Supplementary Materials for Sculpting Holistic 3D Representation in Contrastive Language-Image-3D Pre-training

Yipeng Gao[1] Zeyu Wang[2] Wei-Shi Zheng[1] Cihang Xie[2] Yuyin Zhou[2]
[1]Sun Yat-sen University [2]University of California, Santa Cruz

S1. Appendix

S1.1. Details of the 3D encoder

The tokenization of point cloud. We follow Yu et al. [4] to partition the points into 512 point groups (sub-clouds), with a sub-cloud containing precisely 32 points. Then, a mini-PointNet [2] is adopted to project those sub-clouds into point embeddings.

PointBERT backbone. Following OpenShape [1], we scale up the Point-BERT [4] model. The hyperparameters for scaling up are shown in Table S1.

S1.2. More details and experimental results of the strong baseline

We begin with the baseline developed by Liu et al. [1] and use the "Ensemble" dataset for training. The temperature parameters for scaling the logit in two contrastive losses are unified, and the batchsize is 200. To simplify the analysis, we don't use the "Hard Negative Mining" method utilized by OpenShape.

The setting of temperature for the contrastive loss. The temperature controls the range of logits in the softmax function used in the contrastive loss [3]. We first follow the CLIP, which initializes the learnable temperature parameter to 14.28 and clamps the value if it exceeds 100. In the image-text-3D alignment paradigm, the point cloud encoder is trained to align image and text modalities simultaneously. Intuitively, different modalities may have separately appropriate logit ranges. To this end, we verify the effect of temperature settings (a *unified* one used by two losses or two *separate* ones, each used by a loss) in Table S2 and choose "Clamp+Separate" by default.

The effect of batchsize for different model sizes . We systematically investigate the effect of batchsize across model sizes in Table S3. From the results, increasing the model size or batchsize can obtain a better performance on Objaverse-LVIS whose distribution matches the training set very well. However, the results of the other two datasets are barely satisfactory, indicating the model's generalization ability trained by "Ensemble" dataset still has much room to improve. Considering a good trade-off between datasets and training efficiency, we use a medium model size of "25.9M" and batchsize of 2k for all the following ablation studies by default.

Hyperparameter analysis of EMA decay rate . We analyze the effect of the decay rate used in the Exponential-Moving-Average (EMA) update. From the results shown in Table S4, choosing the decay rate from the range "0.999" to "0.9999" all yield promising results. Based on the results from three datasets, we choose "0.9995" by default.

Training stability. Empirically, we observe that the model's test performance on the ScanObjectNN benchmark is unstable during training on the Objaverse dataset (the blue curve in Figure S1). Our improved baseline (the red curve) can significantly alleviate the training instability. Meanwhile, our proposed MixCon3D further boosts the performance for both the Objaverse-LVIS and ScanObjectNN.

S1.3. Additional Experimental Results

Full results of g^{MV} and view amounts. We list the full results of using various types of g^{MV}, and view amounts in Table S6 and Table S7. Using simple view-pooling as g^{MV} obtains consistent improvement across three datasets. Adding an additional FC layer after the view-pooling or max pooling can enhance the Objaverse-LVIS performance while degrading the generalization ability on ScanObjectNN and ModelNet40. Given the availability of the image modality in the Objaverse-LVIS testing scenario, an increase in the number of views during the training phase yields a consistent enhancement in performance. However, this increment marginally impairs the efficacy of the ScanObjectNN and ModelNet40.

Table S1. Hyperparameters for scaling up PointBERT [4].

# Parameters	# Layers	Width	# Heads	MLP Dim	# Patches	Patch Embed Dim
13.3M	6	512	8	1024	64	128
25.9M	12	512	8	1024	128	128
32.3M	12	512	8	1536	384	256

Table S2. The analysis of settings of temperature for the constrastive losses.

