Supplementary Materials
Bringing Masked Autoencoders Explicit Contrastive Properties for Point Cloud Self-Supervised Learning

Bin Ren[1,2,3], Guofeng Mei[4], Danda Pani Paudel[3], Weijie Wang[2,4], Yawei Li[5], Mengyuan Liu[6⋆], Rita Cucchiara[7], Luc Van Gool[3,5], and Nicu Sebe[2]

[1] University of Pisa, 56127, Pisa, Italy
[2] University of Trento. 38123, Trento, Italy
[3] INSAIT, Sofia University "St. Kliment Ohridski", Sofia, Bulgaria
[4] Fondazione Bruno Kessler, 38123, Trento, Italy
[5] ETH Zürich, 8092, Zürich, Switzerland
[6] Peking University, 518055, Shenzhen, China
[7] University of Modena and Reggio Emilia, 41125, Modena, Italy
bin.ren@unitn.it

1 Additional Implementation Details

Masked Point Modeling Reconstruction Loss: For the masked autoencoder (MAE) loss (*i.e.*, $\mathcal{L}_{\text{recon}}$), we use the ℓ_2 Chamfer-Distance [3] following [7]. Let $\mathcal{R} \equiv \mathrm{RP}(g_\phi^1(z_1))$ and $\mathcal{G} \equiv X$ be the reconstructed point clouds and ground truth point clouds, respectively. The reconstruction loss $\mathcal{L}_{\text{recon}}$ can be written as:

$$\mathcal{L}_{\text{recon}} = \sum \left[\frac{1}{|\mathcal{R}|} \sum_{re \in \mathcal{R}} \min_{gt \in \mathcal{G}} \|re - gt\|_2^2 + \sum_{gt \in \mathcal{G}} \min_{re \in \mathcal{R}} \|re - gt\|_2^2 \right]. \tag{1}$$

Detailed Training configurations: We also provide the detailed training recipes for both the pre-training and downstream fine-tuning of our Point-CMAE in Tab. 1. Similarly to ACT [2] or Recon [10], we adopt two kinds of augmentations (*i.e.*, Scale&Translate and Rotation) in this work for pre-training and classification on ShapeNet [1] and ScanObjectNN [11] datasets, respectively.

2 Discussions and Additional Experiments

Differences to other related work. Integrating classic CL (*i.e.*, MOCO, BYOL) and MAE for point clouds with ViTs is challenging because MAE is transformation-sensitive but CL needs well-designed transformation [10]. Though [12, 14] have explored this integration, our approach uniquely embeds CL within

⋆ Corresponding author. Email:liumengyuan@pku.edu.cn

Table 1: Training recipes for pre-training and downstream fine-tuning.

	Pre-training	Classification		Segmentation
Config	ShapeNet [1]	ScanObjectNN [11]	ModelNet [13]	ShapeNetPart [15]
optimizer	AdamW	AdamW	AdamW	AdamW
learning rate	1e-3	5e-4	5e-4	1e-4
weight decay	5e-2	5e-2	5e-2	5e-2
learning rate scheduler	cosine	cosine	cosine	cosine
training epochs	300	300	300	300
warmup epochs	10	10	10	10
batch size	128	32	32	16
drop path rate	0.1	0.1	0.1	0.1
number of points	1024	2048	1024	2048
number of point patches	64	128	64	128
point patch size	32	32	32	32
augmentation1	Scale&Trans	Scale&Trans	Scale&Trans	-
augmentation2	Rotation	Rotation	Scale&Trans	-
GPU device	1 A100 (40G)	1 A100 (40G)	1 A100 (40G)	1 A100 (40G)

MAE without relying on carefully designed heavy data augmentation, making our method both valuable and distinct. Though Point-CMAE and [12, 14] both use MAE and contrastive learning (CL), their design logic is fundamentally different: CL in [12] relies on heavy data augmentation and point matching to build contrastive pairs, while [14] models contrastive constraints through spatial consistency in unmasked areas. In contrast, Point-CMAE constructs contrastive pairs within the MAE framework itself, ensuring co-masked point patches are as close as possible at the feature level. The CL in [6] and our method differ notably. i) [6] needs different data augmentation for two views for construction contrastive pairs while ours digging out CM within MAE (by ensuring co-masked point patches are close at the feature level). ii) [6] applies CL before the decoder, while we introduce CL after it. These make our method simple and effective, without complex designs or numerous hyper-parameters.

Complex Real-World Performance. To validate the performance of the proposed Point-CMAE in a more complex level real-world setting, we fine-tune the pre-trained Point-CMAE model on the S3DIS (Aera 5) dataset, and the results are shown in Tab. 2. It shows that the proposed method also archives better performance in complex real-world experimental scenes.

Reconstruction Results. We also provide the visual results of the reconstructed point cloud via the MAE strategy.

