Appendix of OneFormer3D

Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, Danila Rukhovich
Samsung Research
{m.kolodiazhn, a.vorontsova, a.konushin, d.rukhovich}@samsung.com

A. Per-category Scores

Since segmentation tasks are severely imbalanced in terms of categories, an averaged score might shadow some crucial performance issues. To provide a complete picture, we report 3D panoptic segmentation scores on the ScanNet validation split and on the S3DIS Area-5 split in Tab. 1 and 2, respectively. Besides, per-category 3D instance segmentation scores on the ScanNet test split are listed in Tab. 3. Evidently, OneFormer3D segments every single category more precisely than competitors on the ScanNet validation split. Panoptic segmentation scores on S3DIS have never been reported so far, so we establish a baseline for future research. On the ScanNet test split, our method outperforms others in segmenting objects of 11 out of 18 categories.

B. Performance

To provide a comprehensive overview of the proposed method, we also conduct a detailed performance analysis. Specifically, we decompose our method into several self-sufficient and replaceable components: creating superpoints, extracting 3D features with a sparse 3D CNN, flexible pooling, and running a query decoder. We run a profiler to measure the time required for each component to proceed. Similarly, we identify components of competing approaches, and report the inference time component-wise in Tab. 4. The runtime is measured on the same RTX 3090 GPU. Compared with the SPFormer baseline, OneFormer3D processes a few additional queries for semantic segmentation, and uses another initialization strategy for instance queries. The computation overhead is though minor, causing a less than 3% increase of inference time. Overall, we can claim, that OneFormer3D is on par of SPFormer, which is the fastest among the profiled approaches.

C. Qualitative Results

To give an intuition on how the segmentation scores relate to actual segmentation quality, we provide additional visualizations of original and segmented point clouds from the ScanNet (Fig. 1) and S3DIS (Fig. 2) datasets.

Method	PQ	wall	floor	cabinet	bed	chair	sofa	table	door	window	bkshf	picture	counter	desk	curtain	fridge	s. cur.	toilet	sink	bath	other
SceneGraphFusion [12]	31.5	67.6	25.4	13.9	22.2	47.2	10.5	16.4	12.6	26.4	56.4	22.9	31.3	28.0	38.3	38.0	32.3	34.8	63.2	30.4	11.7
PanopticFusion [6]	33.5	40.4	76.4	23.8	35.8	46.7	42.1	34.8	18.0	19.3	16.4	26.4	10.4	16.1	16.6	39.5	36.3	76.1	36.7	31.0	27.7
TUPPer-Map [14]	50.2	68.5	74.6	47.1	60.3	45.8	49.6	52.5	38.1	38.7	53.5	42.0	38.8	44.6	32.6	47.5	52.3	74.5	45.5	57.4	39.9
OneFormer3D	**71.2**	**78.9**	**94.9**	**60.9**	**80.4**	**88.8**	**74.4**	**74.4**	**61.5**	**58.9**	**55.2**	**57.1**	**55.8**	**65.7**	**62.5**	**63.3**	**71.7**	**95.9**	**73.7**	**85.5**	**65.2**

Table 1. Per-class 3D panoptic segmentation PQ scores on the ScanNet validation split.

Method	PQ	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	b. case	board	clutter
OneFormer3D	62.2	92.0	96.5	81.5	0.0	40.9	66.2	81.4	43.9	87.0	48.5	46.0	81.3	43.9

Table 2. Per-class 3D panoptic segmentation PQ scores on the S3DIS Area-5 split.

Method	mAP_{50}	bath	bed	bkshf	cabinet	chair	counter	curtain	desk	door	other	picture	fridge	s. cur.	sink	sofa	table	toilet	window
NeuralBF [10]	55.5	66.7	89.6	84.3	51.7	75.1	2.9	51.9	41.4	43.9	46.5	0.0	48.4	85.7	28.7	69.3	65.1	100	48.5
PointGroup [3]	63.6	100	76.5	62.4	50.5	79.7	11.6	69.6	38.4	44.1	55.9	47.6	59.6	100	66.6	75.6	55.6	99.7	51.3
DyCo3D [2]	64.1	100	84.1	89.3	53.1	80.2	11.5	58.8	44.8	43.8	53.7	43.0	55.0	85.7	53.4	76.4	65.7	98.7	56.8
SSTNet [5]	69.8	100	69.7	88.8	55.6	80.3	38.7	62.6	41.7	55.6	58.5	70.2	60.0	100	82.4	72.0	69.2	100	50.9
HAIS [1]	69.9	100	84.9	82.0	67.5	80.8	27.9	75.7	46.5	51.7	59.6	55.9	60.0	100	65.4	76.7	67.6	99.4	56.0
DKNet [13]	71.8	100	81.4	78.2	61.9	87.2	22.4	75.1	56.9	67.7	58.5	72.4	63.3	98.1	51.5	81.9	73.6	100	61.7
TD3D [4]	75.1	100	77.4	86.7	62.1	93.4	40.4	70.6	81.2	60.5	63.3	62.6	69.0	100	64.0	82.0	77.7	100	61.2
ISBNet [7]	75.7	100	90.4	73.1	67.8	89.5	45.8	64.4	67.0	71.0	62.0	73.2	65.0	100	75.6	77.8	77.9	100	61.4
SPFormer [9]	77.0	90.3	90.3	80.6	60.9	88.6	56.8	81.5	70.5	71.1	65.5	65.2	68.5	100	78.9	80.9	77.6	100	58.3
Mask3D [8]	78.0	100	78.6	71.6	69.6	88.5	50.0	71.4	81.0	67.2	71.5	67.9	80.9	100	83.1	83.3	78.7	100	60.2
OneFormer3D	**80.1**	100	97.3	90.9	69.8	92.8	58.2	66.8	68.5	78.0	68.7	69.8	70.2	100	79.4	90.0	78.4	98.6	63.5

