions to produce pseudo labels.

1.4. Per-class AP on the ScanNetV2 dataset

We report the detailed results of the 18 classes on the ScanNetV2 validation set and hidden test set in Tab. 4 and Tab. 5, respectively.

Method	**Sup.**	**AP**	**AP$_{50}$**
PointGroup [5]	Mask	-	64.0
SoftGroup [9]		54.4	68.9
ISBNet [7]		60.8	70.5
GaPro + PointGroup	Box	46.0	60.4
GaPro + SoftGroup		51.4	65.8
GaPro + ISBNet		51.5	66.8

Table 2: Results on S3DIS with 6-fold cross validation.

Method	Sup.	# of params	Training time	AP
PointGroup [5]	Mask	7.7M	32H	34.8
SSTNet [6]		113.2M	57H	49.4
SoftGroup [9]		30.8M	47H	46.0
ISBNet [7]		31.1M	39H	54.5
SPFormer [8]		17.6M	51H	56.3
Box2Mask [2]	Box	37M	101H	39.1
WISGP [4] + PointGroup		-	-	31.3
WISGP [4] + SSTNet		-	-	35.2
GaPro + PointGroup	Box	7.7M	32H	33.4
GaPro + SSTNet		113.2M	57H	43.9
GaPro + SoftGroup		30.8M	48H	41.3
GaPro + ISBNet		31.1M	40H	50.6
GaPro + SPFormer		17.6M	52H	51.1

Table 3: Models' parameters and training time on the ScanNetv2 validation set.

1.5. More Qualitative Results of Our Approach

The predicted instance masks of ISBNet [7] trained with our pseudo labels and GT labels are visualized in Fig. 1 (for ScanNetV2) and Fig. 2 (for S3DIS).

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 1

[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 (ECCV)*. Springer, October 2022. 2, 3

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 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, 3

[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. 1, 2, 3

[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, 3

[7] 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. *arXiv preprint arXiv:2303.00246*, 2023. 1, 2, 3, 4, 5

[8] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. *arXiv preprint arXiv:2211.15766*, 2022. 2, 3

[9] Thang Vu, Kookhoi Kim, Tung M. Luu, Xuan Thanh Nguyen, and Chang D. Yoo. Softgroup for 3d instance segmentation on 3d point clouds. In *CVPR*, 2022. 1, 2, 3

Method	AP	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	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 [9]	46.0	66.8	48.6	32.6	37.9	72.6	14.5	37.8	27.8	35.4	42.2	34.3	56.4	57.6	39.8	47.8	54.3	88.7	33.2
ISBNet [7]	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 [8]	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 [4] + 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 + PointGroup	33.4	46.8	58.1	32.4	31.4	63.1	21.8	26.5	36.2	20.3	27.4	20.6	25.8	20.9	18.5	48.2	41.6	65.0	18.4
GaPro + SSTNet	43.9	70.2	67.0	19.0	38.8	75.4	21.3	36.2	44.1	37.8	45.9	34.5	35.6	32.0	44.8	53.0	54.3	76.0	23.2
GaPro + SoftGroup	41.3	64.4	41.0	22.7	37.2	78.4	7.9	35.9	17.2	33.8	42.4	26.2	50.3	51.8	28.6	47.1	44.4	84.2	29.6
GaPro + 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

Table 4: Per-class AP of 3D instance segmentation on the ScanNetV2 validation set. Our GaPro's versions of 3DIS methods achieve competitive performances with SOTA fully supervised version.

Method	AP	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	s.curtain	sink	sofa	table	toilet	window
PointGroup [5]	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
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
ISBNet [7]	55.9	92.6	59.7	39.0	43.6	72.2	27.6	55.6	38.0	45.0	50.5	58.3	73.0	57.5	45.5	60.3	57.3	97.9	33.2
SPFormer [8]	54.9	74.5	64.0	48.4	39.5	73.9	31.1	56.6	33.5	46.8	49.2	55.5	47.8	74.7	43.6	71.2	54.0	89.3	34.3
Box2Mask [2]	43.3	74.1	46.3	43.3	28.3	62.5	10.3	29.8	12.5	26.0	42.4	32.2	47.2	70.1	36.3	71.1	30.9	88.2	27.2
GaPro + PointGroup	39.4	66.7	42.5	43.4	28.8	61.5	2.3	48.0	9.8	20.7	31.3	17.1	46.1	75.4	26.3	43.5	35.1	81.5	28.9
GaPro + SSTNet	45.8	85.2	47.2	38.2	35.8	66.7	13.1	37.2	19.0	32.3	40.8	28.3	34.3	86.3	43.1	52.6	46.3	92.9	24.5
GaPro + SoftGroup	42.1	55.3	45.6	35.7	28.9	69.0	4.3	47.1	19.7	29.3	37.5	19.4	51.9	71.8	24.8	57.4	44.3	86.9	28.8
GaPro + ISBNet	49.8	73.6	56.1	42.3	38.2	70.1	11.5	42.6	13.7	40.8	43.6	53.7	51.3	72.3	46.6	60.8	45.5	93.7	31.1
GaPro + SPFormer	48.2	73.9	50.5	44.2	38.1	71.5	6.4	41.4	18.9	43.0	43.5	55.0	35.4	70.7	42.7	67.6	43.5	84.2	36.7

