Zero-Shot Point Cloud Segmentation by Semantic-Visual Aware Synthesis – Supplementary Material –

In the supplementary material, we will first show the extra details about the implementation in Appendix A, and then illustrate the class-level IoU performance of different datasets on generalized zero-shot 3D point cloud semantic segmentation in Appendix B. Finally, more ablation studies and analyses are showed in Appendix C.

Appendix A. Extra Implementation Details

We follow 3DGenZ [7] to construct our inductive generalized zero-shot segmentation experiments, the generator is implemented by Generative Moment Matching Network (GMMN) [6], which consists of a conditional multi-layer perceptron and apply maximum mean discrepancy (MMD) as the loss for training. The semantic embeddedings are acquired through GloVe [9] + Word2Vec [8], and the embeddeding dimension is 600, while the visual feature embeddeding dimensions are 64, 128 and 128 respectively according to the backbone FKAConv [4] (for ScanNet [5] dataset), ConvPoint [3] (for S3DIS [1] dataset) and KPConv [10] (for SemanticKITTI [2] dataset). We also employ the class-dependent weighting β and calibrated stacking value ϵ as [7] to reduce bias toward the seen classes and set β to 50, ϵ as 0.6, 0.4 and 0.2 for ScanNet, S3DIS and SemanticKITTI datasets respectively. The evaluation metric Harmonic mean IoU (HmIoU) that we use among seen and unseen classes can be formulated as:

$$\mathrm{HmIoU} = \frac{2 * \mathrm{mIoU}(\mathcal{C}^S) * \mathrm{mIoU}(\mathcal{C}^U)}{\mathrm{mIoU}(\mathcal{C}^S) + \mathrm{mIoU}(\mathcal{C}^U)}, \qquad (1)$$

where $\mathrm{mIoU}(\mathcal{C}^S)$ and $\mathrm{mIoU}(\mathcal{C}^U)$ denote the mean Intersection over Union mIoU performance (%) on seen classes $\mathcal{C}^S$ and unseen classes $\mathcal{C}^U$, respectively. All experiments in the paper are implemented on a single NVIDIA 3090 GPU and in pytorch framework with cuda 11.1.

Appendix B. Class-level Segmentation Results

Tab. 1 $\sim$ Tab. 3 show the numerical comparison of specific class-level IoU performance in zero-shot segmentation between our method and the current state-of-art 3DGenZ [7] on various 3 datasets. We apply the annotated data of both seen and unseen classes to train the fully-supervised segmentation models as the performance upper bound. We further visualize the confusion matrix in Fig. 1 for further detailed comparison and analysis:

Comparisons on ScanNet dataset. Tab. 1 provides the class-level semantic segmentation performance on the ScanNet dataset. From the table, we can notice that the performance of our method is significantly higher than that of 3DGenZ [7] in terms of IoU on all unseen classes and overall HmIoU. Moreover, our method performs best on the seen class *"shower curtain"* and unseen class *"sofa"* compared to 3DGenZ. Especially for unseen class *"sofa"*, we outperform 3DGenZ by 10.5% according to HmIoU. We can see from the confusion matrix in Fig. 1 (a) that this is because other unseen classes (*i.e."bookshelf"* and *"desk"*) for 3DGenZ are more likely to be predicted as *"sofa"*. Our method forms a relatively clear boundary among the prediction of unseen classes, which benefits from the semantic-visual aware modules that we propose. Meanwhile, it fully demonstrates that our model synthesize more real and highly generalized features for seen-to-unseen transfer.

Comparisons on S3DIS dataset. Tab. 2 displays the class-level segmentation quantitative results on the S3DIS dataset. We can see that our method is superior to 3DGenZ [7] in most of the seen and unseen classes, so as to acquire a higher HmIoU. In particular, our model acquires remarkable performance on seen classes *"board"*, *"bookcase"*, *"door"* and unseen class *"beam"*. In addition, it should be noted that both our method and 3DGenZ obtain poor results for unseen *"column"*. The same situation occurs in the fully-supervised training results. This shows that the visual characteristics of the seen class *"column"* are easily confused with other classes (*e.g."wall"*, Fig. 1 (b)), resulting in a low performance.

