Multi-Scale Neighborhood Occupancy Masked Autoencoder for Self-Supervised Learning in LiDAR Point Clouds

Mohamed Abdelsamad[1,2] Michael Ulrich[1] Claudius Gläser[1] Abhinav Valada[2]
[1]Bosch Center for AI [2]University of Freiburg

Abstract

Masked autoencoders (MAE) have shown tremendous potential for self-supervised learning (SSL) in vision and beyond. However, point clouds from LiDARs used in automated driving are particularly challenging for MAEs since large areas of the 3D volume are empty. Consequently, existing work suffers from leaking occupancy information into the decoder and has significant computational complexity, thereby limiting the SSL pre-training to only 2D bird's eye view encoders in practice. In this work, we propose the novel neighborhood occupancy MAE (NOMAE) that overcomes the aforementioned challenges by employing masked occupancy reconstruction only in the neighborhood of non-masked voxels. We incorporate voxel masking and occupancy reconstruction at multiple scales with our proposed hierarchical mask generation technique to capture features of objects of different sizes in the point cloud. NOMAEs are extremely flexible and can be directly employed for SSL in existing 3D architectures. We perform extensive evaluations on the nuScenes and Waymo Open datasets for the downstream perception tasks of semantic segmentation and 3D object detection, comparing with both discriminative and generative SSL methods. The results demonstrate that NOMAE sets the new state-of-the-art on multiple benchmarks for multiple point cloud perception tasks.

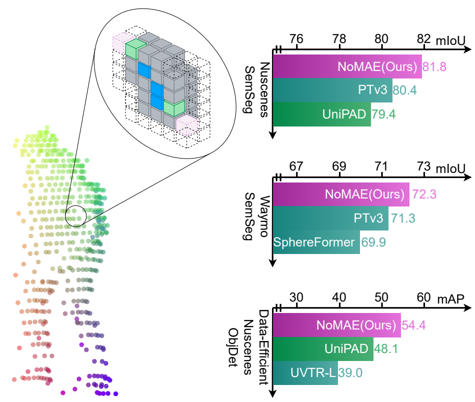


Figure 1. NOMAE enables masking and reconstructing occupancy as a self-supervised pretext task for large-scale point clouds. It limits the reconstruction of masked voxels to the neighborhood of visible voxels and reconstructs the masked occupancy at multiple scales. NOMAE achieves state-of-the-art performance on nuScenes semantic segmentation, Waymo semantic segmentation, and nuScenes object detection tasks, outperforming existing self-supervised methods as well as transformer methods.

1. Introduction

Sensors that generate point clouds, such as LiDARs or radars, have become a cornerstone in automated driving as they provide high-resolution three-dimensional representations of the environment [29]. The rich spatial information represented in point clouds enables vehicles to accurately detect and classify objects, navigate complex environments, and enhance safety through real-time situational awareness. However, annotated point cloud datasets are significantly smaller than their image-based counterparts, which makes learning large-scale perception models extremely challenging. Self-supervised learning (SSL) [2, 5–7, 14–17, 21, 22, 32], through contrastive learning or masked modeling, provides an effective solution to this problem by learning meaningful representations from vast amounts of unlabeled data. SSL also reduces the reliance on arduous annotation processes while improving performance and generalization. Pioneering works [24, 27, 45, 47] have successfully employed SSL to small-scale indoor point clouds, with more recent efforts extending it to large-scale outdoor point clouds [18, 33, 40].

Outdoor point clouds, however, pose a unique challenge for masked modeling as most of the measured 3D volume is empty space. Current approaches resort to reconstructing precise point locations within occupied voxels [18, 33, 40], but this often leaks information to the decoder, signaling that the queried voxel is occupied. Recent methods [26, 43] attempt to overcome this problem by reconstructing the entire scene, but the computational complexity and class imbalance caused by the large number of empty voxels limit these approaches to either 2D bird's-eye-view representations or coarse-grained 3D reconstructions.

In this work, we propose **N**eighborhood **O**ccupancy **MAE** (NOMAE), the first multi-scale sparse self-supervised learning framework for LiDAR point clouds that directly addresses the problem of 3D point cloud sparsity in masked modeling. The novelty in NOMAE lies in the concept that the occupancy of fine-grained 3D voxels is only evaluated (loss) in the neighborhood of visible (not masked) occupied voxels. This is motivated by the fact that LiDAR points are typically clustered in the proximity of other LiDAR points in outdoor driving scenarios. Thereby, we avoid leaking information about masked voxels to the decoder and eliminate the need for dense feature spaces or reconstructing large unoccupied areas. This makes our approach lightweight and usable with more modern sensors that have finer resolutions as well as state-of-the-art 3D transformer architectures. Our framework employs self-supervision at multiple scales, made feasible by the lightweight nature of our reconstruction task. To facilitate this, we introduce a hierarchical mask generation module that is suitable for multi-scale SSL. We perform extensive experiments with NOMAE on the competitive nuScenes [4] and Waymo Open [30] datasets that demonstrate state-of-art pretraining performance for multiple downstream tasks (illustrated in Fig. 1). Our main contributions are as follows:

- The novel *localized reconstruction* self-supervised learning framework for point clouds, **N**eighborhood **O**ccupancy **MAE**.
- A multi-scale SSL strategy, where different feature levels are supervised at different scales.
- A novel mask generating scheme suitable for multi-scale SSL.
- Extensive benchmarking on two standard autonomous driving datasets, achieving state-of-the-art results across two perception tasks.
- Comprehensive ablation studies to highlight the impact of our proposed contributions.

2. Related Work

In this section, we review the existing works related on point cloud SSL and Automotive LiDAR SSL.

