npj | heritage science

Article

https://doi.org/10.1038/s40494-025-01571-8

PointMoment: a mixed-moment self-supervised learning approach for 3D Terracotta Warriors

Check for updates

Xin Cao[1], Xinxin Han[1], Wenlong Tang[1], Yong Ren[1], Kang Li[1] ✉, Ping Zhou[2] & Linzhi Su[1] ✉

The Terracotta Warriors, a hallmark of China's cultural heritage, frequently exhibit fragmentation and deformation due to natural factors like earthquakes and human activities. Accurate classification and segmentation of their 3D models are essential for effective restoration. However, the irregularity of the fragmented terracotta pieces renders manual annotation time-consuming and labor-intesńsive. To address this challenge, we propose a self-supervised learning method utilizing high-order mixed moments for 3D point clouds. It employs a high-order mixed moment loss function instead of the traditional contrastive loss function and does not require special techniques like asymmetric network architectures or gradient stopping. Our method involves calculating the high-order mixed moment of feature variables and forcing them to decompose into individual moments, enhancing variable independence and minimizing feature redundancy. Additionally, we integrate a contrastive learning approach to maximize feature invariance across different augmentations of the same point cloud. Experiments demonstrate that our method outperforms previous unsupervised learning techniques in the downstream tasks of 3D point cloud classification and segmentation. Additionally, our method shows strong performance in the specific tasks related to the Terracotta Warriors. We hope this success can pave the way for new avenues in the virtual protection and restoration of cultural relics. Code is available at https://github.com/caoxin918/PointMoment.

As a significant material cultural heritage of Chinese civilization, the Terracotta Warriors possess considerable academic value and social significance for an in-depth understanding of China's extensive history and culture. However, due to long-term natural erosion and human activities, most of the Terracotta Warriors have been unearthed as fragments, posing substantial challenges to their protection and restoration[1]. Traditional manual restoration methods are cumbersome and time-consuming, making it difficult to meet the demands of large-scale restoration projects. In recent years, the rapid development of 3D laser scanning technology has significantly advanced virtual restoration techniques, offering new possibilities for the digital preservation and restoration of the Terracotta Warriors[2]. Within the virtual restoration process, the classification and segmentation of Terracotta fragments are crucial steps. Accurate fragment classification directly impacts the effectiveness of subsequent matching and assembly, while precise segmentation not only addresses the issue of insufficient calibration data but also enables semantic annotation of the Terracotta Warriors, facilitating an in-depth analysis of the relics' structure and features. Thus, obtaining an effective feature representation is essential for improving the accuracy of downstream classification and segmentation tasks, thereby enhancing the overall restoration efficiency.

Traditional heritage feature extraction methods primarily rely on expert-designed feature descriptors. Rasheed et al.[3] proposed a method to extract texture features based on RGB color features and the gray-level co-occurrence matrix. Lin et al.[4] proposed a multi-scan point cloud hierarchical registration method for the 3D reconstruction of ancient buildings, which effectively solved the problem of digitally reconstructing complete and real models of complex architectural structures without damage. Although effective, these methods often depend on expert prior knowledge, and the expressive power of hand-designed feature extraction techniques is limited, which constrains the generalization performance of classification models. With the rapid advancement of deep learning technology and the proliferation of 3D data acquisition devices, point cloud-based representation learning algorithms, such as PointNet[5], PointNet + +[6], and DGCNN[7], have been proposed, providing novel approaches for feature extraction from cultural relic fragments. Zhou et al.[8] introduced the Multi-scale Local Geometric Transformer-Mamba (MLGTM) method to improve the

[1]School of Information Science and Technology, Northwest University, Xi'an, Shaanxi, 710127, China. [2]Emperor Qin Shihuang's Mausoleum Site Museum, Key Scientific Research Base of Ancient Polychrome Pottery Conservation, Xi'an, Shaanxi, 710600, China. ✉e-mail: likang@nwu.edu.cn; sulinzhi029@163.com

accuracy and robustness of Terracotta Warriors point cloud classification tasks by effectively capturing complex local morphology and handling data sparsity and irregularity. Zhu et al.[9] developed a transfer learning-based method to recover the three-dimensional shape of cultural relics faces from a single old photo, effectively addressing the issue of limited cultural relic samples. However, these methods typically require labeled data, posing challenges for their application in the field of cultural heritage. Self-supervised representation learning (SSRL), which has shown promise in fields like computer vision and natural language processing (NLP), may provide new solutions to these challenges[10].

SSRL seeks to enhance the performance of various downstream tasks by learning general and robust feature representations from unlabeled data[11]. Recent research in the field of image processing has shown that representations obtained through SSDL can be as effective as those achieved through supervised learning methods[12]. However, the learned embedded features often contain redundant information, leading to a decrease in semantic representation ability and impacting downstream task performance. To address these challenges, innovative approaches such as Barlow Twins[13] and HOME[14] have been proposed. These methods focus on feature redundancy to of SSDL in image-processing tasks. In recent years, several algorithms also have been developed for self-resolve the aforementioned issues, enhancing the overall efficacy supervised learning with point clouds, including self-reconstruction, adversarial generation, and completion[15,16]. However, methods like Latent GAN[17], FoldingNet[18], and OcCo[19] require significant computing resources and time, and all these algorithms are highly sensitive to rotational and translational variations. Contrastive learning, successful in video and image domains, is increasingly being applied to point cloud understanding[20]. For instance, PointContrast uses contrastive learning to achieve viewpoint-invariant point cloud representations, facilitating high-level scene understanding[21]. CrossPoint enhances this by leveraging multiple modalities for contrastive learning, extracting richer signals[22]. These methods involve pre-training with a pretext task designed to bring similar samples closer and push dissimilar ones further apart in the feature space. Yet, these methods commonly face model collapse, where the learned representation vector shrinks to a constant or low-dimensional subspace, underutilizing the full representational capacity. Existing contrast-based learning methods often mitigate this by using complex mechanisms, such as memory banks[23] or asymmetric networks with gradient stopping, predictor networks, and momentum update strategies[24]. Notably, in the field of cultural heritage protection, particularly in the digital preservation of Terracotta Warriors, significant strides have been made. Wang et al.[25] proposed a transformer-based method to enhance point cloud registration, effectively improving the accuracy of point cloud alignment in Terracotta Warriors preservation. Additionally, Xu et al.[26] developed CPDC-MFNet, a conditional point diffusion completion network with multi-scale feedback refinement for repairing damaged 3D models of Terracotta Warriors, which improves generation speed while maintaining diverse outputs. Despite these advancements in the preservation of Terracotta Warriors, the open challenge of preventing model collapse without relying on complex designs still remains, underscoring the need for further research in this area.

Motivated by the above analysis and inspiration from[14], we introduce a self-supervised contrastive learning method for terracotta fragment classification and segmentation that leverages high-order mixed moments. This approach effectively prevents model collapse by minimizing redundancy among feature variables without relying on complex mechanisms. Specifically, we utilize high-order mixed moments to minimize redundancy among arbitrary feature variables, aiming to learn meaningful point cloud representations. It is known that pairwise independence of each variable does not guarantee their mutual independence, and that the total correlation among all variables is minimized only if multiple variables are independent. This implies that the mixed moment of multiple features can be decomposed into the product of their individual moments. Drawing from statistical theory and the contrastive learning paradigm, we design a loss function based on high-order mixed moments. This function reduces redundancy among multiple variables, allowing the high-dimensional features learned through self-supervised learning to be rich in information and independent. It also ensures maximum consistency in representation across different augmented point clouds. Our extensive experiments on various datasets show that our method achieves state-of-the-art accuracy.