Clamp	Temperature Setting	Objaverse-LVIS				ScanObjectNN				ModelNet40			
		Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
✗	Unified	46.5	34.0	69.0	76.8	52.0	53.2	77.5	87.5	84.2	**84.9**	95.9	97.4
✓	Unified	46.5	34.1	69.0	76.8	52.2	53.3	77.5	**87.7**	**84.4**	**84.9**	96.0	**97.6**
✗	Separate	46.4	34.0	69.0	76.8	52.5	53.7	77.2	87.2	84.3	84.6	96.0	97.5
✓	Separate	**46.8**	**34.4**	**69.2**	**77.1**	**52.8**	**54.0**	**77.6**	87.4	**84.4**	84.6	**96.1**	97.4

A unified view of multi-modal inference We analyze various multi-modal feature ensemble methods under four views ($M = 4$) in the main text. To further analyze the combined impact of view amount and inference schemes, we perform in-depth analysis in Table S5, including individual modality inference (point cloud y_i^P and image y_i^P) and modality ensemble inference (using g^{MV} to obtain y_i^{3D} and $y_i^P + y_i^I$). From the results, the ensemble scheme of the point cloud and the image modalities significantly improves performance. Moreover, benefitting from the large-scale pretrained CLIP model, the $y_i^P + y_i^I$ scheme further boosts the performance on Objaverse-LVIS when using multi-view images for inference.

References

[1] 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. *NeurIPS*, 2024. 1

[2] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 1

[4] 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 *CVPR*, 2022. 1, 2

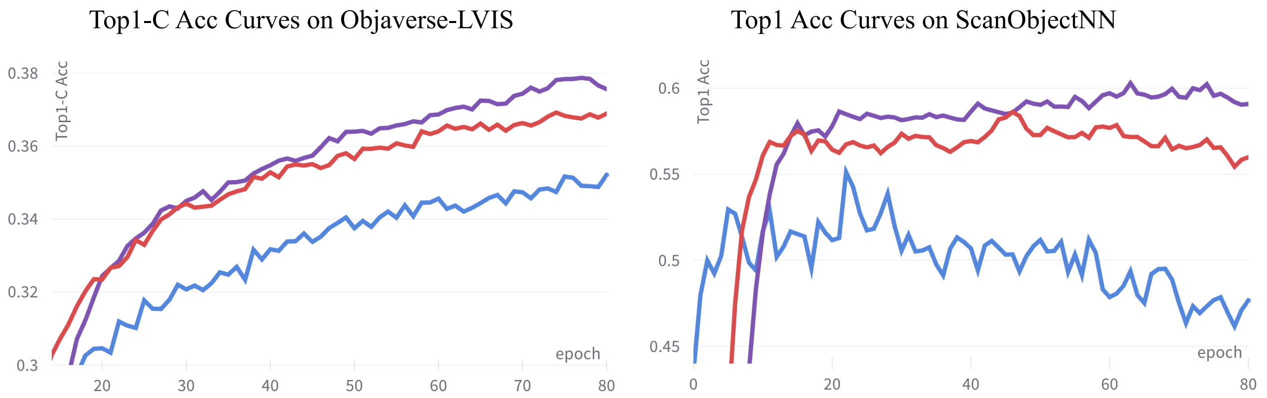

Figure S1. The zero-shot Top1 accuracy curve comparisons between the baseline, the improved strong baseline and our MixCon3D. Our improved baseline can not only perform better on the Objaverse-LVIS benchmark (the left sub-figure) but also stabilize the generalization performance (the right sub-figure).

Table S3. The analysis of batchsize across different model sizes.

Para.	Batchsize	Objaverse-LVIS				ScanObjectNN				ModelNet40			
		Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
32.3M	256	47.0	34.9	69.2	76.9	50.2	52.6	77.9	87.5	85.0	85.2	96.4	97.4
	512	47.7	36.1	69.8	77.2	49.9	52.1	75.7	85.5	**84.6**	**84.6**	95.8	97.1
	1024	49.1	36.9	70.9	78.1	52.1	**55.0**	76.0	**85.9**	83.3	83.7	**96.8**	**98.3**
	2048	**49.6**	37.4	**71.1**	**78.3**	**53.2**	**55.0**	75.3	85.7	**84.6**	83.7	95.2	96.9
	4096	**49.6**	**37.9**	70.9	78.1	52.7	54.1	**76.1**	85.3	83.7	82.3	96.1	97.7
25.9M	256	46.8	34.4	69.2	77.1	52.8	54.0	77.6	87.4	84.4	84.6	96.1	97.4
	512	47.3	34.7	69.6	77.1	52.5	55.6	77.2	87.4	84.3	84.6	96.3	**98.1**
	1024	47.8	35.0	69.9	77.2	52.9	**56.2**	77.7	87.5	84.4	85.3	96.3	98.0
	2048	48.0	35.3	70.1	77.4	**53.5**	55.5	**78.0**	87.7	**84.8**	85.3	**96.4**	97.7
	4096	**48.5**	**35.6**	**70.4**	**77.6**	52.9	55.1	77.9	**87.8**	84.3	**85.4**	95.9	97.5
13.3M	512	45.2	33.7	66.7	74.5	**54.7**	**56.6**	77.3	87.0	83.7	83.7	94.7	96.8
	1024	45.8	34.3	67.1	**74.8**	54.2	56.4	76.3	86.4	**85.2**	**84.0**	**95.6**	**97.7**
	2048	**46.3**	**35.1**	**67.3**	**74.8**	53.1	54.4	**78.5**	**87.4**	83.5	83.4	95.4	97.5

