2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Re-MAE: Rethinking Masked Autoencoders towards Geometry-Aware Self-Supervised LiDAR-based 3D Object Detection

Youngho Cheon, Jae-Keun Lee, Soon Kwon, Jin-Hee Lee*, Yongseob Lim*

***Abstract*— Self-supervised pre-training with masked autoencoders has shown promise for 3D perception, yet most approaches treat LiDAR point clouds in a geometry-agnostic manner. In this paper, we introduce Re-MAE, a geometry-aware self-supervised learning framework for LiDAR-based 3D object detection that explicitly encodes core properties of LiDAR point clouds: occlusion, distance-driven sparsity, and occupied-empty voxel structure. Re-MAE rethinks the geometric characteristics of LiDAR point clouds from the perspectives of "what to learn" and "how to learn", and introduces three components: (i) Geometry-Aware Masking, which realistically simulates occlusions in LiDAR scans and enables learning complete object representations from partial observations; (ii) Reconstruction-Contextual BCE loss, which effectively guides a multi-scale occupancy prediction task to mitigate distance-dependent point sparsity and the strong occupied-empty voxel imbalance, improving detection of both large vehicles and small, distant pedestrians; and (iii) Realistic Object Augmentation, a label-free foreground augmentation strategy that promotes object-centric representation learning and yields consistent gains across categories. Experiments on ONCE and Waymo Open Dataset validate the effectiveness of Re-MAE, delivering 2.83 mAP and 1.53 L2 mAP respectively over baselines. These results demonstrate that explicitly incorporating the geometric characteristics of LiDAR point clouds enhances the effectiveness of self-supervised learning. The code[1] will be released.**

I. Introduction

Recent advances in deep learning for LiDAR-based 3D object detection [1]–[10] have delivered substantial gains and established a leading paradigm for autonomous driving perception. However, these improvements depend heavily on large-scale annotated 3D datasets such as the Waymo Open Dataset (WOD) [11] and nuScenes [12]. Creating high-quality labels for sparse, irregular point clouds is notably more expensive than for images [13]–[15]; for example, annotating a single 3D object instance takes on average 114 seconds [16]. This annotation bottleneck motivates self-supervised learning (SSL), which pre-trains models on large-scale unlabeled data. Among SSL paradigms, the Masked Autoencoder (MAE) framework, which reconstructs masked portions of the input, has been actively explored for learning robust and generalizable 3D representations from LiDAR data [17]–[21]. Yet, as illustrated in Fig. 1(a), existing LiDAR MAE methods remain insufficiently geometry-aware, and detectors still struggle with heavily occluded or distant objects.

* Corresponding authors. All authors are with the Daegu Gyeongbuk Institute of Science & Technology, Daegu 42988, Republic of Korea (email: {yhcheon, lejk8104, soonyk, jhlee07, yslim73}@dgist.ac.kr)

[1] https://github.com/JH-Research/Re-MAE

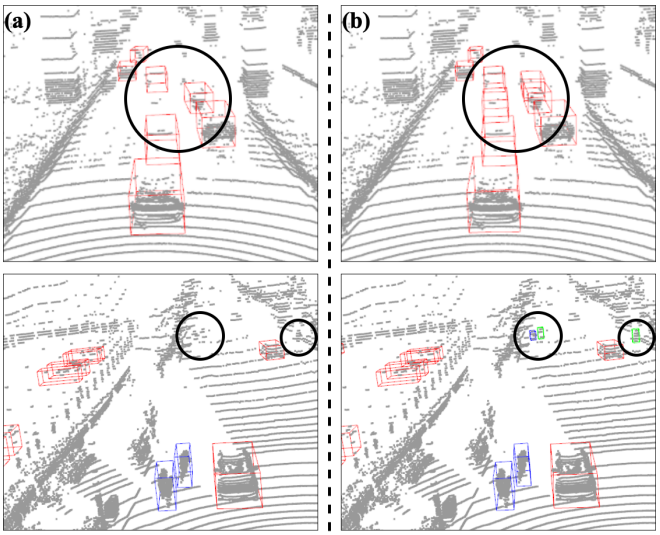

Fig. 1. **Qualitative comparison of 3D object detection on the ONCE validation split.** Both models use the SECOND detector initialized with different self-supervised pre-training methods: (a) Occupancy-MAE, (b) Re-MAE (ours). Bounding boxes denote vehicles (red), pedestrians (green), and cyclists (blue).

These limitations motivate a principled rethinking of LiDAR-based MAE frameworks, particularly regarding the design principles of "what to learn" and "how to learn", which have yet to incorporate the unique geometric properties of LiDAR point clouds. From the perspective of "what to learn," voxel masking strategies [17], [20], [22] randomly remove points on object surfaces in small voxel units (Fig. 2(b)), resulting in trivial reconstruction tasks where the model learns only local neighborhood information rather than global object structures. Conversely, BEV masking [21] removes entire regions based on bird's-eye view (BEV) grids (Fig. 2(c)), risking complete removal of all points from objects and thus providing insufficient reconstruction cues.

In addition, the "how to learn" aspect pertains to guiding the reconstruction task defined by masking strategies through suitable loss functions to enable effective representation learning. In this regard, prior works typically overlook crucial geometric characteristics of LiDAR point clouds. For instance, despite distance-dependent point sparsity, previous methods enforce single-scale reconstruction, uniformly requiring the same target resolution across all regions. This approach forces the model to reconstruct both dense near-range and sparse far-range data under the same criteria, complicating effective representation learning. Furthermore, these methods employ conventional loss functions, *e.g.*, BCE

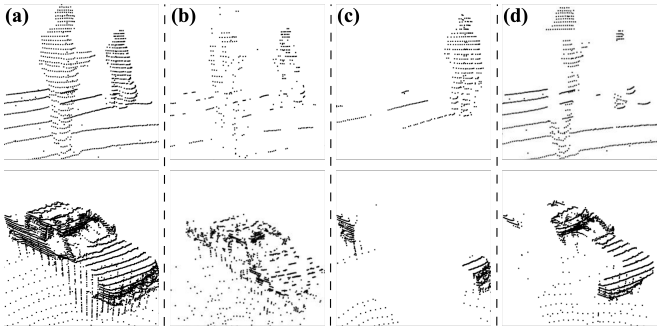

Fig. 2. **Comparison of different masking strategies.** (a) Original point cloud, (b) Voxel Masking, (c) BEV Masking, (d) Geometry-Aware Masking (ours).

loss, that treat all voxels equally, disregarding the severe imbalance between occupied and empty voxels. Consequently, the model tends to prioritize learning the numerous easily reconstructible empty voxels over the fewer, more critical occupied ones.

