Occlusion-aware Text-Image-Point Cloud Pretraining for Open-World 3D Object Recognition

Supplementary Material

8. Additional Discussion on Existing Works

Discussion on Occlusion Methods. Ren et al. [13] simplified occlusion by treating it as a form of corruption, referred to as "Drop Local," where k-NN clusters are randomly removed from point clouds. They then proposed an architecture and an augmentation strategy (based on deforming and mixing objects) to address *general* corruptions rather than focusing on occlusion. Hamdi et al. [8] introduced a viewpoint prediction module as a component for multi-view 3D recognition (which rely on 3D-to-2D projection). By predicting 'good' views to render images from point clouds, indirectly, the recognition model becomes more robust to occlusion (empirically simulated by randomly cropping the object point clouds along canonical directions). In contrast, our OccTIP method more realistically simulates self-occlusion through the rendering process and integrates single-view point clouds during pretraining, improving occlusion robustness for *any* point cloud encoders.

Comparison with VisionMamba (Vim). While Vim [22] also has a two-stream design, it has two key limitations: (1) reliance on one-directional neighborhood aggregation (CausalConv1D) and (2) only able to utilize a *single* neighborhood structure due to its simple forward and backward scanning strategy. In contrast, DuoMamba uses Conv1D for bidirectional local aggregation and can flexibly process two diverse orderings (e.g., Hilbert, Trans-Hilbert) simultaneously within a single block to fully exploit 3D geometry of the point clouds. These technical enhancements lead to improved performance as shown in Table 6.

Dataset	Vim [22]	Vim [22] + Hilbert	DuoMamba
ModelNet40-P	65.3	63.8	**67.7**
ScanObjectNN	61.1	62.7	**63.5**

Table 6. Zero-shot accuracy of Vim and DuoMamba.

9. Implementation Details

Triplet Generations. We render RBG images with a resolution of 512×512 and a transparent background. Similar to OpenShape [11], descriptions for each object come from three sources: (1) raw texts from the dataset's metadata, (2) captions generated by BLIP [10] and Azure Cognitive Services, (3) retrieved captions from visually similar images in the LAION-5B [14] dataset. The first source of captions (created from metadata) includes three texts: (a) object name, (b) object category, and (c) concatenation of the subcategory name.

Training Details. During pretraining, we use a batch size of 32 and randomly replace point colors with a constant value of 0.4 with a probability of 0.5. During testing, we assign the same constant value to point clouds that do not have color information, such as those in the ScanObjectNN [17] dataset. For more efficient training, we precompute and cache text and image features from CLIP [12] and directly use them as inputs to the text and image projection heads. Since there is significant fluctuation when training with partial point clouds, we follow [7] to employ Exponential Moving Average (EMA) [16] with a decay factor of 0.9995 to stabilize the training process. We use a cosine learning rate scheduler with a base learning rate of $7\mathrm{e}{-4}$.

10. Comparisons with Previous Works Pretrained on Larger Datasets

We further compare our method (pretrained on 52K ShapeNetCore [1] objects) with previous works pretrained on a significantly larger ensemble of 880K 3D objects from four datasets: ShapeNetCore [1], ABO [3], 3D-FUTURE [6], and Objaverse [5]. We use the official results reported in previous papers and evaluate all approaches on the real-world ScanObjectNN [17] dataset to assess their recognition performance in practical scenarios.

Model Size and Zero-Shot Object Classification Performance. We compare the parameter counts of various point cloud encoders and their zero-shot performance in Figure 5. Despite only being pretrained on ShapeNetCore [1], our DuoMamba outperforms all existing models of comparable size that are pretrained on 880K 3D objects – 17 times more data. Notably, the zero-shot accuracy gap between our model and the best-performing model Uni3D-giant [21] is just 1.8%, even though our model is only 1/35 its size. This highlights DuoMamba's superior size-to-performance efficiency. Scaling up the model and pretraining on larger datasets is likely to further enhance performance, which we leave as future work.

Few-Shot Linear Probing. We perform a few-shot experiment similar to the one in Section 6.2 (main paper), this time comparing our approach against models pretrained on the ensemble of 880K 3D objects. As illustrated in Figure 6, our method consistently outperforms all other works across all few-shot settings, highlighting our pretraining framework's data efficiency and effectiveness in learning robust and generalizable features for real-world recognition.

