U3DS^3: Unsupervised 3D Semantic Scene Segmentation Supplementary Material

Jiaxu Liu[1] Zhengdi Yu[1] Toby P. Breckon[1,2] Hubert P. H. Shum[1]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK
{jiaxu.liu, zhengdi.yu, toby.breckon, hubert.shum}@durham.ac.uk

1. Introduction

In this supplementary, we provide more ablation studies on ScanNet [1] and SemanticKITTI [2]. Moreover, we demonstrate the two-pathway approach facilitates quicker convergence. Additionally, we provide more qualitative results for different scenes to validate the effectiveness and generalization ability of our method on both indoor (S3DIS [3], ScanNet [1]) and outdoor(SemanticKITTI [2]) datasets. Furthermore, we provide a demo video for improved visualization (https://www.youtube.com/watch?v=X_NLmoh5Q0c.). Tables S1 and S2 present the ablation studies on ScanNet [1] and SemanticKITTI [2] datasets, respectively, demonstrating the efficacy of our method. Both tables confirm that each component of our approach performs effectively and that the choice of superpoint number significantly influences the final outcomes. Figure S1 illustrates that two-pathway training not only enhances performance but also expedites convergence. Figures S2 to S4 showcase the qualitative results on the SemanticKITTI [2], ScanNet [1] and S3DIS [3] datasets. We compare our results with two classical clustering methods and the current existing work GrowSP [4].

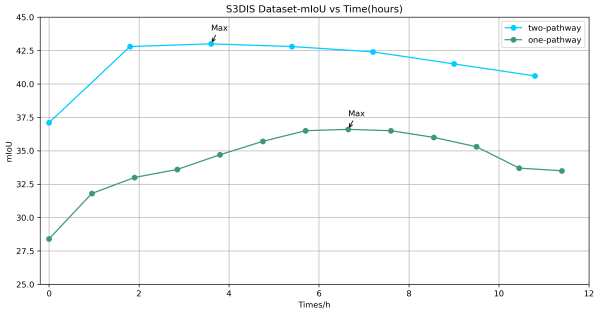

Figure S1. Curve of mIoU(%) changes over time(hour). The figure illustrates the expedited convergence achieved by the two-pathway approach.

Baseline	Eqv	Inv	γ_{sp}	mIoU	mAcc	oAcc
✓				11.2	25.5	36.3
✓		✓		12.0	26.3	37.5
✓	✓			14.3	29.3	40.4
✓	✓	✓		15.6	30.7	41.5
✓	✓	✓	80	22.7	41.5	52.8
✓	✓	✓	60	25.2	44.3	55.6
✓	✓	✓	40	**27.3**	**46.8**	**60.1**
✓	✓	✓	20	26.2	45.2	58.2

Table S1. Ablation study on ScanNet: Eqv denotes equivariant voxelized feature transformation; Inv denotes invariant colour transformation. γ_{sp} denotes the final superpoint number.

Baseline	Eqv	γ_{sp}	mIoU	mAcc	oAcc
✓			6.9	13.2	18.9
✓	✓		8.1	15.4	21.2
✓	✓	80	11.4	19.8	29.5
✓	✓	60	13.5	21.9	32.8
✓	✓	40	**14.2**	**23.1**	**34.8**
✓	✓	20	13.2	22.1	33.6

Table S2. Ablation study on SemanticKITTI: Eqv denotes equivariant voxelized feature transformation. However, there is no Inv due to the lack of color information in this dataset. γ_{sp} denotes the final superpoint number.