Motivated by these insights, we propose Re-MAE, which reinterprets the conventional MAE by considering the geometric characteristics of LiDAR point clouds throughout the entire process of deciding "what and how to learn." As shown in Fig. 2(d), Re-MAE adopts Geometry-Aware Masking that realistically simulates occlusions in LiDAR point clouds, thereby guiding the model to learn robust representations of complete object shapes from partial observations. Additionally, Realistic Object Augmentation increases the number and diversity of foreground objects in unlabeled data, fostering more meaningful object-centric representation learning. For the "how to learn" aspect, multi-scale occupancy prediction coupled with Reconstruction-Contextual BCE (ReCon BCE) loss effectively addresses challenges such as variable point density with distance and severe occupied-empty voxel imbalance, enabling the model to focus on crucial geometric information.

By systematically incorporating LiDAR-specific geometric considerations into both the "what to learn" and "how to learn", Re-MAE effectively learns rich 3D representations during the pre-training stage. Consequently, backbones pre-trained with Re-MAE demonstrate superior performance in downstream tasks, clearly observable from qualitative comparisons in Fig. 1(b), where models pre-trained with Re-MAE successfully detect challenging objects, even under severe occlusions or at distant ranges. Quantitative experiments on the ONCE [23] and WOD [11] further validate the superiority of our approach.

Our main contributions are as follows. First, we reinterpret LiDAR-specific geometric characteristics from the perspectives of "what to learn" and "how to learn," and propose Re-MAE, an effective SSL framework composed of three core modules: (i) Geometry-Aware Masking, (ii) ReCon BCE loss, and (iii) Realistic Object Augmentation. Second, experiments on ONCE and WOD demonstrate state-of-the-art performance, with improvements of 2.83 mAP and 1.53 L2 mAP, respectively, over baseline methods.

II. Related Work

A. LiDAR-based 3D Object Detection

LiDAR-based 3D object detection approaches can generally be categorized into point-based and voxel-based methods, based on how they process and represent point clouds. Early studies, such as PointNet [1] and PointNet++ [2] propose directly processing point clouds. These methods demonstrate the potential of deep learning-based approaches but suffer from significant computational overhead due to processing massive amounts of points and the difficulty of effectively capturing local features.

To address these limitations, voxel-based approaches have emerged as mainstream solutions by converting point clouds into regularized voxel representations. VoxelNet [3] and its improved version, SECOND [4], efficiently apply 3D sparse convolutions [24], [25], achieving notable advancements in both speed and accuracy. Subsequent research integrates the strengths of point and voxel-based representations [6] or introduces more sophisticated anchor-free detection heads [5] to further enhance performance. Recently, transformer architectures have been adopted as 3D backbones, leveraging larger receptive fields to better encode geometric information, as exemplified by methods such as Voxel-Transformer [7], SST [8], DSVT [9], and ScatterFormer [10]. However, all these supervised learning methods share a common limitation: their performance heavily relies on large-scale labeled 3D datasets [11], [12], highlighting the necessity of SSL for effective representation learning without labels.

B. LiDAR-based Self-Supervised Learning

SSL seeks to learn informative representations from large-scale unlabeled data and has emerged as a promising approach in 3D perception, where labeling costs are particularly high. SSL primarily follows two paradigms: Contrastive Learning (CL) [26], [27] and MAE [28]–[30]. CL methods such as PointContrast [31], ProposalContrast [32], and BEVContrast [33] focus on learning global representations by maximizing similarity across augmented views of the same input point cloud. However, these approaches typically require two separate networks to process multiple augmented views simultaneously, making them potentially less efficient compared to MAE methods, which rely on a single network.

In contrast, MAE methods learn local and structural representations by masking and reconstructing parts of the input point cloud. Recent studies such as Voxel-MAE [17], Occupancy-MAE [18], GeoMAE [20], and BEV-MAE [21] successfully apply this paradigm to LiDAR, yet differ methodologically in defining their reconstruction tasks. Some works [17], [21] directly regress the original point coordinates of masked regions. However, this approach tends to emphasize redundant low-level details [22], causing the model to focus excessively on trivial point-wise accuracy rather than global structural features. Conversely, binary occupancy prediction, adopted in this paper, reconstructs the occupancy status in 3D space [18], [20], allowing the model to concentrate on volumetric structures rather than individual point locations. Integrated with multi-scale occupancy