The main contributions of this paper are as follows:

- We propose a self-supervised learning method for point clouds using high-order mixed moments. Our approach reduces feature redundancy, enhancing representational capability while inherently avoiding model collapse without additional techniques. We also successfully applied it to Terracotta warrior analysis.
- Our approach offers a plug-and-play solution with symmetric network architecture, enabling flexible integration with various point cloud processing methods.
- We evaluated our approach on object classification and segmentation tasks. PointMoment significantly outperformed existing unsupervised learning methods. Notably, its application to the Terracotta Warriors dataset yielded exceptional results, further validating our method's effectiveness and superiority in point cloud analysis.

Related work

In this section, we provide a brief overview of recent advancements in two pertinent areas: supervised and self-supervised representation learning for point clouds.

Supervised representation learning on point cloud

We can divide supervised representation learning methods for point clouds into two main categories: structure-based and point-based. Structure-based methods often transform point clouds into 2D images or regular structured data like voxels for feature extraction. MVCNN[27] uses 2D convolutional networks with max-pooling to create global shape descriptors from multi-view features. However, it may result in information loss and overlooks view relationships. To address this, View-GCN[28] employs a Graph Convolutional Network (GCN) on the view graph for hierarchical feature aggregation. VoxNet[29] uses 3D convolutional networks to extract features from voxel grids but can be computationally intensive and memory-demanding with dense 3D data, which grows cubically with resolution. Overall, these methods may struggle to capture fine-grained geometric details and can be memory-intensive.

Point-based methods utilize raw data without conversion to other formats and can be divided into four types: MLP-based, Transformer-based, Graph-based, and CNN-based. MLP-based methods, such as the pioneering PointNet[5], as a pioneering work, has inspired many subsequent methods, such as PointWeb[30]. PointNet + +[6] improves upon PointNet by using set abstraction to capture local features and generate global representations. CNN-based methods use specially designed convolution kernels for point cloud feature extraction. PointCNN[31] introduces the X-Conv operator for direct processing of point clouds, while SpiderCNN[32] employs step functions and Taylor expansions to capture complex local geometries. Graph-based methods treat each point as a graph vertex and construct edges to reflect relationships with neighbors. DGCNN[7] reconstructs local graphs and uses EdgeConv for local feature extraction but may not fully capture edge features. Recently, Transformer-based methods, leveraging self-attention mechanisms successful in image analysis and NLP, have been explored for point cloud data, as seen in PCT[33] and PointTransformer[34]. However, the primary challenge for these supervised methods is the scarcity of large annotated datasets.

Self-supervised representation learning on point cloud

Self-supervised representation learning is a prominent research area in point cloud analysis, adept at extracting effective features without labeled data. It can be divided into two main types: contrastive and generative.

Generative methods use models to approximate the underlying data or feature distribution, revealing intrinsic point cloud characteristics. Common models include variational autoencoders and generative adversarial

networks (GANs). Latent-GAN[17] is a pioneering model that applies GAN to raw point cloud data and embedded features to generate high-quality samples. Yang et al.[18] developed a decoder based on the concept that 3D object surfaces can be folded from 2D planes, enhancing 3D reconstruction quality. As transformers gain popularity in image and NLP fields, transformer-based generative methods like Point-BERT[35] and Point-MAE[36] are emerging, with Point-MAE extending Point-BERT. OcCo[19], introduced by Wang et al. is a novel approach to point cloud completion that learns representations by reconstructing missing data. However, these methods can struggle with accurate reconstruction and are computationally demanding.

Contrastive learning methods train a feature encoder to learn representations that are similar to positive examples and dissimilar to negatives. Info3D[23] merges mutual information with contrastive learning to enhance 3D object representation by optimizing the mutual information between local structures and the global shape. PointGLR[20] integrates normal estimation, self-reconstruction, and contrastive learning into a single framework for bidirectional inference between global and local 3D object features, in an unsupervised manner. PointContrast[21] unifies the contrastive learning paradigm for 3D point clouds to promote high-level scene understanding through multi-view learning. DepthContrast[37] introduces a joint strategy for voxel and point cloud contrastive learning to improve downstream task performance. STRL[24] adapts BYOL[10] for point cloud representation learning, while CrossPoint[22] uses 2D images to enhance 3D point cloud understanding with a cross-modal contrastive method, however, obtaining 2D-rendered images can be challenging.

Unlike methods that rely on complex strategies to prevent model collapse, such as memory banks or asymmetric networks, we propose a versatile contrastive learning framework with a novel loss function for ideal point cloud representation.

Methodology

In this section, we introduce PointMoment, a self-supervised approach for learning meaningful point cloud representations using high-order mixed moments. We first explain the concept of high-order mixed moments (section High-order mixed-moment). Then, we present a loss function (section Formulation of high-order mixed-moment as a loss) designed to pre-train the feature extractor $f_\theta()$ in a self-supervised manner. Figure 1 shows the network framework of our approach. Finally, we detail an application of our approach to third-order mixed moment-based self-supervised representation learning (section Instantiating three-order mixed-moment for contrastive learning), as depicted in Fig. 2.

High-order mixed-moment

Moments are numerical characteristics used in probability theory and statistics to describe the distribution of random variables. The most common moments are the first-order moment (mean) and the second-order central moment (variance), which measure the average and dispersion of a single random variable, respectively. For a more comprehensive assessment of multiple random variables, mixed moments are utilized.

Mixed moments are statistical measures that describe the fundamental characteristics of a multivariate random variable's distribution. They are typically defined as follows:

For any positive integer k_i, the mathematical expectation $E[X_1^{k_1}...X_n^{k_n}]$ of multiple random variables is called the k-order mixed moment, where $k = k_1 + ...+k_n$. $E[(X_1 - E[X_1])^{k_1}...(X_n - E[X_n])^{k_n}]$ is then called the k-order central mixed moment. If there are only two random variables and both k_i are 1, then the mixed moment $E[(X_1 - E[X_1])(X_2 - E[X_2])]$ is called the covariance, which describes the degree of correlation between the two random variables[13]. just start from the perspective of covariance and reduce the feature redundancy by reducing the correlation between any two feature variables.

After introducing mixed moments, we'll explain their utility. In high-dimensional feature spaces, redundant information within embedded features can impair their utility. To mitigate this, we apply mutual information to measure the interdependence among random variables. Let $X_1, X_2, ..., X_d$ be the random variables representing each dimension (denoted as d) of the embedding feature. We can use the mutual information (MI) formula to measure the shared information between variables:

$$I(X_1, X_2, ..., X_d) = \int_{X_1}\int_{X_2}\cdots\int_{X_d} p(x_1, x_2, ..., x_d)\log\frac{p(x_1, x_2, ..., x_d)}{p(x_1)p(x_2)...p(x_d)}dx_1dx_2...dx_d \quad (1)$$

where $p(x_1, x_2, ..., x_d)$ is the joint density function of $X_1, X_2, ...X_d$ and $p(x_1), p(x_2), ..., p(x_d)$ is the marginal probability density function of $X_1, X_2, ...X_d$ respectively.