	Method	Mean	Cab	Bed	Chair	Sofa	Tabl	Door	Wind	Bksf	Pic	Cntr	Desk	Curt	Fridg	ShwrCurt	Toil	Sink	Bath	Bin
AP_{25}	PointCLIP [9]	6.00	3.99	4.82	45.16	4.82	7.36	4.62	2.19	-	-	1.02	4.00	-	-	-	-	13.40	6.46	-
	PointCLIP V2 [23]	18.97	**19.32**	20.98	61.89	15.55	23.78	13.22	**17.42**	-	-	**12.43**	21.43	-	-	-	-	14.54	16.77	-
	OpenShape* [11]	20.40	9.63	38.62	73.05	57.28	37.00	29.52	5.74	23.94	2.07	3.37	16.25	1.25	4.45	0.84	9.00	22.76	16.21	16.23
	MixCon3D† [7]	24.11	11.55	43.21	79.33	63.97	**42.91**	29.94	4.85	25.26	**3.98**	1.49	25.58	2.00	4.95	0.81	13.23	20.58	38.03	22.25
	TAMM* [20]	23.07	10.03	32.68	75.16	55.73	36.72	32.44	5.26	24.82	2.52	2.04	22.53	2.11	3.26	1.23	17.83	23.87	**46.50**	20.48
	OccTIP	**28.92**	12.85	**56.43**	**80.41**	**68.78**	40.11	**37.68**	7.09	**30.51**	3.21	2.46	**31.55**	**5.18**	**8.54**	**2.14**	**29.89**	**35.64**	41.93	**26.24**
AP_{50}	PointCLIP [19]	4.76	1.67	4.33	39.53	3.65	5.97	2.61	0.52	-	-	0.42	2.45	-	-	-	-	5.27	1.31	-
	PointCLIP V2 [23]	11.53	**10.43**	13.54	41.23	6.60	15.21	6.23	**11.35**	-	-	**6.23**	10.84	-	-	-	-	11.43	10.14	-
	OpenShape* [11]	16.12	3.78	36.99	62.48	49.48	33.05	17.40	2.12	21.97	0.61	1.34	11.97	0.45	4.18	0.59	8.38	10.68	16.16	8.55
	MixCon3D† [7]	19.09	3.61	41.90	67.67	51.13	**38.22**	17.34	1.56	23.44	**1.56**	0.36	18.63	0.59	4.71	0.43	12.07	9.18	37.69	13.51
	TAMM* [20]	18.11	3.10	31.64	64.35	42.51	30.82	20.55	2.11	21.26	0.85	0.50	17.71	0.80	3.09	0.81	17.00	10.44	**46.27**	12.26
	OccTIP	**22.73**	5.44	**54.77**	**68.91**	**55.53**	34.55	**22.55**	2.92	**25.71**	0.98	0.84	**22.91**	**2.34**	**8.36**	**1.31**	**27.27**	**16.86**	41.65	**16.27**

Table 7. Zero-shot 3D object detection results on ScanNetV2 [4]. Our method OccTIP achieves the highest mAP and consistently has the highest or second-highest AP scores across most categories, showing the superiority of the proposed approach in complex real-world recognition. (*: results obtained using released pretrained weights, †: results reproduced using the authors' public code.)

	Method	Mean	Bed	Table	Sofa	Chair	Toilet	Desk	Dresser	Night Stand	Bookshelf	Bathtub
AP_{25}	OpenShape* [11]	18.61	33.09	24.18	28.96	45.51	10.42	13.58	2.75	**11.77**	11.13	4.71
	MixCon3D† [7]	18.69	28.25	26.75	**34.44**	47.77	6.05	15.76	2.31	11.56	6.91	7.14
	TAMM* [20]	18.91	18.15	27.78	27.67	47.00	**21.41**	14.54	2.43	10.81	11.14	8.20
	OccTIP	**24.37**	**43.45**	**29.21**	34.22	**51.19**	12.78	**18.16**	**3.76**	11.14	**13.96**	**25.90**
AP_{50}	OpenShape* [11]	9.78	23.71	9.01	20.85	24.37	7.74	3.02	1.00	5.47	1.77	0.89
	MixCon3D† [7]	9.63	17.97	10.22	24.53	26.00	3.80	3.38	0.51	**6.30**	1.73	1.86
	TAMM* [20]	9.96	12.37	11.01	20.36	25.41	**17.96**	3.22	0.81	4.87	1.71	1.90
	OccTIP	**13.01**	**32.67**	**11.21**	**25.46**	**28.04**	8.50	**4.33**	**1.71**	5.11	**1.92**	**11.18**

Table 8. Zero-shot 3D object detection results on SUN RGB-D [15]. Our method OccTIP achieves the highest mAP and consistently has the highest or second-highest AP scores across most categories, showing the superiority of the proposed approach in complex real-world recognition. (*: results obtained using released pretrained weights, †: results reproduced using the authors' public code.)

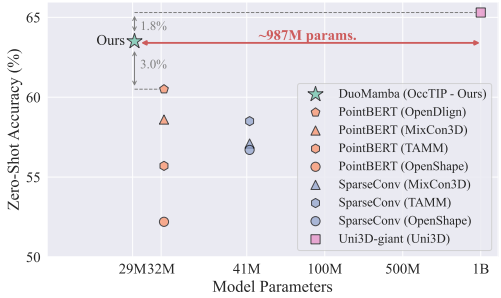

Figure 5. **Comparisons of model size and zero-shot accuracy on ScanObjectNN** [17]. Our model is pretrained on 52K ShapeNetCore [1] objects, whereas all other approaches are pretrained on an ensemble of 880K objects from four datasets: Objaverse [5], ABO [3], 3D-FUTURE [6], and ShapeNetCore [1]. Despite being pretrained on a less diverse set of objects and having the smallest size, DuoMamba demonstrates competitive performance. Among models with fewer than 50M parameters (DuoMamba, PointBERT [18], SparseConv [2]), our model outperforms all others by a significant margin of 3% in zero-shot accuracy. While Uni3D-giant [21] achieves a slightly higher accuracy with a gap of 1.8%, it comes at the cost of a substantially larger model size, with 1016.5M parameters – 35 times the size of DuoMamba. This highlights the optimal balance between model size and performance offered by our method compared to existing approaches.