Comparisons on SemanticKITTI dataset. Tab. 3 shows the zero-shot segmentation results for each individual class on the SemanticKITTI dataset, which is more challenging in outdoor large-scale complex point cloud scenes. Our method performs relatively best on seen classes *"other vehicle"*, *"trunk"* and unseen class *"motorcycle"*, leading to a superior HmIoU performance. In contrast, 3DGenZ [7] incorrectly predicts more seen *"other-vehicle"*, unseen *"motorcycle"* as unseen *"truck"* and seen *"trunk"* as unseen *"traffic-sign"* (see Fig. 1 (c)). Moreover, it is worth noting that there exists a high degree of feature similarities among the seen *"bicycle"*, *"motorcyclist"*, *"person"* and unseen *"bicyclist"*, *"motorcycle"*, which is more likely to lead to the confusion in the predicted results of both our model and 3DGenZ (shown in Fig. 1 (c)).

Table 1. Class-level IoU performance (%) comparison of 3D point cloud generalized zero-shot semantic segmentation with state-of-art 3DGenZ [7] on the ScanNet dataset. "Full-Sup" denotes the fully-supervised training (upper bound) on both seen and unseen data with annotations. Both 3DGenZ [7] and our method are based on feature synthesis and employ GloVe [9]+Word2Vec [8] as word embeddings. "HmIoU" (%) represents the Harmonic mean IoU among seen and unseen classes. The best numerical results are in bold.

ScanNet	seen classes															**unseen classes**				
	bathtub	bed	cabinet	chair	counter	curtain	door	floor	other furniture	picture	refrigerator	shower curtain	sink	table	wall	window	**bookshelf**	**desk**	**sofa**	**toilet**	Hm IoU
Full-Sup	58.0	67.5	21.2	75.5	12.0	35.2	13.6	96.5	20.6	10.7	39.9	63.3	34.2	59.5	81.1	4.8	56.9	30.0	57.4	63.4	47.2
3DGenZ	**64.9**	44.0	16.9	63.2	**15.3**	**33.8**	**10.4**	**91.0**	10.1	**4.3**	26.1	0.2	27.5	43.1	71.3	2.8	6.3	3.3	13.1	8.1	12.5
Ours	61.0	**46.2**	**18.6**	**63.3**	14.2	31.1	4.6	90.7	**11.2**	0.9	**27.2**	**30.9**	**29.0**	**46.5**	**72.4**	**3.7**	**11.1**	**9.9**	**23.6**	**12.6**	**20.2**

Table 2. Class-level IoU performance (%) comparison of 3D point cloud generalized zero-shot semantic segmentation with state-of-art 3DGenZ [7] on the S3DIS dataset. "Full-Sup" denotes the fully-supervised training (upper bound) on both seen and unseen data with annotations. Both 3DGenZ [7] and our method are based on feature synthesis and employ GloVe [9]+Word2Vec [8] as word embeddings. "HmIoU" (%) represents the Harmonic mean IoU among seen and unseen classes. The best numerical results are in bold.

S3DIS	seen classes									**unseen classes**				
	board	bookcase	ceiling	chair	clutter	door	floor	table	wall	**beam**	**column**	**sofa**	**window**	Hm IoU
Full-Sup	53.9	54.4	96.5	75.9	66.0	78.7	96.0	70.3	74.1	63.1	10.2	54.1	72.4	59.6
3DGenZ	19.1	34.1	92.8	56.3	39.2	25.4	91.5	**57.3**	62.3	13.9	**2.4**	4.9	8.1	12.9
Ours	**33.0**	**48.3**	**96.0**	**57.2**	**44.3**	**40.4**	**91.9**	54.2	**64.8**	**22.3**	1.2	**6.2**	**9.3**	**16.7**

Table 3. Class-level IoU performance (%) comparison of 3D point cloud generalized zero-shot semantic segmentation with state-of-art 3DGenZ [7] on the SemanticKITTI dataset. "Full-Sup" denotes the fully-supervised training (upper bound) on both seen and unseen data with annotations. Both 3DGenZ [7] and our method are based on feature synthesis and employ GloVe [9]+Word2Vec [8] as word embeddings. "HmIoU" (%) represents the Harmonic mean IoU among seen and unseen classes. The best numerical results are in bold.