To make the information contained in the embedding features rich and compact, we need to minimize Eq.1 to reduce the redundant information between feature variables. According to statistics, Eq.1 can be minimized when $X_1, X_2, ...X_d$ are independent of each other, that is, $p(x_1, x_2, ..., x_d) = p(x_1)p(x_2)...p(x_d)$. However, this minimization process faces two key challenges. One challenge is that pairwise independence of random variables does not necessarily ensure that the total redundancy of all variables is minimized, unless $p(x_1, x_2, ..., x_d)$ follows a multivariate normal

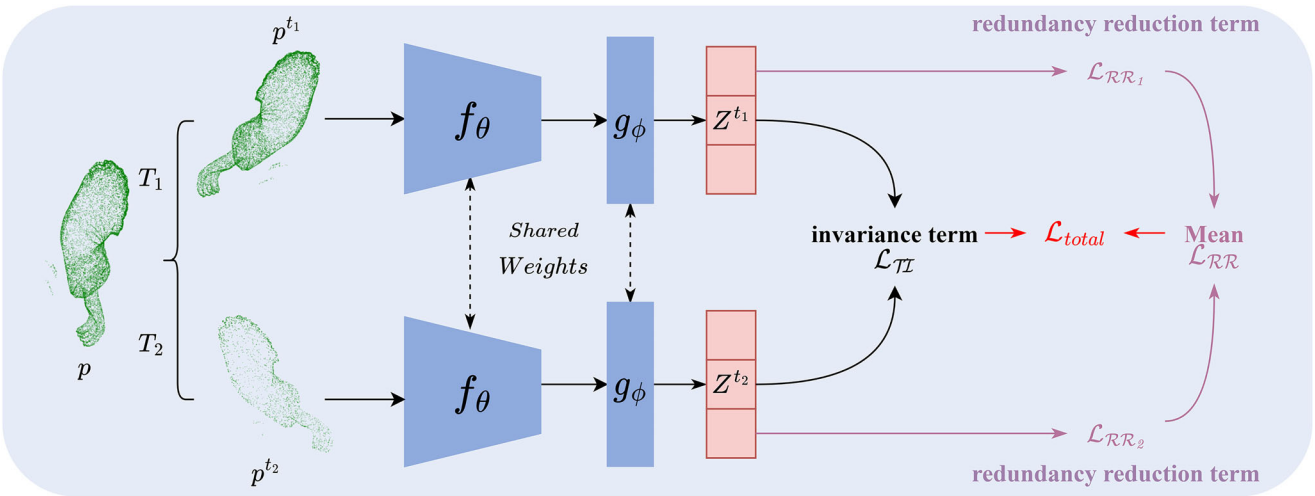

Fig. 1 | The overall architecture of our approach PointMoment. The loss is composed of two important parts, one is the loss based on invariance, and the other is the loss based on redundancy reduction. The latter involves constraints of order two or higher.

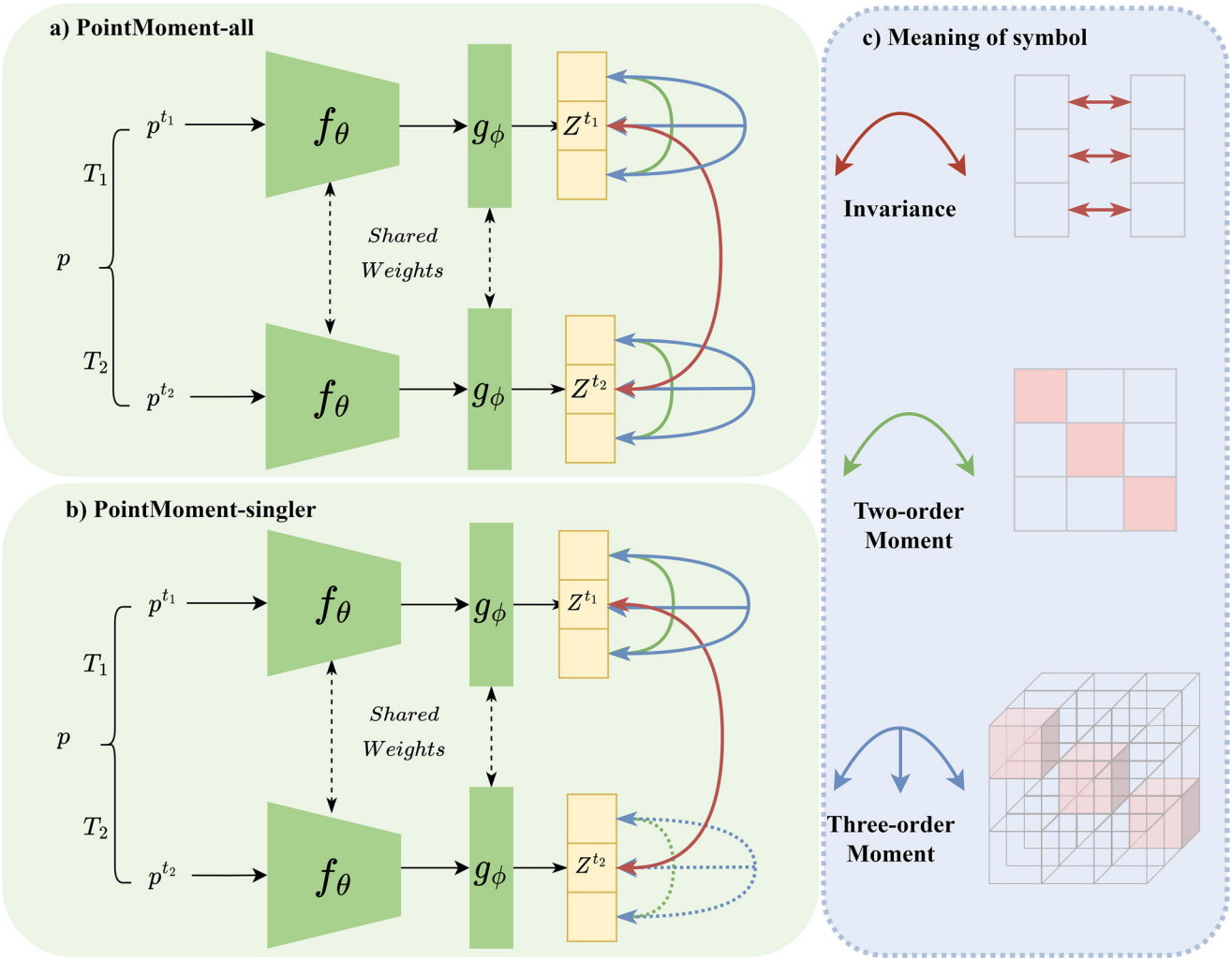

Fig. 2 | Self-supervised representation learning based on third-order mixed moment. a Indicates that mixed moment are calculated for the output of each branch, and **b** indicates that mixed moment are calculated for the output of a particular branch, where the dashed line indicates that the corresponding output is not constrained by the mixed moment. **c** Visualization shows the specific meaning of each curve, and the red square part does not take into account the calculation of the mixed moment.

distribution, which is often hard to satisfy in practice. Another challenge is that directly modelling the probability distribution of continuous variables can be difficult. Therefore, a common approach is to use statistical moment to model the probability distribution. We introduce high-order mixed moment. We exploit the fact that all random variables are independent of each other if and only if the expectation of all random variables is equal to the product of their individual expectations, which can be mathematically expressed as follows:

$$E[\prod_{d=1}^{D} X_d] = \prod_{d=1}^{D} E[X_d] \tag{2}$$

where $E[\prod_{d=1}^{D} X_d]$ is the mixed moment of order D. We can minimize Eq.1 by modelling and enforcing the random variables to satisfy the condition of Eq.2, which induces more independence among them.