3D Self-Supervised Learning: The success of generative self-supervised learning in natural language processing and computer vision has inspired several works [24, 27, 34, 46] to explore masked auto-encoders for 3D point clouds. These approaches are typically tailored towards small-scale single object recognition tasks, where a standard ViT [13] architecture suffices to encode the point cloud. For example, PointMAE [27] explicitly reconstructs point cloud patches using the Chamfer distance. MaskPoint [24] discriminates between the reconstructed points and noise points. In OcCo [34], point cloud completion is performed on occluded regions. Most prominently, Point2Vec [46] reconstructs the encoded features of a teacher model for the masked patches. Alternatively, contrastive methods can be employed with point clouds to distinguish multiple partial views [39] or point-level correspondences [19]. These pioneering works achieve promising results on object-scale and room-scale 3D point clouds but are not usable for large-scale automotive LiDAR point clouds due to their inefficient scaling with the size of the point cloud. In contrast, our work focuses on large-scale automotive LiDAR point clouds and employs efficient hierarchical architectures. Additionally, our work proposes self-supervision for multiple feature levels, contrary to the single-scale supervision employed in these works.

Automotive LiDAR Self-Supervised Learning: The focus of existing works on automotive large-scale point clouds is computational efficiency. One common technique is to reduce 3D point clouds to a 2D bird's-eye-view (BEV) grid, using pillar architectures [3, 18, 33, 40, 43]. [18] reconstructs 3D points inside masked 2D pillars and [40] additionally predicts the order of the 2D BEV pillars. GeoMAE [33] predicts centroids and 3D sub-occupancy in the pretraining. GD-MAE [43] utilizes a generative decoder to reconstruct the whole scene to alleviate the leakage of positional information to the decoder. ALSO [3] reconstructs the surface occupancy of the point cloud. UniPAD [44] is a pioneering work that generates coarse 3D features of the whole scene and uses a neural rendering approach for supervision. Occupancy-MAE [26] proposes to utilize the occupancy as a compressed representation of the point cloud and reconstruct the occupancy in the 3D space using a masked point cloud. Despite the significant performance achieved by them, reconstructing occupancy or features over the entire 3D volume is an expensive task, which allows supervision only on a coarse-scale. Furthermore, the large computational cost prohibits employing modern 3D scene understanding architectures with high voxel resolutions. As a result, the state of the art for self-supervised learning lags behind the performance of a fully supervised training of transformer architectures from scratch. In contrast, this paper addresses fully sparse SSL as a remedy, scaling well with higher 3D voxel resolutions.

3. Technical Approach

Fig. 2 presents an overview of our proposed framework for self-supervised representation learning. The network consists of an encoder (to be trained), a token upsampling module, and multiple decoders for the hierarchical masked voxel reconstruction. We employ PTv3 [37] as the encoder. In contrast to earlier works, we maintain a sparse feature space and generate fine-grained features only for the visible voxels using an upsampling module. We employ a sparse decoder for masked voxel reconstruction. This decoder is designed to be simple and lightweight, as described in Sec. 3.2, allowing us to deploy a separate decoder instance at each feature scale in the multi-scale pretext (MSP), which is further detailed in Sec. 3.3. The input point clouds are first

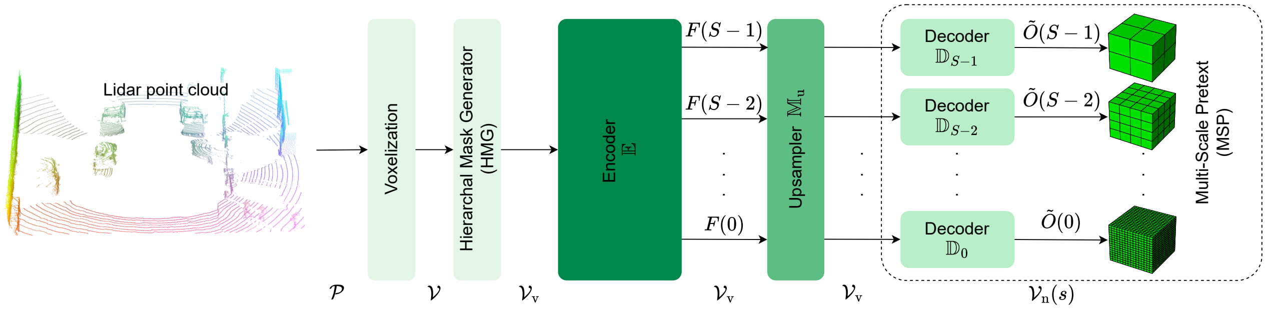


Figure 2. Overview of the proposed NOMAE approach. The input point cloud is first voxelized and masked by the hierarchal mask generator. The encoder $\mathbb{E}$ processes the visible voxels $\mathcal{V}_v$ to yield a hierarchical representation. The upsampler $\mathbb{M}_u$ then fuses the multi-scale representations to capture high-level features at each scale. For each feature scale, a separate neighboring decoder predicts occupancy in $\mathcal{V}_n$, corresponding to the immediate neighborhood of the visible voxels. The combination of independent learning tasks across multiple feature scales and the localized predictions by the neighboring decoders enables learning representations that are well-suited for 3D point clouds.

voxelized and then masked before being fed into the encoder. Our masking strategy ensures that there is adequate masking coverage while also maintaining a sufficient number of occupied voxels in the reconstructed neighborhoods at multiple hierarchical scales, as explained in Sec. 3.4.

3.1. Encoder and Token Upsampling

This input point cloud $\mathcal{P}$ is first voxelized to obtain the set of all occupied voxels $\mathcal{V}$, which is then split into a set of visible voxels $\mathcal{V}_v$ and masked voxels $\mathcal{V}_m$, as detailed in Sec. 3.4. A sparse transformer encoder $\mathbb{E}$ based on PTv3 [37] is used to encode the features V at the positions of $\mathcal{V}_v$ to generate the set of tokens

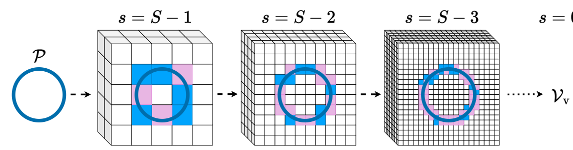
$$F(s) = \mathbb{E}(V)(s). \tag{1}$$

PTv3 employs partition-based pooling on the tokens to generate more abstract representations for coarser resolutions, similar to pooling in CNNs. $F(s)$ is the features (tokens) at the s-th scale level of PTv3 and $s \in \{0, ..., S-1\}$.

