Information Sciences 760 (2027) 124117

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

PointSiRE: Siamese representation learning with relative geometry and evolved masking for point cloud

Xin Cao, Linrong Ye, Xinmeng Hu, Tianyi Liu, Linzhi Su*, Xingxing Hao, Kang Li*

College of Computer Science, Northwest University, Xi'an, China

HIGHLIGHTS

- Unifies contrastive-generative learning for self-supervised 3D learning.
- First to introduce Masked Dynamic Prototype Alignment for point cloud SSL.
- Employs synergistic dual-alignment objectives for robust representations.
- Features Teacher-Student architecture with Evolving Prototype Memory.

ARTICLE INFO

Keywords:
Point cloud
Self-supervised learning
Masked representation learning
Geometric consistency
Hierarchical masking

ABSTRACT

Masked self-supervised representation learning has achieved promising results in point cloud pre-training. However, existing paradigms largely regard geometric transformations (e.g., rotation and scaling) only as stochastic augmentations for learning invariant representations and adopt fixed, content-agnostic masking strategies, overlooking the structural cues embedded in inter-view geometric relationships and the evolving learning capacity of the model. In this paper, we propose PointSiRE, a Siamese self-supervised framework that reformulates masked point cloud pre-training as a relative geometry-conditioned cross-view prediction task, where inter-view transformations are explicitly exploited as supervisory signals rather than merely augmentation operations. Specifically, we introduce a Relative Geometry-aware Position Encoding (RGPE) mechanism that models the relative rotation, translation, and scale between augmented views to guide feature alignment across coordinate systems. To further enhance semantic learning, we propose an Attention-guided Evolved Hierarchical Masking (AEHM) strategy, which progressively evolves masking patterns from fine-grained random drops to structurally coherent regions according to the teacher's attention. Extensive experiments demonstrate that PointSiRE learns robust and transferable representations and achieves competitive or state-of-the-art performance across classification, few-shot learning, and dense prediction tasks.

1. Introduction

Point clouds, as a fundamental representation of 3D geometry, capture the raw spatial structure of scenes and objects, offering a compact and precise description of the physical world [1]. Despite their ability to capture fine-grained geometric details, the inherent irregularity and unorderedness of point clouds make them unsuitable for direct processing by standard convolutional neural

* Corresponding authors.
Email addresses: sulinzhi029@nwu.edu.cn (L. Su), likang@nwu.edu.cn (K. Li).

https://doi.org/10.1016/j.ins.2026.124117
Received 27 March 2026; Received in revised form 3 September 2026; Accepted 4 September 2026
Available online 5 September 2026

networks designed for regular grids. Early deep learning approaches such as PointNet struggled to effectively model local regions [2], a limitation addressed by PointNet++ through hierarchical sampling and grouping strategies that better capture local geometric structures [3]. Supervised methods such as DGCNN [4], PointCNN [5], and KPConv [6] have since achieved impressive results on benchmark datasets, but collecting large-scale, high-quality 3D annotations remains costly and labor-intensive, limiting their scalability. These limitations have spurred growing interest in self-supervised learning (SSL), which learns transferable 3D representations directly from unlabeled point clouds, thereby eliminating the dependence on costly manual annotation while exploiting the large volumes of unlabeled 3D data that are increasingly available [7].

Early self-supervised learning for point clouds mainly relied on reconstruction or completion objectives to recover missing geometry from partial observations [8]. While effective in capturing local geometric details, these methods often assume data completeness, which limits their generality. Contrastive learning was later introduced to learn discriminative 3D representations via instance discrimination [9]. Subsequent works further explored cross-view or cross-modal representation learning. DepthContrast demonstrates that effective representations can be learned from single-view 3D data via contrastive learning and further improves performance by jointly leveraging multiple input modalities such as point clouds and voxels [10]. STRL exploits spatio-temporal cues from sequential point clouds and learns invariant representations by aligning temporally correlated frames under spatial augmentations in a self-supervised manner [11]. Further developments, such as OcCo [12] and CrossPoint [13], enhance representation transferability by enforcing view-level consistency, while CrossPoint additionally aligns representations across modalities. However, these contrastive approaches typically depend on large batch sizes, negative sampling, or hard-negative mining, which may compromise training stability and efficiency.

BERT pre-trains deep bidirectional representations by predicting masked tokens conditioned on both left and right context [14]. In computer vision, MAE extends this paradigm by masking a large proportion of image patches and reconstructing the missing content with an asymmetric encoder–decoder architecture [15]. Motivated by these advances, masked representation learning has become a scalable paradigm for point cloud pre-training: by masking a subset of tokens or local regions and predicting the missing content from visible context, this paradigm provides an efficient self-supervision signal and consistently improves performance across downstream tasks. Despite these advances, most existing frameworks treat geometric transformations (e.g., rotation, scaling) merely as random augmentations that encourage invariance, while neglecting the underlying spatial relations between views. Yet 3D point clouds inherently encode meaningful geometric dependencies across views, and disregarding them limits pre-training to low-level pattern completion rather than reasoning about structural variation. A robust 3D learner should capture both invariance and relative geometric relationships. Prior works on orientation estimation [16], SE(3)-equivariant networks [17], and unsupervised pose learning [18] highlight the importance of explicit geometric reasoning, yet this remains largely unexplored within masked point cloud pre-training frameworks. This motivates our first design: a position encoding that explicitly models relative geometry between views, rather than treating it as a nuisance to be averaged away.

Beyond geometric awareness, the design of the pretext task itself also plays a critical role in determining what semantic structures can be learned during pre-training. However, existing masking strategies remain largely static and content-agnostic. Random or patch-based masking strategies are widely used for their simplicity. Point-BERT randomly masks local patches and reconstructs discrete point tokens for pre-training [19]. MaskPoint instead formulates masking as a discriminative task by distinguishing masked object points from sampled noise points [20]. However, such strategies ignore the intrinsic structure of point clouds, and the standard reconstruction loss (e.g., Chamfer Distance) mainly restores low-level coordinates rather than enforcing semantic consistency [21]. As shown in Fig. 1, random masking often degenerates into local interpolation, while block masking may remove entire semantic parts, posing an excessively difficult objective in early training stages. Although multi-ratio masking strategies [22] adjust task difficulty, they still lack adaptation to the model's evolving structural awareness. Consequently, masking remains weakly aligned with geometric and semantic cues, leaving the encoder's relational reasoning underutilized. This motivates our second design: a masking strategy that evolves jointly with the model's own structural understanding.

Motivated by these two limitations, we revisit masked representation learning for point clouds from a geometric and structural perspective and propose PointSiRE, a relative geometry-aware framework that integrates evolved masking with cross-view prediction. Instead of reconstructing the input view, PointSiRE performs geometry-conditioned cross-view representation prediction within a

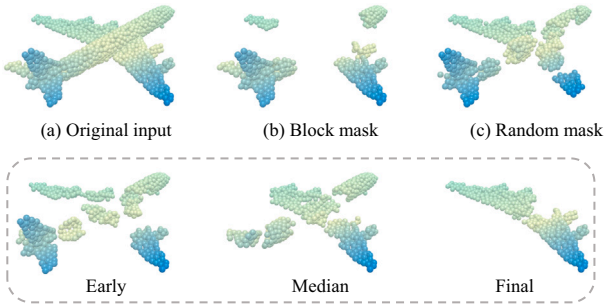

Fig. 1. Comparison of different masking strategies for point cloud pre-training. (a) original input point cloud. (b) block masking. (c) random masking. (d) the proposed attention-guided evolved hierarchical masking (AEHM) at different training stages (early, median, and final).

Siamese student–teacher architecture: the teacher provides stable, structure-aware targets, while the student is explicitly conditioned on the geometric relation between the two views, allowing it to reason about transformations rather than merely tolerate them. Concretely, relative transformations—scale, rotation, and translation—are explicitly encoded via a Relative Geometry-aware Position Encoding (RGPE) to enable spatial alignment between views. Guided by this geometry, the student predicts the teacher's latent features at corresponding masked positions, encouraging the model to reason about structural transformations. To adapt masking to object structure, an Attention-guided Evolved Hierarchical Masking (AEHM) progressively transitions from fine-grained to larger, semantically coherent regions, leveraging attention maps from the momentum teacher. For example, in Fig. 1(d), the evolved mask at later epochs covers large portions of the fuselage and tail while leaving the wing region unmasked. The framework is trained end-to-end with a UniGrad-based objective that aligns the student's predicted features with the teacher's latent representations at the corresponding masked positions while preventing representational collapse. Experiments on ScanObjectNN and few-shot classification confirm the effectiveness of this design in learning robust, geometry-aware representations.

