UniPre3D: Unified Pre-training of 3D Point Cloud Models with Cross-Modal Gaussian Splatting

Supplementary Material

1. Additional Experiments

1.1. Object Detection Fine-tuning

For the scene-level object detection task, we leverage UniPre3D to pre-train the backbone of the classical VoteNet [6] using the ScanNet20 [3] dataset. Subsequently, we fine-tune the model for the ScanNet20 detection task within the MMDetection3D [2] framework. As shown in Table 1, our UniPre3D significantly outperforms prior methods, achieving state-of-the-art performance and delivering substantial improvements, particularly on the challenging mAP@50 metric. These results strongly validate our claim in the main paper that UniPre3D serves as an efficient and effective unified 3D pre-training approach.

Table 1. **Object detection on the scene-level ScanNet20 [3].** We report the mean average precision on the validation set.

Model	Pre-train	mAP@50	mAP@25
VoteNet [6]	✗	33.5	58.6
	RandomRooms [7]	36.2	61.3
	PointContrast [8]	38.0	59.2
	PC-FractalDB [9]	38.3	61.9
	STRL [5]	38.4	59.5
	DepthContrast [11]	39.1	62.1
	IAE [10]	39.8	61.5
	Ponder-RGBD [4]	41.0	63.6
	UniPre3D	**43.3**	**64.0**

1.2. Ablations on Reference Views

We conduct additional ablation studies to evaluate the impact of reference view selection strategies and the number of reference views in scene-level experiments. For these ablations, we pre-train the SparseUNet [1] model on the ScanNet20 [3] dataset and fine-tune it on the downstream semantic segmentation task on the ScanNet dataset. The mean Intersection over Union (mIoU) results on the validation set are presented in Table 2.

Table 2. **Ablation studies on reference view selection and number choices.** We report the PSNR metric for the pre-training stage and mean IoU for the semantic segmentation fine-tuning task on the ScanNet20 [3] dataset. The backbone is SparseUNet [1].

Reference View		Metric Results	
Restrict	Number	Pre-train PSNR	ScanNet20 mIoU
✗	8	16.80	75.04
✓	8	**16.82**	**75.76**
✓	2	16.81	75.37
✓	4	16.80	74.95
✓	8	**16.82**	**75.76**
✓	12	16.71	75.18

Reference View Restriction. In the main paper, we discuss imposing a restriction on the perspective gap between reference and rendered images. To explicitly analyze its necessity, we present results in the first two rows of Table 2. For the experiments without this restriction, reference and rendering view angles are randomly selected across the scene. The quantitative results demonstrate that applying the restriction enhances both pre-training effectiveness and fine-tuning performance. This improvement can be attributed to the fact that, without the restriction, the supplementary image information becomes too weak and irrelevant, failing to appropriately balance the pre-training task complexity.

Number of Reference Views. In our implementation, we select eight reference views to provide supplementary texture and color information from the pre-trained image model. In this ablation study, we examine the effect of varying the number of reference views. Specifically, we conduct experiments with 2, 4, 8, and 12 reference views, with quantitative results indicating that eight is the optimal choice. These findings suggest that the supplementary information should neither be too sparse nor too dense. When the number of reference views is too low, the pre-training task remains overly complex. Conversely, when too many reference views are used, the pre-training task becomes overly simplistic, limiting the ability of the backbone model to learn effectively.

2. Supplementary Visualizations

Figures 1 and 2 present additional visualization results from the pre-training stage. For each object sample, we provide the original point cloud alongside one reference image, while for each scene sample, we include the original point cloud and two reference images. The rendered outputs comprise multiple images from varying perspectives to comprehensively illustrate the predicted Gaussian primitives. The object samples highlight how color information from a single view is effectively propagated to other views through the learned geometric structures. The scene samples demonstrate that the backbone model successfully captures complex geometric relationships during pre-training, although some details remain blurred due to the limited number of Gaussian primitives.