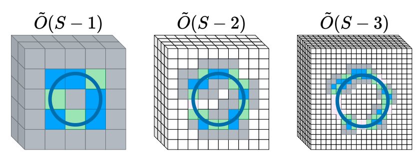
NOMAE uses an upsampling module $\mathbb{M}_u$ to propagate the abstract encoding of coarser resolution tokens to the tokens of finer resolution while keeping the representation sparse. This is similar to a feature pyramid network for CNNs. $\mathbb{M}_u$ consists of a single PTv3 transformer block at every scale, which is very lightweight.

3.2. Neighboring Decoder

Prior work on self-supervised learning for large-scale point clouds reconstruct exact locations of points [18], geometrical properties [33] or voxel ordering [40], for each masked voxel $\mathcal{V}_m$. This requires passing $\mathcal{V}_m$ to the decoder, causing information leakage. [26, 43] avoid this well-known information leakage by reconstructing the scene as a whole, which is computationally expensive. In contrast, NOMAE reconstructs the occupancy $O(v_n, s)$ of all voxels $v_n \in \mathcal{V}_n$ within a certain neighborhood n of visible voxels $\mathcal{V}_v$ at scale



(a) Hierarchical mask generation renders $\mathcal{P}$ to the coarsest scale $s = S-1$ and then applies a random mask to divide occupied voxels into visible and masked voxels. For the subsequent scales, only the visible voxels of the previous scale undergo random masking. Masked voxels of a coarser scale always correspond to masked or empty voxels at a finer scale.



(b) Multiscale pretext reconstructs the cells around visible voxels at multiple scales. The neighborhood $\mathcal{V}_n$ contains voxels that are reconstructed but empty as well as masked and recovered voxels. Some masked voxels which are too distant from visible voxels are not recovered and do not contribute to the loss. The predicted occupancy $\tilde{O}$ at coarser scales covers a wider region, while finer scales reconstruct more detail.

Figure 3. Illustration of multiscale pretext (MSP) and hierarchical mask generation (HMG).

s. To achieve this, we employ a decoder $\mathbb{D}_s$ consisting of sparse convolution layers.

$$\tilde{O}(v_n, s) = \mathbb{D}_s(\mathbb{M}_u(F)(s)), \tag{2}$$

where $\tilde{O}$ is the networks prediction of O. Visible voxels of $\mathcal{V}_v$ are excluded from $\mathcal{V}_n$. An example is depicted in Fig. 3.

We note that $\mathcal{V}_n$ will not cover masked voxels in $\mathcal{V}_m$ that are not nearby visible voxels. This is an approximation that we make in our approach and we observed that LiDAR points in outdoor point clouds typically lie in the proximity of other points on the surfaces of objects. Hence, sufficient number of masked voxels in $\mathcal{V}_m$ are included in $\mathcal{V}_n$. Faraway isolated points do not contribute to the loss, which improves performance, as evaluated in the ablation study presented in Sec. 4.6. Our interpretation is that such isolated points belong to strongly occluded or masked objects and are infeasible to reconstruct, hence affecting the pretraining. This approximation allows us to avoid reconstructing large volumes of unoccupied space without leaking information about $\mathcal{V}_\mathrm{m}$ to the decoder $\mathbb{D}_\mathrm{s}$ at the same time.

3.3. Multi-Scale Pretext

GeoMAE [33] is the only prior work that exploits multiple hierarchical scales in the reconstruction during self-supervised training. Most other works [3, 18, 26, 33, 43] use a single scale for their pretraining tasks. This is unexpected since it is common practice in automated driving perception models to attach task heads to feature representations at different scales. The intuition is that finer resolutions are more suitable for small objects, such as pedestrians, while coarser resolutions are more suitable for larger objects, such as trucks. However, the multi-scale reconstruction in GeoMAE [33] is derived from a single feature map. In our approach, we use s instances $\mathbb{D}_\mathrm{s}$ of the decoder architecture $\mathbb{D}$ with separate weights. Coupled with the neighborhood size being scale-dependent, this implicitly encourages the coarse-grained features to contain information from a larger area, while the fine-grained features contain more localized details.

3.4. Hierarchal Mask Generator

Generally, the scale of the masking can be different from the scale of reconstruction. For example, a random mask can be generated on a coarse scale and then upsampled to match the resolution of the reconstruction. However, experiments in [18, 33, 43] showed that random masking on the same scale as the reconstruction scale performs best, which is intuitive. Consequently, we generate random masks for multiple scales in the MSP. These masks should be consistent to avoid information leakage, i.e., a voxel that is masked on a coarser scale should not be visible on a finer scale [47]. A straightforward manner to generate consistent masks would be to create a random mask for the finest scale and then derive the masks of coarser scales by defining a coarse-scale voxel as masked if all its corresponding fine-scale sub-voxels are masked. However, this would lead to rapidly decreasing masking ratios when moving to coarser scales because a single visible sub-voxel is sufficient to mask a coarser-scale voxel visible. Another alternative is masking at the coarsest scale and upsampling the mask [47] to finer scales which leads to a more consistent masking ratio across scales. However, we found that this approach rapidly decreased the size of reconstructed neighborhoods n moving to finer scales.

Hence, we propose the masking scheme depicted in Fig. 3a. We mask the coarsest scale first, using a random sampling of all occupied voxels $\mathcal{V}$, and the probability that a voxel is masked equals masking ratio r. Next, we take all voxels at scale $s-1$ within visible voxels of the previous, coarser, scale s and repeat the sampling of additional masked voxels at scale $s-1$ with probability r.Consequently, the total masking ratio at scale $r_\mathrm{t}(s)$ is approximately

$$r_\mathrm{t}(s) \approx 1 - (1-r)^{S-s+1}. \quad (3)$$

By doing so, we ensure that coarser scales have a sufficient number of masked voxels without reducing the size of reconstructed neighborhoods at finer scales. Only the visible voxels $\mathcal{V}_\mathrm{v}$ of the finest scale are fed to the encoder backbone $\mathbb{E}$. An ablation study presented in Sec. 4.6 quantifies the positive effect of HMG on the pretraining and downstream task performance.