In summary, the main contributions of this work are as follows:

- We propose PointSiRE, a Siamese self-supervised framework that addresses two limitations in masked point cloud pre-training—geometric transformations used merely as augmentation and static, content-agnostic masking—through geometry-conditioned supervision and structure-evolving masking.
- We introduce a Relative Geometry-aware Position Encoding (RGPE) that explicitly models the relative rotation, scaling, and translation between augmented views, serving as a geometric navigator for cross-view feature alignment.
- We propose an Attention-guided Evolved Hierarchical Masking (AEHM) strategy that leverages the teacher's attention to evolve masking from fine-grained to semantically coherent regions, progressively increasing pretext task difficulty.
- We conduct experiments across classification, few-shot learning, and dense prediction, demonstrating that PointSiRE achieves competitive or state-of-the-art performance across diverse downstream tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on self-supervised learning, geometric modeling, and masked representation learning for point clouds. Section 3 presents the proposed methodology, including the overall pre-training framework, structure-aware patch embedding, relative geometry-aware position encoding, evolved hierarchical masking, and the corresponding learning objectives. Section 4 presents the experimental results and ablation studies. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related work

2.1. Self-supervised learning for point clouds

Self-supervised learning (SSL) aims to learn transferable representations from unlabeled 3D data. Early reconstruction-based approaches train autoencoders to reconstruct full shapes or complete partial observations, yielding meaningful latent spaces for recognition and generative modeling [23]. Representative models like FoldingNet [8] reconstruct point clouds in deterministic or probabilistic manners. While effective for low-level geometry, these methods often overemphasize coordinate-level recovery, limiting semantic abstraction. Inspired by 2D vision, contrastive methods enforce feature consistency across views. PointContrast [9] utilizes point-level invariance, while CrossPoint [13] aligns 3D data with 2D images. Although variants like ReCon [24], CrossNet [25] and TriCI [26] improve alignment, contrastive paradigms typically face optimization challenges related to large batch sizes and hard-negative mining.

Recently, masked representation learning has become a dominant paradigm. Point-BERT [19] and Point-McBERT [27] adopt BERT-style masked token prediction with offline tokenizers, while MaskPoint formulates masked discrimination without explicit tokenization. Point-MAE [28] extends masked autoencoding to irregular point clouds and achieves strong generalization, with subsequent works such as Point-M2AE [29] and PCP-MAE [30] further improving hierarchical modeling, local enhancement, and pre-training strategies.

2.2. Geometric modeling and transformation-aware representation learning

A long-standing line of research focuses on achieving robustness to rotations and rigid transformations through rotation-invariant or rotation-equivariant representations. Prior work addresses this through rotation-invariant or equivariant representations via handcrafted operators (RIConv [31], RIConv++ [32]), and local reference frames. Recent methods like LGR-Net [33] combine local rotation-invariant geometry with global topology, while PaRot [34] achieves patch-wise invariance through feature disentanglement. SE(3)-equivariant neural networks explicitly incorporate group symmetries into architectures, providing theoretical guarantees for rigid transformation equivariance [17]. In masked modeling, RI-MAE [35] incorporates rotation-invariant embeddings into autoencoding. More recently, HFBRI-MAE [36] integrates handcrafted rotation-invariant local and global features into the masked autoencoding pipeline and further aligns the reconstruction target to a canonical pose, representing a recent instantiation of the invariance-pursuing paradigm: rotational variation is treated as a nuisance to be removed prior to reconstruction. However, most masked point cloud pre-training treats geometric transformations merely as data augmentation to encourage invariance, rarely integrating explicit cross-view relative relations into prediction. Unlike methods pursuing strict rotation invariance or explicit pose prediction, our method explicitly infers relative scale, rotation, and translation between augmented views, injecting them

as conditioning signals for masked prediction to enable transformation-aware cross-view consistency while preserving token-level correspondence.

2.3. Masked modeling and structure-aware pretext design

Masking strategy design plays a crucial role in masked representation learning, as it directly determines the difficulty and semantic nature of the pretext task. Most existing masked point cloud methods adopt static and content-agnostic masking strategies, such as random or block-wise masking. Building upon the foundational random masking strategy of Point-MAE [28], recent studies have primarily refined the paradigm from two dimensions. Regarding mask distribution, Point-M2AE [29] employs multi-scale masking and hybrid global-local masking strategies, respectively, to enhance the modeling of multi-level geometric features. In terms of masking mechanisms, PCP-MAE [30] identifies that explicitly providing masked patch centers allows the decoder to bypass the encoder via location cues; thus, it proposes predicting these centers to enforce semantic learning. Moreover, many masked point cloud approaches couple masking with low-level coordinate reconstruction objectives, which emphasize geometric recovery rather than high-level structural or semantic consistency.

In the 2D vision domain, [37] proposes constructing hierarchical structures from model responses and progressively evolving masking patterns from fine-grained textures to semantically coherent regions during training. Siamese Image Modeling further demonstrates that dense masked prediction can be enhanced by explicitly aligning representations across views using relative positional information [38]. Unlike Siamese Image Modeling, which mainly considers relative positional relationships induced by 2-D image cropping operations, PointSiRE addresses the more challenging geometric variations in 3-D point clouds, where transformations involve coupled rotation, translation, and scaling. Therefore, instead of encoding simple spatial offsets, PointSiRE explicitly derives the relative transformation between two views and represents rotation through quaternion-based encoding, enabling geometry-conditioned masked prediction under general 3-D transformations. Within the point cloud domain, recent efforts have also begun to explore structured and progressive masking. Beyond Random Masking [39] combines 3D spatial grid masking with attention-driven semantic masking through a dynamically weighted curriculum that shifts emphasis during training. Sonata [40] applies a progressive scheduler over mask size and ratio to mitigate the “geometric shortcut” in point self-distillation. Compared with Beyond Random Masking [39], which combines spatial grid masking with attention-guided weighting, AEHM constructs a hierarchical clustering tree based on teacher attention similarity and dynamically adjusts the cluster granularity through epoch-aware hierarchy cutting. Compared with Sonata [40], which mainly increases masking difficulty by scheduling mask size and ratio, AEHM keeps the masking ratio unchanged and evolves only the structural organization of masked regions. These advances suggest that masking should not be treated as a fixed heuristic, but rather as an integral part of pretext task design that can co-evolve with the learned representations [41]. Motivated by these insights, PointSiRE introduces an Attention-guided Evolved Hierarchical Masking (AEHM) strategy for point clouds, where the teacher’s self-attention induces an adaptive hierarchy over patches and the masking pattern evolves along training epochs. Combined with explicit relative geometric conditioning, this design aligns the masked pretext task with both the structural semantics of point clouds and their intrinsic geometric transformations.

3. Method

3.1. Overall framework

We aim to learn point cloud representations that are aware of geometric transformations and intrinsic structural semantics via masked representation learning with multi-view consistency. To this end, we propose a geometry-conditioned masked representation learning framework for point clouds. As illustrated in Fig. 2, the framework integrates two key designs: Relative Geometry-aware Position Encoding (RGPE) for explicit cross-view alignment, and Attention-guided Evolved Hierarchical Masking (AEHM) for structure-consistent masked modeling.

Given an input point cloud $\mathcal{P}$, two augmented views $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are generated by applying stochastic geometric transformations (scaling, rotation, and translation). To ensure token-level correspondence across views, farthest point sampling (FPS) is applied once on the original point set. The resulting center indices are shared between both views. Neighborhoods are then constructed around the same centers. This produces G paired local patches with one-to-one correspondence between $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$. A mini-PointNet patch embedding network maps each patch to a latent token, yielding token sequences $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$. G denotes the number of local patches, d denotes the token embedding dimension, and $\mathbf{Z}^{(1)}, \mathbf{Z}^{(2)} \in \mathbb{R}^{G\times d}$ denote the token sequences of the two views.

The teacher network processes $\mathbf{Z}^{(1)}$ and outputs high-level target representations, which provide supervision for pre-training. The student network takes $\mathbf{Z}^{(2)}$ as input and, guided by the explicit relative geometric encoding, predicts the momentum teacher’s latent representations corresponding to the masked positions of view 1.

The relative transformation from view 2 to view 1 is explicitly computed and encoded as a pose embedding. This embedding is added to both visible tokens and learnable mask tokens before the predictor. Since the two views share the same FPS centers, token slots are index-aligned across views. The student predictions on the masked slots are thus supervised using the corresponding momentum-teacher targets from view 1.

To avoid the limitations of random or block-wise masking, we adopt an evolved hierarchical masking strategy driven by the teacher’s self-attention. Specifically, the last-layer attention map of the teacher encoder induces a hierarchical grouping over tokens, from which an epoch-dependent mask is generated. As training proceeds, the masking pattern evolves toward more structured, cluster-consistent masking while keeping a fixed mask ratio.

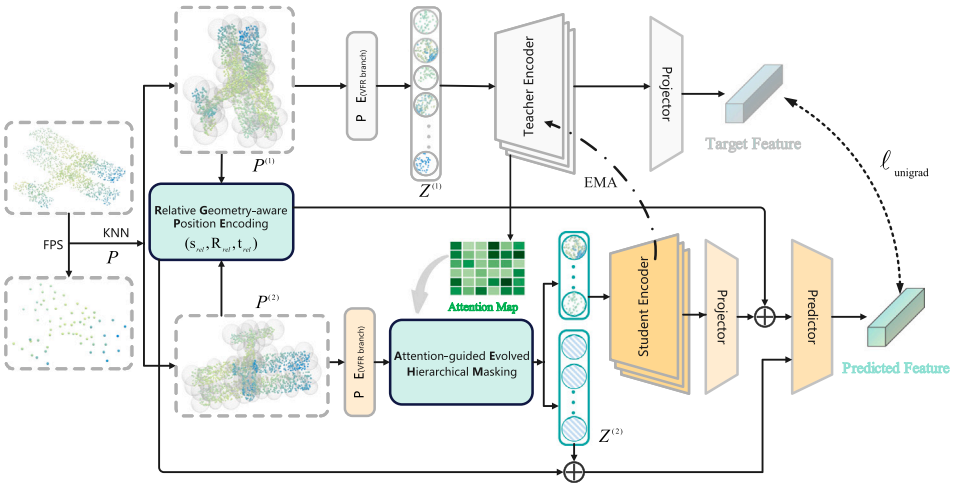

Fig. 2. Overview of the proposed pre-training framework. Given an input point cloud $\mathcal{P}$, two augmented views $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are generated via stochastic transformations. Shared FPS centers yield G paired local patches embedded into tokens. The teacher encodes view 1 tokens and produces target features, while its self-attention map drives AEHM to generate structured masks. The student encodes visible tokens from view 2, adds relative geometry-aware encodings derived from $(s_{\text{rel}}, \mathbf{R}_{\text{rel}}, \mathbf{t}_{\text{rel}})$, and predicts the momentum teacher's latent representations from view 1 at the corresponding masked token slots via the UniGrad loss. The teacher is updated by EMA of the student parameters.