Table 3. Per-class 3D instance segmentation mAP_{50} scores on the ScanNet hidden test split at 17 Nov. 2023.

Method	Component	Device	Component time, ms	Total time, ms	mAP_{50}
PointGroup [3]	Backbone	GPU	48	372	56.7
	Grouping	GPU+CPU	218		
	ScoreNet	GPU	106		
SSTNet [5]	Superpoint extraction	CPU	168	400	64.3
	Backbone	GPU	26		
	Tree Network	GPU+CPU	148		
	ScoreNet	GPU	58		
HAIS [1]	Backbone	GPU	50	256	64.4
	Hierarchical aggregation	GPU+CPU	116		
	Intra-instance refinement	GPU	90		
SoftGroup [11]	Backbone	GPU	48	266	67.6
	Soft grouping	GPU+CPU	121		
	Top-down refinement	GPU	97		
Mask3D [8] w/o clustering	Backbone	GPU	106	221	73.0
	Mask module	GPU	100		
	Query refinement	GPU	15		
Mask3D [8]	Backbone	GPU	106	19851	73.7
	Mask module	GPU	100		
	Query refinement	GPU	15		
	DBSCAN clustering	CPU	19630		
SPFormer [9]	Superpoint extraction	CPU	168	215	73.9
	Backbone	GPU	26		
	Superpoint pooling	GPU	4		
	Query decoder	GPU	17		
OneFormer3D	Superpoint extraction	CPU	168	221	78.1
	Backbone	GPU	26		
	Superpoint pooling	GPU	4		
	Query decoder	GPU	23		

Table 4. The inference time and instance segmentation accuracy on the ScanNet validation split. We show comparable inference time to the fastest SPFormer [9], being significantly more accurate than all existing methods.