3.5. Pretrainging Loss

We use the Binary Cross Entropy loss (BCE) as the occupancy loss per scale. The final loss L is the average of all single scale losses $L(s)$:

$$L = \frac{1}{S}\sum_{s=0}^{S-1}\frac{1}{|\mathcal{V}_\mathrm{n}(s)|}\sum_{v\in\mathcal{V}_\mathrm{n}(s)}\mathrm{BCE}(\tilde{O}(v,s), O(v,s)), \quad (4)$$

with the ground truth occupancy $O(v,s) \in \{0,1\}$.

4. Experiments

In this section, we discuss the datasets, metrics, and the evaluation protocol that we use for benchmarking. We compare our proposed approach with state-of-the-art methods in the benchmarks and present extensive ablation studies to demonstrate the novelty of our contributions.

4.1. Datasets and Evaluation Metrics

The **nuScenes** dataset [4] is a challenging dataset due to the sparsity of the LiDAR point cloud. It consists of 700 driving sequences for training, 150 for validation, and 150 for testing, with annotations for a variety of tasks. We evaluate both semantic segmentation and object detection on the nuScenes dataset. We use the mean intersection over union (mIoU) as the main evaluation metric for semantic segmentation and the nuScenes detection score (NDS) and mean average precision (mAP) for 3D object detection.

The **Waymo Open Dataset** [30] is a large-scale autonomous driving dataset. It consists of 798 driving sequences for training, 202 validation sequences, and 150 test sequences. We use the mean intersection over union (mIoU) and mean accuracy (mAcc) as the main metrics for evaluating the semantic segmentation performance.

4.2. Task Heads

This section discusses the methods to evaluate the effectiveness of the pretraining and the quality of the learned representation.

Table 1. Comparison of LiDAR semantic segmentation performance on the nuScenes and Waymo Open datasets. For the first time, an SSL pretraining method outperforms strong supervised learning models. Methods marked with * are our implementation.

Method	SSL pretraining	nuScenes				Waymo		
		val mIoU	val mAcc	test mIoU	test fwIoU	val mIoU	val mAcc	test mIoU
MinkUNet [9]	-	73.3	-	-	-	65.9	76.6	69.8
SPVNAS [31]	-	-	-	77.4	89.7	-	-	68.0
Cylinder3D [49]	-	76.1	-	77.2	89.9	-	-	-
AF2S3Net [8]	-	62.2	-	78.0	88.5	-	-	-
2DPASS [41]	-	-	-	80.8	-	-	-	-
SphereFormer [20]	-	78.4	-	81.9	-	69.9	-	-
PTv2 [36]	-	80.2	-	82.6	-	70.6	80.2	-
PTv3 [37]	-	80.4	87.3	**82.7**	91.1	71.3	80.5	-
UniPAD [44]	✓	79.4	-	81.1	-	-	-	-
GEO-MAE [33]	✓	78.6	-	-	-	-	-	-
GEO-MAE [33] + PTv3 [37]*	✓	78.9	84.7	-	-	-	-	-
Occupancy-MAE [26]	✓	72.9	-	-	-	-	-	-
Occupancy-MAE [26] + PTv3 [37]*	✓	80.0	86.1	-	-	-	-	-
NOMAE + MinkUnet (ours)	✓	80.1	86.2	-	-	-	-	-
NOMAE + PTv3 (ours)	✓	**81.8**	**87.7**	82.6	**91.5**	**72.3**	**82.5**	**70.3**

Fine-tuning: In our comparisons with state-of-the-art methods, fine-tuning follows the self-supervised pre-training of the encoder. For this purpose, a task-specific head is added with randomly initialized weights, and both the encoder $\mathbb{E}$ and the task-specific head are trained using the annotated dataset. We use the same head as PTv3 [37] for the semantic segmentation tasks and the same head as our baseline UVTR [23] for the object detection task. We use layer-wise learning rate decay (LLRD) [17] to avoid forgetting the SSL representations in the encoder.

Non-linear Probing: The purpose of the ablation study is to evaluate the learned representation from our SSL approach on the downstream semantic segmentation task. Therefore, the encoder is kept frozen after pre-training, i.e., no fine-tuning. Following Probe3D [1] and the insights of earlier works [6, 17], we use a multi-scale non-linear probe (NonLP) instead of the commonly used linear probing protocol. NonLP aggregates the feature tokens of all scales after up-sampling tokens of coarser scales before passing them to a voxel-wise small MLP. NonLP avoids that the representation learning happens in the head. Still, it is probing the stronger but non-linear features, correlating better with transfer performance [6, 17]. Similar to the commonly used linear probe, NonLP is trained for a few epochs using annotated data.

4.3. Implementation Details

We perform the experiments for semantic segmentation in the Pointcept [11] framework and in the MMDetection3D [10] framework for the object detection task. We use PTv3 [37] as the encoder $\mathbb{E}$, unless stated otherwise. We use a single NVIDIA A100 GPU for the pretraining. We use $S = 4$ for the reconstruction and masking scales, corresponding to target voxel sizes of $\{0.05, 0.10, 0.20, 0.40\}$ meters. The input voxel dimension is 0.05 meters. The masking ratio r_s of the finest scale is 70% for nuScenes and 85% for Waymo. In the self-supervised pre-training, the decoder $\mathbb{D}$ consists of sparse convolution layers [12] of kernel size 5, followed by a single sparse submanifold convolution layer to generate $\tilde{O}$. We use common augmentation techniques such as rotation, scaling, and jittering from the semantic segmentation literature [9, 31, 37] during pretraining.