The teacher is updated by exponential moving average (EMA) of the student parameters, producing stable and gradually refined targets. From a representation learning perspective, PointSiRE learns geometry-aware and structure-aware representations through geometry-conditioned cross-view prediction and evolved hierarchical masking. Through the synergy of relative-geometry conditioning and evolved hierarchical masking, the proposed framework encourages the learned representations to preserve geometric correspondence while maintaining semantic consistency across views, thereby learning transformation-aware and structure-sensitive point cloud representations.

3.2. Structure-aware patch embedding

To obtain patch-level representations that preserve local geometric structures, we adopt a structure-aware patch embedding scheme to convert raw point clouds into token sequences. Given the input point cloud $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^{N}$, where N denotes the total number of points, G patch centers $\{\mathbf{c}_j\}_{j=1}^{G}$ are sampled by farthest point sampling (FPS). The FPS indices are shared across augmented views to preserve one-to-one patch correspondence. For each shared patch center $\mathbf{c}_j$, a local patch $\mathcal{G}_j = \{\mathbf{p}_{j,l}\}_{l=1}^{K}$ is constructed by K-nearest-neighbor (KNN) grouping, where K denotes the number of points in each local patch and $\mathbf{p}_{j,l}$ is the l-th neighboring point associated with center $\mathbf{c}_j$. To better capture local geometry, we introduce a lightweight Variation-aware Feature Refinement (VFR) branch into the Mini-PointNet encoder. Specifically, VFR extracts intra-patch variance features, which are concatenated with standard global and point-wise features. The aggregated features are projected to form the token sequence $\mathbf{Z} = \{\mathbf{z}_j\}_{j=1}^{G}$.

3.3. Relative geometry-aware position encoding (RGPE)

1) Relative Transformation Derivation. Two augmented views $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ are generated from the input point cloud $\mathcal{P}$ via stochastic geometric transformations. For view $k \in \{1, 2\}$, we denote the matrix form of the input point cloud as $\mathbf{P} = [\mathbf{p}_1; \cdots; \mathbf{p}_N] \in \mathbb{R}^{N\times 3}$. The transformed point coordinates are defined as

$$\mathbf{V}^{(k)} = s_k \, \mathbf{P} \, \mathbf{R}_k^{\top} + \mathbf{t}_k^{\top}, \tag{1}$$

where $\mathbf{V}^{(k)} \in \mathbb{R}^{N\times 3}$ denotes the coordinate matrix of view k, $s_k \in \mathbb{R}^{+}$ is the isotropic scaling factor, $\mathbf{R}_k \in SO(3)$ is the rotation matrix, and $\mathbf{t}_k \in \mathbb{R}^3$ is the translation vector, which is broadcast to all points during the transformation.

Since the student network takes view 2 as input to predict the teacher's latent representations of view 1, it must explicitly reason about the spatial misalignment between the two views. To bridge this gap, we derive the exact relative geometric transformation $(s_{\text{rel}}, \mathbf{R}_{\text{rel}}, \mathbf{t}_{\text{rel}})$ that aligns the coordinate system of view 2 to view 1:

$$\begin{cases} s_{\text{rel}} = s_1 / s_2, \\ \mathbf{R}_{\text{rel}} = \mathbf{R}_1 \mathbf{R}_2^{\top}, \\ \mathbf{t}_{\text{rel}} = \mathbf{t}_1 - \left(s_{\text{rel}} \, \mathbf{t}_2\right) \mathbf{R}_{\text{rel}}^{\top}. \end{cases} \tag{2}$$

where s_{rel}, $\mathbf{R}_{\text{rel}}$, and $\mathbf{t}_{\text{rel}}$ denote the relative scale, rotation, and translation that transform coordinates from view 2 to view 1, respectively. Accordingly, a point coordinate $\mathbf{x}^{(2)}$ in view 2 can be mapped to the coordinate system of view 1 as

$$\mathbf{x}^{(1)} = s_{\text{rel}} \, \mathbf{x}^{(2)} \mathbf{R}_{\text{rel}}^{\top} + \mathbf{t}_{\text{rel}}, \tag{3}$$

where $\mathbf{x}^{(2)}, \mathbf{x}^{(1)} \in \mathbb{R}^3$ denote the same point represented in the coordinate systems of view 2 and view 1, respectively.

Each component of this relative transformation follows directly from the composition of rigid-body transformations. Since $\mathbf{R}_k$ and $\mathbf{t}_k$ describe the transformation from the original point cloud to view k, composing the inverse of view 2's transformation with view 1's transformation yields the rotation $\mathbf{R}_{\text{rel}} = \mathbf{R}_1\mathbf{R}_2^\top$ and the scale ratio $s_{\text{rel}} = s_1/s_2$. Crucially, $\mathbf{t}_{\text{rel}}$ is not simply $\mathbf{t}_1 - \mathbf{t}_2$: because the coordinate origin itself is first rescaled by s_{rel} and rotated by $\mathbf{R}_{\text{rel}}$ before translation is applied, the offset term $(s_{\text{rel}}\mathbf{t}_2)\mathbf{R}_{\text{rel}}^\top$ explicitly compensates for this coupling, ensuring that $(s_{\text{rel}}, \mathbf{R}_{\text{rel}}, \mathbf{t}_{\text{rel}})$ exactly maps view 2's coordinate system onto view 1's, rather than merely approximating it. These parameters thus provide explicit geometric conditions, enabling the student network to perform geometry-aware spatial alignment in the feature space rather than enforcing static invariance.

2) Geometric Representation and Encoding. To inject these geometric priors into the Transformer architecture, we propose the Relative Geometry-aware Position Encoding (RGPE). For numerical stability and representational compactness, the relative rotation $\mathbf{R}_{\text{rel}}$ is first converted into a unit quaternion $\mathbf{q}_{\text{rel}} \in \mathbb{R}^4$, which avoids the redundant nine-parameter, orthogonality-constrained representation of $\mathbf{R}_{\text{rel}}$ and provides a compact, continuous parameterization suitable for sinusoidal encoding.

Let $\mathbf{c}_j^{(2)}$ denote the absolute patch center coordinates in view 2. We concatenate $\mathbf{c}_j^{(2)}$ with the relative transformation parameters to form a unified geometric tuple $(\mathbf{c}_j^{(2)}, \log s_{\text{rel}}, \mathbf{t}_{\text{rel}}, \mathbf{q}_{\text{rel}})$. Multi-frequency sinusoidal functions are applied to this tuple to preserve spatial continuity and periodicity. The resulting embedding is linearly projected to match the token dimension and is added to both visible and learnable mask tokens in the student branch. This explicit RGPE acts as a geometric navigator, enabling the predictor to perform transformation-aware cross-view feature alignment. Unlike conventional positional encodings that mainly provide spatial location cues, RGPE embeds relative geometric relationships directly into token representations. Consequently, each token is represented not only by its local geometry but also by its geometric state relative to another view, allowing the model to learn transformation-aware representations that preserve geometric correspondence while remaining robust to view variations.

3.4. Transformer-based student-teacher learning with attention-guided evolved hierarchical masking (AEHM)

1) Transformer Encoder Backbone. Both student and teacher networks utilize standard Transformer encoders to aggregate global context via multi-head self-attention (MSA). Given the patch token sequence $\mathbf{Z} = \{\mathbf{z}_j\}_{j=1}^G$, positional encodings are added and fed into a stack of Transformer encoder layers. The teacher provides stable target representations and structural guidance (via attention maps); its parameters are updated via an exponential moving average (EMA) of the student parameters.