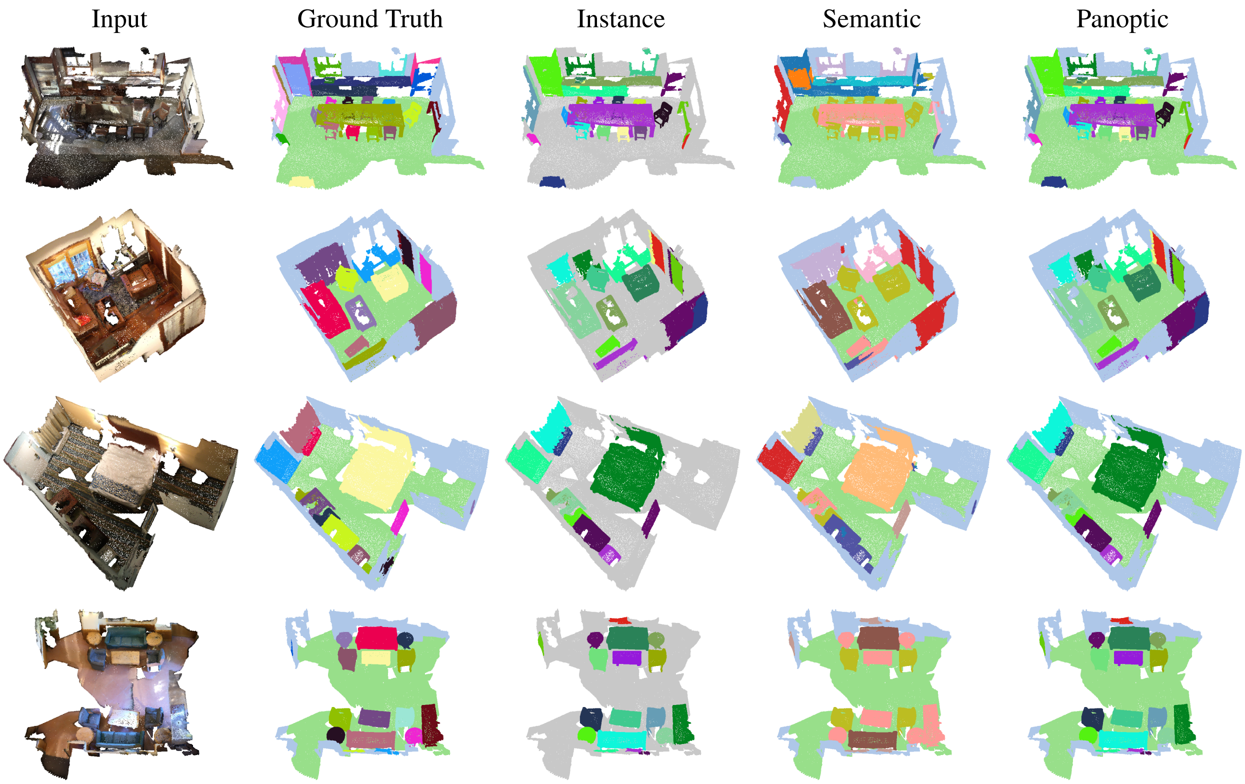

Figure 1. OneFormer3D predictions on ScanNet validation split. Left to right: an input point cloud, a ground truth panoptic mask, predicted 3D instance, 3D semantic, and 3D panoptic segmentation masks.

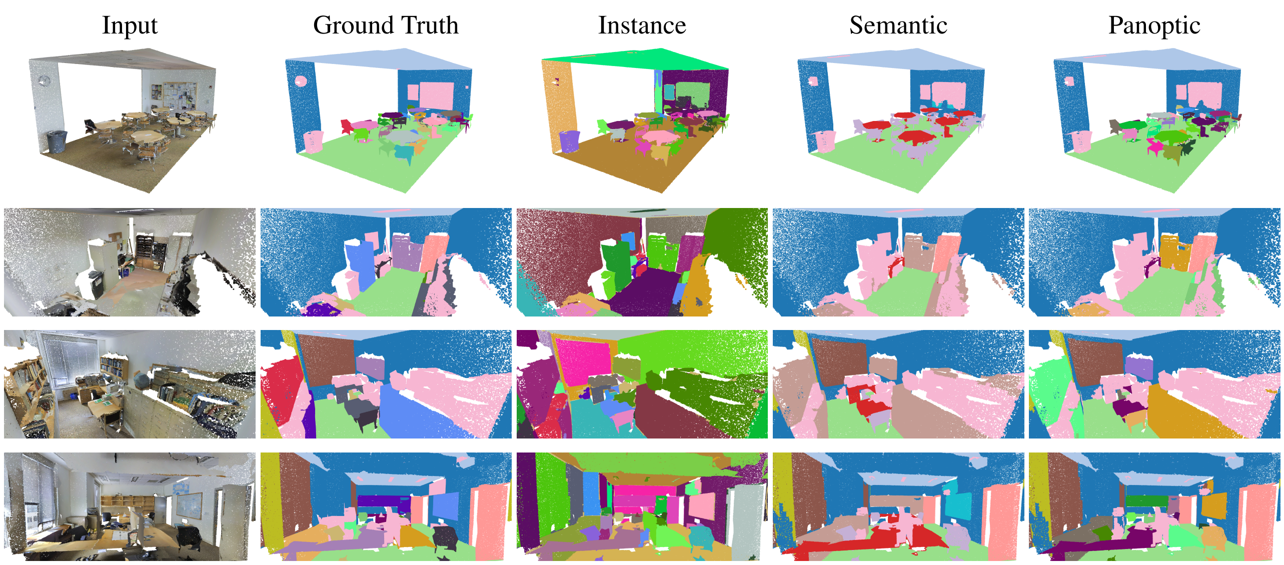

Figure 2. OneFormer3D predictions on the S3DIS Area-5 split. Left to right: an input point cloud, a ground truth panoptic mask, predicted 3D instance, 3D semantic, and 3D panoptic segmentation masks.

References

[1] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 2, 3

[2] Tong He, Chunhua Shen, and Anton Van Den Hengel. Dyco3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 354–363, 2021. 2

[3] 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, 3

[4] Maksim Kolodiazhnyi, Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Top-down beats bottom-up in 3d instance segmentation. *arXiv preprint arXiv:2302.02871*, 2023. 2

[5] 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

[6] Gaku Narita, Takashi Seno, Tomoya Ishikawa, and Yohsuke Kaji. Panopticfusion: Online volumetric semantic mapping at the level of stuff and things. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4205–4212. IEEE, 2019. 2

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13550–13559, 2023. 2

[8] 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, 3

[9] 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. 2, 3

[10] Weiwei Sun, Daniel Rebain, Renjie Liao, Vladimir Tankovich, Soroosh Yazdani, Kwang Moo Yi, and Andrea Tagliasacchi. Neuralbf: Neural bilateral filtering for top-down instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 551–560, 2023. 2

[11] 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. 3

[12] Shun-Cheng Wu, Johanna Wald, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scenegraphfusion: Incremental 3d scene graph prediction from rgb-d sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7515–7525, 2021. 2

[13] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *European Conference on Computer Vision*, pages 235–252. Springer, 2022. 2

[14] Zhiliu Yang and Chen Liu. Tupper-map: Temporal and unified panoptic perception for 3d metric-semantic mapping. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1094–1101. IEEE, 2021. 2