Table 2. Comparison of object detection performance on the nuScenes dataset with state-of-the-art point-based pre-training methods. Following the evaluation protocol of [43, 44], the methods are finetuned using 20% labeled frames, without CBGS and copy-paste augmentation.

Methods	NDS	mAP
UVTR-L (Baseline)	46.7	39.0
+ALSO [3]	48.2	41.2
+GD-MAE [43]	48.8	42.6
+Learning from 2D [25]	49.2	48.8
+UniPAD [44]	55.8	48.1
+noMAE (ours)	**60.9**	**54.4**

4.4. Benchmarking Results

In this section, we present the benchmarking results for both 3D semantic segmentation and 3D object detection.

3D Semantic Segmentation: Tab. 1 summarizes the best-performing methods on the nuScenes and Waymo Open Dataset leaderboards for the LiDAR semantic segmentation task. The results include the most performant self-supervised pretraining methods. We fine-tuned our model as described in Sec. 4.2. We observe that our proposed self-supervised pretraining achieves a mIoU score of 81.8 on the nuScenes validation set, adding 1.4 mIoU points over the baseline and setting the state-of-the-art. NOMAE is an SSL method for any architecture, MinkUnet [9] pretrained with NOMAE out-

Table 3. Comparison of different SSL methods using the same backbone with NonLP. The iteration time (iter.time) is computed using a batch size of 1. Sup.res is the finest resolution of pretraining supervision.

Model	mIoU	mACC	iter.time	sup.res
GEO-MAE [33] + PTv3 [37]	53.8	67.7	40.1ms	0.20m
Occupancy-MAE [26] + PTv3 [37]	59.0	73.4	74.0ms	0.10m
NOMAE (Ours)	**74.8**	**85.0**	**39.8ms**	**0.05m**

Table 4. Ablation study on the various components in NOMAE for semantic segmentation on the nuScenes validation set. Lines with * are with fine-tuning, and all the other results are with NonLP. For more details refer to Sec. 4.6

Model	mIoU	mACC	ACC
Occupancy-MAE + PTv3	59.0	73.4	91.0
+ reconstruct only $\mathcal{V}_n$	66.7	78.2	92.4
+ MSP			
naive masking	70.2	82.3	93.5
Point-M2AE [47] masking	70.5	82.3	93.6
+ HMG	72.6	83.9	93.7
+ $n = 9$	73.3	84.3	94.1
+ batch size 8	74.8	85.0	94.0
+ fine-tuning = noMAE*	**81.8**	**87.7**	**94.9**

of-the box achieves mIoU of 80.1 which is on par with SOTA architectures. For more results using other architectures see Supplementary Sec. 8 and Sec. 11.

The results on the nuScenes semantic segmentation test set are on par with the strong PTv3 baseline for the mIoU score and outperforms it in the frequency-weighted IoU (fwIoU) score by 0.4%. The improvement on the test set is lesser than the validation dataset because NOMAE has relatively poor performance for the minority class *bicycle*, and we did not make special adaptations for the test set submission.

On the Waymo Open dataset val set, our method achieves a mIoU score of 72.3 and mAcc of 75.2. NOMAE improves by 1.0 and 2.0 points in the mIoU and mAcc, respectively, over the baseline PTv3 [37]. The current version of the semantic segmentation challenge is relatively new for the Waymo Open dataset, with only a few submissions. With 70.3% mIoU score on the test set, NOMAE sets a new state-of-the-art on the Waymo Open Dataset single frame semantic segmentation challenge.

3D Object Detection: We present results for object detection on the nuScenes validation set using the fine-tuning approach described in Sec. 4.2. We follow the experiment setup of GD-MAE [43] and UniPAD [44], utilizing only 20% of the annotated frames during fine-tuning, without the use of CBGS [48] or Copy-and-Paste (object sample)[42] augmentation. We adopt the same training settings as the baseline UVTR [23] and do not use test-time augmentation or model ensembling.

Tab. 2 shows that our approach achieves 60.9 and 54.4 NDS and mAP scores respectively, improving by 14.2 in NDS points and 15.4 in mAP points over the UVTR-L [37] baseline, and by 5.1 NDS points and 5.6 mAP over the closest contrastive SSL method Learning-from-2D [44]. The significant improvement demonstrates the effectiveness of our proposed localized multi-scale SSL for 3D object detection with limited annotated data.

4.5. Comparison with SSL Methods

In this experiment, we compare the performance of our proposed method with the self-supervised pretraining methods of Occupancy-MAE [26] and GeoMAE [33], with the same encoder architecture of PTv3 [37]. Tab. 1 shows that our re-implementation of Occupancy-MAE [26] and GeoMAE [33] with the state-of-the-art PTv3 [37] backbones (marked with *) outperforms the results reported in the original papers by 0.3 and 7.1 mIoU points respectively for semantic segmentation with fine-tuning on the nuScenes dataset.

Tab. 3 shows the results of non-linear probing (NonLP). We observe that the NonLP performance in Tab. 3 correlates with the fine-tuning results in Tab. 1. Furthermore, the relative performance improvement of NOMAE over the baselines is higher for NonLP in comparison to fine-tuning, which indicates richer representation learning. Additionally, the time of a single training step (iteration time) is lowest for NOMAE, despite the multi-scale pretraining and a much finer resolution of the pretraining supervision. We note that the iteration time of NOMAE is 10ms for the pretraining with a single scale.

4.6. Ablation Study

In this section, we present ablation studies on the nuScenes semantic segmentation validation set to investigate the design choices of the proposed method. We performed the experiments using the pre-trained frozen encoder using NonLP. Please refer to Sec. 4.2 for further details.