2) Attention-guided Evolved Hierarchical Masking (AEHM). To align the pretext task with intrinsic object structures, we propose AEHM (illustrated in Fig. 3). Let $\mathbf{A}^{(h)} \in \mathbb{R}^{G\times G}$ denote the patch-to-patch attention matrix of the h-th attention head in the last teacher layer, where each element $\mathbf{A}_{ij}^{(h)}$ measures the attention affinity from patch i to patch j. The attention maps from all H heads are averaged to obtain a global patch similarity matrix $\mathbf{A} \in \mathbb{R}^{G\times G}$:

$$\mathbf{A} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{A}^{(h)}. \tag{4}$$

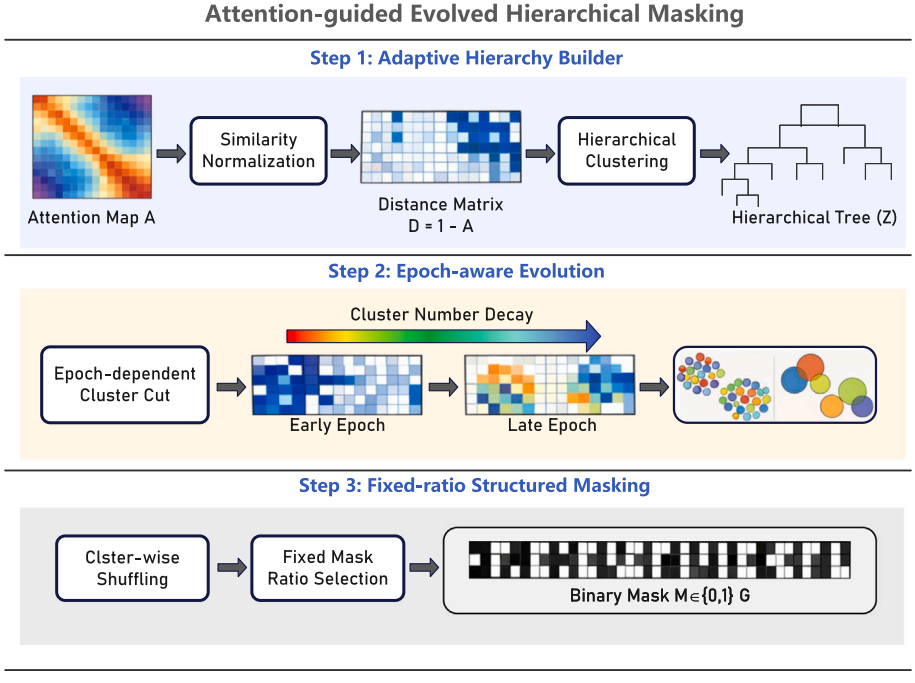

Fig. 3. Illustration of the proposed attention-guided evolved hierarchical masking (AEHM).

Algorithm 1 Attention-guided evolved hierarchical masking (AEHM).

Require: Patch tokens $\mathbf{Z} = \{\mathbf{z}_j\}_{j=1}^{G}$; teacher last-layer attention maps $\{\mathbf{A}^{(h)}\}_{h=1}^{H}$; mask ratio ρ; current epoch e; total epochs E; minimum cluster number $K_{\min}$.
Ensure: Binary mask $\mathbf{m} \in \{0,1\}^G$.
$\mathbf{A} \leftarrow \frac{1}{H}\sum_{h=1}^{H}\mathbf{A}^{(h)}$
$\mathbf{D}_{ij} \leftarrow 1 - \frac{\mathbf{A}_{ij}-\min(\mathbf{A})}{\max(\mathbf{A})-\min(\mathbf{A})+\epsilon}$
Construct a hierarchical clustering tree using agglomerative clustering with average linkage based on $\mathbf{D}$
$K(e) \leftarrow \max\left(K_{\min},\ G-(G-K_{\min})\cdot\frac{e}{E}\right)$
Cut the hierarchy to obtain $K(e)$ clusters and assign labels $\{c_j\}_{j=1}^{G}$
$\mathbf{m} \leftarrow \mathbf{0}, \quad N_m \leftarrow \lfloor \rho G \rfloor$
Randomly permute clusters and select patches cluster by cluster until $\sum_j \mathbf{m}_j = N_m$
return m

To construct a hierarchical clustering tree via agglomerative clustering (average linkage), we convert $\mathbf{A}$ into a distance matrix $\mathbf{D}$ via min-max normalization:

$$\mathbf{D}_{ij} = 1 - \frac{\mathbf{A}_{ij} - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A}) + \epsilon}, \tag{5}$$

where $\mathbf{D}_{ij}$ denotes the normalized distance between patches i and j, and ϵ is a small constant used for numerical stability. This hierarchy captures multi-scale structural relations among point patches.

Evolved Mask Generation: We introduce an epoch-aware strategy to control the masking granularity. Let e and E denote the current and total epochs. We define a dynamic target cluster number $K(e)$ that decays linearly:

$$K(e) = \max\left(K_{\min},\ G - (G - K_{\min}) \cdot \frac{e}{E}\right). \tag{6}$$

where $K_{\min}$ is the predefined minimum number of clusters. By cutting the hierarchy at $K(e)$ clusters, we obtain a grouping that evolves from scattered (large K) to structurally coherent (small K). We then randomly select entire clusters until the number of masked patches reaches $N_m = \lfloor \rho G \rfloor$, where N_m denotes the target number of masked patches and $\rho \in (0,1)$ is the mask ratio. This process yields a binary mask $\mathbf{m} \in \{0,1\}^G$. The masked and visible index sets are defined as $\mathcal{M} = \{j \mid \mathbf{m}_j = 1\}$ and $\mathcal{V} = \{j \mid \mathbf{m}_j = 0\}$, respectively. The detailed procedure is summarized in Algorithm 1.

Note that the evolution affects only the spatial structure of masked regions rather than the mask ratio or loss weighting. Compared with conventional random masking strategies that independently remove points without considering underlying structures, AEHM preserves structural continuity by progressively masking semantically coherent regions guided by the teacher's attention distribution. From a representation learning perspective, AEHM serves as a curriculum mechanism that gradually increases the structural reasoning requirement during pre-training, shifting the learning objective from local geometric completion toward high-level semantic understanding. During early training stages, fine-grained masking encourages the acquisition of local geometric patterns from visible contexts, while later structured masking introduces more coherent occlusions that require the model to infer semantic object organization and long-range dependencies. This progressive evolution enables the model to learn richer structural priors and more transferable representations.

3) Relative Geometry-aware Masked Prediction. Learnable mask tokens are introduced for masked positions. Relative geometric encodings derived in Section 3.3 are added to both visible and masked tokens. The predictor, implemented as a deeper Transformer in the student network, takes the combined sequence as input. Let $\mathbf{H}_{\text{vis}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ denote the encoded visible student tokens, and let $\mathbf{T}_{\text{mask}} \in \mathbb{R}^{|\mathcal{M}|\times d}$ denote the learnable mask tokens inserted at the masked positions. The relative geometry-aware encodings for visible and masked positions are denoted as $\mathbf{E}_{\text{rel}}^{\text{vis}} \in \mathbb{R}^{|\mathcal{V}|\times d}$ and $\mathbf{E}_{\text{rel}}^{\text{mask}} \in \mathbb{R}^{|\mathcal{M}|\times d}$, respectively. The predictor input is constructed as

$$\mathbf{X}_{\text{pred}} = \left[\mathbf{H}_{\text{vis}} + \mathbf{E}_{\text{rel}}^{\text{vis}},\ \mathbf{T}_{\text{mask}} + \mathbf{E}_{\text{rel}}^{\text{mask}}\right], \tag{7}$$

where $\mathbf{X}_{\text{pred}}$ denotes the predictor input sequence. Here, $[\cdot,\cdot]$ denotes concatenation along the token dimension. The concatenated sequence is fed into the student predictor, which produces predictions for the masked patches in the token space of view 1. These predictions are explicitly conditioned on the encoded relative transformation from view 2 to view 1, and are supervised by the corresponding momentum-teacher representations from view 1 at the same token slots, thereby enforcing cross-view consistency conditioned on relative geometric transformations.

3.5. Training objective and optimization

Let $\hat{\mathbf{y}}_j \in \mathbb{R}^d$ denote the student prediction for masked patch $j \in \mathcal{M}$ after the predictor, and let $\mathbf{y}_j^{(1)} \in \mathbb{R}^d$ denote the corresponding target feature extracted by the momentum teacher from view 1 at the same token slot. To balance representation alignment and prevent feature collapse, the training loss adopts a UniGrad formulation consisting of two complementary terms:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda\,\mathcal{L}_{\text{decorr}}, \tag{8}$$

where λ controls the strength of the decorrelation regularization. In our implementation, λ corresponds to the decorrelation weight `neg_weight` in the UniGrad loss, which is set to 0.02 by default during pre-training.

The alignment term is defined as

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \left\| \hat{\mathbf{y}}_j - \mathbf{y}_j^{(1)} \right\|_2^2. \tag{9}$$

To prevent representation collapse, a decorrelation regularization term is introduced. Before computing the feature correlation matrix, teacher features are ℓ_2-normalized along the feature dimension. The correlation matrix $\mathbf{C} \in \mathbb{R}^{d \times d}$ is estimated from the normalized teacher representations:

$$\mathbf{C} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \mathbf{y}_j^{(1)} \mathbf{y}_j^{(1)\top}. \tag{10}$$

where $\mathbf{C}_{ij}$ measures the correlation between the i-th and j-th feature dimensions.

The decorrelation loss encourages orthogonality among feature dimensions:

$$\mathcal{L}_{\text{decorr}} = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} \hat{\mathbf{y}}_j^{\top} \mathbf{C}\, \hat{\mathbf{y}}_j. \tag{11}$$

The teacher target $\mathbf{y}_j^{(1)}$ is treated as a stop-gradient target, and gradients are back-propagated only through the student network. The teacher parameters θ_t are updated as an exponential moving average (EMA) of the student parameters θ_s:

$$\theta_t \leftarrow \lambda_{\text{ema}}\, \theta_t + (1 - \lambda_{\text{ema}})\, \theta_s. \tag{12}$$

where $\lambda_{\text{ema}} \in [0, 1)$ is the EMA momentum coefficient, which is set to 0.996 in our experiments. A larger λ_{ema} makes the teacher evolve more slowly and provides more stable targets for self-supervised training.

The entire framework is optimized end-to-end using the AdamW optimizer. By combining explicitly geometry-conditioned masked prediction with alignment and decorrelation objectives, the proposed training strategy enables stable self-supervised learning without contrastive sampling.