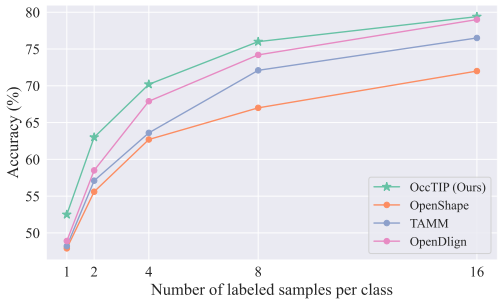

Figure 6. **Few-shot linear probing on ScanObjectNN** [17]. Our method is pretrained on 52K ShapeNetCore [1] objects, whereas other models are pretrained on 880K objects. Despite using significantly less data, our framework OccTIP outperforms all existing methods across all few-shot settings, demonstrating the data efficiency and the high-quality latent space learned by our approach.

11. Additional Quantitative Results

Evaluate Pretrained DuoMamba on ModelNet40. To evaluate DuoMamba (pretrained with OccTIP) on complete point clouds, we generate partial point clouds from 12 views (as in pretraining) and use majority voting for class prediction. Figure 7 shows that on ModelNet40, we perform competitively with previous works pretrained on full point clouds and even **surpass OpenShape** by 1.3%.

Complete Results for Zero-Shot 3D Object Detection. The average precision (AP) for each class and the mean Average Precision (mAP) for the zero-shot 3D object detection experiments (Section 6.4 in the main paper) are provided in Table 7 (for ScanNetV2 [4] benchmark) and Table 8 (for

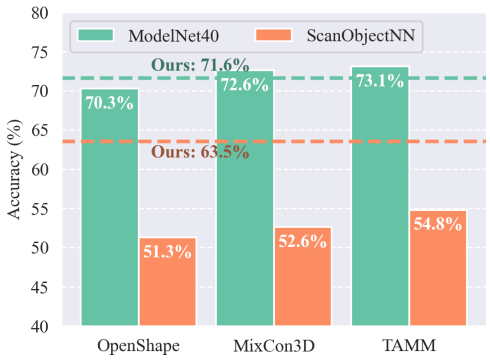

Figure 7. Comparison with methods pretrained on *complete* point clouds.

SUN RGB-D [15] benchmark). Our method OccTIP consistently achieves the best or second-best AP across most categories and achieves the highest mAP, with a significant margin over existing techniques on both datasets. These results highlight the effectiveness of OccTIP and its applicability to complex, real-world recognition tasks.

Pretraining with Complete vs. Partial Point Clouds. Table 9 shows that our synthetic partial data consistently improves all models' accuracy on real-world ScanObjectNN, with DuoMamba performing best in both settings.

Pretraining data	SparseConv	PointBERT	DuoMamba
Complete	56.0	55.5	**57.5**
Partial (OccTIP)	61.7 (+5.7)	60.6 (+5.1)	**63.5 (+6.0)**

Table 9. ScanObjectNN accuracy when pretraining with full vs partial data.

Architecture Influence on Object Detection Performance. Table 10 compares object detection performance of DuoMamba and PointBERT pretrained with OccTIP against PointBERT's best performance by previous pretraining baselines. OccTIP consistently enhances PointBERT's performance, and its combination with DuoMamba achieves the best results.

Pretraining method	Encoder	ScanNetV2		SUN RGB-D	
		mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
Best current	PointBERT	24.1	19.1	18.9	10.0
OccTIP		25.4	19.3	21.9	11.7
OccTIP	DuoMamba	**28.9**	**22.7**	**24.4**	**13.0**

Table 10. Detection results of different models and pretraining methods.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1, 2

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2

[3] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21126–21136, 2022. 1, 2

[4] 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*, pages 5828–5839, 2017. 2

[5] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 1, 2

[6] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 1, 2

[7] Yipeng Gao, Zeyu Wang, Wei-Shi Zheng, Cihang Xie, and Yuyin Zhou. Sculpting holistic 3d representation in contrastive language-image-3d pre-training. In *CVPR*, 2024. 1, 2

[8] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. Mvtn: Multi-view transformation network for 3d shape recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2021. 1

[9] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22157–22167, 2023. 2

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1

[11] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

[13] Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning*, pages 18559–18575. PMLR, 2022. 1

[14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[15] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 2, 3

[16] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1

[17] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1588–1597, 2019. 1, 2

[18] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 2

[19] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8552–8562, 2022. 2

[20] Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21413–21423, 2024. 2

[21] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[22] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, 2024. 1

[23] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2639–2650, 2023. 2