GeoMAE: Masked Geometric Target Prediction for Self-supervised Point Cloud Pre-Training Supplementary Materials

Xiaoyu Tian[1] Haoxi Ran[2] Yue Wang[3] Hang Zhao[1*]

[1]IIIS, Tsinghua University [2]CMU [3]Nvidia

A. Implementation Details

We reproduce four previous self-supervised learning methods, including two contrastive learning methods tailored to point clouds (PointContrast [6] and STRL [4]), as well as two general methods (BYOL [3] and SwAV [2]).

General Configurations. We adopt the standard SST as the backbone. For the Waymo Open Dataset [5], the point cloud range is set to [-74.88m, 74.88m] for X-axes and Y-axes, [-2m, 4m] for Z-axes, and the voxel size is set to (0.32m, 0.32m, 6m). For nuScenes Dataset [1], the point cloud range is set to [-51.2m, 51.2m] for X-axes and Y-axes, [-5m, 3m] for Z-axes, and the voxel size is set to (0.256m, 0.256m, 8m). For all the methods, the pretaining learning rate is initialized as 1e-5, and the fine-tuning learning rate is initialized as 1e-4. We use the adam optimizer and the cosine annealing learning scheme. The models are trained with the batch size 64.

PointContrast. We first transform the original point cloud into two augmented views by random geometric transformations, which include random flip, random scaling with a scale factor sampled uniformly from [0.95, 1.05] and random rotation around vertical yaw axis by an angle between [-15, 15] degrees. The scenes will be passed through the SST backbone to obtain voxel-wise features. We randomly select half of the voxel features and then embed them into latent space by using a two-layer MLP (with BatchNorm and ReLU, and the dimensions are 128, 64). The latent space feature will be concatenated with initial features and passed through a one-layer MLP with dimension 64. The concatenated features are used for comparative learning as in the original PointContrast.

BYOL. BYOL consists of two networks, an online network and a target network. It iteratively bootstraps the outputs of the target network to serve as targets without using negative pairs. We train its online network to predict the target network's representation of the other augmented view of the same 3D scene. We pass the voxel-wise features through a two-layer MLP (with dimensions 512, 2048). After that, a two-layer MLP (with dimensions 4096, 256) predictor in the online network will project the embeddings into a latent space as the final representation of the online network. The target network is updated by a slow-moving averaging of the online network with parameter 0.999. For other configurations, we follow the settings in the original paper.

SwAV. Different from contrastive learning methods, SwAV does not directly compare embedding features by introducing prototypes and swapped predictions. Similar to the implementation of PointContrast, we apply the same view generation module and obtain voxel-vise features of different views. We adopt a two-layer MLP projection head with dimensions 512 and 128. We then compute "codes" by assigning features to prototype vectors. Note that we do not adopt multi-crop strategy proposed in the original paper due to the differences between images and point clouds.

STRL. STRL learns invariant representations from two augmented views, which are obtained by spatial aumentation and temporal sampling. For spatial data augmentation, we adopt the same generation approach in PointContrast. For temporal sampling, we follow the settings in the original paper. We add a max-pooling layer at the end of the backbone to obtain the global features. The global features are passed through a projector and a predictor for contrastive learning.

B. Visualizations of Centroid Prediction

In this section, we provide examples of a centroid prediction visualization on the nuScenes validation set. Figure. 1 shows the original point clouds. Figure. 2, 4 and 6 show the centroid of points inside each non-empty voxels in the bottom, middle, and top level separately. The blue dots stand for the visible (unmasked) centroids and the gray dots stand for the masked ones (we only predict and supervise the masked centroids). Figure. 3, 5 and 7 show the centroid prediction outputs of our GeoMAE in the bottom,