4. Experiments

We conduct all experiments on raw point cloud data without converting to mesh or voxel representations. Each point cloud is normalized by centering it at the origin and scaling it into a unit sphere. To construct cross-view supervision signals, we generate two augmented views from the same input using random isotropic scaling, translation, and probabilistic SO(3) rotation ($p = 0.5$). Importantly, we establish explicit token-level correspondence between the two views by performing farthest point sampling (FPS) once before augmentation, ensuring shared geometric anchors across views. Local patch features are then constructed via K-nearest neighbors (KNN) in each transformed view, enabling consistent local structure comparison under different geometric transformations. This design ensures that the learning objective focuses on relative geometric reasoning rather than view-specific appearance.

4.1. Pre-training settings

We pre-train PointSiRE on the ShapeNet55 dataset [42], which contains approximately 52,500 clean 3D models covering 55 categories. The pre-training is conducted using a self-supervised learning paradigm without any category labels. All experiments utilize 8192 points sampled from each object via Farthest Point Sampling (FPS).

PointSiRE employs a Siamese student–teacher architecture, where the teacher network is updated via an exponential moving average (EMA) of the student weights with a momentum of 0.996. To construct the pretext task, we employ the proposed *Attention-guided Evolved Hierarchical Masking (AEHM)* strategy with a masking ratio of 60%. A key component of our framework is the relative geometry-aware modeling. We apply random scaling in the range of $[0.8, 1.2]$, translation in $[-0.2, 0.2]$, and probabilistic rotation to generate augmented views. Specifically, rotation augmentation is applied with a probability of 0.5, forcing the model to explicitly reason about relative geometric transformations rather than merely memorizing canonical poses.

The network is optimized using the AdamW optimizer with a weight decay of 0.05 and an initial learning rate of 1×10^{-3}, following a cosine learning rate decay schedule. We set the batch size to 128 and train for 300 epochs. All experiments are implemented in PyTorch and conducted on a single NVIDIA TITAN RTX GPU. After pre-training, the student encoder weights are transferred to downstream tasks and fine-tuned end-to-end.

4.2. Object classification on synthetic and real-world data

We evaluate the transferability and robustness of the learned representations on both synthetic (ModelNet40 [43]) and real-world (ScanObjectNN [44]) benchmarks. For fair comparison, Point-MAE results are reproduced under the exact same settings.

Synthetic Data: On ModelNet40, we follow standard fine-tuning and voting protocols. As shown in Table 1, PointSiRE achieves an accuracy of 93.7%, outperforming representative baselines such as Point-BERT (93.2%) and Point-MAE (93.3%) under identical settings. Although ModelNet40 objects are largely canonically aligned and thus exhibit limited global pose variation, PointSiRE still provides consistent gains over existing masked modeling methods. This indicates that the improvement does not solely stem from

Table 1
Accuracy (%) of synthetic object classification on ModelNet40 (8 K points). "FSL" and "SSL" indicate fully-supervised and self-supervised methods, respectively. "ST" denotes the standard transformer architecture.

Method	ST	FS/SSL	Acc
PointNet [2]	–	FSL	89.2
PointNet++ [3]	–	FSL	90.7
PointCNN [5]	–	FSL	92.5
DGCNN [4]	–	FSL	92.9
KPConv [6]	–	FSL	92.9
Point Transformer [7]	N	FSL	92.8
DGCNN + FoldingNet [8]	–	SSL	93.1
DGCNN + STRL [11]	–	SSL	93.1
DGCNN + OcCo [12]	–	SSL	93.0
Point-BERT [19]	–	SSL	93.2
Point-LGMask [22]	Y	SSL	93.1 ± 0.1
Point-MAE [28]	Y	SSL	93.3 ± 0.1
Point-MPP [21]	Y	SSL	93.4 ± 0.2
PointSiRE (Ours)	Y	SSL	**93.7 ± 0.1**

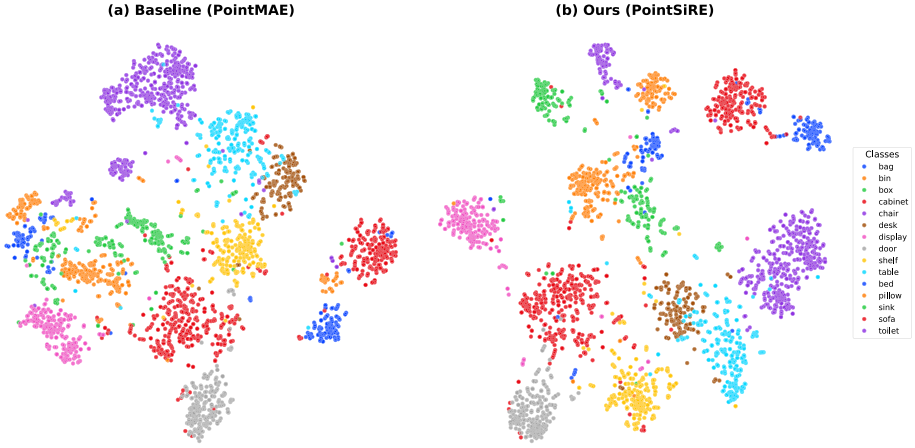

Fig. 4. T-SNE visualization of learned representations on the ScanObjectNN dataset (PB_T50_RS split). Compared to the baseline point-MAE (a), our PointSiRE (b) produces more compact clusters with clearer separation boundaries between semantically similar classes, demonstrating the robustness of the learned features.

handling rotation invariance, but rather from learning more structured geometric dependencies through explicit relative transformation conditioning. In contrast to methods that rely on implicit invariance learning, our approach forces the model to reason about cross-view geometric relationships even in near-aligned settings, which leads to more robust encoding of intrinsic object structure. This demonstrates that RGPE improves representation quality beyond simply addressing pose variation.

Real-scanned Data: To assess robustness against background noise and non-canonical poses, we conduct experiments on ScanObjectNN. We report results on three variants (*OBJ_BG*,*OBJ_ONLY*, and *PB_T50_RS*) and also evaluate a version with simple rotation augmentation (denoted by †). As detailed in Table 2, PointSiRE demonstrates superior performance across all metrics. In the standard setting, our method achieves 92.25% on *OBJ_BG*, significantly surpassing Point-MAE (88.46%). When rotation augmentation is applied, PointSiRE establishes new state-of-the-art results with 94.15%, 92.43%, and 88.45% on the three splits, respectively. Compared to Point-MAE†, consistent improvements (+0.6% to +1.4%) validate that our geometry-aware pre-training equips the model with a deeper spatial understanding to handle real-world perturbations effectively. Notably, the performance gap between PointSiRE and Point-MAE is substantially larger in the standard setting (e.g., +3.79% on OBJ_BG) than in the rotation-augmented setting (+1.38% on OBJ_BG). This suggests that a significant portion of robustness is already acquired during pre-training through explicit geometric conditioning, reducing reliance on downstream augmentation strategies. We attribute this trend to the complementary roles of pre-training and fine-tuning in acquiring rotation robustness. Since RGPE explicitly conditions the pre-training objective on the relative geometric transformation between views, PointSiRE acquires a degree of rotational reasoning ability directly during pre-training. In the standard fine-tuning setting, where no explicit rotation augmentation is introduced, this pre-trained geometric reasoning capability becomes the primary source of robustness against the arbitrary orientations present in ScanObjectNN, which explains the larger margin over Point-MAE in this setting. When rotation augmentation is additionally introduced during fine-tuning, Point-MAE can partially compensate for its lack of geometric conditioning during pre-training by directly observing rotated samples, narrowing the gap; however, PointSiRE still retains a consistent advantage, indicating that the geometric reasoning ability acquired during pre-training and the robustness gained from fine-tuning-time augmentation are complementary rather than substitutive.

Table 2
Accuracy (%) of real-world object classification on ScanObjectNN. † indicates that a simple rotation augmentation strategy is used during fine-tuning.

Method	OBJ_BG	OBJ_ONLY	PB_T50_RS
PointNet [2]	73.3	79.2	68.0
PointNet++ [3]	82.3	84.3	77.9
DGCNN [4]	82.8	86.2	78.1
PointCNN [5]	86.1	85.5	78.5
PointMLP [45]	–	–	85.2
Point-BERT [19]	87.43	88.12	83.07
Point-McBERT [27]	89.00	90.00	84.30
Point-LGMask [22]	89.80	89.30	85.30
Point-MAE [28]	88.46	87.95	84.90
Point-MAE [28]†	92.77	91.22	87.82
Point-MPP [21]	90.90	88.60	84.10
PointSiRE (Ours)	92.25	91.04	85.74
PointSiRE (Ours)†	**94.15**	**92.43**	**88.45**

Table 3
Accuracy (%) of few-shot object classification on the ModelNet40 dataset.