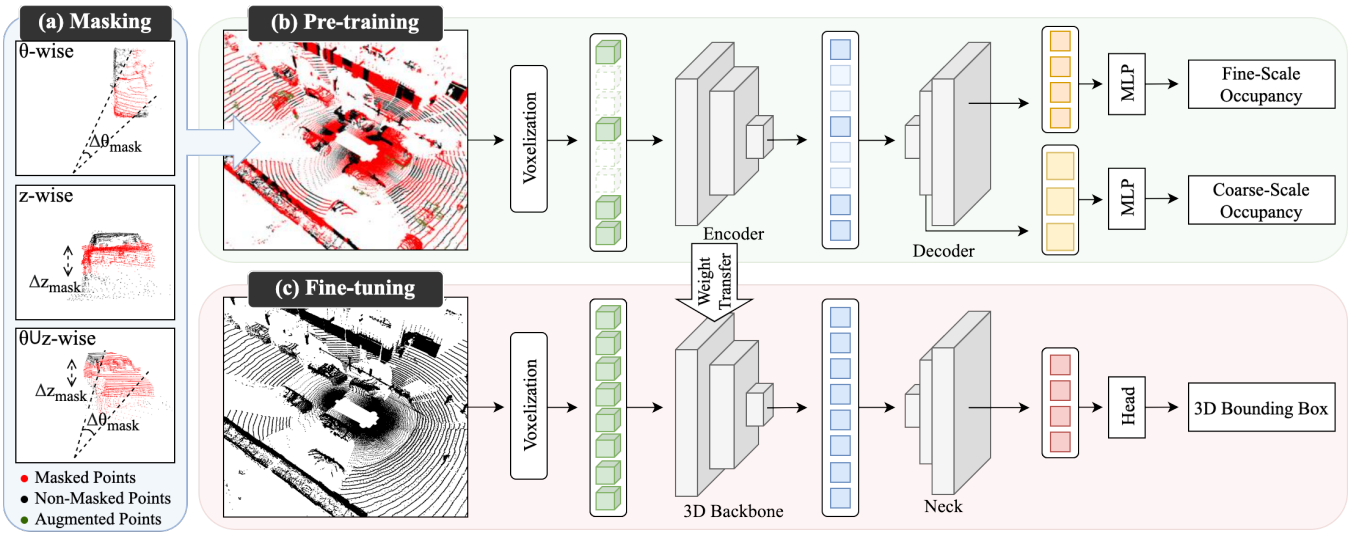

Fig. 3. **Overall architecture of the Re-MAE framework.** In the pre-training stage, the input point cloud is augmented via Realistic Object Augmentation and partially masked by Geometry-Aware Masking. The encoder-decoder network is trained through a multi-scale occupancy reconstruction task using the proposed ReCon BCE loss. The pre-trained encoder weights are then transferred to initialize the 3D backbone for the 3D object detection task.

prediction and our proposed ReCon BCE loss, this strategy optimizes representation learning by explicitly considering LiDAR-specific geometric characteristics. Recently, a hybrid approach combining CL and MAE, CMAE-3D [22], has also been proposed. However, this method adopts a teacher-student framework, making it inherently less efficient compared to the single-network MAE approach.

III. Methods

A. Overall Architecture

We propose Re-MAE, an SSL framework designed to leverage the geometric structure of LiDAR point clouds for downstream 3D object detection. As shown in Fig. 3, the pre-training pipeline consists of three main stages. First, Realistic Object Augmentation increases the number and diversity of objects without manual labels. Second, Geometry-Aware Masking realistically simulates occlusions by masking spatially contiguous regions of the point cloud. Finally, an encoder-decoder network is trained to reconstruct the masked regions using a multi-scale occupancy prediction task. Detailed explanations of these core components are provided in the subsequent subsections (Sec. III-B, Sec. III-C, and Sec. III-D). Through pre-training, the encoder learns rich, geometry-aware representations from LiDAR point clouds. The pre-trained encoder then initializes the backbone of a downstream 3D object detector, enabling effective fine-tuning with limited labeled data.

B. Geometry-Aware Masking

In MAE-based SSL, the masking strategy is critical because it defines the pretext task of "what to learn." However, prior strategies often yield suboptimal objectives: voxel masking [17], [20], [22] tends to induce trivial, highly local reconstructions, whereas BEV masking [21] can remove large regions and thus become overly challenging. To address these limitations, we propose Geometry-Aware Masking, which realistically simulates the occlusions commonly observed in LiDAR point clouds. Concretely, our strategy begins by partitioning the input point cloud into a regular grid on the BEV plane and randomly selecting a proportion, *e.g.*, 70%, of grid cells for masking. To account for the distance-dependent decay of point density, our approach incorporates a distance-aware masking ratio r_{occ} that gradually decreases with sensor distance d. This prevents excessively challenging reconstruction at long ranges where points are already sparse, thereby ensuring a more stable and effective learning process. The masking ratio r_{occ} is defined as:

$$r_{occ} = r_{max} - (r_{max} - r_{min})\frac{d}{d_{\max}} + \delta_{noise} \cdot \varepsilon, \quad (1)$$

where r_{max} and r_{min} are the maximum and minimum masking ratios, d is the distance between the grid center and the sensor, d_{max} is the maximum detection range, δ_{noise} is the noise magnitude, and $\varepsilon \sim \mathcal{U}(-1, 1)$ is a random noise term. To mimic LiDAR occlusion patterns, we model three primary occlusion types: azimuthal (θ-wise), vertical (z-wise), and combined ($\theta \cup z$-wise), as illustrated in Fig. 3(a). For each grid cell selected for masking, one of these three types is randomly sampled. $[\theta_{min}, \theta_{max}]$ denote the azimuth span and $[z_{min}, z_{max}]$ the vertical range of points within the cell. The lengths of the masking intervals are set as:

$$\Delta\theta_{occ} = (\theta_{max} - \theta_{min}) \times r_{occ}, \quad (2)$$

$$\Delta z_{occ} = (z_{max} - z_{min}) \times r_{occ}. \quad (3)$$

Under θ-wise occlusion, points within a randomly sampled, contiguous azimuthal interval $\Delta\theta_{occ}$ from the cell's azimuth span are removed. Similarly, under z-wise occlusion, points within a randomly sampled, contiguous vertical interval Δz_{occ} from the cell's vertical range are removed. For $\theta \cup z$-wise occlusion, the two intervals are sampled independently and their union is removed. This masking process yields realistic, distance-aware occlusions while preserving a solvable reconstruction task.

C. ReCon BCE Loss

Previous LiDAR-based MAE approaches learn representations through a reconstruction pretext task but often fail to incorporate LiDAR-specific geometric properties into the learning objective. We therefore focus on "how to learn" by designing a training strategy that addresses three key challenges: (i) the distance-dependent decay of point density, (ii) the severe imbalance between occupied and empty voxels, and (iii) the need to prioritize reconstruction in masked regions. As shown in Fig. 3(b), we implement this strategy with multi-scale occupancy prediction coupled with the proposed ReCon BCE loss. The loss assigns voxel-level weights based on spatial context, thereby addressing these challenges and guiding effective representation learning.