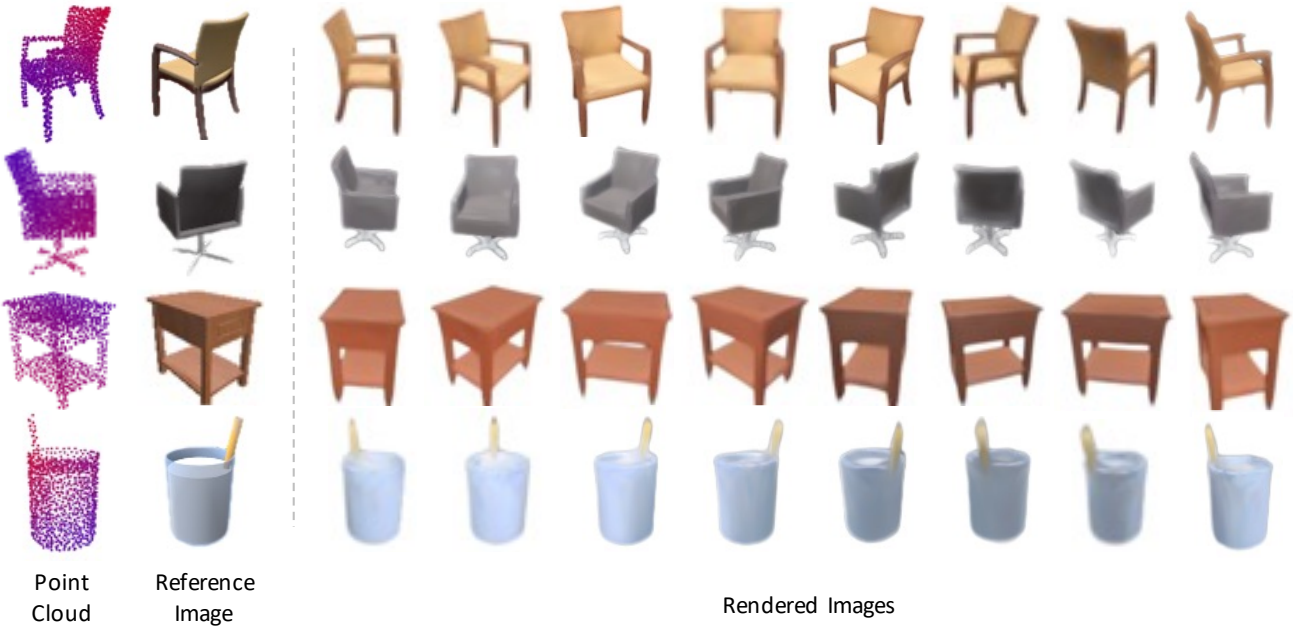

Figure 1. **Visualization of UniPre3D pre-training outputs on object-level experiments.** The first column presents the input point clouds, followed by the reference view images in the second column. The remaining rows display the rendered images.

Figure 2. **Visualization of UniPre3D pre-training outputs on scene-level experiments.** The first column presents the input point clouds, followed by the reference view images in the second and third columns. The remaining columns display the rendered images (upper rows) and their ground truths (lower rows).

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 1

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[4] Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, and Wanli Ouyang. Ponder: Point cloud pre-training via neural rendering. In *ICCV*, 2023. 1

[5] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *ICCV*, 2021. 1

[6] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 1

[7] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *ICCV*, 2021. 1

[8] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 1

[9] Ryosuke Yamada, Hirokatsu Kataoka, Naoya Chiba, Yukiyasu Domae, and Tetsuya Ogata. Point cloud pre-training with natural 3d structures. In *CVPR*, 2022. 1

[10] Siming Yan, Zhenpei Yang, Haoxiang Li, Chen Song, Li Guan, Hao Kang, Gang Hua, and Qixing Huang. Implicit autoencoder for point-cloud self-supervised representation learning. In *ICCV*, 2023. 1

[11] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *ICCV*, 2021. 1