Methods	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
PointNet [2]	52.0 ± 3.8	57.8 ± 4.9	46.6 ± 4.3	35.2 ± 4.8
PointNet++ [3]	38.5 ± 4.4	42.4 ± 4.5	23.1 ± 2.2	18.8 ± 1.7
PointCNN [5]	65.4 ± 2.8	68.6 ± 2.2	46.6 ± 1.5	50.0 ± 2.3
DGCNN [4]	31.6 ± 2.8	40.8 ± 4.6	19.9 ± 2.1	16.9 ± 1.5
FoldingNet [8]	33.4 ± 4.1	35.8 ± 5.8	18.6 ± 1.8	15.4 ± 2.2
DGCNN + OcCo [12]	90.6 ± 2.8	92.5 ± 1.9	82.9 ± 1.3	86.5 ± 2.2
Point-BERT [19]	94.6 ± 3.1	96.3 ± 2.7	91.0 ± 5.4	92.7 ± 5.1
Point-MAE [28]	96.3 ± 2.5	97.8 ± 1.8	$\mathbf{92.6 \pm 4.1}$	95.0 ± 3.0
Point-MPP [21]	96.6 ± 2.2	$\mathbf{97.9 \pm 1.7}$	92.5 ± 5.0	$\mathbf{95.1 \pm 3.3}$
PointSiRE (Ours)	$\mathbf{97.2 \pm 2.0}$	$\mathbf{97.9 \pm 1.8}$	92.1 ± 5.0	$\mathbf{95.1 \pm 3.3}$

Overall, these results validate that RGPE provides an explicit mechanism for modeling pose-dependent feature alignment, which is particularly beneficial in real-world scenarios where objects exhibit arbitrary orientations, occlusions, and background clutter.

Feature Visualization: We further visualize the feature distributions of the challenging *PB_T50_RS* split using t-SNE in Fig. 4. Compared to the baseline Point-MAE, the representations learned by PointSiRE exhibit more compact intra-class clustering and clearer inter-class separation. This confirms that our framework generates more discriminative features, effectively mitigating the entanglement often caused by complex real-world noise and geometric variations. This qualitative observation is consistent with the quantitative results in Table 2, where *PB_T50_RS* is the split exhibiting the most severe background clutter and pose variation, and is also where the standard-setting margin over Point-MAE (+0.84%) is accompanied by the largest gain after rotation augmentation (+0.63%) among the three splits, corroborating that the improved feature separability visualized here translates directly into the classification gains reported above.

4.3. Few-shot object classification

We further evaluate the data efficiency and generalization capability of our model through few-shot classification on ModelNet40. We conduct few-shot experiments on ModelNet40 following the standard K-way N-shot protocol ($K \in \{5, 10\}, N \in \{10, 20\}$). The results are detailed in Table 3. PointSiRE demonstrates robust generalization under data-scarce conditions. PointSiRE achieves the best result in the *5-way 10-shot* setting (97.2%), surpassing Point-MAE (96.3%) and Point-MPP (96.6%) by 0.9% and 0.6%, respectively, and matches or slightly exceeds the strongest baselines in the *5-way 20-shot* and *10-way 20-shot* settings (97.9% and 95.1%). In the *10-way 10-shot* setting, PointSiRE obtains $92.1 \pm 5.0\%$, nominally 0.4–0.5% below Point-MAE ($92.6 \pm 4.1\%$) and Point-MPP ($92.5 \pm 5.0\%$). Given that the episode-level standard deviation in this setting (4–5%) is substantially larger than the gap between methods, the performance differences among the three methods are small relative to the observed episode-level variability. Overall, PointSiRE performs on par with the strongest existing self-supervised baselines across all four few-shot settings, with a measurable edge in the most data-scarce configuration. These results indicate that the representations learned by PointSiRE transfer effectively to low-data regimes, without a systematic advantage or disadvantage emerging across different shot/way configurations.

Table 4
Part segmentation results on the ShapeNetPart dataset.

Method	$mIoU_c$	$mIoU_I$	aero	bag	cap	car	chair	earph.	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skateb.	table
PointNet [2]	80.4	83.7	83.4	78.7	82.5	74.9	89.6	73.0	91.5	85.9	80.8	95.3	65.2	93.0	81.2	57.9	72.8	80.6
PointNet++ [3]	81.9	85.1	82.4	79.0	87.7	77.3	90.8	71.8	91.0	85.9	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
DGCNN [4]	82.3	85.2	84.0	83.4	86.7	77.8	90.6	74.7	91.2	87.5	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
Transformer [47]	83.4	85.1	85.4	82.9	87.7	78.8	90.5	80.8	91.1	87.7	85.3	95.6	73.9	94.9	83.5	61.2	74.9	80.6
Point-BERT [19]	84.1	85.6	84.3	84.8	88.0	79.8	91.0	81.7	91.6	87.9	85.2	95.6	75.6	94.7	84.3	63.4	76.3	81.5
Point-MAE [28]	84.2	86.1	84.3	85.0	88.3	80.5	91.3	78.5	92.1	87.4	86.1	96.1	75.2	94.6	84.7	63.5	77.1	82.4
Point-MPP [21]	84.7	86.1	84.8	84.7	89.0	81.2	91.5	80.8	91.9	87.5	86.0	96.1	77.3	94.9	85.9	63.5	77.6	81.7
PointSiRE (Ours)	84.7	86.1	84.6	85.1	89.4	81.3	91.7	80.6	92.0	87.9	86.1	96.3	76.9	94.9	85.3	64.5	76.1	81.9

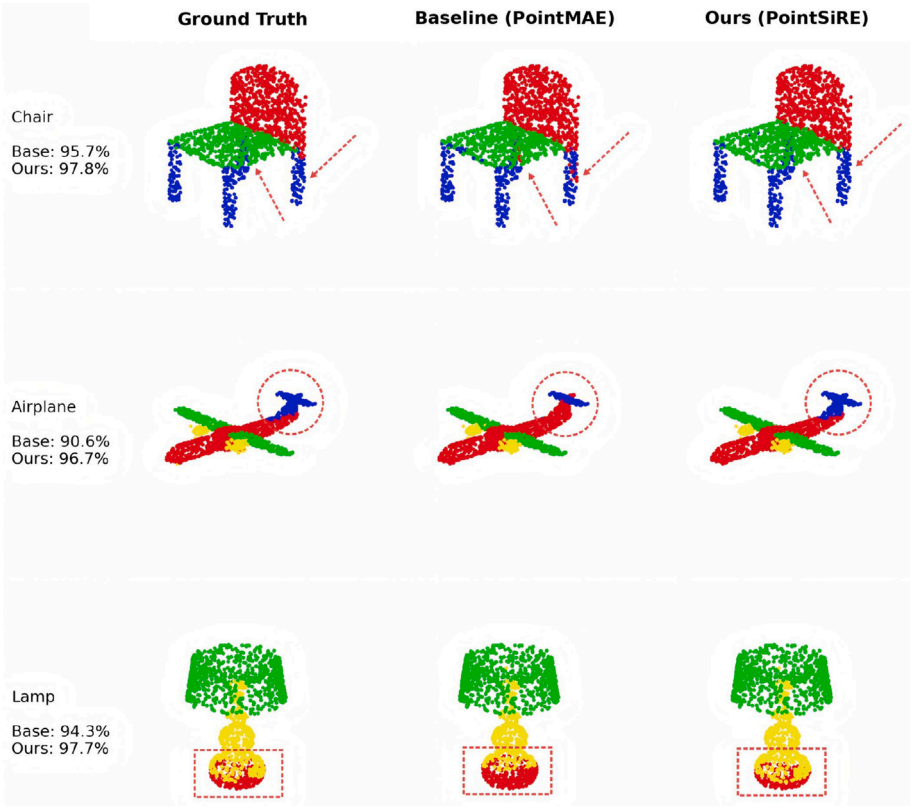

Fig. 5. Qualitative results of object part segmentation on ShapeNetPart. Red dashed boxes indicate regions where PointSiRE generates more precise predictions than the baseline. Our method effectively captures fine-grained geometric details, such as the legs of the chair and the tail of the airplane, which are blurred or misclassified by point-MAE.

4.4. Object part segmentation on ShapeNetPart

We assess the capability of PointSiRE in fine-grained dense prediction tasks using the ShapeNetPart dataset [46]. This dataset contains 16,881 objects from 16 categories, annotated with 50 part labels. Following standard protocols, we sample 2048 points for each object and use the mean Intersection-over-Union (mIoU) for class-wise ($mIoU_c$) and instance-wise ($mIoU_I$) evaluation.

Table 4 reports the quantitative results. PointSiRE achieves an instance mIoU of 86.1% and a class mIoU of 84.7%. These results are highly competitive, matching the performance of Point-MPP and outperforming the Point-MAE baseline (84.2% $mIoU_c$). PointSiRE attains the highest IoU among all compared methods on 7 of the 16 categories, including *Cap*, *Chair*, and *Rocket*, which contain fine-grained boundaries involving multiple spatially related components. On the remaining categories, performance is comparable to or marginally below the strongest baseline. Overall, these results indicate that PointSiRE performs on par with the best existing self-supervised method on this benchmark, with its more pronounced advantages observed on ScanObjectNN, where larger pose and background variation give the geometry-aware pre-training more room to contribute. This indicates that our pre-training strategy not only learns global semantic features but also preserves local geometric details essential for dense prediction.

Qualitative Analysis on Part Segmentation: To intuitively demonstrate the fine-grained understanding of PointSiRE, we visualize part segmentation results on ShapeNetPart in Fig. 5. Compared to the baseline Point-MAE, our method exhibits significantly sharper boundaries and higher semantic consistency, particularly in transitional regions. For instance, in the *Chair* and *Airplane* examples,

Table 5
Semantic segmentation results on the S3DIS area 5.