Formulation of high-order mixed-moment as a loss

Preliminary. Given a batch of randomly selected point clouds $B = \{p_i\}_{i=1}^{|B|}$, where $p_i \in R^{N\times 3}$, $|B|$ denotes batch size and N denotes the total number of points in each point cloud. We apply random data augmentation techniques such as rotation, jittering and scaling to each point cloud p_i, generating two distinct enhanced versions of the original data, $p_i^{t_1}$ and $p_i^{t_2}$. Next, we feed both enhanced datasets into a network architecture. Specifically, the feature extractor f_θ maps $p_i^{t_1}$ and $p_i^{t_2}$ into a feature representation space, followed by a projection head g_φ, which projects resulting feature vectors onto an embedding feature space. We denote the final embedding vectors produced from this process as $z_i^{t_1}$ and $z_i^{t_2}$, respectively, where $z_i^t = g_\varphi(f(p_i^t;\theta_f);\varphi_g)$ and $z_i^t \in R^D$, θ_f and φ_g denote the trainable parameters of the encoder and projector, respectively, and D denotes the dimension of the embedding vector. Our goal is to learn compact, meaningful feature vectors without the need for complex network architectures or optimization processes. Furthermore, our loss function can be used in combination with a range of other network frameworks. Figure 1 provides a general representation of our self-supervised learning approach for point clouds.

Loss based on high-order mixed-moment. We argue that an ideal embedding feature should possess two key attributes: invariance to random augmentations of samples and minimal total correlation among the vector's variables. To achieve these, we've crafted distinct loss functions for each property.

Redundancy reduction based on high-order mixed-moment: Minimizing redundancy prompts the network to learn a compact embedding where each variable conveys unique semantic information. As outlined in Section "High-order mixed-moment," we enforce Eq. 2 to reduce total

correlation among variables to a certain extent. We apply the Law of Large Numbers to estimate a variable's expectation using sample moments, which are then used to calculate the expectation in Eq. 2. For ease, we standardize each feature variable z across the batch dimension to have a mean of zero and a variance of one, as follows:

$$\hat{z}^{t}_{b,d} = \frac{z^{t}_{b,d} - \frac{1}{B}\sum_{b=1}^{B} z^{t}_{b,d}}{\sqrt{\sum_{b=1}^{B}(z^{t}_{b,d} - \frac{1}{B}\sum_{b=1}^{B} z^{t}_{b,d})^{2}/B}} \tag{3}$$

Now for each dimension satisfies $E[\hat{Z}_d] = 0$, so we only need to set $E[\prod_{d=1}^{D} \hat{Z}_d] = 0$ to make Eq.2 hold, that is, $E[\prod_{d=1}^{D} \hat{Z}_d] = \prod_{d=1}^{D} E[\hat{Z}_d] = 0$. Therefore, we propose the redundancy reduction loss based on the high-order mixed moment as follows:

$$\mathscr{L}_{RR} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{2M}\sum_{K=2}^{D}\left(\sum_{d_1}^{D}\sum_{d_2\neq d_1}^{D}\cdots\sum_{d_K\neq\cdots\neq d_2\neq d_1}^{D}\left(E_{d_1,d_2,\cdots,d_K}\right)^{2}\right)\right] \tag{4}$$

where K denotes the order of the mixed moment and M denotes the total number of combinations of mixed moment of all orders, i.e. $M = \sum_{K=2}^{D}\frac{D!}{(D-K)!K!}$. $E_{d_1,d_2,\cdots,d_k}$ is the matrix of mixed moment of order K calculated along the batch dimension:

$$E_{d_1,d_2,\cdots,d_K} = \frac{1}{B}\sum_{b=1}^{B}\prod_{i=1}^{K}\hat{z}^{t}_{b,d_i} \tag{5}$$

where $d_1, d_2, \cdots, d_K$ etc. denote the index of the dimension of the embedding vector, respectively, for any K variables with $K \le D$ and with $1 \le d_K \le D$ for any d_K. $E_{d_1,d_2,\cdots,d_k}$ denotes the value of the K-order mixed moment of the random variables with index value $d_1, d_2, \cdots, d_k$, respectively.

Transformation invariance based on co-correlation: Invariance is the property that ensures semantically similar point cloud data are mapped to close regions in the embedding feature space. While many methods calculate cosine similarity between different augmented point cloud versions and build contrastive losses around it, we utilize the co-correlation matrix from[13] to bolster invariance, proposing the following loss function:

$$\mathcal{L}_{TI} = \frac{1}{D}\sum_{d=1}^{D}(1 - C_{dd})^{2} \tag{6}$$

where C is the co-correlation matrix, calculated from the outputs of two identical networks along the batch dimension:

$$C_{i,j} = \frac{\sum_{b=1}^{B}\hat{z}^{t_1}_{b,i}\hat{z}^{t_2}_{b,j}}{\sqrt{\sum_{b=1}^{B}\left(\hat{z}^{t_1}_{b,i}\right)^{2}}\sqrt{\sum_{b=1}^{B}\left(\hat{z}^{t_2}_{b,j}\right)^{2}}} = \frac{1}{B}\sum_{b=1}^{B}\hat{z}^{t_1}_{b,i}\hat{z}^{t_2}_{b,j} \tag{7}$$

where i and j denote the dimensional index of the embedding vector, C is a square matrix of size D and $-1 \le C_{i,j} \le 1$, -1 denotes negative correlation and 1 denotes positive correlation. It is analytically easy to see that Eq.7 can eventually be reduced to the same expression as Eq.5, which essentially satisfies the same underlying logic. Equation 6 maximizes the correlation of the different augmented embedding features by forcing the diagonal elements of the co-correlation matrix to be 1, thus satisfying the transformation invariance.

Based on the above analysis, we give the final loss:

$$\mathscr{L}_{total} = \mathscr{L}_{TI} + \lambda\mathscr{L}_{RR} \tag{8}$$

where λ is a parameter used to trade off the first and second terms.

Instantiating three-order mixed-moment for contrastive learning

Following the discussions in sections "High-order mixed-moment" and "Formulation of high-order mixed-moment as a loss," we introduce a self-supervised representation learning approach for point clouds that utilizes third-order mixed moments, as depicted in Fig. 2a. Additionally, we present a more computationally efficient version shown in Fig. 2b, which alleviates the requirement for all network branches to be active. It randomly selects a branch to achieve the desired outcome, thereby reducing computational demands. The experimental results validating this approach are detailed in the subsequent section. The subsequent experiments are all based on k = 3, that is, the third-order mixing moment.

Figure 2 illustrates our approach utilizing third-order mixed moments to compute the total loss $\mathscr{L}_{total}$ for K = {2,3}. The green curve represents the second-order mixed moment calculated for the embedding features of the branch output. Similarly, the blue curve represents the third-order mixed moment computed for the branch output. These two constraints are combined, i.e. $\mathscr{L}_{RR}$, to eliminate redundancy in the embedding features. Finally, the red curve signifies the consistency of data after undergoing two transformations, i.e. $\mathscr{L}_{TI}$, which completes the transformation invariance constraint.

Experiments

This section presents a comprehensive evaluation of our proposed method. First, we elucidate our pre-training strategy and describe the datasets employed. Second, we assess the transferability of our approach through two prevalent downstream tasks: object classification and segmentation. Third, to validate the efficacy of our loss function and parameter selection, we conduct thorough ablation studies. Finally, we demonstrate the practical utility of our method by applying it to the Terracotta Warriors dataset, showcasing its performance in a real-world scenario.

Pre-training

Dataset. For pre-training, we utilized the ShapeNet[38] dataset, a comprehensive repository of synthetic 3D shapes comprising over 50,000 unique models across 55 common object categories. To ensure comparability, we adhered to the training protocol established by STRL[24]. This procedure involved randomly sampling 2048 points from each model in the dataset. Subsequently, we applied a series of data augmentation techniques, including random rotation, translation, scaling, clipping, and jittering, followed by normalization. These augmented samples were then fed into the network for pre-training, maintaining consistency with established methodologies in the field.