Semantic KITTI	seen classes														**unseen classes**				
	bicycle	building	car	fence	motorcyclist	other ground	other vehicle	parking	person	pole	road	sidewalk	terrain	trunk	vegetation	**bicyclist**	**motorcycle**	**traffic sign**	**truck**	Hm IoU
Full-Sup	42.0	88.6	93.6	65.8	0.0	2.7	41.1	28.9	69.7	63.7	89.4	77.1	70.5	70.7	87.5	74.4	58.6	26.7	41.6	54.5
3DGenZ	0.0	87.3	86.9	**61.8**	0.0	0.0	0.0	18.6	0.0	0.0	88.8	**78.6**	73.6	38.2	87.8	**28.0**	11.5	0.9	2.6	17.1
Ours	0.0	**89.1**	**91.7**	61.6	0.0	0.0	**26.9**	**26.7**	0.0	0.0	**89.5**	77.8	**73.8**	**71.3**	**88.2**	26.8	**16.4**	**1.5**	**6.6**	**20.1**

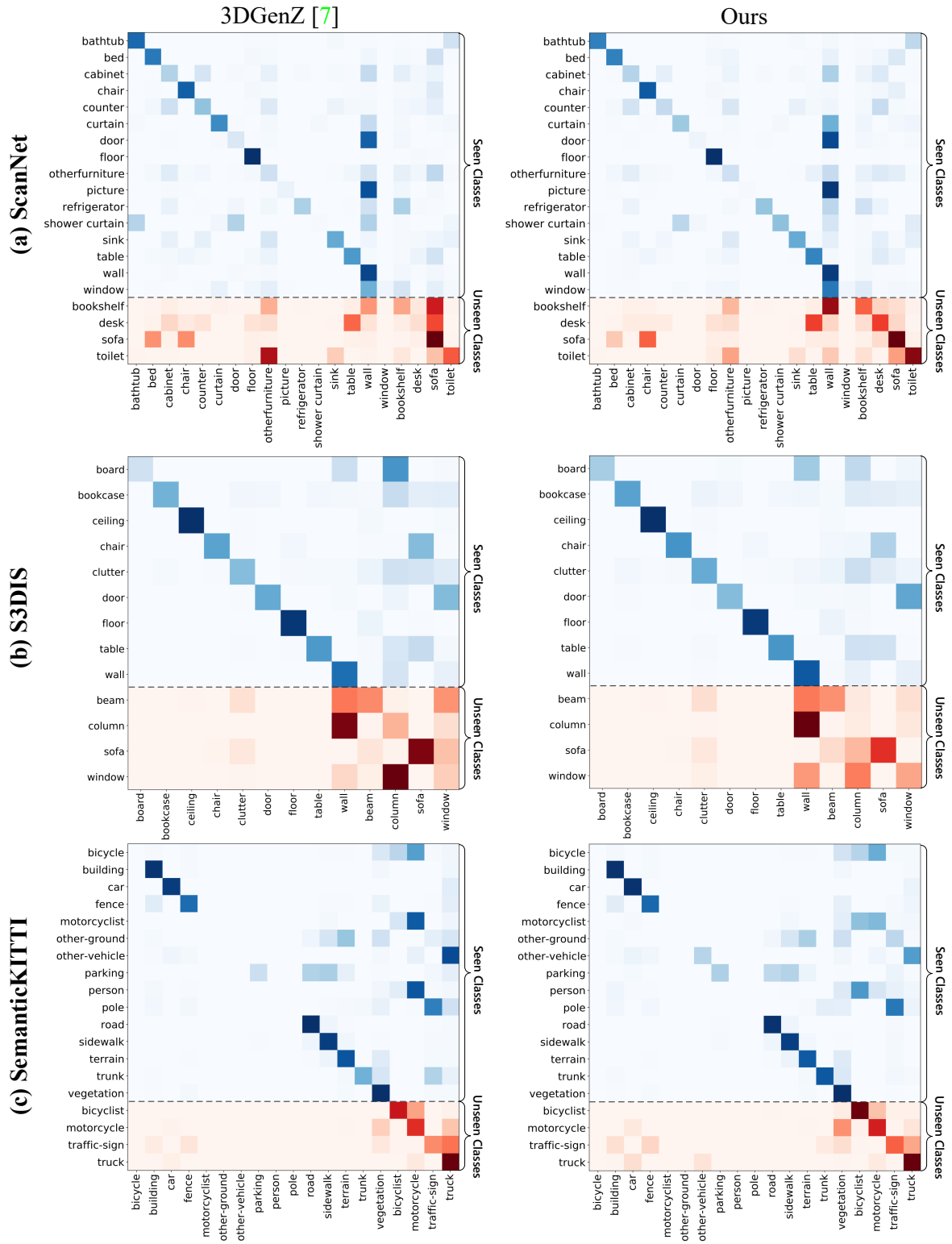

Figure 1. Visualization comparison of confusion matrices for generalized zero-shot point cloud segmentation on various datasets: (a) ScanNet [5], (b) S3DIS [1] and (c) SemanticKITTI [2]. The seen and unseen classes are in the blue and red color map respectively. We apply the confusion matrix to show the distribution of the model predicted results. The darker the color is, the more points of this class are predicted to be the class of the column. Therefore, the darker colors on the diagonal represent more correct predictions. Our method makes a significant improvement over the current state-of-the-art 3DGenZ [7] in both the seen and unseen classes.

Appendix C. More Ablation Studies

Results under vanilla ZSL setting: We show mAcc (%) and mIoU (%) of unseen classes under *vanilla* ZSL setting (rather than GZSL) on 3 datasets in Tab. 4. We reproduce the 3DGenZ [7] method for comparison. Results show quite significant gains of mIoU in ZSL setting, *i.e.,* ScanNet (↑ **12.2**), S3DIS (↑ **7.0**) and SemanticKITTI (↑ **7.5**).