Methods	Input	mAcc (%)	mIoU (%)
PointNet [2]	xyz + rgb	49.0	41.1
PointNet++ [3]	xyz + rgb	67.1	53.5
PointCNN [5]	xyz + rgb	63.9	57.3
Transformer [47]	xyz	68.6	60.0
Point-BERT [19]	xyz	69.7	60.5
Point-MAE [28]	xyz	69.9	60.8
Point-MPP [21]	xyz	69.8	61.4
Point-LGMask [22]	xyz	70.3	61.3
PointSiRE (Ours)	xyz	**70.7**	**61.8**

Point-MAE struggles to distinguish legs from backrests or fuselages from tails, resulting in noisy predictions. In contrast, PointSiRE accurately delineates these intricate components. Note that these examples are selected to highlight qualitative differences in boundary precision and topological correctness around fine-grained, multi-part junctions, rather than to represent the largest quantitative margins in Table 4; *Chair* and *Lamp* are among the categories where PointSiRE attains the highest per-class IoU, while the *Airplane* example illustrates that even in a category where the two methods are quantitatively close, the boundary-level structural consistency of PointSiRE's predictions is visibly improved. Furthermore, in the *Lamp* case, where the baseline misinterprets the connecting tube, our model successfully recovers the correct topology. These results confirm that our relative geometry-aware pre-training effectively transfers robust structural priors to downstream dense prediction tasks.

4.5. Semantic segmentation on S3DIS

To further validate the scalability of PointSiRE to large-scale indoor scenes, we conduct semantic segmentation experiments on the Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset [48]. This dataset comprises 271 rooms from 6 separate areas, annotated with 13 semantic categories. Unlike object-level datasets, S3DIS presents significant challenges due to varying point densities, clutter, and occlusion in complex scene layouts. Following standard practice [28], we evaluate our method on Area 5 while training on the other areas. We report the mean Accuracy (mAcc) and mean Intersection-over-Union (mIoU). The quantitative results are presented in Table 5. PointSiRE achieves 70.7% mAcc and 61.8% mIoU, outperforming both supervised baselines and state-of-the-art self-supervised methods. Our method surpasses Point-MAE (60.8% mIoU) by 1.0% and Point-LGMask (61.3% mIoU) by 0.5%. This performance gain is attributed to our geometry-aware pre-training, which enables the model to robustly handle the diverse geometric variations and complex spatial structures found in real-world indoor scenes.

We note that the margin of improvement on S3DIS is comparatively smaller than that observed on ScanObjectNN. This is consistent with the nature of the benchmark: indoor scenes in S3DIS are conventionally aligned along the gravity direction and exhibit considerably less global rotational and scale variation than object-level point clouds, and the overall gap among competing self-supervised methods on this benchmark is itself narrow (within 0.6% mIoU among Point-MAE, Point-MPP, and Point-LGMask). Nevertheless, the consistent improvement of PointSiRE over all baselines indicates that the geometric and structural priors learned during pre-training transfer beneficially even to this more challenging, scene-level setting.

4.6. Ablation studies

In this section, we perform extensive ablation studies on the ScanObjectNN dataset to assess the effectiveness of the designed modules and analyze the impact of key hyperparameters. We choose the most challenging *PB_T50_RS* split as the primary testbed, as its complex background and arbitrary rotations provide a rigorous evaluation of our geometry-aware framework.

1) Effect of Core Design Components: To investigate the individual contributions of the Relative Geometry-aware Position Encoding (RGPE) and the Attention-guided Evolved Hierarchical Masking (AEHM), we conduct a component-wise analysis. We establish three variants: one using RGPE with random masking, one using AEHM without relative geometry, and the full model. As shown in Table 6, removing RGPE (Model 2) leads to a substantial performance drop of 1.78% (from 88.45% to 86.67%). This sharp decline indicates that in scenarios with severe pose variations, the student network struggles to align features with the teacher's coordinate system without explicit relative geometric guidance. Furthermore, replacing AEHM with random masking (Model 1) results in a 0.84% decrease. This suggests that the standard random masking strategy, which is agnostic to object structure, may disrupt semantic continuity. Our evolved, structure-aware masking strategy effectively guides the model to learn richer semantic representations than structure-agnostic random masking. The superior performance of the full model (Model 3) demonstrates that geometric reasoning and structural understanding are complementary, jointly empowering robust representation learning.

2) Sensitivity Analysis on Masking Ratio: The masking ratio is a critical hyperparameter governing the difficulty of the pretext task: a low ratio may allow the model to cheat via local interpolation, while an excessively high ratio may destroy essential geometric topology. We evaluate the performance under mask ratios $\rho \in \{0.4, 0.6, 0.75, 0.9\}$. As presented in Table 7, the performance improves as the ratio increases from 0.4 to 0.6, confirming that a moderately difficult task encourages the learning of more discriminative features. Increasing the ratio further to 0.75 produces a dataset-dependent effect rather than a uniform gain: performance on the heavily corrupted *PB_T50_RS* split improves slightly (88.55% vs. 88.45%), remains essentially unchanged on *OBJ_BG* (94.15%, identical to

Table 6
Ablation study on different module combinations on ScanObjectNN (PB_T50_RS). RGPE: relative geometry-aware position encoding; AEHM: attention-guided evolved hierarchical masking.

Variant	RGPE	AEHM	PB_T50_RS
Model 1	✓	–	87.61
Model 2	–	✓	86.67
Model 3	✓	✓	**88.45**

Table 7
Effect of different mask ratios on ScanObjectNN classification accuracy (%).

Mask Ratio	OBJ_BG	OBJ_ONLY	PB_T50_RS
0.4	93.46	92.08	87.96
0.6	**94.15**	**92.43**	88.45
0.75	**94.15**	91.39	**88.55**
0.9	91.22	90.53	87.78

Table 8
Effect of different predictor depths on ScanObjectNN classification accuracy (%).

Predictor Depth	OBJ_BG	OBJ_ONLY	PB_T50_RS
2	91.05	91.22	87.61
3	91.22	91.91	87.99
4	**94.15**	**92.43**	**88.45**
5	90.19	91.56	87.54
6	90.88	90.19	88.34

the 0.6 setting), but decreases on the clean *OBJ_ONLY* split (91.39% vs. 92.43%). We interpret this pattern as reflecting a trade-off between context sparsity and input degradation rather than a uniformly stronger shape prior. On *OBJ_ONLY*, where objects are clean and well-segmented, the 0.6 ratio already provides sufficient visible context for stable geometric representation learning; raising the ratio to 0.75 further reduces the available local information without a corresponding degradation in the test-time input to offset this loss, resulting in a net drop in performance. On *PB_T50_RS*, where inputs are themselves corrupted by background noise and pose perturbations, a higher mask ratio encourages the model to rely more on global context under sparse observations, which better matches the degraded input distribution and yields a modest gain. *OBJ_BG*, whose corruption level falls between the two, shows a correspondingly neutral effect. This explanation is drawn from the empirical trend across the three splits; a more systematic validation under controlled noise levels would be needed to confirm it directly, which we leave to future work. Overall, the ratio of 0.6 still achieves the best aggregate performance across splits and is adopted as our default. An aggressive ratio of 0.9 leads to a performance collapse, as the remaining points are insufficient to infer the complete shape. To balance robustness and generalization across diverse scenarios, we set the default masking ratio to 0.6.

3) Impact of Predictor Depth: In our Siamese framework, the student predictor serves as a "geometric navigator", responsible for digesting the relative geometric conditions and predicting the target features. To explore the optimal capacity for this reasoning process, we vary the depth of the predictor transformer from 2 to 6 layers. As shown in Table 8, the performance peaks at a depth of 4. Shallow predictors (depth 2 or 3) underperform, likely because they lack sufficient capacity to model the complex interactions between semantic tokens and high-dimensional geometric embeddings (rotation quaternions and translation vectors). Conversely, increasing the depth to 5 or 6 leads to performance saturation or even degradation. We attribute this to the difficulty in optimizing deeper networks with limited pre-training data, or potential overfitting to the pretext task rather than learning transferable features. Therefore, we adopt a 4-layer predictor to achieve an optimal trade-off between reasoning capability and training efficiency.

Unified analysis. The ablation results consistently demonstrate that both Relative Geometry-aware Position Encoding (RGPE) and Attention-guided Evolved Hierarchical Masking (AEHM) contribute complementary benefits. RGPE enhances the model's ability to explicitly reason about cross-view geometric transformations, while AEHM improves the structural awareness of the masked prediction task by introducing data-driven hierarchical masking. The full model consistently outperforms either component used in isolation, confirming that the two designs provide complementary rather than redundant benefits, and that effective point cloud representation learning requires both transformation-aware modeling and structure-aware supervision.

5. Conclusion

In this work, we proposed PointSiRE, a geometry-aware Siamese self-supervised framework for point cloud representation learning. Different from existing masked point cloud pre-training approaches that mainly treat geometric transformations as random augmentations, PointSiRE reformulates masked modeling as a relative geometry-conditioned cross-view prediction task. By intro-

ducing Relative Geometry-aware Position Encoding (RGPE), the model explicitly exploits inter-view scale, rotation, and translation relationships as supervisory signals, enabling transformation-aware feature alignment. Furthermore, the proposed Attention-guided Evolved Hierarchical Masking (AEHM) dynamically evolves the masking strategy according to the structural understanding of the teacher network, encouraging the model to progressively learn from local geometric patterns to higher-level semantic structures. Extensive experiments on classification, few-shot learning, and dense prediction tasks demonstrate that PointSiRE learns robust and transferable representations and achieves competitive or state-of-the-art performance compared with existing self-supervised point cloud methods. Detailed ablation studies further verify that RGPE and AEHM provide complementary benefits by jointly improving geometric reasoning and structural representation learning. Despite these promising results, several challenges remain. In future work, we will investigate the extension of PointSiRE to large-scale real-world LiDAR scenarios with complex environments and long-range spatial dependencies. We also plan to explore the integration of relative geometry reasoning with multimodal learning frameworks, where point clouds can be jointly modeled with images and other sensory modalities. Moreover, evaluating RGPE under realistic transformation estimation errors generated by registration algorithms will be an important direction toward practical deployment.