The encoder processes input voxel features obtained by voxelizing the point cloud. These features are then compressed into 2D BEV features, which are subsequently processed by the decoder through gradual 2D upsampling. At each upsampling stage, a lightweight MLP recovers the Z axis information from the 2D features to predict occupancy, yielding predictions at both coarse and fine scales. This design is computationally more efficient than methods that perform direct 3D occupancy upsampling [18]. Concretely, coarse-scale occupancy is generated by an MLP from the features of the second upsampling layer, while fine-scale predictions are produced similarly from the third upsampling layer. This multi-scale occupancy reconstruction is guided by the proposed ReCon BCE loss. The overall loss $\mathcal{L}$ is defined as the sum of the two scale-specific losses:

$$\mathcal{L} = \mathcal{L}_{\text{fine}} + \mathcal{L}_{\text{coarse}}. \tag{4}$$

Each scale-specific loss $\mathcal{L}_s$ for $s \in \{\text{fine}, \text{coarse}\}$ is defined as voxel-wise weighted binary cross-entropy (BCE) loss:

$$\mathcal{L}_s = \frac{\sum_{i=1}^{N_v} w_s(v_i) \cdot \ell_{\text{BCE}}(p_s(v_i), t_s(v_i))}{\sum_{i=1}^{N_v} w_s(v_i)}, \tag{5}$$

where ℓ_{BCE} denotes the BCE between predicted occupancy $p_s(v_i)$ and ground-truth occupancy $t_s(v_i)$ at voxel v_i. The voxel weights $w_s(v_i)$ provide contextual modulation by dynamically weighting voxels according to LiDAR geometry and reconstruction priority. They factorize into three terms:

$$w_s(v_i) = d_s(v_i) \cdot b_s(v_i) \cdot m_s(v_i), \tag{6}$$

where d_s is the distance weight, b_s is the boundary weight, and m_s is the masking weight. Each term is detailed below.

Distance Weights. To address the distance-dependent variation in point density, fine-scale predictions are prioritized at short ranges to capture fine-grained geometric details. Conversely, coarse-scale predictions are prioritized at long ranges to learn the overall object structure. This prioritization is implemented through the following distance weights:

$$d_{\text{fine}}(v_i) = \alpha_{\text{fine}} - (\alpha_{\text{fine}} - \alpha_{\text{coarse}}) \frac{d(v_i)}{d_{\text{max}}}, \tag{7}$$

$$d_{\text{coarse}}(v_i) = \alpha_{\text{coarse}} + (\alpha_{\text{fine}} - \alpha_{\text{coarse}}) \frac{d(v_i)}{d_{\text{max}}}, \tag{8}$$

where $d(v_i)$ is the horizontal distance from the center of voxel v_i to the sensor, d_{max} is the maximum detection distance. The hyper-parameters α_{fine} and α_{coarse} adjust the relative importance between the two scales.

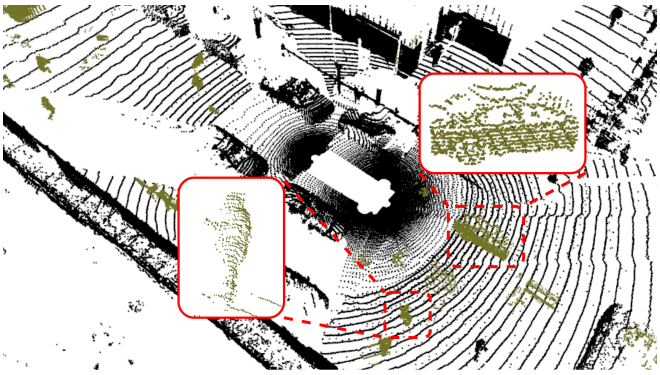

Fig. 4. **Visualization of the proposed Realistic Object Augmentation.** Black points denote the original point cloud, and green points indicate augmented objects placed on object-free road areas by our strategy.

Boundary Weights. To mitigate the severe imbalance between occupied and empty voxels, higher weights are assigned to boundary voxels, which are crucial for object shape delineation. This strategy guides the representation learning to focus on these critical boundary regions, rather than the less informative empty space. Specifically, the process begins with the ground-truth binary occupancy map, where voxels containing points are set to 1 (occupied) and others to 0 (empty). To identify not only the occupied voxels but also their immediate surroundings, we apply a dilation operation with a $3\times3\times3$ positive-valued kernel. Voxels in this dilated region are considered the boundary and are assigned a higher weight, while all other voxels are assigned a weight of 1.

Masking Weights. In the MAE framework, the supervision for representation learning is derived from the reconstruction of masked regions. To guide the learning process to focus on this crucial source of supervision, masked occupied voxels are assigned a higher weight, while all unmasked voxels are assigned a weight of 1.

D. Realistic Object Augmentation

Foreground objects account for only a small fraction of LiDAR points in autonomous driving datasets, which hinders object-centric representation learning in SSL. To remedy this, we introduce Realistic Object Augmentation, a label-free strategy that increases foreground frequency and diversity while preserving LiDAR geometry, as illustrated in Fig. 4. The Realistic Object Augmentation process begins by partitioning the input point cloud into a regular BEV grid. Based on the height distribution within each cell, the process then identifies two specific types of regions: potential object regions and object-free road regions. Potential object regions are identified as cells where the vertical extent of points falls within a predefined object-height range, *e.g.*, 1.5-2.0m. This criterion helps filter out most non-object background structures, such as buildings and walls. Conversely, object-free road regions are identified as cells containing only points that form a nearly flat surface within a ground-height range. The scene is then augmented by duplicating the points from the identified object regions and placing them onto

TABLE I

PERFORMANCE COMPARISON ON THE ONCE VALIDATION SPLIT. "FROM SCRATCH" DENOTES THE BASELINE WITHOUT PRE-TRAINING. * DENOTES OUR RE-IMPLEMENTATION USING OFFICIAL CODE.