Pre-training Detail. For a fair comparison with existing methods, we adopt the same method as STRL[24], OCCO[19], etc., using PointNet[5] and DGCNN[7] as feature extractors for point clouds and a two-layer multilayer perceptron as the projection head to map feature vectors into a 512-dimensional embedding space. We train the model in an end-to-end manner for 200 rounds using an Adam optimizer with a weight decay of 1×10^{-6} and an initial learning rate of 1×10^{-3}. Additionally, we also adjust the learning rate by implementing a decay strategy based on cosine annealing. The batch size is 16. After pre-training, we discarded the projection head $g_\phi()$ and retained $f_\theta()$ for the following downstream task.

Downstream tasks

3D Object classification. The 3D object classification task involves categorizing point cloud data to identify the specific class of each point cloud. We assess the shape understanding and generalization of pre-trained models using two benchmark datasets: ScanObjectNN[39] and ModelNet40[40]. ScanObjectNN, a challenging dataset of real-world 3D point clouds from indoor scenes, contains 15 categories with 2880 objects, Among these, 2304 for training and 576 for testing. ModelNet40, featuring synthetic objects, includes 12,311 CAD models across 40 categories, and 2468 for testing and 9843 for training, allowing us to evaluate classification performance on synthetic data.

Table 1 | Comparison of the linear SVM classification on ModelNet40

Method	ModelNet40
3D-GAN[42]	83.3
Latent-GAN[17]	85.7
SO-Net[43]	87.3
FoldingNet[18]	88.4
MRTNet[44]	86.4
3D-PointCapsNet[45]	88.9
MAP-VAE[46]	88.4
DepthContrast[37]	85.4
Jigsaw[47]+PointNet	87.3
Rotation[11]+PointNet	88.6
OcCo[19]+PointNet	88.7
STRL[24]+PointNet	88.3
PointMoment-all(Ours)+PointNet	88.8
PointMoment-singler(Ours)+PointNet	88.8
Jigsaw[47] + DGCNN	90.6
Rotation[11] + DGCNN	90.8
STRL[24] + DGCNN	90.9
OcCo[19] + DGCNN	89.2
PointMoment-all(Ours)+DGCNN	90.9
PointMoment-singler(Ours)+DGCNN	91.0

The linear classifier is fitted on the training set of ModelNet40 using the pre-trained model, and the model performance is evaluated on the test set.

Table 2 | Comparison of classification on ScanObjectNN

Encoder	Method	Acc.
PointNet	Jigsaw[47]	55.2
	OcCo[19]	69.5
	STRL[24]	74.2
	PointMoment(Ours)	79.0
DGCNN	Jigsaw[47]	59.5
	OcCo[19]	78.3
	STRL[24]	77.9
	PointMoment(Ours)	80.5

PointMoment achieves improvements compared to other self-supervised methods on both PointNet and DGCNN, which illustrates the effectiveness of our method in real-world scene classification.

Linear classification is a common method to evaluate the migration and generalization ability of self-supervised models in classification tasks. We follow the standard protocols of ref. 24 and ref. 19 to test the accuracy of our network model in object classification. On the classification data set, a linear Support Vector Machine (SVM) classifier is employed. This classifier is trained on features extracted from the training set using a pre-trained feature extractor, whose parameters remain fixed during this process. Subsequently, the trained SVM is applied to predict classifications based on the 3D features extracted from the test set. This methodology is widely adopted in the field for assessing the effectiveness of learned feature representations in downstream classification tasks. For our experiments, we employ two commonly used backbone networks, PointNet and DGCNN, as feature extractors.

Table 1 presents the accuracy of PointMoment for linear classification on ModelNet40. To conserve computational resources, we randomly chose one network branch for the high-order mixed moment constraint, achieving comparable results to using all branches, and adopted this single-branch constraint for subsequent experiments. Our method surpasses other state-of-the-art (SOTA) unsupervised and self-supervised algorithms when employing PointNet or DGCNN as the backbone network. Notably, our approach uses a basic network architecture without the complex features of STRL, such as asymmetric networks or gradient stopping. Specifically, our method outperforms STRL by 0.5% and 0.1% when using PointNet and DGCNN, respectively, highlighting the effectiveness of our approach. We assessed the generalizability of PointMoment in real-world scenarios by testing on ScanObjectNN with an SVM classifier.

Table 2 compares the linear classification accuracy of other self-supervised methods on ScanObjectNN. Our method outperformed the previous SOTA approaches by 4.8% and 2.6% when using PointNet and DGCNN as feature extractors, respectively. This result underscores the generalizability of representations learned from synthetic data, confirming the effectiveness of our approach.

Table 3 | Part segmentation results on ShapeNetPart dataset

Category	Method	mIoU
Supervised	PointNet[5]	83.7
	PointNet + +[6]	85.1
	DGCNN[7]	85.1
Self-Supervised	PointContrast[21]	85.1
	Jigsaw[47]	84.3
	OcCo[19]	85.0
	PointMoment(Ours)	85.4

Our method outperforms supervised learning methods with random initial weights and other self-supervised learning methods with pre-trained weights.

3D Object part segmentation. Object part segmentation, a complex and crucial task in 3D recognition, involves categorizing each point of an object into specific part classes, such as a table's leg or a car's tire. We conducted experiments using the ShapeNetPart[41] dataset, which includes 16,991 objects across 16 categories with 50 distinct parts, ranging from 2 to 6 parts per object. As a benchmark, ShapeNetPart effectively measures object part segmentation performance. Following the approach of previous studies[24] and[19], we pre-trained our model using DGCNN as the backbone network, followed by fine-tuning to enhance performance. Specifically, we conducted fine-tuning experiments on the ShapeNetPart dataset in an end-to-end manner. For the fine-tuning process, we employed SGD as the optimizer, with an initial learning rate of 0.1 and weight decay of 1×10^{-4}. The momentum was set to 0.9, with a batch size of 8. The model was trained for 300 epochs. We selected mean Intersection over Union (mIoU) as the evaluation metric, given its precision and widespread use in the field.

Table 3 compares segmentation outcomes for supervised learning methods and various self-supervised approaches on ShapeNetPart. Our pre-training approach offered better initial weights for DGCNN than those from random initialization by supervised learning, increasing mIoU by 0.3%. Additionally, our model outperformed the current SOTA self-supervised method by 0.3% in mIoU, indicating that our use of high-order mixed moments yields more discriminative and less redundant features. The visualization results are presented in Fig. 3 Our segmentation outcomes demonstrate a high degree of similarity to the ground truth, indicating that our method effectively captures fine-grained information within the point cloud. This close correspondence between our results and the actual segmentation underscores the capability of our approach to discern and represent detailed structural features in point cloud data.