References

[1] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017. 1, 4

[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 *Int. Conf. Comput. Vis. (ICCV)*, October 2019. 1, 2, 3

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2016. 1, 5

[4] Zihui Zhang, Bo Yang, Bing Wang, and Bo Li, "Growsp: Unsupervised semantic segmentation of 3d point clouds," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2023, pp. 17619–17629. 1

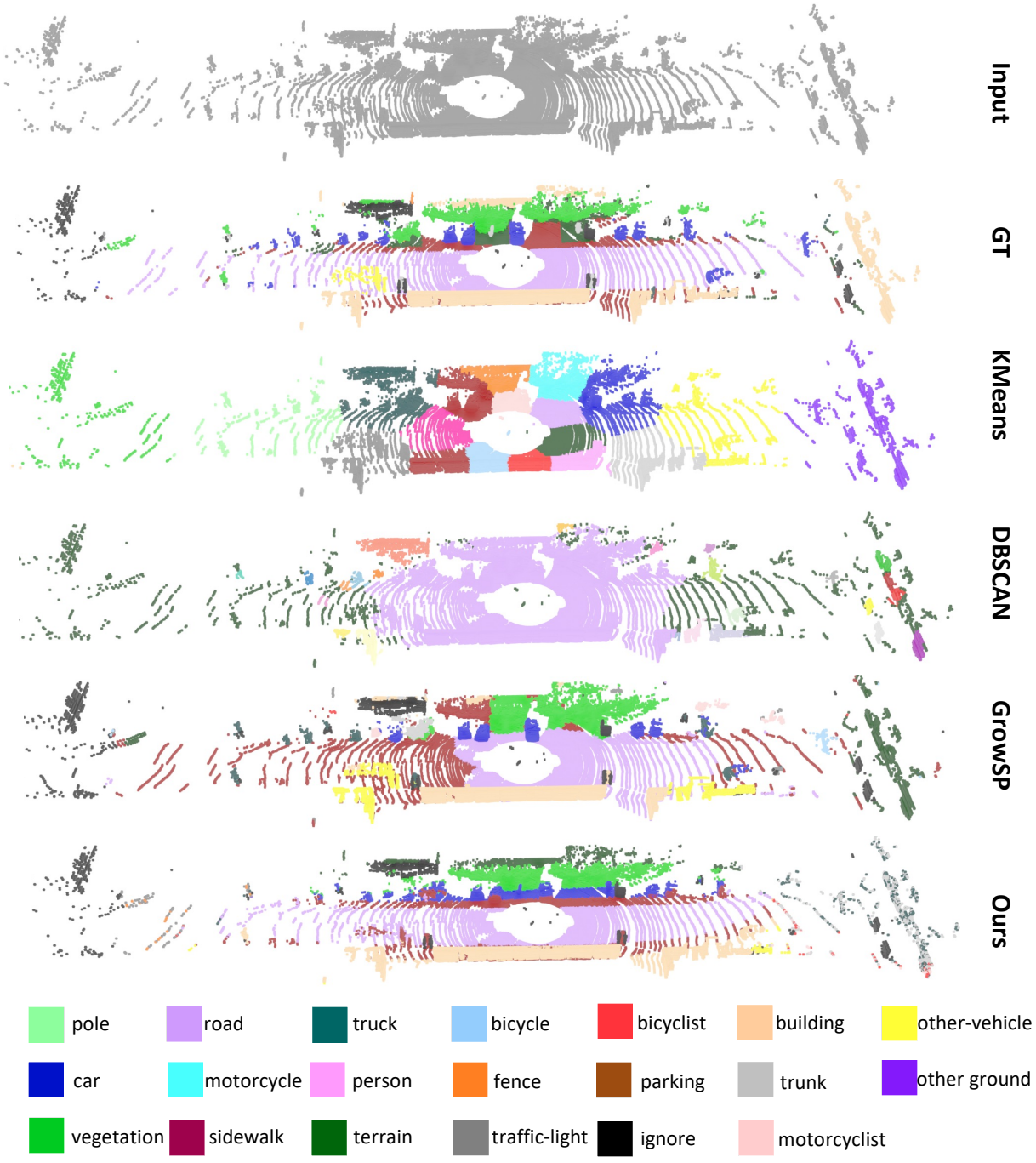

Figure S2. Qualitative example (a) on SemanticKitti [2]. Where each class is assigned to a colour (as per legend, bottom). This illustration shows superior performance compared to the baseline