Method	Reference	mAP	Vehicle				Pedestrian				Cyclist			
			Overall	0-30m	30-50m	50m-inf	Overall	0-30m	30-50m	50m-inf	Overall	0-30m	30-50m	50m-inf
From scratch [4]	-	51.89	71.19	84.04	63.02	47.25	26.44	29.33	24.05	18.05	58.04	69.96	52.43	34.61
SwAV [34]	NeurIPS'20	51.96 (+0.07)	72.71	83.68	65.91	50.10	25.13	27.77	22.77	16.36	58.05	69.99	52.23	34.86
ProposalContrast [32]	ECCV'22	52.33 (+0.44)	72.99	84.41	65.92	50.11	25.77	27.95	23.74	18.06	58.23	69.99	53.03	35.48
Occupancy-MAE [18]	T-IV'23	52.51 (+0.62)	72.78	83.77	66.01	50.26	27.49	30.54	25.28	16.11	57.26	69.71	52.31	33.51
BEV-MAE* [21]	AAAI'24	53.08 (+1.19)	72.78	83.74	65.51	52.28	28.10	31.13	25.40	17.52	58.37	70.38	52.52	35.01
Re-MAE	**Ours**	**54.72 (+2.83)**	73.78	84.64	67.35	52.00	32.07	36.68	28.56	17.51	58.32	70.82	51.71	34.20

TABLE II

PERFORMANCE COMPARISON ON THE WOD VALIDATION SPLIT. "FROM SCRATCH" DENOTES THE BASELINE WITHOUT PRE-TRAINING. "PRETRAIN" INDICATES THE PERCENTAGE OF THE TRAINING SPLIT USED FOR THE PRE-TRAINING STAGE. † DENOTES RESULTS RE-IMPLEMENTED BY THE AUTHORS OF BEV-MAE.

Method	Reference	Pretrain	L2 mAP / mAPH	Vehicle	Pedestrian	Cyclist
				L2 AP / APH	L2 AP / APH	L2 AP / APH
From scratch [5]	-	-	65.60 / 63.21	64.18 / 63.69	65.22 / 59.68	67.41 / 66.25
Occupancy-MAE [18]	T-IV'23	20%	65.85 (+0.25) / 63.23 (+0.02)	64.05 / 63.53	65.78 / 59.62	67.76 / 66.53
BEV-MAE [21]	AAAI'24	20%	66.70 (+1.10) / 64.25 (+1.04)	64.71 / 64.22	66.21 / 60.59	69.11 / 67.93
CMAE-3D [22]	IJCV'25	20%	66.33 (+0.73) / 63.86 (+0.65)	64.36 / 63.88	66.08 / 60.36	68.56 / 67.35
Re-MAE	**Ours**	20%	**66.80 (+1.20) / 64.34 (+1.13)**	64.74 / 64.24	65.95 / 60.27	69.70 / 68.52
PointContrast [31]	ECCV'20	100%	65.88 (+0.28) / 63.23 (+0.07)	63.81 / 63.33	66.67 / 60.51	67.17 / 66.00
DepthContrast [35]	ICCV'21	100%	65.84 (+0.24) / 63.33 (+0.12)	64.45 / 63.95	65.61 / 59.86	67.43 / 66.22
Point-M2AE [36]	NeurIPS'22	100%	66.10 (+0.50) / 63.59 (+0.38)	64.26 / 63.77	65.64 / 60.00	68.20 / 67.01
ProposalContrast [32]	ECCV'22	100%	66.42 (+0.82) / 63.85 (+0.64)	65.03 / 64.53	65.93 / 59.95	68.26 / 67.04
GD-MAE† [19]	CVPR'23	100%	66.98 (+1.38) / 64.53 (+1.32)	65.64 / 64.95	66.39 / 61.12	68.92 / 67.52
BEV-MAE [21]	AAAI'24	100%	67.02 (+1.42) / 64.55 (+1.34)	65.01 / 64.53	66.58 / 60.87	69.46 / 68.25
Re-MAE	**Ours**	100%	**67.13 (+1.53) / 64.65 (+1.44)**	64.94 / 64.46	66.51 / 60.75	69.93 / 68.76

TABLE III

PERFORMANCE COMPARISON ON THE WOD DATA-EFFICIENT BENCHMARK. "FROM SCRATCH" DENOTES THE BASELINE WITHOUT PRE-TRAINING. "DATA AMOUNT" DENOTES THE PERCENTAGE OF THE TRAINING SPLIT (SEQUENCE-LEVEL SPLIT) USED FOR FINE-TUNING.

Method	Reference	Data amount: 5%	Data amount: 10%	Data amount: 20%	Data amount: 50%	Data amount: 100%
		L2 mAP / mAPH	L2 mAP / mAPH	L2 mAP / mAPH	L2 mAP / mAPH	L2 mAP / mAPH
From scratch [8]	-	44.41 / 40.34	54.31 / 50.46	60.16 / 56.78	66.43 / 63.36	68.50 / 65.54
PointContrast [31]	ECCV'20	45.32 / 41.30	53.69 / 49.94	59.35 / 55.78	65.51 / 62.21	68.06 / 64.84
ProposalContrast [32]	ECCV'22	46.62 / 42.58	53.89 / 50.13	59.52 / 55.91	65.76 / 62.49	68.17 / 65.01
MV-JAR [37]	CVPR'23	50.52 / 46.68	57.44 / 54.06	62.28 / 59.15	66.70 / 63.69	69.16 / 66.20
BEV-MAE [21]	AAAI'24	51.63 / 47.77	58.16 / 54.75	**62.88** / 59.97	67.16 / 64.07	69.35 / 66.46
Re-MAE	**Ours**	**52.78 / 49.99**	**59.08 / 56.46**	62.82 / **60.24**	**67.63 / 65.11**	**69.51 / 67.08**

the object-free road regions. To enhance diversity, these augmented points are randomly rotated by multiple discrete angles, *e.g.*, 0°, 90°, 180°, and 270°. Furthermore, to closely mimic real-world scans, the point density of each augmented object is adjusted to follow the natural density decay with its distance from the LiDAR sensor. Finally, these augmented objects undergo the same Geometry-Aware Masking process, followed by the occupancy reconstruction task.