*Corresponding to: hangzhao@mail.tsinghua.edu.cn

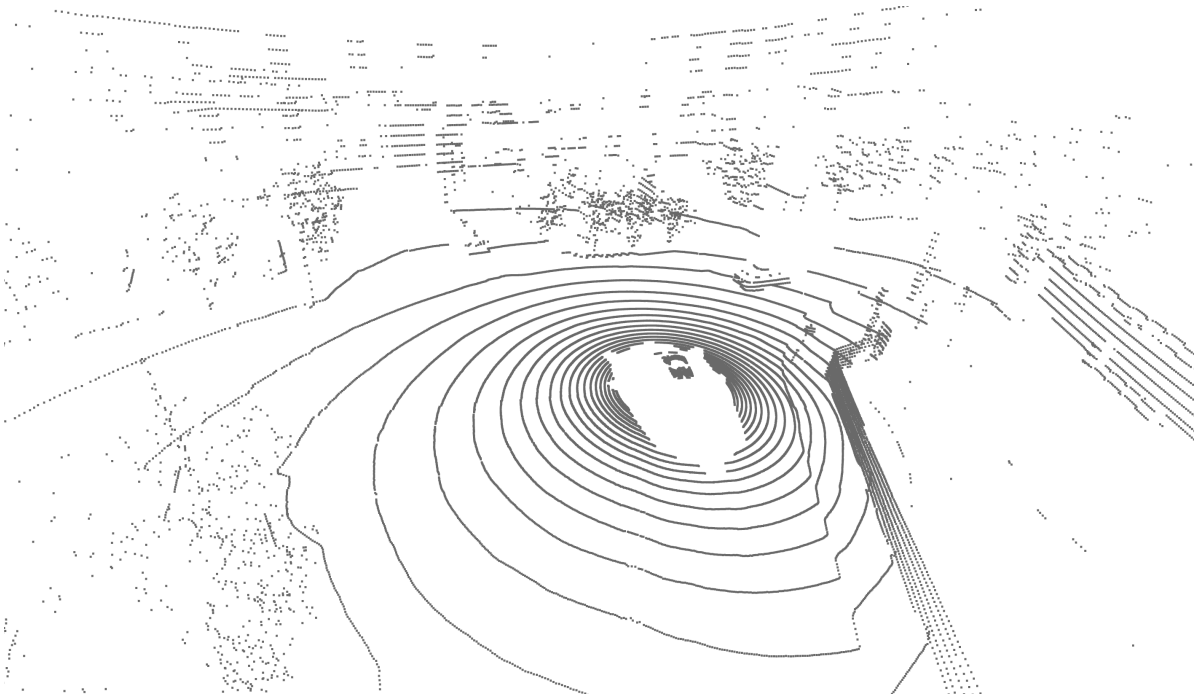

Figure 1. An original point cloud scene in the nuScenes validation set.

middle, and top level respectively. These figures show that compared to the middle and top levels, the GeoMAE reconstructs more accurate centroids in the bottom level. For instance, the gray dots in the blue boxes are masked voxel centroids of point clouds whose shape is an occluded vehicle. The centroids we predicted in Figure. 3 (bottom level) are closer to the ground truth than those in Figure. 7 (top level), especially in the height dimension. This is because the bottom level has a more fine-grained voxel sub-division (144 sub-voxels) than the top level (1 voxel). All those visualizations also indicate the essentials of the pyramid prediction strategy, which encourages the model to capture the geometry features coarse to fine.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1

[3] 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. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1

[4] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 1

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1

[6] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pages 574–591. Springer, 2020. 1

