Masked Scene Contrast: A Scalable Framework for Unsupervised 3D Representation Learning – Appendix

Xiaoyang Wu Xin Wen Xihui Liu Hengshuang Zhao
The University of Hong Kong
https://github.com/Pointcept/Pointcept

A. Implementation Details

This section introduces the implementation details of our proposed *Masked Scene Contrast* (MSC), which is crucial to making these novel designs work.

A.1. Backbone Architecture

We adopt *SparseUNet* [2], which is widely applied by previous works as ablation studies and result comparisons. *SparseUNet* adopt a U-Net style architecture, and the config details follow previous works [5–7]. The main config is available in Table 1, and the name of the backbone is marked in gray.

Config	Value
backbone	SparseUNet34
patch embed depth	1
patch embed channels	32
patch embed kernel size	5
encode depths	[2, 3, 4, 6]
encode channels	[32, 64, 128, 256]
encode kernel size	3
decode depths	[2, 2, 2, 2]
decode channels	[256, 128, 64, 64]
decode kernel size	3
pooling stride	[2, 2, 2, 2]

Table 1. **Backbone setting.**

A.2. View Generation Pipeline.

The specific constitution of our generation pipeline is concluded in Table 2. For a given point cloud input, we first dedicate two copies of the original point cloud for separated random view generation. Then we apply the augmentation sequence in Table 2 to produce differentiated views of a single scene. The original coordinates (w/o rotation) are saved for both views, and both grid sampling and point matching are performed on this original coordinate system. Spatial augmentations, photometric augmentations, and sampling augmentations are marked in green, yellow and blue.

Spatial augmentations. We simulate different orientations of point cloud scenes by randomly rotating around the z-axis. Slight rotations around the x-axis and y-axis are also applied to simulate the unavoidable slope of the ground. Additional random flipping also adds geometric diversity to objects in the scenes and is thus also applied.

Photometric augmentations. Our photometric augmentations contain brightness, contrast, saturation, and hue adjusting from 2D images to 3D point clouds. These augmentations enhance the chromatic augmentations scheme introduced by Choy et al. [2] three years ago, and we hope these advanced photometric augmentations can also benefit future works. As for the augmentation parameters, we follow BYOL [4], a reputed unsupervised representation learning framework for 2D images. We shrink the boundary of hue adjustment since the hue diversity of 3D indoor scenes is limited compared with image datasets. These stochastic photometric augmentations can effectively simulate diverse light conditions. A visualization of these augmentations is available in Figure 1.

Augmentation	Value
random rotate	angle=[-1, 1], axis='z', p=1
random rotate	angle=[-1/64, 1/64], axis='x', p=1
random rotate	angle=[-1/64, 1/64], axis='y', p=1
random flip	p=0.5
random coord jitter	sigma=0.005, clip=0.02
random color brightness jitter	ratio=0.4, p=0.8
random color contrast jitter	ratio=0.4, p=0.8
random color saturation jitter	ratio=0.2, p=0.8
random color hue jitter	ratio=0.02, p=0.8
random color gaussian jitter	std=0.05, p=0.95
grid sample	grid size=0.02
random crop	ratio=0.6
center shift	n/a
color normalze	n/a

Table 2. **View generation pipeline.**

Sampling augmentations. Grid sampling is a necessary process that both reduces point redundancy and increases data diversity. Combined with random rotation, the grid sampling is applied to different grids and points from the original point cloud, which adds to the data diversity. Further, random cropping is also applied to simulate the occlu-

Config	Value
optimizer	SGD
scheduler	cosine decay
learning rate	0.1
weight decay	1e-4
optimizer momentum	0.8
batch size	32
datasets	ScanNet, ArkitScene
warmup epochs	6
epochs	600

Table 3. **Pre-training setting.**

Config	Value
optimizer	SGD
scheduler	cosine decay
learning rate	0.05
weight decay	1e-4
optimizer momentum	0.9
batch size	48
warmup epochs	40
epochs	800

Table 4. **Fine-tuning setting.**

sion relationship and enforce the model to differentiate the visible region of contrastive views, which is also an important component.

A.3. Training Setting.

Pre-training. The default setting is in Table 3. We only utilize ScanNet point cloud scene data for efficient pre-training. And we adopt both ScanNet [3] and ArkitScene [1] for large-scale pretraining.

Fine-tuning. The default setting for fine-tuning on ScanNet semantic segmentation is in Table 4. It is worth noting that good fine-tuning results rely on higher batch size. And the conclusion holds for most of our experimented downstream tasks. We use the same setting proposed by CSC [5] and adopt 48 as the fine-tuning batch size for downstream tasks.

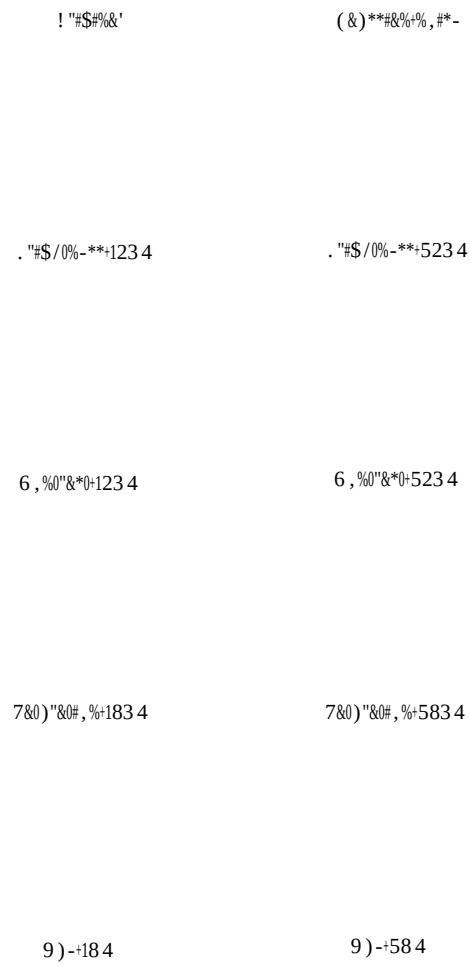

Figure 1. **Photometric augmentation.**

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS Workshops*, 2021. 2

[2] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 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. 2

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 1

[5] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 1, 2

[6] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. In *NeurIPS*, 2022. 1

[7] 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