Detailed Study of NOMAE: This experiment evaluates the improvement due to our proposed contributions, and the results are presented in Tab. 4. We start from our implementation of Occupancy-MAE [26] as in Tab. 3. Reconstructing only the local neighborhood $\mathcal{V}_n$ of visible voxels $\mathcal{V}_v$ increases the NonLP mIoU from 59.0% to 66.7%. This requires replacing the Occupancy-MAE decoder with our proposed upsampling module and neighborhood decoder. Adding the multi-scale reconstruction (MSP) from Sec. 3.3 further improves the mIoU to 70.1 for naive mask construction and to 70.5 for masking strategy from Point-M2AE [47], as opposed to single-scale reconstruction in Occupancy-MAE. Our proposed hierarchical mask generation from Sec. 3.4 yields an improvement of 72.46 mIoU, as further investigated in Sec. 4.6. Moreover, increasing the reconstruction neighborhood size from 5 to 9 improves the mIoU to 74.8, as further investigated in Sec. 4.6. Reducing the batch size

Table 5. Reconstruction and masking on different single scales, with multi-scale pretext (MSP) and hierarchical mask generation (HMG). Results are reported on the nuScenes val set with the encoder frozen after SSL pretraining, with nonlinear probing. IoU(**p**) and IoU(**t**) are the IoU for the classes pedestrian and truck, respectively.

Model	**mIoU**	**mACC**	**ACC**	**IoU(p)**	**IoU(t)**
Single scale $2^s = 1$	63.8	74.4	91.5	72.6	68.8
Single scale $2^s = 2$	66.7	78.2	92.4	75.6	74.3
Single scale $2^s = 4$	68.0	80.0	92.8	76.5	75.5
Single scale $2^s = 8$	68.3	80.5	93.0	75.8	76.2
Single scale $2^s = 16$	67.3	80.3	92.7	73.4	76.6
MSP $2^s \in \{1, 2, 4, 8\}$					
naive masking	70.2	82.3	93.5	83.2	79.6
Point-M2AE [47]	70.5	82.3	93.6	82.5	81.2
HMG (ours)	**72.6**	**83.9**	**93.7**	**85.2**	**83.6**

to 8 (further investigated in Sec. 6.1 of the supplementary material) yields the final NonLP performance of 74.8% and fine-tuning performance of 81.8% mIoU score.

Mask Block Size, MSP and HMG: This experiment investigates different reconstruction and mask scales s for our SSL task. Tab. 5 compares the performance of reconstructing different *single* scales with the proposed multi-scale pretext (MSP) from Sec. 3.3. The experiment of MSP without hierarchical mask generation (HMG) uses either a random mask at the finest scale, which is pooled to generate masks of coarser scales, as described in the naive solution in Sec. 3.4, or the method of Point-M2AE [47].

It can be observed that no single-scale occupancy task is suitable for all object types. For example, trucks benefit from a coarser occupancy reconstruction, while pedestrians prefer a finer resolution in the pretraining task, except for the very fine scales of $2^s \in \{1, 2\}$. MSP combines coarse and fine tasks, thereby maximizing the overall mIoU. We observe that HMG achieves an additional improvement of 2.4 and 2.1% mIoU over random mask at the finest scale and the mask generation proposed in Point-M2AE [47] respectively. This underlines the importance of proper training examples at all scales.

Neighborhood Size: Fig. 4 investigates the effect of the neighborhood size (number of voxels) to generate $\mathcal{V}_n$ from $\mathcal{V}_v$. For example, $n = 5$ indicates that a total of 5 voxels are covered in all 3 (x,y,z) dimensions for every scale. Experiments use MSP and HMG. Since every scale has a different voxel size, the reconstructed volume depends on the scale. We observe a maximum of the NonLP performance for $n = 9$. Our interpretation is that a smaller neighborhood does not cover sufficient LiDAR measurements for reconstruction, while a larger neighborhood hinders local representations in the pretraining. Further, we observe that limiting the reconstruction size performs consistently better than reconstructing the whole space, which would correspond to the method of Occupancy-MAE [26].

Masking Ratio: Fig. 5 investigates the effect of different

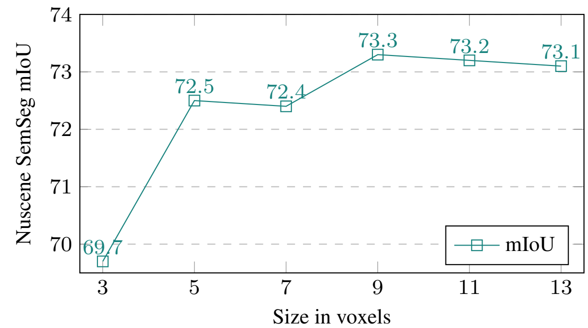


Figure 4. Size (number of voxels) of the reconstructed neighborhood n around visible voxels $\mathcal{V}_v$, to create $\mathcal{V}_n$ in the proposed pretext task. We observe that the downstream NonLP semantic segmentation peaks at $n = 9$. Note that $n \to \infty$ corresponds to the method of [26].

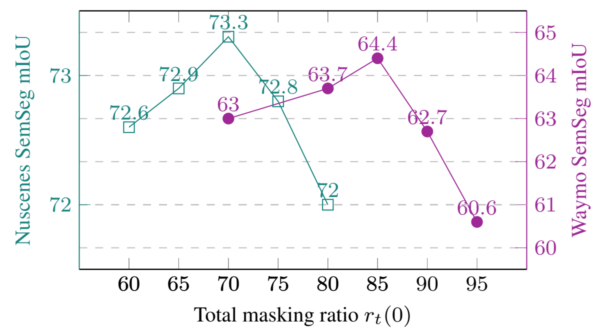


Figure 5. NonLP performance over the masking ratio r_t on the nuScenes and Waymo datasets. We observe that the optimal $r_t(0)$ is 70% for the nuScenes and 85% for the Waymo Open Dataset. Our interpretation is that Waymo requires a higher masking ratio due to the higher density of the LiDAR point cloud.

masking ratios on the quality of the representations for semantic segmentation on the Waymo and nuScenes datasets. This experiment uses a constant $r(s) = r$ to achieve the total masking ratio r_t according to Eq. 3. We observe that the optimal total masking ratio is 70% for the nuScenes and 85% for the Waymo Open dataset. This is intuitive since the point cloud density in Waymo (ca. 180k points per frame) is higher than the relatively sparse nuScenes (ca. 34k points per frame) LiDAR point clouds.