IV. EXPERIMENTS

A. Implementation Details

We evaluate our Re-MAE framework on two representative autonomous driving datasets, ONCE [23] and WOD [11]. For the ONCE dataset, we adopt SECOND [4] as the baseline detector and pre-train for 10 epochs. Similarly, for the WOD dataset, we use CenterPoint [5] and pre-train for 20 epochs. These detector choices are consistent with those of the prior works we compare against. In accordance with SSL protocols, we do not use any ground truth annotations during pre-training. For fine-tuning, we strictly follow the from scratch training recipe, initializing only the 3D backbone with the pre-trained weights. All experiments are conducted on 8 NVIDIA RTX A5000 GPUs using the Adam optimizer with a batch size of 4. The learning rate is set to 0.0003 for pre-training and 0.003 for fine-tuning. For Geometry-Aware Masking in Eq. (1), r_{max} and r_{min} are set to 0.75 and 0.35, respectively, and δ_{noise} is set to 0.15. For ReCon BCE loss, α_{fine} and α_{coarse} are set to 0.75 and 0.25, respectively.

B. Comparison with State-of-the-art Methods

ONCE Results. We evaluate Re-MAE on ONCE [23] by pre-training on the unlabeled small split (100k scenes) and fine-tuning on the labeled training split (5k scenes). As shown in Tab. I, Re-MAE achieves 54.72 mAP, outperforming prior SOTA SSL methods. Specifically, it surpasses

the training from scratch by 2.83 mAP, and outperforms Occupancy-MAE [18] and BEV-MAE [21] by 2.21 and 1.64 mAP, respectively. The gains are especially pronounced for the Pedestrian class, with improvements of 5.63 mAP over the baseline, 4.58 mAP over Occupancy-MAE, and 3.97 mAP over BEV-MAE.

WOD Results. We also evaluate on WOD [11] under two uniform-sampling setups (Tab. II). In "Pretrain 20%", both pre-training and fine-tuning use the same 20% subset uniformly sampled from the training split. In "Pretrain 100%", pre-training uses the entire training split, whereas fine-tuning uses a uniformly sampled 20% subset. Performance is evaluated using Level-2 (L2) mean Average Precision (mAP) and heading-aware mAP (mAPH), which consider objects with at least one LiDAR point. In the first setup, Re-MAE attains 66.80 L2 mAP and 64.34 L2 mAPH, improving over the training from scratch baseline by 1.20 L2 mAP and 1.13 L2 mAPH. It also outperforms CMAE-3D [22] and BEV-MAE [21] by 0.47 and 0.10 L2 mAP. In the second setup, Re-MAE reaches 67.13 L2 mAP and 64.65 L2 mAPH, improving over the baseline by 1.53 L2 mAP and 1.44 L2 mAPH, and exceeding GD-MAE [19] and BEV-MAE by 0.15 and 0.11 L2 mAP.

WOD Data Efficiency Results. We further evaluate the data efficiency of Re-MAE using the data-efficient benchmark proposed by MV-JAR [37]. In this setting, we pre-train on the entire training split and fine-tune on sequence-level subsets of the training split at 5%, 10%, 20%, 50%, and 100% (Tab. III). Re-MAE consistently outperforms the training from scratch baseline and prior SSL baselines across all data percentages. At the 5% setting, Re-MAE attains 52.78 L2 mAP and 49.99 L2 mAPH, improving over the baseline by 8.37 and 9.65 and exceeding BEV-MAE [21] by 1.15 and 2.22, respectively. It also achieves the best results at 10%, 50%, and 100%, and is on par with BEV-MAE in L2 mAP at 20% while achieving higher L2 mAPH.

TABLE IV

ABLATION STUDIES ON DIFFERENT MASKING STRATEGIES. RE-MAE* DENOTES A SIMPLIFIED VERSION WITH SINGLE-SCALE OCCUPANCY PREDICTION AND A STANDARD BCE LOSS.

Method	Masking Strategy	mAP
From scratch		51.89
Re-MAE*	Voxel masking	52.93 (+1.04)
	BEV masking	53.01 (+1.12)
	Geometry-Aware Masking (θ)	53.56 (+1.67)
	Geometry-Aware Masking (θ, z)	53.80 (+1.91)
	Geometry-Aware Masking (θ, z, $\theta \cup z$)	**53.87 (+1.98)**

C. Ablation Study

We conduct ablation studies on the ONCE [23] validation split to analyze the impact of the key components in our proposed Re-MAE framework.

Masking Strategies. To verify the effectiveness of our proposed Geometry-Aware Masking, we compare it against various masking strategies. The results, presented in Tab. IV, show that conventional random masking strategies such as voxel masking and BEV masking yield only limited improvements of 1.04 mAP and 1.12 mAP, respectively, over the baseline. In contrast, our Geometry-Aware Masking significantly boosts performance. Even the simplest version (θ-wise) achieves a substantial gain of 1.67 mAP because it creates a pretext task that is both challenging and solvable, forcing the model to infer the global shape of an object from a partially occluded view. Furthermore, performance progressively increases with the introduction of more complex occlusion types, *e.g.*, azimuthal (θ-wise), vertical (z-wise), and their combination ($\theta \cup z$-wise). This demonstrates that as the masking strategy incorporates a more diverse set of occlusion types, it better simulates the variety of partial observations found in real-world LiDAR point clouds. Exposing the model to this richer and more realistic data distribution compels it to learn a more robust and generalizable representation of complete object shapes. The final configuration, which randomly applies all three types, achieves the highest gain of 1.98 mAP.

TABLE V

ABLATION STUDIES ON COMPONENTS OF RE-MAE. GAM: GEOMETRY-AWARE MASKING; MOP: MULTI-SCALE OCCUPANCY PREDICTION; RCL: RECON BCE LOSS; ROA: REALISTIC OBJECT AUGMENTATION.