CRediT authorship contribution statement

Xin Cao: Writing – original draft, Methodology. **Linrong Ye:** Writing – original draft, Validation, Software, Methodology. **Xinmeng Hu:** Formal analysis. **Tianyi Liu:** Resources. **Linzhi Su:** Supervision, Formal analysis. **Xingxing Hao:** Validation, Investigation. **Kang Li:** Writing – review & editing, Project administration, Funding acquisition.

Funding

This work was supported in part by the Key Research and Development Program of Shaanxi Province (2024SF-YBXM-681), in part by the National Natural Science Foundation of China (62572394).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this study are publicly available datasets. The original sources for these datasets have been cited in the references.

References

[1] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, M. Bennamoun, Deep learning for 3D point clouds: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (12) (2020) 4338–4364.
[2] C.R. Qi, H. Su, K. Mo, 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.
[3] C.R. Qi, L. Yi, H. Su, L.J. Guibas, Pointnet++: deep hierarchical feature learning on point sets in a metric space, Adv. Neural Inf. Process. Syst. 30 (2017).
[4] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein, J.M. Solomon, Dynamic graph CNN for learning on point clouds, ACM Trans. Graph. 38 (5) (2019) 1–12.
[5] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, B. Chen, Pointcnn: convolution on x-transformed points, Adv. Neural Inf. Process. Syst. 31 (2018).
[6] H. Thomas, C.R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, L.J. Guibas, Kpconv: flexible and deformable convolution for point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6411–6420.
[7] H. Zhao, L. Jiang, J. Jia, P.H. Torr, V. Koltun, Point transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16259–16268.
[8] Y. Yang, C. Feng, Y. Shen, D. Tian, Foldingnet: point cloud auto-encoder via deep grid deformation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 206–215.
[9] S. Xie, J. Gu, D. Guo, C.R. Qi, L. Guibas, O. Litany, Pointcontrast: unsupervised pre-training for 3D point cloud understanding, in: European Conference on Computer Vision, Springer, 2020, pp. 574–591.
[10] Z. Zhang, R. Girdhar, A. Joulin, 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. 10252–10263.
[11] S. Huang, Y. Xie, S.-C. Zhu, Y. Zhu, Spatio-temporal self-supervised representation learning for 3D point clouds, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6535–6545.
[12] H. Wang, Q. Liu, X. Yue, J. Lasenby, M.J. Kusner, Unsupervised point cloud pre-training via occlusion completion, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9782–9792.
[13] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, R. Rodrigo, Crosspoint: self-supervised cross-modal contrastive learning for 3D point cloud understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9902–9912.
[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
[16] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, V.G. Kim, Self-supervised learning of point clouds via orientation estimation, in: 2020 International Conference on 3D Vision (3DV), IEEE, 2020, pp. 1018–1028.
[17] E. Chatzipantazis, S. Pertigkiozoglou, E. Dobriban, K. Daniilidis, SE(3)-equivariant attention networks for shape reconstruction in function space, in: The Eleventh International Conference on Learning Representations, 2023.
[18] C. Löwens, T. Funke, A. Wagner, A.P. Condurache, Unsupervised point cloud registration with self-distillation, in: 35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25–28, 2024, BMVA, 2024.
[19] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, J. Lu, Point-bert: pre-training 3D point cloud transformers with masked point modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 19313–19322.

[20] H. Liu, M. Cai, Y.J. Lee, Masked discrimination for self-supervised learning on point clouds, in: European Conference on Computer Vision, Springer, 2022, pp. 657–675.
[21] S. Fan, W. Gao, G. Li, Point-MPP: point cloud self-supervised learning from masked position prediction, IEEE Trans. Neural Netw. Learn. Syst. 36 (2024) 12964–12976.
[22] Y. Tang, X. Li, J. Xu, Q. Yu, L. Hu, Y. Hao, M. Chen, Point-lgmask: local and global contexts embedding for point cloud pre-training with multi-ratio masking, IEEE Trans. Multimed. 26 (2023) 8360–8370.
[23] P. Achlioptas, O. Diamanti, I. Mitliagkas, L. Guibas, Learning representations and generative models for 3D point clouds, in: International Conference on Machine Learning, PMLR, 2018, pp. 40–49.
[24] Z. Qi, R. Dong, G. Fan, Z. Ge, X. Zhang, K. Ma, L. Yi, Contrast with reconstruct: contrastive 3D representation learning guided by generative pretraining, in: International Conference on Machine Learning, PMLR, 2023, pp. 28223–28243.
[25] Y. Wu, J. Liu, M. Gong, P. Gong, X. Fan, A.K. Qin, Q. Miao, W. Ma, Self-supervised intra-modal and cross-modal contrastive learning for point cloud understanding, IEEE Trans. Multimed. 26 (2023) 1626–1638.
[26] D. Shao, X. Lu, W. Wang, X. Liu, A.S. Mian, Trici: triple cross-intra branch contrastive learning for point cloud analysis, IEEE Trans. Vis. Comput. Graph. 31 (2024) 5321–5333.
[27] K. Fu, M. Yuan, S. Liu, M. Wang, Boosting point-BERT by multi-choice tokens, IEEE Trans. Circuits Syst. Video Technol. 34 (1) (2023) 438–447.
[28] Y. Pang, E.H.F. Tay, L. Yuan, Z. Chen, Masked autoencoders for 3D point cloud self-supervised learning, World Scientific Annual Review of Artificial Intelligence 1 (2023) 2440001.
[29] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, H. Li, Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training, Adv. Neural Inf. Process. Syst. 35 (2022) 27061–27074.
[30] X. Zhang, S. Zhang, J. Yan, Pcp-mae: learning to predict centers for point masked autoencoders, Adv. Neural Inf. Process. Syst. 37 (2024) 80303–80327.
[31] Z. Zhang, B.-S. Hua, D.W. Rosen, S.-K. Yeung, Rotation invariant convolutions for 3D point clouds deep learning, in: 2019 International Conference on 3D Vision (3DV), IEEE, 2019, pp. 204–213.
[32] Z. Zhang, B.-S. Hua, S.-K. Yeung, RIConv++: effective rotation invariant convolutions for 3D point clouds deep learning: Z. zhang et al, Int. J. Comput. Vis. 130 (5) (2022) 1228–1243.
[33] C. Zhao, J. Yang, X. Xiong, A. Zhu, Z. Cao, X. Li, Rotation invariant point cloud analysis: where local geometry meets global topology, Pattern Recogn. 127 (C) (2022).
[34] D. Zhang, J. Yu, C. Zhang, W. Cai, Parot: patch-wise rotation-invariant network via feature disentanglement and pose restoration, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 3418–3426.
[35] K. Su, Q. Wu, P. Cai, X. Zhu, X. Lu, Z. Wang, K. Hu, RI-MAE: rotation-invariant masked AutoEncoders for self-supervised point cloud representation learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, 2025, pp. 7015–7023.
[36] X. Yin, D. Zhang, J. Yu, W. Cai, HFBRI-MAE: handcrafted feature based rotation-invariant masked autoencoder for 3D point cloud analysis, in: 2025 International Joint Conference on Neural Networks (IJCNN), IEEE, 2025, pp. 1–9.
[37] Z. Feng, S. Zhang, Evolved hierarchical masking for self-supervised learning, IEEE Trans. Pattern Anal. Mach. Intell. 47 (2024) 1013–1027.
[38] C. Tao, X. Zhu, W. Su, G. Huang, B. Li, J. Zhou, Y. Qiao, X. Wang, J. Dai, Siamese image modeling for self-supervised vision representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2132–2141.
[39] X. Yin, D. Zhang, Y. Feng, S. Mao, J. Yu, W. Cai, Beyond random masking: a dual-stream approach for rotation-invariant point cloud masked autoencoders, in: 2025 International Conference on Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2025, pp. 1–8.
[40] X. Wu, D. DeTone, D. Frost, T. Shen, C. Xie, N. Yang, J. Engel, R. Newcombe, H. Zhao, J. Straub, Sonata: self-supervised learning of reliable point representations, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 22193–22204.
[41] R. Zhu, Y. Bai, T. Yao, J. Liu, Z. Sun, T. Mei, C.W. Chen, Teaching masked autoencoder with strong augmentations, IEEE Trans. Neural Netw. Learn. Syst. 36 (2024) 9550–9564.
[42] A.X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al., Shapenet: an information-rich 3D model repository, arXiv preprint arXiv:1512.03012, 2015.
[43] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D shapenets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1912–1920.
[44] M.A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, S.-K. Yeung, Revisiting point cloud classification: a new benchmark dataset and classification model on real-world data, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1588–1597.
[45] X. Ma, C. Qin, H. You, H. Ran, Y. Fu, Rethinking network design and local geometry in point cloud: a simple residual MLP framework, in: International Conference on Learning Representations, 2022.
[46] L. Yi, V.G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, L. Guibas, A scalable active framework for region annotation in 3D shape collections, ACM Trans. Graph. 35 (6) (2016) 1–12.
[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[48] I. Armeni, O. Sener, A.R. Zamir, H. Jiang, I. Brilakis, M. Fischer, S. Savarese, 3D semantic parsing of large-scale indoor spaces, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1534–1543.