Data Efficient nuScenes: Tab. 6 analyzes the performance under limited annotated data for semantic segmentation on the nuScenes dataset. The sub-sampling of the data is performed sequence-wise, meaning that all frames of a sequence are either included or excluded. For example, 0.1% indicates that only one of the 1000 scenes of the nuScenes dataset is used, namely *scene-0392*. In every experiment, all sequences of the training set are used for the self-supervised pertaining. We observe that NOMAE (fine-tune) consistently outperforms training from scratch, which demonstrates the method's ability to benefit from unannotated data. Further-

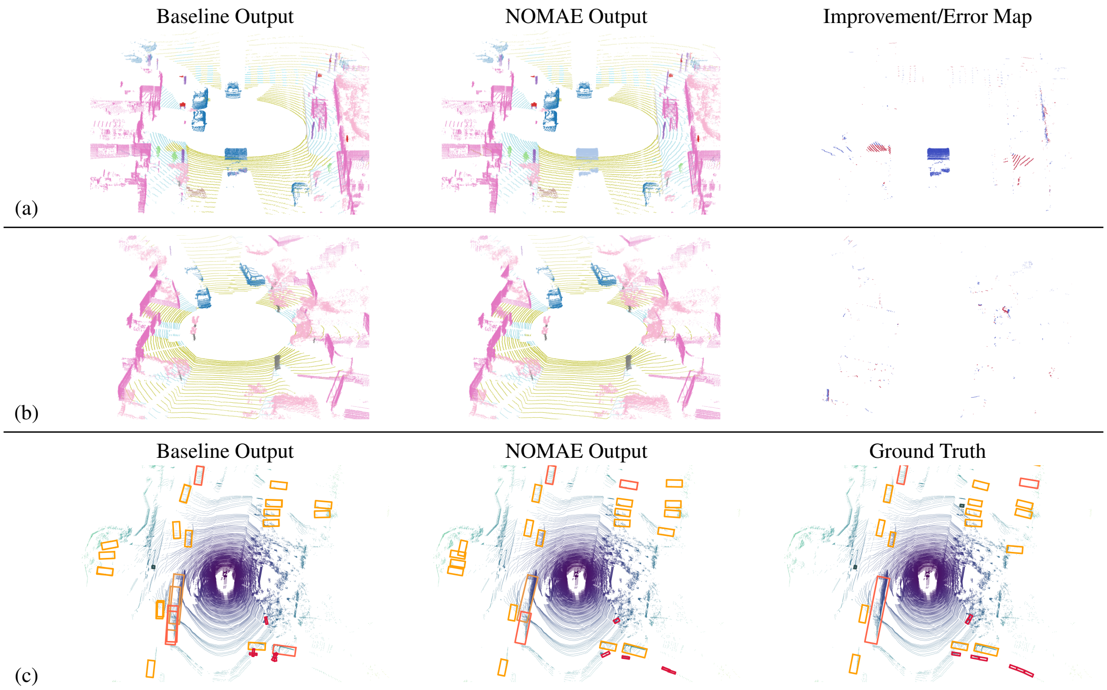


Figure 6. Qualitative comparison of semantic segmentation performance with Ptv3 [37] and object detection performance with UVTR [23].

Table 6. Results with varying amounts of annotated data, evaluated for semantic segmentation mIoU on the nuScenes val set.

Method	**Annotated Scenes**				
	0.1%	**1%**	**10%**	**50%**	**100%**
PTv3 [37]	28.2	41.6	68.7	78.7	80.4
NOMAE (NonLP)	35.7	47.5	67.7	73.8	74.2
NOMAE (fine-tune)	**35.8**	**48.1**	**69.9**	**80.1**	**81.8**

more, we observe that in experiments with very little data, NonLP performs similarly to fine-tuning. This suggests that NOMAE can benefit from the unannotated data to learn a sufficiently strong representation in the encoder, such that the fine-tuning of the encoder has little benefit.

4.7. Qualitative Evaluations

We present qualitative comparisons in Fig. 6. In (a) and (b) for semantic segmentation, we observe that NOMAE improves the accuracy by reducing class mix-up and by improving the boundaries between objects. In (a), we observe that the baseline mis-segments the truck while NOMAE accurately recognizes it. In (b), we see that NOMAE recognizes the smaller object missed by the baseline (the bottom left pole). We can also see that it fails to recognize some drivable areas in (a). In Example (c), we visualize the detections from NOMAE. We see that compared to the baseline, NOMAE is able to more accurately estimate the orientation of the objects and has higher true positive detections. We also see both models hallucinating in further away regions (on the left), and NOMAE fails to detect the pedestrian on the left. More qualitative results are presented in Sec. 12 of the supplementary material.

5. Conclusion

In this work, we proposed NOMAE, a novel multi-scale self-supervised learning framework for large-scale point clouds. Observing the large-scale nature of LiDAR point clouds, NOMAE reconstructs only local neighborhoods, keeping the computation tractable at higher voxel resolutions, avoiding information leakage, and learning a localized representation suitable for diverse downstream perception tasks. Enabled by its efficiency, NOMAE utilizes multiple scales in the pre-training, enabling the model to learn both coarse and fine representations. A novel hierarchical mask generation scheme balances the pre-training of coarse and fine features, which is important for objects of different sizes, such as pedestrians and trucks We presented experimental results that underline the benefit of our proposed contributions, achieving state-of-the-art performance on multiple benchmarks.

Limitations: NOMAE is sensitive to the density of the point cloud and future work will investigate the proposed method for sparse 3D sensors such as radar. Additionally, NOMAE does not utilize the temporal nature of LiDAR data which can open the door for further performance improvement. Furthermore, the application of high-resolution sparse 3D representations in the encoders should be further investigated for the object detection task, where 2D bird's eye view and low-resolution 3D approaches are still dominant.