Table S4. The analysis of settings of temperature for the constrastive losses.

Decay Rate	Objaverse-LVIS				ScanObjectNN				ModelNet40			
	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
w/o EMA	48.5	36.0	70.6	77.7	54.1	56.3	78.2	87.9	85.0	85.0	96.4	97.9
0.99	49.2	36.5	71.1	78.3	54.8	57.1	78.9	88.2	85.7	85.9	96.8	**98.4**
0.999	49.3	36.5	71.2	78.3	55.3	58.3	79.4	88.8	**86.4**	**86.3**	**96.9**	**98.4**
0.9995	49.8	36.9	**71.7**	**78.7**	**55.6**	**58.9**	**79.3**	**88.6**	86.1	86.2	96.8	98.3
0.9999	**50.1**	**37.0**	71.3	78.6	55.4	58.5	78.9	88.4	85.7	85.3	96.9	98.2
0.99999	0.3	0.1	0.5	1.1	17.1	11.2	25.2	42.6	5.3	5.4	13.5	21.6

Table S5. The ablations of inference schemes under different settings of views (M).

Inference Scheme	$M = 1$				$M = 4$				$M = 8$				$M = 12$			
	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
y_i^P	51.1	37.9	73.2	80.0	50.4	37.4	72.2	79.1	51.1	38.4	73.1	79.8	51.5	39.4	73.7	80.5
y_i^I	45.1	34.6	64.3	70.8	51.9	38.5	73.1	79.4	52.0	41.1	73.1	79.5	52.5	41.5	73.8	80.1
y_i^{3D}	**51.6**	**38.2**	**73.7**	**80.6**	52.5	38.8	74.5	81.2	52.8	39.1	74.7	81.5	53.2	39.5	75.4	82.1
$y_i^P + y_i^I$	51.2	37.8	73.1	79.6	**53.8**	**40.9**	**75.5**	**81.9**	**54.8**	**43.1**	**76.3**	**82.7**	**55.3**	**43.8**	**77.1**	**83.4**

Table S6. The analysis of variants of g^{MV}.

Function g^{MV}	Objaverse-LVIS				ScanObjectNN				ModelNet40			
	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
-	51.6	38.2	73.7	80.6	58.1	61.9	**80.3**	**89.2**	86.6	86.6	96.4	98.1
View-pooling	52.5	38.8	74.5	81.2	**58.6**	**62.3**	**80.3**	**89.2**	**86.8**	**86.8**	**96.9**	**98.3**
View-pooling + FC	**52.7**	**39.1**	**74.8**	**81.4**	52.4	54.1	75.2	86.5	84.5	84.0	95.1	96.5
Max pooling	52.1	38.4	74.1	80.4	56.7	60.0	79.3	89.1	85.9	85.6	96.9	98.1
Max pooling + FC	51.6	38.0	73.2	80.3	55.8	58.7	77.1	87.6	85.2	85.6	96.0	97.6

Table S7. Ablation studies for the amount (M) of the view.

Multi-View	Objaverse-LVIS				ScanObjectNN				ModelNet40			
	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5	Top1	Top1-C	Top3	Top5
1	51.6	38.2	73.7	80.6	58.1	61.9	**80.3**	**89.2**	86.6	86.6	96.4	98.1
2	52.3	38.9	74.1	80.0	57.0	60.5	77.8	88.0	86.2	86.7	96.2	97.8
4	52.5	38.8	74.5	81.2	**58.6**	**62.3**	**80.3**	**89.2**	**86.8**	**86.8**	**96.9**	**98.3**
8	52.7	39.3	74.7	81.7	58.1	61.7	78.9	88.5	86.2	85.5	96.8	98.1
12	**53.2**	**39.5**	**75.4**	**82.1**	54.2	56.1	77.8	86.7	83.3	83.6	95.1	96.8