Table 5: Per-class AP of 3D instance segmentation on the ScanNetV2 hidden test set. Our GaPro's versions of 3DIS methods achieve competitive performances with SOTA fully supervised versions.

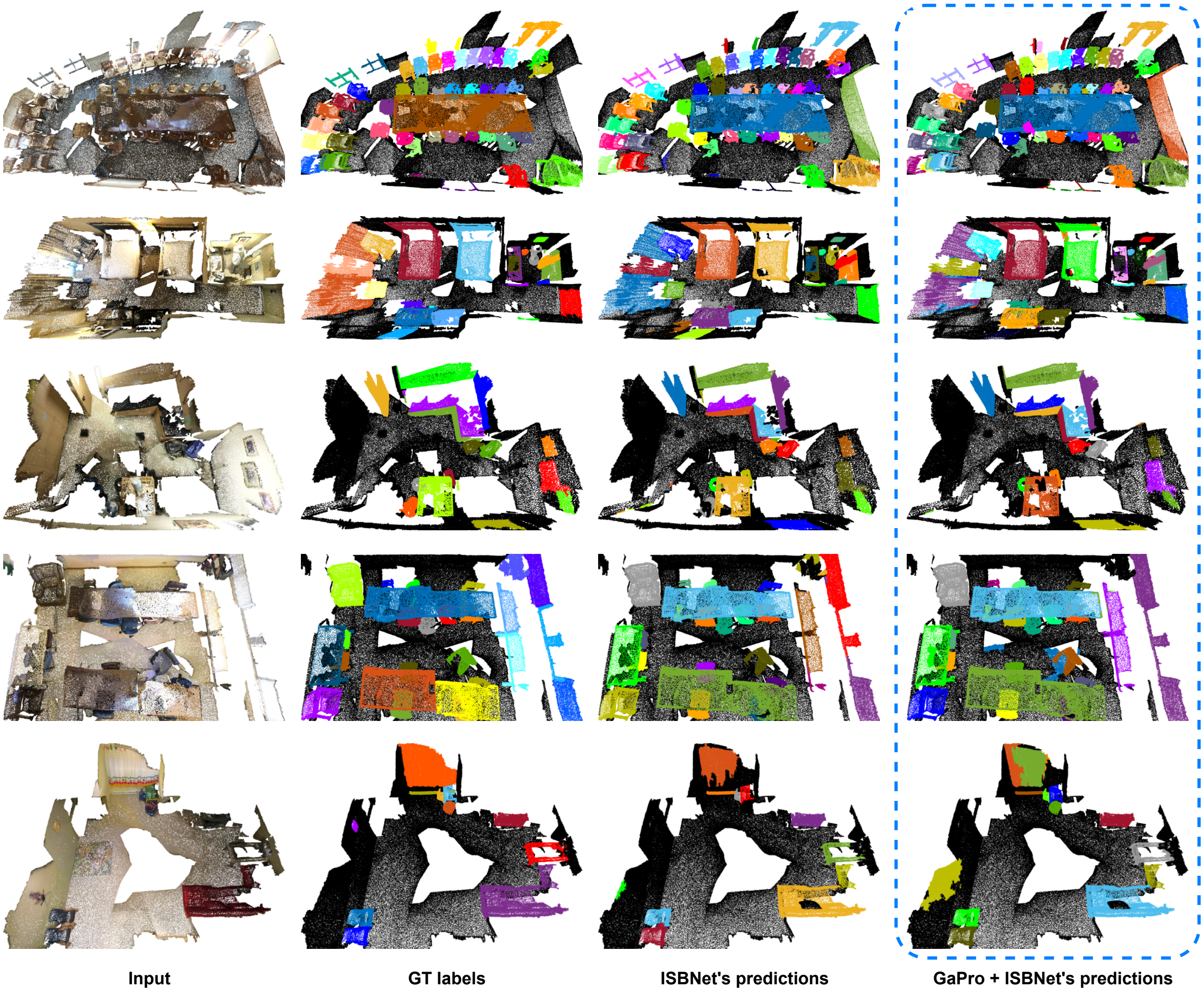

Figure 1: The predicted instance masks of ISBNet [7] trained with our pseudo labels and GT labels on ScanNetV2. Each row shows one example including the input point cloud, GT labels, predictions of ISBNet trained with GT labels, and predictions of ISBNet trained with our pseudo labels (dash box). The ISBNet trained with our labels gives comparable results to the fully supervised counterpart, except for the last row.