3D semantic segmentation. Semantic segmentation is a challenging task that aims to assign a semantic label to each point in a point cloud,

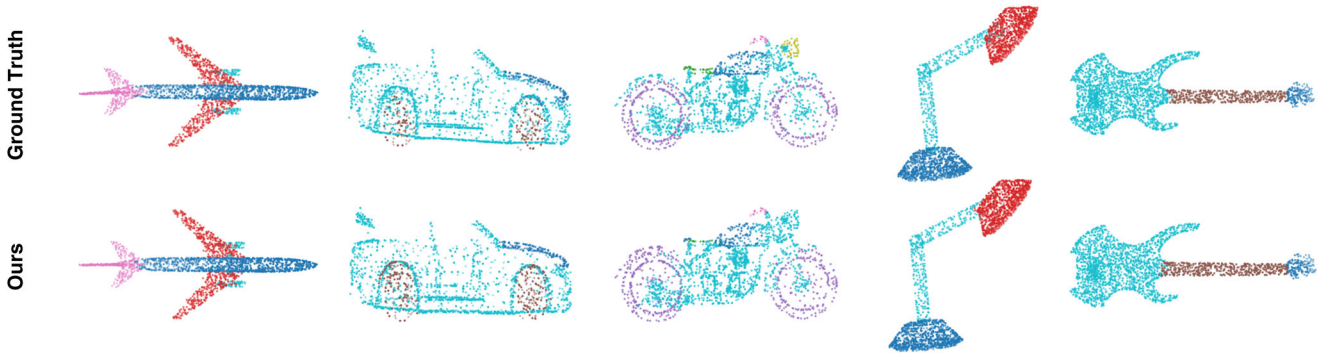

Fig. 3 | The visualization of part segmentation results on ShapeNetPart. The first row is the ground truth, and the second row is our method.

Table 4 | Semantic segmentation results on S3DIS dataset, evaluated on Area 5

Category	Method	mIoU	mAcc
Supervised	PointNet[5]	45.4	49.0
	PointNet + +[6]	53.5	–
Self-Supervised	Multi-view rendering[48]	46.7	85.0
	Jigsaw[47]	43.6	82.5
	OcCo[19]	44.5	83.6
	PointMoment(Ours)	46.9	85.5

enabling the grouping of regions with meaningful significance. This task is particularly important in complex indoor and outdoor scenes, which are often characterized by substantial background noise. To evaluate the representational capacity and generalization capability of our model, we conducted semantic segmentation experiments on the Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset. S3DIS is a widely used 3D indoor scene dataset that comprises scanned data from 272 rooms across 6 zones, covering a total area of approximately 6000 square meters. The dataset defines 13 semantic categories and provides fine-grained, point-wise semantic labels, where each point is annotated with comprehensive 9-dimensional feature information, including spatial coordinates (XYZ), color attributes (RGB), and normalized positional coordinates.

In our experiments, we fine-tuned the pre-trained model on all areas except Area 5 (the largest region in the dataset) and evaluated it on Area 5. The backbone network of our model is PointNet. To ensure experimental fairness, we strictly adhered to the experimental protocol proposed by Qi et al.[5] and Wang et al.[7] Specifically, we divided each room into small blocks of 1 m × 1 m and randomly sampled 4096 points from each block as inputs to the model, using only geometric features (XYZ coordinates) for each point.

The experimental results are summarized in Table 4. Our method demonstrates significant performance improvements compared to existing supervised and self-supervised learning approaches. Specifically, our method achieves a mIoU (mean Intersection-over-Union) improvement of 0.2 over the Multi-view rendering method. Furthermore, by applying self-supervised pre-training on PointNet, our approach achieves a 1.5-point gain in mIoU compared to PointNet trained from scratch. These results clearly highlight the effectiveness of self-supervised pre-training in learning robust and transferable features, particularly in scenarios with limited labeled data.

Ablations and analysis

Impact of high-order mixed moment. To assess the impact of high-order mixed moments, we conducted ablation studies with three loss functions: i) an invariance-based loss as a baseline; ii) the baseline plus a second-order mixed moment loss to evaluate its effectiveness; iii) the second-order loss plus a third-order mixed moment loss to understand the third-order's impact. Table 5 presents the comparative results. Relying only on the invariance-based loss led to an uneven feature distribution and reduced classification accuracy. Adding the second-order mixed moment mitigated model collapse and redundancy, significantly improving the model's representational power for both PointNet and DGCNN across datasets. The third-order mixed moment further minimized redundancy. It increased classification accuracy by 0.8% for PointNet and 1.7% for DGCNN on ModelNet40 compared to using only the second-order moment. These findings underscore the importance of incorporating higher-order mixed moments.

Figure 4 visualizes features from the ModelNet10[40] test set using t-SNE, extracted with the pre-trained PointNet. Incorporating second- and third-order mixed moments improved the separability of different classes. Notably, the addition of the third-order mixed moment allowed for clearer differentiation of objects with less distinct boundaries, including sofas, beds, and bathtubs.

Sensitivity analysis of λ. Intuitively, the parameter λ significantly influences pre-training and, consequently, the performance on downstream tasks. Our study examines how varying λ affects classification performance. We tested λ values from 0.001 to 5, using PointNet as the backbone network for pre-training, and conducted linear classification on both ScanObjectNN and ModelNet40 datasets. As Table 6 shows, an optimal classification performance on both datasets was achieved when λ was set to 0.5.

Application on the Terracotta Warrior Dataset

Terracotta Warriors Dataset. The Terracotta Warriors, renowned as one of the Seven Wonders of the World, represent a significant ceramic cultural relic in China. Their virtual restoration holds great importance for cultural heritage preservation and transmission. This study focuses on the 3D digitization and processing of Terracotta Warrior fragments for neural network analysis. Our dataset was acquired using a Creaform VIU 718 handheld 3D scanner in the Visualization Laboratory. Due to the high resolution of the resulting point clouds, which poses challenges for direct neural network input, we employed a preprocessing step. The Clustering Decimation method, available in the Meshlab tool, was utilized to downsample the point cloud data. This approach effectively preserves structural information while reducing each fragment to a uniform 2048 points. The dataset was categorized according to the anatomical parts of the Terracotta Warriors: arms, heads, legs, and bodies (as illustrated in Fig. 5). The sample distribution across these categories is presented in Table 7. For our experimental protocol, we adopted an 80–20 split, allocating 80% of the data for training and the remaining 20% for testing.

Classification of Terracotta Warrior Dataset. To validate the efficacy of our method on real-world 3D Terracotta Warrior fragments, we conducted classification experiments using our dataset. We fine-tuned a pre-trained DGCNN model on the Terracotta Warrior dataset, with the

Table 5 | The accuracy of linear SVM classification using retrained embedding on ModelNet40 and ScanObjectNN for PointMoment

Encoder	invariance	two-order mixed moment	three-order mixed moment	Acc.	
				ModelNet40	ScanObjectNN
PointNet	√			40.5	40.6
	√	√		88.0	73.4
	√	√	√	88.8	75.4
DGCNN	√			77.3	56.6
	√	√		89.3	79.3
	√	√	√	91.0	80.5

Fig. 4 | The T-SNE feature visualization on the ModelNet10 test set, post the training of PointNet as the self-supervised backbone network. The feature learned by three-order mixed moment(right) provides better discrimination of classes (e.g., sofas, beds and bathtubs) than using only invariance(left) or two-order mixed moment(middle).

Table 6 | Linear classification results for different λ parameters on ModelNet40 and ScanObjectNN datasets after pre-training using PointNet

λ	Acc.	
	ModelNet40	ScanObjectNN
0.001	72.5	57.3
0.005	78.5	62.9
0.01	81.2	66.6
0.05	84.4	71.4
0.1	88.5	74.8
0.5	88.8	79.0
1	88.0	76.4
5	87.8	74.1

results presented in Table 8. It's worth noting that research on self-supervised representation learning methods based on Terracotta Warrior point cloud data is scarce. Consequently, our comparisons primarily involve traditional and supervised methods. The highest accuracy achieved by existing traditional methods is 87.64%. Our approach significantly outperforms this benchmark, demonstrating an improvement of 4.46%. Moreover, our method yields competitive results when compared to supervised learning approaches. These experimental outcomes indicate that our technique effectively bridges the gap between supervised and self-supervised learning. It provides a robust set of initial model parameters for the task of Terracotta Warrior fragment classification, thus contributing a valuable research methodology for the virtual restoration of these artifacts. These results not only demonstrate the potential of our method for the specific task of Terracotta Warrior restoration but also suggest its applicability to broader cultural heritage preservation efforts involving 3D artifact reconstruction.