Method	Components				mAP
	GAM	MOP	RCL	ROA	
From scratch					51.89
Re-MAE	✓				53.87 (+1.98)
	✓	✓			53.69 (+1.80)
	✓	✓	✓		54.40 (+2.51)
	✓	✓	✓	✓	**54.72 (+2.83)**

Component-wise Analysis. We further evaluate the contribution of each component, as detailed in Tab. V. First, using Geometry-Aware Masking alone provides a substantial 1.98 mAP improvement over the baseline. Interestingly, adding multi-scale occupancy prediction results in a slight drop in performance, suggesting that more complex targets require more sophisticated guidance. The subsequent introduction of the ReCon BCE loss not only recovers this drop but also boosts performance to 2.51 mAP improvement, highlighting the critical synergy between our reconstruction target and loss function, validating our design for "how to learn". Finally, incorporating Realistic Object Augmentation achieves the best overall performance with a 2.83 mAP gain, as it directly addresses foreground data scarcity by increasing object diversity and promoting object-centric representation learning.

D. Real-Vehicle Integration and On-Road Validation

ROS-based Re-MAE. A key practical advantage of Re-MAE is that it improves detection performance while maintaining the low latency required for real-time onboard perception in autonomous vehicles. To facilitate practical deployment, we integrate a detector initialized with Re-MAE pre-trained weights into the open-source Robot Operating System (ROS) and design a ROS-based processing pipeline. Concretely, the system comprises three modular components:

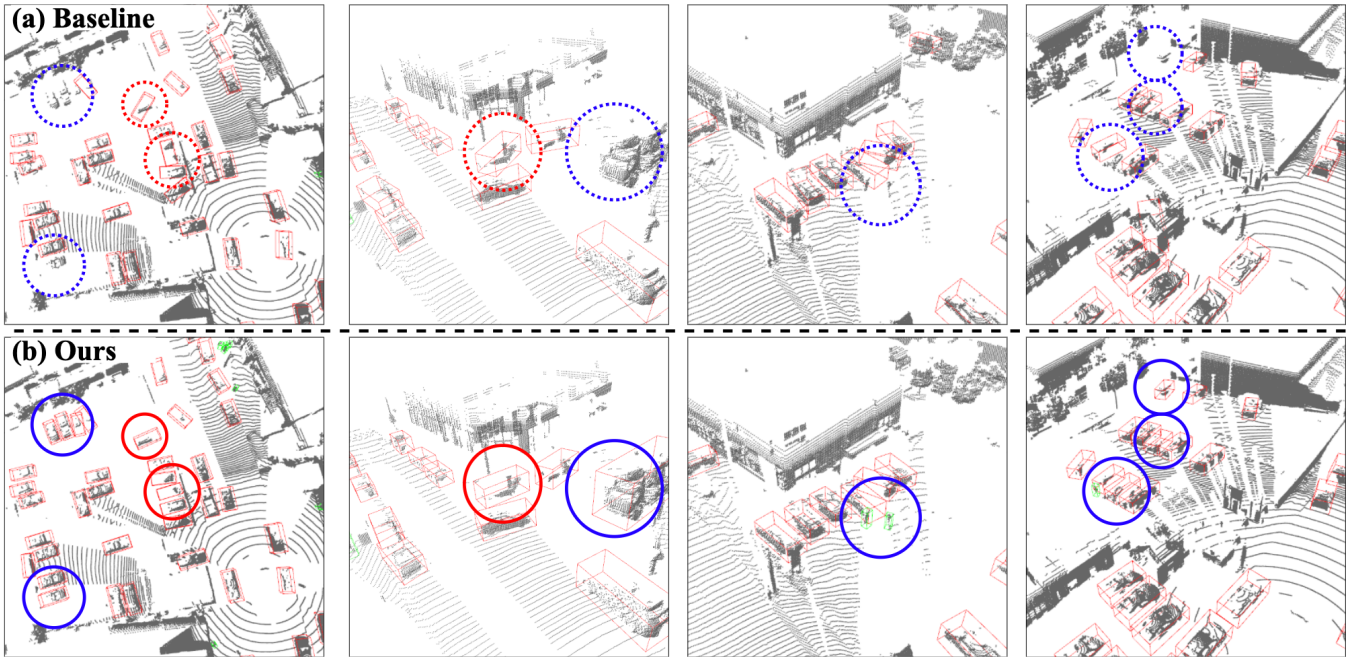

Fig. 5. **Road-tested qualitative comparison of 3D object detection.** Both models use the CenterPoint detector. (a) The baseline model is trained from scratch and (b) our model is pre-trained with Re-MAE. Red and blue dashed circles denote false positives and false negatives, respectively.

a subscriber node, a 3D detection node, and a publisher node. The pipeline begins by subscribing to a point cloud topic (*e.g.*, from a vehicle-mounted LiDAR sensor). The incoming messages are parsed and preprocessed to form model-ready inputs, which are then passed to the detector, whose backbone is initialized with Re-MAE pre-trained weights. The detector's outputs, such as object classes and 3D bounding boxes, are packaged and published on dedicated ROS topics. These messages are readily consumed by downstream modules integrated in the ROS ecosystem, such as tracking and planning, enabling seamless end-to-end operation on the real-vehicle platform.

Road-Tested Results. In this subsection, we evaluate the practical effectiveness of Re-MAE in on-road settings. This setting reflects a common real-world challenge: acquiring unlabeled data is feasible (10k scenes), whereas obtaining extensive annotations is difficult, resulting in a limited labeled set (500 scenes). Fig. 5 illustrates a qualitative comparison on unseen test sequences, contrasting a baseline detector trained from scratch with a detector initialized with Re-MAE pre-trained weights. For this experiment, Re-MAE is pre-trained for 20 epochs and then fine-tuned for 80 epochs on the small labeled subset, while the baseline detector is trained from scratch on the same labeled subset for 80 epochs. The baseline detector, trained solely on the limited labeled data, not only often fails to detect challenging objects, such as small or partially occluded ones, but also frequently predicts inaccurate headings for them, as shown in Fig. 5(a). In contrast, the detector initialized with Re-MAE pre-trained weights demonstrates significantly superior detection capabilities, as depicted in Fig. 5(b). It successfully identifies these challenging instances and accurately estimates their orientation, reflecting the robust geometric representations learned from the vast unlabeled data. These results highlight the practical utility of Re-MAE, showing that it can substantially boost performance in real-world applications where large-scale public datasets are unsuitable due to differences in sensors or environments.