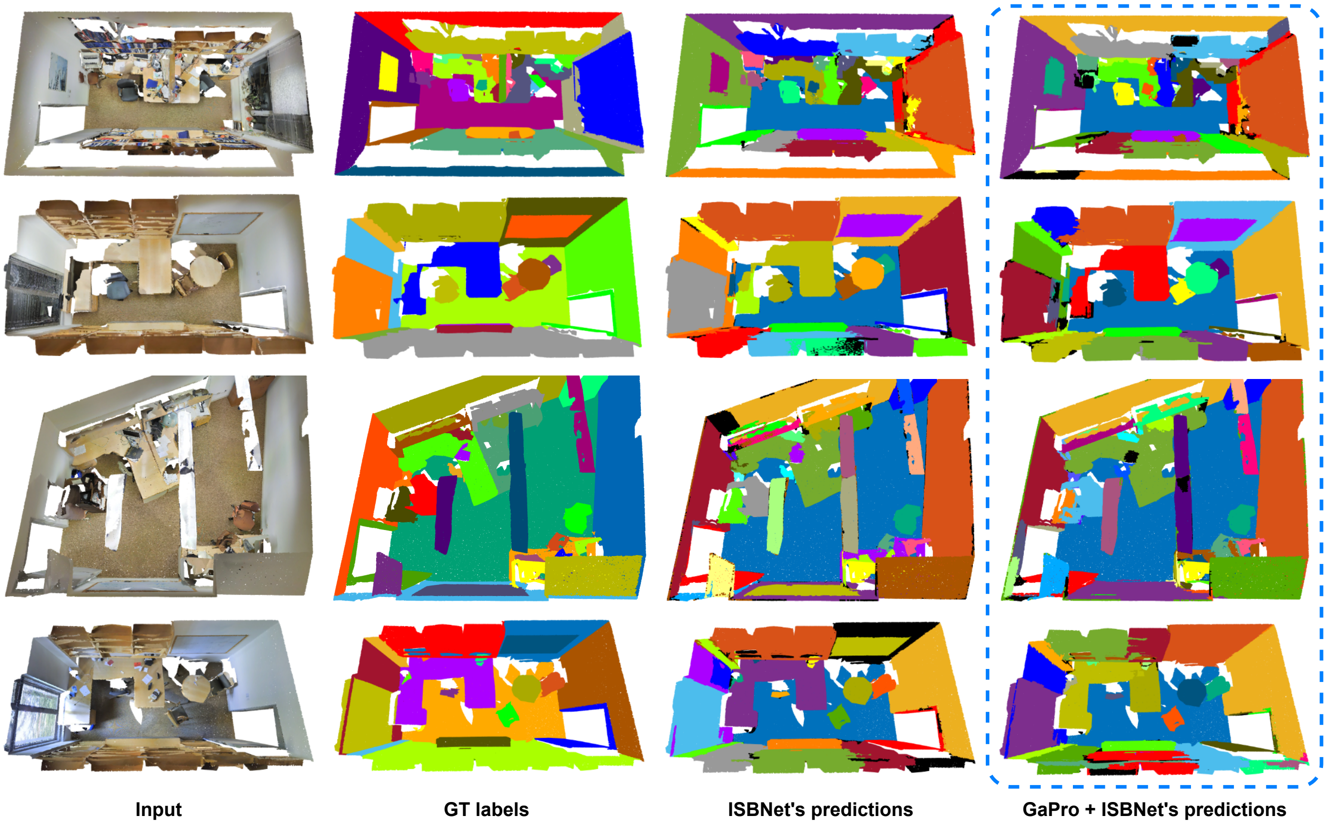

Figure 2: The predicted instance masks of ISBNet [7] trained with our pseudo labels and GT labels on S3DIS. Each row shows one example including the input point cloud, GT labels, predictions of ISBNet trained with GT labels, and predictions of ISBNet trained with our pseudo labels (dash box). The ISBNet trained with our labels gives comparable results to the fully supervised counterpart, except for the last row