Segmentation of Terracotta Warrior Dataset. Segmentation of Terracotta Warriors plays a crucial role in the effective and accurate restoration of cultural relics, particularly in the virtual reconstruction of ceramic artifacts. Unlike the Terracotta fragment classification task, our segmentation study utilizes complete Terracotta Warrior models. We compiled a dataset of 150 complete Terracotta models using 3D scanners and data augmentation techniques, and all terracotta warriors models are uniformly downsampled into 4096 point clouds. Traditionally, the three-dimensional model of a Terracotta Warrior is divided into six parts: head, body, left arm, right arm, left leg, and right leg. However, to enhance the restoration process and rigorously evaluate our method's performance, we manually annotated the original Terracotta models into eight distinct segments: head, body, left hand, left arm, right hand, right arm, left leg, and right leg. We employed an 8-2 split for our dataset, allocating 80% for training and 20% for testing. To validate the effectiveness of our approach in segmenting Terracotta Warriors, we fine-tuned a pre-trained Dynamic Graph CNN (DGCNN) on our Terracotta Warriors dataset. The segmentation results are presented in Table 9. The empirical evidence demonstrates that our approach significantly outperforms existing unsupervised segmentation methods for Terra-cotta warriors. Specifically, our method achieves improvements of 6.8% and 3.8% in segmentation accuracy compared to SRG(DGCNN) and EGG(DGCNN), respectively. The resulting segmentation outcomes are illustrated in Fig. 6. The visual results demonstrate that our method achieves high-quality segmentation, effectively distinguishing between the eight predefined parts of the Terracotta Warriors. These results not only showcase the capability of our method in accurately segmenting Terracotta Warrior models but also highlight its potential in facilitating more precise virtual restoration processes. The improved granularity of segmentation (eight parts instead of six) allows for more detailed analysis and reconstruction, potentially leading to more accurate and comprehensive restoration outcomes.

Conclusion

The digital preservation of cultural heritage has become increasingly crucial in our technologically advancing world. This paper has explored the application of point cloud self-supervised learning technology to the Terracotta Warriors, introducing high-order mixed moments as an innovative

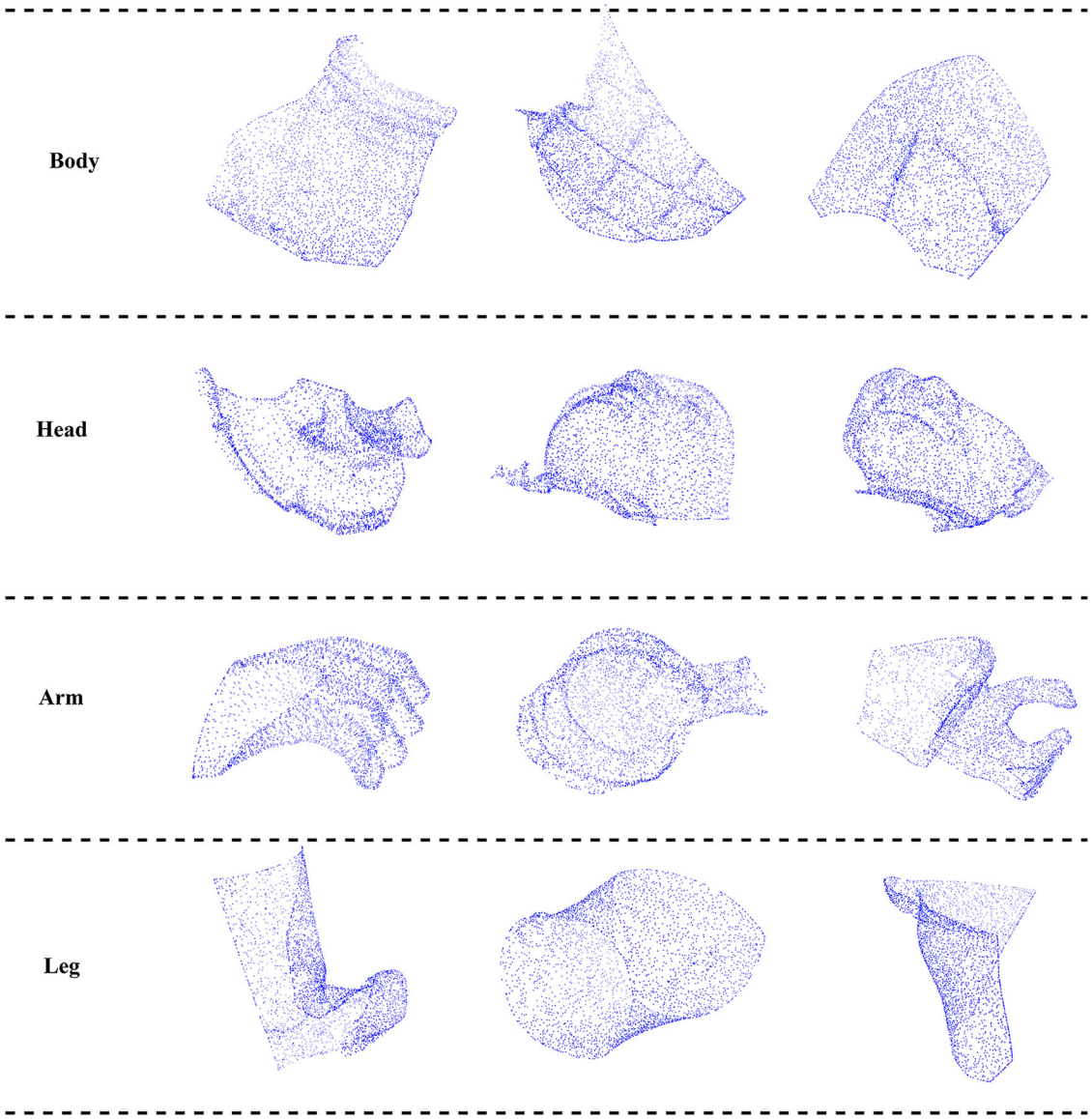

Fig. 5 | Illustration of the Terracotta Warriors fragments.