V. Conclusion

In this work, we present Re-MAE, a self-supervised learning framework designed to leverage the geometric properties of LiDAR point clouds. By comprehensively addressing the "what to learn" and "how to learn" aspects, Re-MAE incorporates Geometry-Aware Masking, which realistically simulates frequent occlusions in LiDAR point clouds and enables robust learning of complete object representations from partial observations. It further employs the proposed ReCon BCE loss to effectively address challenges posed by distance-dependent point density variations and severe occupied-empty voxel imbalance. Finally, it leverages Realistic Object Augmentation to significantly increase the number and diversity of foreground objects without annotations, thereby promoting object-centric representation learning. Experimental evaluations on the ONCE and WOD demonstrate that Re-MAE achieves substantial improvements of up to 2.83 mAP and 1.53 L2 mAP over baseline methods, respectively. These results highlight that explicitly accounting for LiDAR geometry substantially improves the effectiveness of SSL. This work underscores the value of sensor-aware SSL design, and we hope it inspires further extensions to complex multi-modal and cross-modal settings.

Acknowledgment

This work was supported by the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (RS-2024-00432495), the National Research Foundation of Korea

(NRF) grant funded by the Korea government (MSIT) (RS-2025-16065352), and Korea Innovation Foundation through the Ministry of Science and ICT (RS-2025-02219242 and RS-2025-02634821).

REFERENCES

[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[3] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.

[4] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[5] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 784–11 793.

[6] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 529–10 538.

[7] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel transformer for 3d object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3164–3173.

[8] L. Fan, Z. Pang, T. Zhang, Y.-X. Wang, H. Zhao, F. Wang, N. Wang, and Z. Zhang, "Embracing single stride 3d object detector with sparse transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8458–8468.

[9] H. Wang, C. Shi, S. Shi, M. Lei, S. Wang, D. He, B. Schiele, and L. Wang, "Dsvt: Dynamic sparse voxel transformer with rotated sets," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 520–13 529.

[10] C. He, R. Li, G. Zhang, and L. Zhang, "Scatterformer: Efficient voxel transformer with scattered linear attention," in *European Conference on Computer Vision*. Springer, 2024, pp. 74–92.

[11] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. 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*, 2020, pp. 2446–2454.

[12] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[14] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[15] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[16] Q. Meng, W. Wang, T. Zhou, J. Shen, L. Van Gool, and D. Dai, "Weakly supervised 3d object detection from lidar point cloud," in *European Conference on computer vision*. Springer, 2020, pp. 515–531.

[17] G. Hess, J. Jaxing, E. Svensson, D. Hagerman, C. Petersson, and L. Svensson, "Masked autoencoder for self-supervised pre-training on lidar point clouds," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 350–359.

[18] C. Min, L. Xiao, D. Zhao, Y. Nie, and B. Dai, "Occupancy-mae: Self-supervised pre-training large-scale lidar point clouds with masked occupancy autoencoders," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 7, pp. 5150–5162, 2023.

[19] H. Yang, T. He, J. Liu, H. Chen, B. Wu, B. Lin, X. He, and W. Ouyang, "Gd-mae: generative decoder for mae pre-training on lidar point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9403–9414.

[20] X. Tian, H. Ran, Y. Wang, and H. Zhao, "Geomae: Masked geometric target prediction for self-supervised point cloud pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 570–13 580.

[21] Z. Lin, Y. Wang, S. Qi, N. Dong, and M.-H. Yang, "Bev-mae: Bird's eye view masked autoencoders for point cloud pre-training in autonomous driving scenarios," in *Proceedings of the AAAI conference on artificial intelligence*, 2024.

[22] Y. Zhang, J. Chen, and D. Huang, "Cmae-3d: Contrastive masked autoencoders for self-supervised 3d object detection," *International Journal of Computer Vision*, vol. 133, no. 5, pp. 2783–2804, 2025.

[23] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li *et al.*, "One million scenes for autonomous driving: Once dataset," *arXiv preprint arXiv:2106.11037*, 2021.

[24] B. Graham and L. Van der Maaten, "Submanifold sparse convolutional networks," *arXiv preprint arXiv:1706.01307*, 2017.

[25] B. Graham, M. Engelcke, and L. Van Der Maaten, "3d semantic segmentation with submanifold sparse convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9224–9232.

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[28] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[30] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9653–9663.

[31] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany, "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding," in *European conference on computer vision*. Springer, 2020, pp. 574–591.

[32] J. Yin, D. Zhou, L. Zhang, J. Fang, C.-Z. Xu, J. Shen, and W. Wang, "Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection," in *European conference on computer vision*. Springer, 2022, pp. 17–33.

[33] C. Sautier, G. Puy, A. Boulch, R. Marlet, and V. Lepetit, "Bevcontrast: Self-supervision in bev space for automotive lidar point clouds," in *2024 International Conference on 3D Vision (3DV)*. IEEE, 2024, pp. 559–568.

[34] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[35] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra, "Self-supervised pretraining of 3d features on any point-cloud," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 252–10 263.

[36] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li, "Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training," *Advances in neural information processing systems*, vol. 35, pp. 27 061–27 074, 2022.

[37] R. Xu, T. Wang, W. Zhang, R. Chen, J. Cao, J. Pang, and D. Lin, "Mv-jar: Masked voxel jigsaw and reconstruction for lidar-based self-supervised pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 445–13 454.