Table 4. Experimental results of vanilla ZSL on three benchmarks.

Methods	ScanNet [5]		S3DIS [1]		SemanticKITTI [2]	
	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU
3DGenZ [7]*	62.3	40.5	24.7	14.9	53.3	39.9
Ours	**74.7**	**52.7**	**36.1**	**21.9**	**62.1**	**47.4**

Quantify separation in MCL module: We use the trained generators to synthesize 4 unseen classes features (500 samples per class) on ScanNet and measure the average Maximum Mean Discrepancy (MMD) between each two classes. We get MMD of **0.32** (w/o MCL) vs **1.87** (w/ MCL), suggesting that MCL enhances seperation between classes.

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 (CVPR)*, pages 1534–1543, 2016. 1, 3, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019. 1, 3, 4

[3] Alexandre Boulch. Convpoint: Continuous convolutions for point cloud processing. *Computers & Graphics*, 88:24–34, 2020. 1

[4] Alexandre Boulch, Gilles Puy, and Renaud Marlet. Fkaconv: Feature-kernel alignment for point cloud convolution. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 1

[5] 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 (CVPR)*, pages 5828–5839, 2017. 1, 3, 4

[6] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning (ICML)*, pages 1718–1727. PMLR, 2015. 1

[7] Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pages 992–1002. IEEE, 2021. 1, 2, 3, 4

[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013. 1, 2

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 1, 2

[10] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6411–6420, 2019. 1