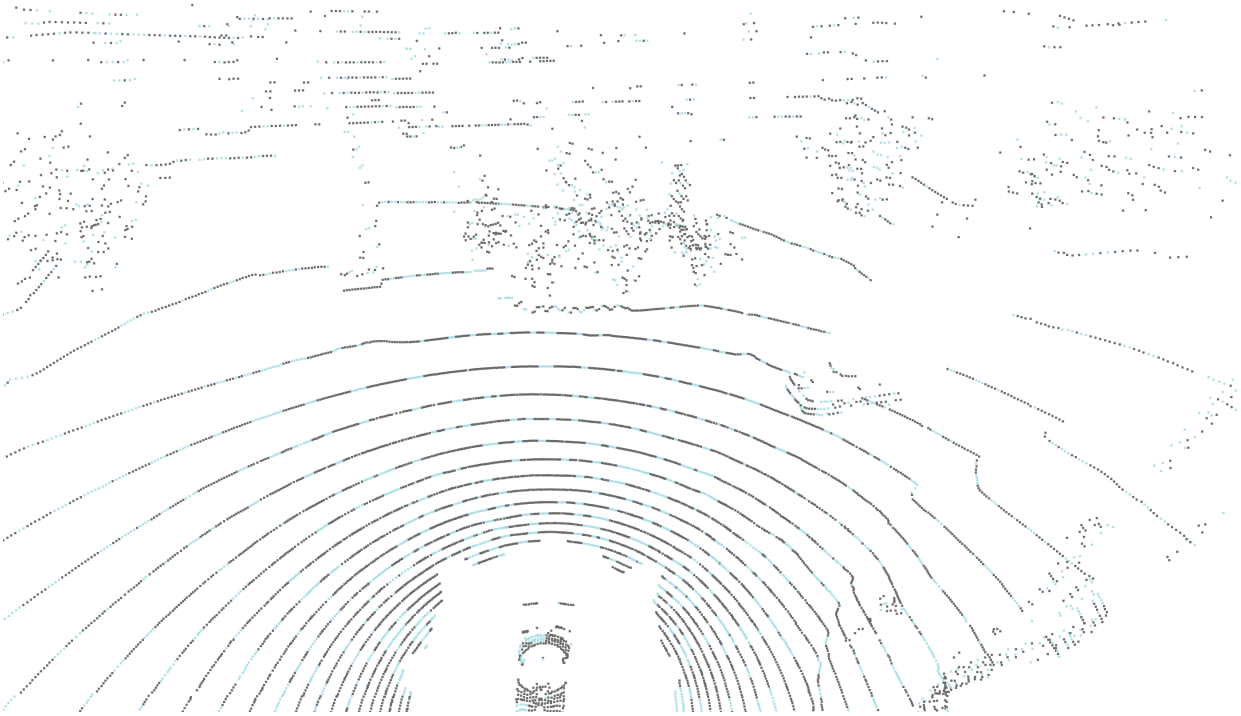

Figure 2. Voxel centroids in the bottom level. Blue dots stand for the visible (unmasked) ones and gray dots stand for the masked ones.