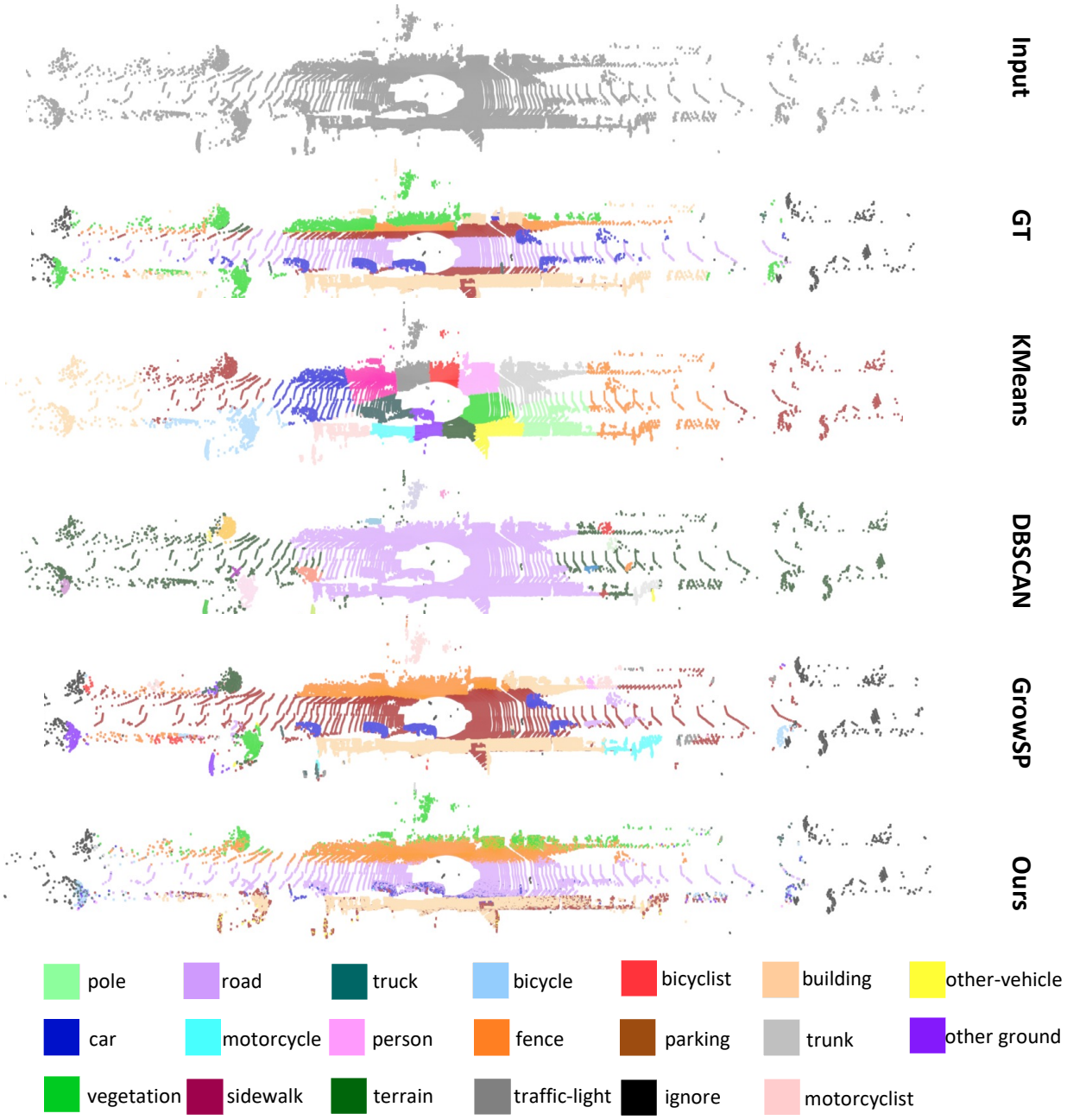

Figure S2. Qualitative example (b) on SemanticKitti [2]. Where each class is assigned to a colour (as per legend, bottom). This illustration shows superior performance compared to the baseline

Figure S3. Qualitative results on Scannet [1]. Evaluated with 20 categories exclude the ignored label, and each label is assigned to a colour.