3 Limitation and Future Works

While the proposed Point-CMAE achieves competitive results across various downstream tasks, it does not incorporate the multi-modal information commonly used in recent research. Future work should explore extending Point-CMAE to integrate additional modalities, such as images and depth maps, to potentially

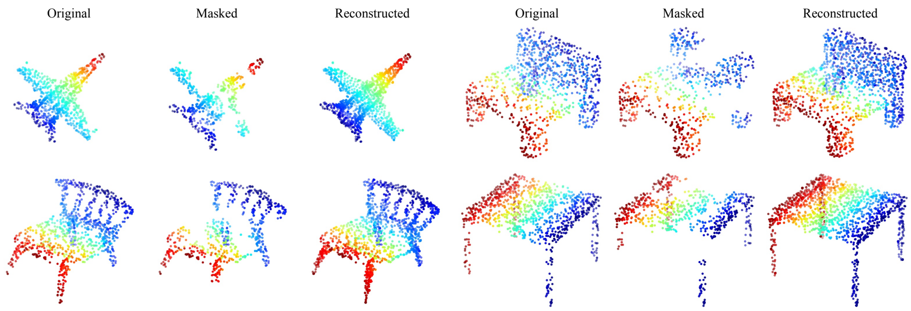

Fig. 1: The reconstructed visual results of the proposed Point-CMAE.

Table 2: Semantic segmentation results on S3DIS Area 5. We report the mean IoU(%) and mean Accuracy(%).

Methods	Pretraining	mIoU (%) ↑	mAcc (%) ↑
PointNet [8]	×	41.1	49.0
PointNet++ [9]	×	53.5	-
PointCNN [4]	×	57.3	63.9
Point-BERT [16]	✓	68.9	76.1
MaskPoint [5]	✓	68.6	74.2
Point-MAE [7]	✓	68.4	76.2
Point-CMAE (Ours)	✓	**69.8**	**77.0**

enhance its performance. Additionally, the current work employs the same mask ratio for both masks. Investigating the impact of varying the mask ratios could provide deeper insights into the invariance properties of Vision Transformers (ViTs) for point cloud representation learning. This could further refine our understanding and improve the model's robustness and generalization capabilities. Besides, exploring SSL with large-scale point clouds is an interesting related area that we would like to explore as well as one of the future directions.

4 Broader Impact

Our research introduces Point-CMAE, an innovative approach integrating contrastive learning with masked autoencoder pre-training for Vision Transformers in 3D point cloud data. This advancement enhances 3D object recognition and segmentation, improving applications like autonomous driving and robotics while promoting data efficiency, and making robust models feasible even with limited labeled data. The techniques developed can inspire innovations in other domains, fostering cross-disciplinary advancements. By releasing our code, we contribute to the open-source community, facilitating collaboration and reproducibility. Point-CMAE also serves as a valuable educational resource, while its ethical application can enhance societal benefits.

References

1. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 1, 2
2. Dong, R., Qi, Z., Zhang, L., Zhang, J., Sun, J., Ge, Z., Yi, L., Ma, K.: Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? In: The Eleventh International Conference on Learning Representations (2022) 1
3. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 605–613 (2017) 1
4. Li, Y., Bu, R., Sun, M., Wu, W., Di, X., Chen, B.: Pointcnn: Convolution on x-transformed points. Advances in neural information processing systems **31** (2018) 3
5. Liu, H., Cai, M., Lee, Y.J.: Masked discrimination for self-supervised learning on point clouds. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II. pp. 657–675. Springer (2022) 3
6. Mishra, S., Robinson, J., Chang, H., Jacobs, D., Sarna, A., Maschinot, A., Krishnan, D.: A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. arXiv preprint arXiv:2210.16870 (2022) 2
7. Pang, Y., Wang, W., Tay, F.E., Liu, W., Tian, Y., Yuan, L.: Masked autoencoders for point cloud self-supervised learning. In: European conference on computer vision. pp. 604–621. Springer (2022) 1, 3
8. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017) 3
9. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017) 3
10. Qi, Z., Dong, R., Fan, G., Ge, Z., Zhang, X., Ma, K., Yi, L.: Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In: International Conference on Machine Learning. pp. 28223–28243. PMLR (2023) 1
11. Uy, M.A., Pham, Q.H., Hua, B.S., Nguyen, T., Yeung, S.K.: 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. pp. 1588–1597 (2019) 1, 2
12. Wu, X., Wen, X., Liu, X., Zhao, H.: Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9415–9424 (2023) 1, 2
13. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1912–1920 (2015) 2
14. Xu, M., Xu, M., He, T., Ouyang, W., Wang, Y., Han, X., Qiao, Y.: Mm-3dscene: 3d scene understanding by customizing masked modeling with informative-preserved reconstruction and self-distilled consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4380–4390 (2023) 1, 2

15. Yi, L., Kim, V.G., Ceylan, D., Shen, I.C., Yan, M., Su, H., Lu, C., Huang, Q., Sheffer, A., Guibas, L.: A scalable active framework for region annotation in 3d shape collections. ACM Transactions on Graphics (ToG) **35**(6), 1–12 (2016) 2
16. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J., Lu, J.: 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. pp. 19313–19322 (2022) 3