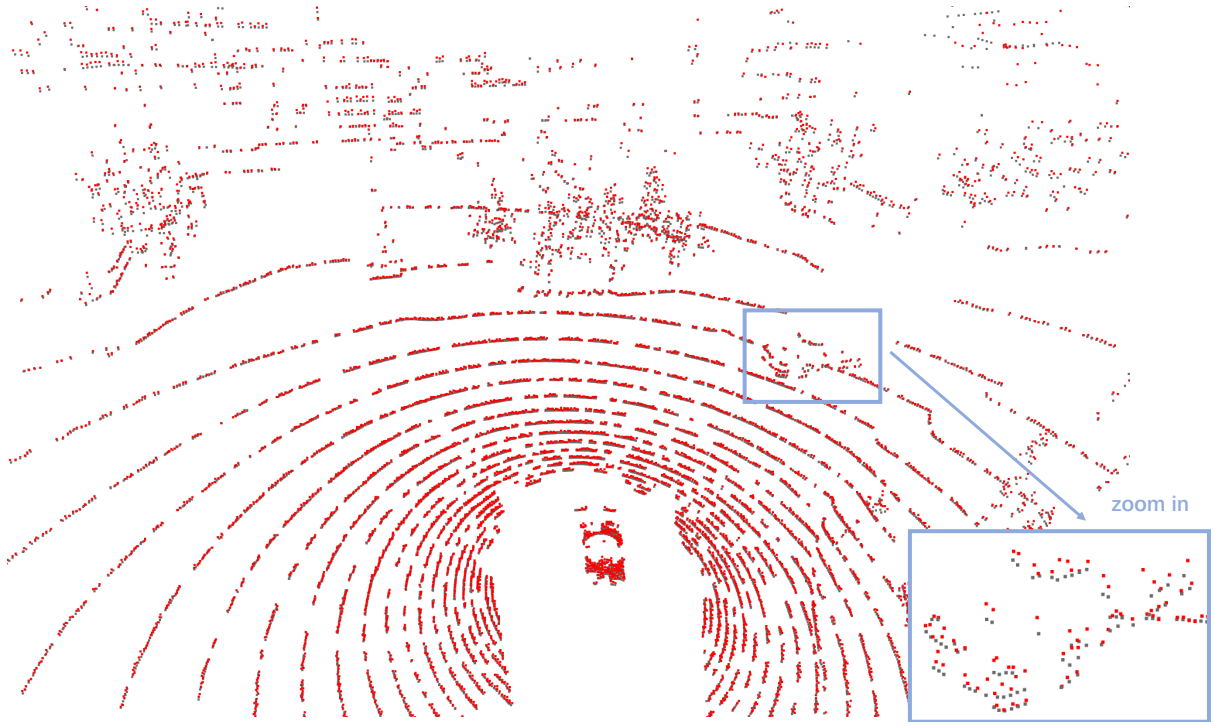

Figure 3. The prediction outputs of the masked centroids in the bottom level. Gary dots stand for the ground truth and red dots stand for the prediction results.

Figure 4. Voxel centroids in the middle level. Blue dots stand for the visible (unmasked) ones and gray dots stand for the masked ones.

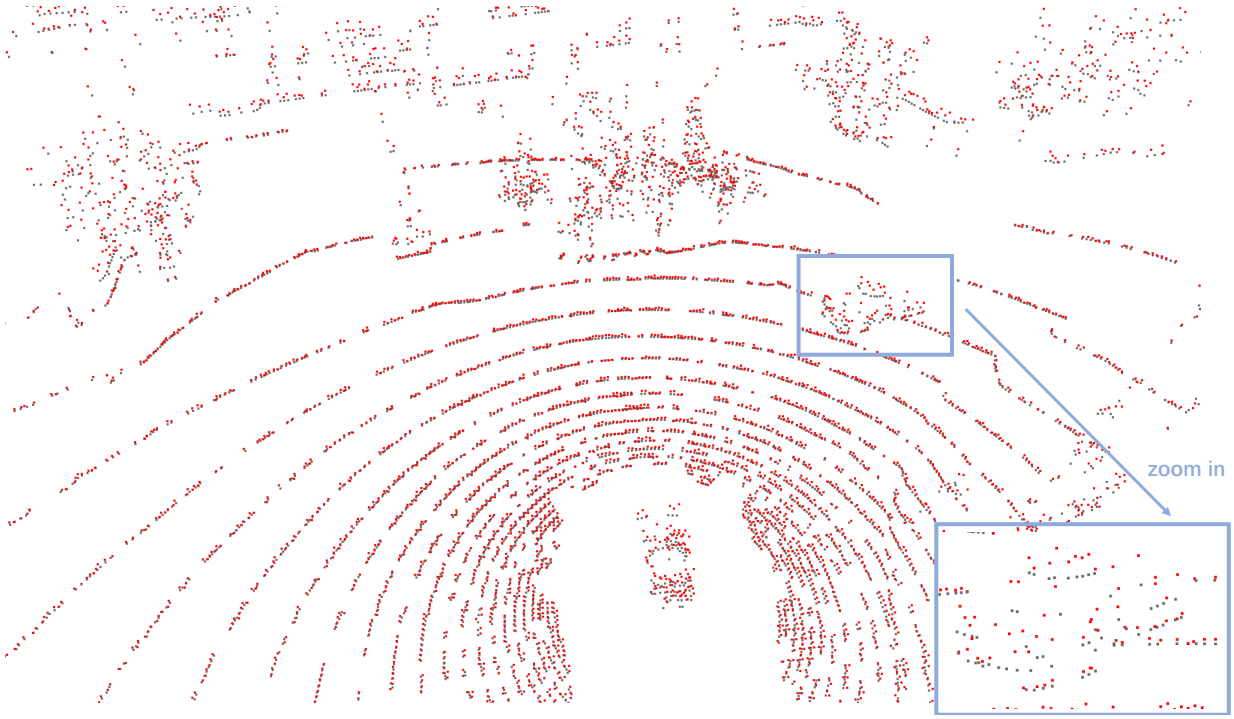

Figure 5. The prediction outputs of the masked centroids in the middle level. Gary dots stand for the ground truth and red dots stand for the prediction results.