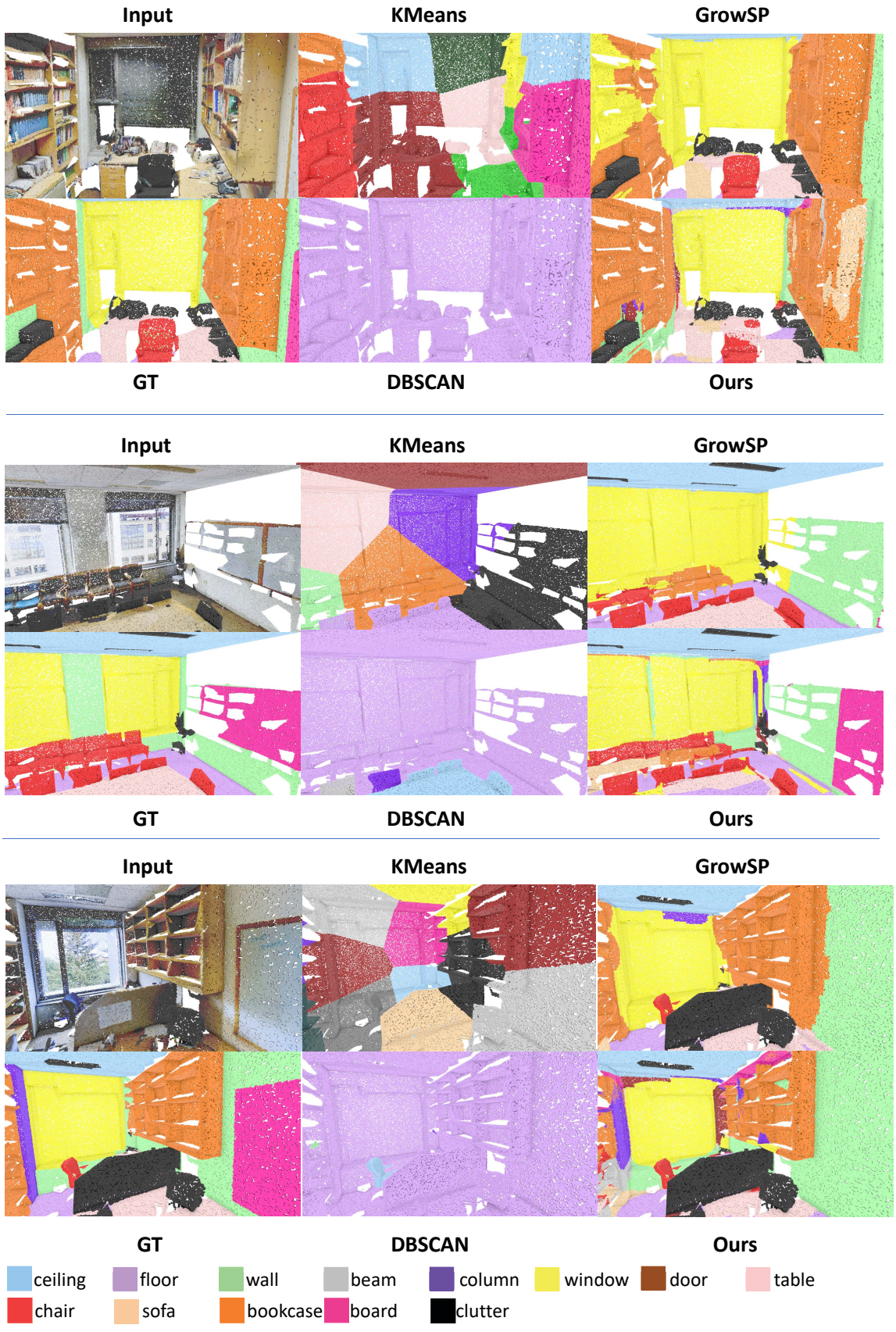

Figure S4. Qualitative results on S3DIS [3]. Each label at the bottom denotes one class, and this figure shows promising results.