References

[1] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas J. Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024. 5

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *Int. Conf. on Learning Representations*, 2022. 1

[3] Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, and Renaud Marlet. ALSO: automotive lidar self-supervision by occupancy estimation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023. 2, 4, 5

[4] 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 *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2020. 2, 4

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Int. Conf. on Machine Learning*, 2020. 1

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2021. 5

[7] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 1

[8] Ran Cheng, Ryan Razani, Ehsan Moeen Taghavi, Enxu Li, and Bingbing Liu. (af)2-s3net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 12542–12551, 2021. 5

[9] Christopher B. Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2019. 5, 13

[10] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d`, 2020. 5, 11

[11] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. `https://github.com/Pointcept/Pointcept`, 2023. 5, 11

[12] Spconv Contributors. Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, 2022. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Int. Conf. on Learning Representations*, 2021. 2

[14] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Jiao Qiao. Mcmae: Masked convolution meets masked autoencoders. In *Proc. of the Conf. on Neural Information Processing Systems*, 2022. 1

[15] Nikhil Gosala, Kürsat Petek, Paulo LJ Drews-Jr, Wolfram Burgard, and Abhinav Valada. Skyeye: Self-supervised bird's-eye-view semantic mapping using monocular frontal view images. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 14901–14910, 2023.

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2022. 1, 5, 11

[18] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christoffer Petersson, and Lennart Svensson. Masked autoencoder for self-supervised pre-training on lidar point clouds. In *Proc. of the IEEE winter Conf. on applications of computer vision*, 2023. 1, 2, 3, 4

[19] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 15582–15592, 2020. 2

[20] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 17545–17555, 2023. 5

[21] Christopher Lang, Alexander Braun, Lars Schillingmann, and Abhinav Valada. Self-supervised multi-object tracking for autonomous driving from consistency across timescales. *IEEE Robotics and Automation Letters*, 8(11):7711–7718, 2023. 1

[22] Christopher Lang, Alexander Braun, Lars Schillingmann, Karsten Haug, and Abhinav Valada. Self-supervised representation learning from temporal ordering of automated driving sequences. *IEEE Robotics and Automation Letters*, 9 (3):2582–2589, 2024. 1

[23] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. In *Proc. of the Conf. on Neural Information Processing Systems*, 2022. 5, 6, 8, 11, 15

[24] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *Proc. of the Europ. Conf. on Computer Vision*, 2022. 1, 2

[25] Yueh-Cheng Liu, Yu-Kai Huang, HungYueh Chiang, Hung-Ting Su, Zhe Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. *arXiv preprint arXiv:2104.04687*, 2021. 5

[26] Chen Min, Xinli Xu, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles*, 9: 5150–5162, 2022. 1, 2, 3, 4, 5, 6, 7

[27] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point

cloud self-supervised learning. In *Proc. of the Europ. Conf. on Computer Vision*, 2022. 1, 2

[28] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 21305–21315, 2024. 13

[29] Jonas Schramm, Niclas Vödisch, Kürsat Petek, B Ravi Kiran, Senthil Yogamani, Wolfram Burgard, and Abhinav Valada. Bevcar: Camera-radar fusion for bev map and object segmentation. In *Int. Conf. on Intelligent Robots and Systems*, pages 1435–1442, 2024. 1

[30] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2020. 2, 4

[31] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *Proc. of the Europ. Conf. on Computer Vision*, 2020. 5

[32] Keyu Tian, Yi Jiang, Qishuai Diao, Chen Lin, Liwei Wang, and Zehuan Yuan. Designing BERT for convolutional networks: Sparse and hierarchical masked modeling. In *Int. Conf. on Learning Representations*, 2023. 1

[33] Xiaoyu Tian, Haoxi Ran, Yue Wang, and Hang Zhao. Geomae: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 5, 6, 13

[34] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J. Kusner. Unsupervised point cloud pre-training via occlusion completion. *Int. Conf. on Computer Vision*, pages 9762–9772, 2020. 2

[35] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics*, 42:1 – 11, 2023. 13

[36] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer V2: grouped vector attention and partition-based pooling. In *Advances in Neural Information Processing Systems*, 2022. 5

[37] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4840–4851, 2023. 2, 3, 5, 6, 8, 11, 12, 13, 14

[38] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2024. 11

[39] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Proc. of the Europ. Conf. on Computer Vision*, 2020. 2

[40] Runsen Xu, Tai Wang, Wenwei Zhang, Runjian Chen, Jinkun Cao, Jiangmiao Pang, and Dahua Lin. MV-JAR: masked voxel jigsaw and reconstruction for lidar-based self-supervised pre-training. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3

[41] Xu Yan, Jiantao Gao, Chaoda Zheng, Chaoda Zheng, Ruimao Zhang, Shenghui Cui, and Zhen Li. 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In *Proc. of the Europ. Conf. on Computer Vision*, 2022. 5

[42] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 6, 16

[43] Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, and Wanli Ouyang. GD-MAE: generative decoder for MAE pre-training on lidar point clouds. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 5, 6

[44] Honghui Yang, Sha Zhang, Di Huang, Xiaoyang Wu, Haoyi Zhu, Tong He, Shixiang Tang, Hengshuang Zhao, Qibo Qiu, Binbin Lin, Xiaofei He, and Wanli Ouyang. Unipad: A universal pre-training paradigm for autonomous driving. *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 15238–15250, 2023. 2, 5, 6

[45] 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 *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2022. 1

[46] Karim Abou Zeid, Jonas Schult, Alexander Hermans, and Bastian Leibe. Point2vec for self-supervised representation learning on point clouds. In *DAGM German Conference on Pattern Recognition*, pages 131–146. Springer, 2023. 2

[47] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *Proc. of the Conf. on Neural Information Processing Systems*, 2022. 1, 4, 6, 7

[48] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 6, 16

[49] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2021. 5