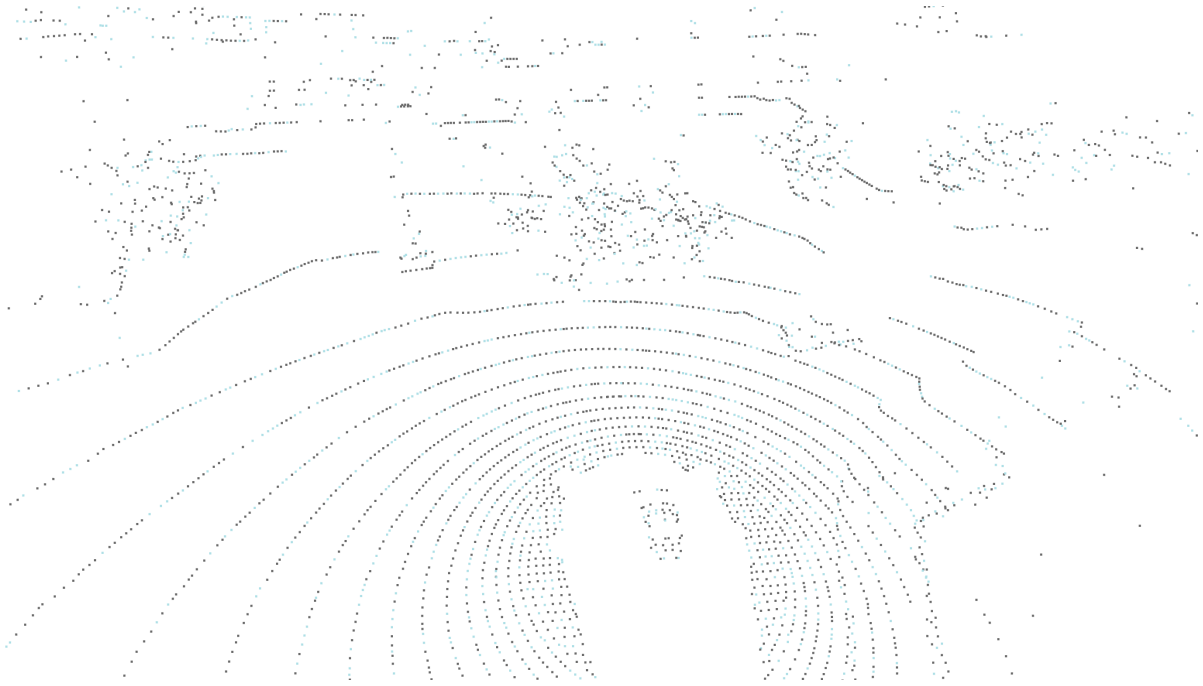

Figure 6. Voxel centroids in the top level. Blue dots stand for the visible (unmasked) ones and gray dots stand for the masked ones.

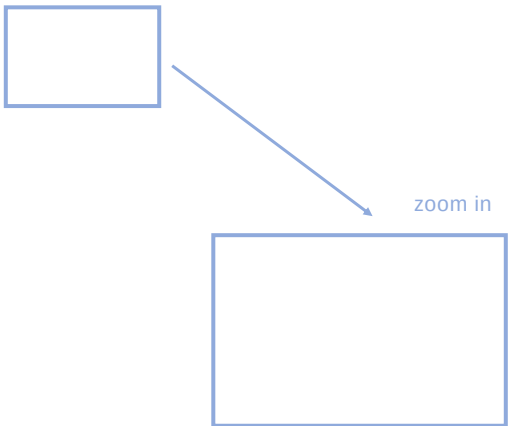

Figure 7. The prediction outputs of the masked centroids in the top level. Gary dots stand for the ground truth and red dots stand for the prediction results.