Table 7 | Number of fragments for each class in the Terracotta Warriors fragments dataset

Label	Arm	Body	Head	Leg	Total
Train	4178	4738	2430	4274	15,620
Test	1045	1185	607	1068	3905

Table 8 | Compared with other methods on the 3D Terracotta Warrior fragment datasets

Method	Acc.
Method in[2]	87.64
Method in[1]	91.41
PointNet[5]	88.93
PointMoment(Ours)	92.10

approach to enhance feature characterization while reducing redundant information in high-dimensional embedded features. Firstly, our research demonstrates the significant potential of high-order mixed moments in feature redundancy reduction. By effectively minimizing redundant information in high-dimensional embedded features, we have developed a feature extractor with superior representational capabilities. This approach results in more independent and compact representational information, which is crucial for accurate analysis and restoration of complex artifacts like the Terracotta Warriors. Secondly, a key advantage of our method is its ability to address the model collapse problem inherent in self-supervised learning without resorting to complex techniques such as asymmetric network frameworks. This simplification in implementation, while maintaining robust performance, represents a significant step forward in the field of self-supervised learning for 3D data. Furthermore, extensive testing on multiple downstream tasks using existing public datasets has validated the versatility and effectiveness of our technique. The competitive results achieved across various applications underscore the broad applicability of our approach beyond the specific context of cultural heritage preservation. Notably, when applied to the Terracotta Warriors dataset, our method has shown remarkable performance, outperforming existing approaches in both fragment classification accuracy and segmentation precision. These results are particularly significant given the complexity and historical importance of the Terracotta Warriors. The improved accuracy in

Table 9 | Comparison of different methods on Terracotta Warrior dataset

Method	mIoU
SRG[49] + DGCNN	65.6
SRG[49]+PointNet	54.4
EGG[50] + DGCNN	68.6
EGG[50]+PointNet	62.4
PointMoment(Ours)	72.4

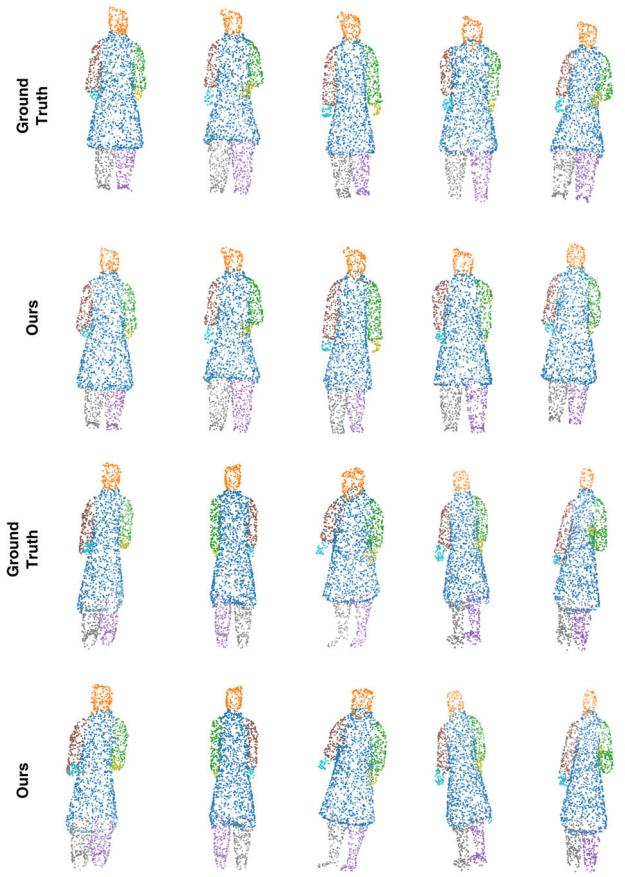

Fig. 6 | PointMoment segmentation results on Terracotta Warrior Datasets.

classification and segmentation can potentially lead to more precise virtual restorations and deeper insights into the manufacturing techniques and artistic styles of ancient China. Looking ahead, we recognize that the computational complexity of calculating high-order mixed moments is a challenge that requires further improvement. To address this, we plan to explore more efficient algorithms to approximate or estimate high-order mixed moments. Our initial idea is to utilize feature decomposition techniques to decompose high-order mixed moments into combinations of lower-order moments, thereby reducing computational complexity. Additionally, we can employ block processing techniques to divide large-scale data into smaller chunks, compute each separately, and then combine the results. This approach would reduce memory usage and improve computational efficiency. Furthermore, we intend to perform random sampling on the final result matrix based on existing high-order mixed-moment algorithms. By avoiding the inclusion of all matrix elements in the loss calculation, we can significantly reduce computational complexity. Specifically, we plan to adopt Monte Carlo sampling techniques to randomly select a subset of matrix elements for estimation, approximating the true values. We will then apply weighted sampling according to the importance of feature variables to ensure that the impact of key features is fully considered. Finally, during backpropagation, we will compute gradients only for the important sampled points, further reducing computational load. This continued research aims to optimize the performance of our methods and push the boundaries of what's possible in digital cultural heritage preservation. Additionally, we anticipate that our approach could be extended to other types of 3D cultural artifacts, potentially opening up new possibilities in the field of virtual museums and digital archiving. In conclusion, our work not only presents a novel technical approach but also demonstrates the transformative potential of interdisciplinary research combining computer science and archaeology. As we continue to refine and expand these techniques, we move closer to a future where our cultural heritage is not only preserved but also made more accessible and understandable through advanced digital technologies. The success of our method in both general point cloud tasks and specific Terracotta Warrior applications underscores the potential of high-order mixed moments as a powerful technique for point cloud representation learning, opening up new avenues for research in self-supervised learning for 3D data across various domains.

Data availability

Data underlying the results presented in this paper can be obtained from the internet. The Terracotta Warriors data will be available upon reasonable request.

Received: 27 August 2024; Accepted: 28 December 2024;
Published online: 10 June 2025

References

1. Yang, K., Cao, X., Geng, G., Li, K. & Zhou, M. Classification of 3D terracotta warriors fragments based on geospatial and texture information. *J. Vis.* **24**, 251–259 (2021).
2. Du, G., Zhou, M., Yin, C., Wu, Z. & Shui, W. Classifying fragments of terracotta warriors using template-based partial matching. *Multimed. Tools Appl.* **77**, 19171–19191 (2018).
3. Rasheed, N. A. & Nordin, M. J. Classification and reconstruction algorithms for the archaeological fragments. *J. King Saud. Univ.-Computer Inf. Sci.* **32**, 883–894 (2020).
4. Lin, X., Xue, B. & Wang, X. Digital 3D Reconstruction of Ancient Chinese Great Wild Goose Pagoda by TLS Point Cloud Hierarchical Registration. *ACM J. Comput. Cult. Herit.* **17**, 1–16 (2024).
5. Qi, C. R., Su, H., Mo, K. & Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 652–660 (IEEE, Honolulu, HI, 2017).
6. Qi, C. R., Yi, L., Su, H. & Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv. Neural. Inf. Process Syst*. https://doi.org/10.48550/arXiv.1706.02413 (2017).
7. Wang, Y. et al. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph. (tog)* **38**, 1–12 (2019).
8. Zhou, P., An, L., Wang, Y. & Geng, G. MLGTM: Multi-Scale Local Geometric Transformer-Mamba Application in Terracotta Warriors Point Cloud Classification. *Remote Sens.* **16**, 2920 (2024).
9. Zhu, J., Fang, B., Chen, T. & Yang, H. Face repairing based on transfer learning method with fewer training samples: application to a Terracotta Warrior with facial cracks and a Buddha with a broken nose. *Herit. Sci.* **12**, 186 (2024).
10. Grill, J.-B. et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
11. Poursaeed, O., Jiang, T., Qiao, H., Xu, N. & Kim, V. G. Self-supervised learning of point clouds via orientation estimation. In *Proc. International Conference on 3D Vision (3DV)* 1018–1028 (IEEE, 2020).
12. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning* 1597–1607 (PMLR, 2020).

13. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proc. International Conference on Machine Learning* (12310–12320) (PMLR, 2021).
14. Niu, C. & Wang, G. HOME: High-order mixed-moment-based embedding for representation learning. Preprint at *arXiv:220707743*. https://doi.org/10.48550/arXiv.2207.07743 (2022).
15. Xiao, A. et al. Unsupervised point cloud representation learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 11321–11339 (2023).
16. Sohail, S. S., Himeur, Y., Kheddar, H., Amira, A., Fadli, F., Atalla, S. et al. Advancing 3D point cloud understanding through deep transfer learning: A comprehensive survey. *Inf. Fusion*. **113**, 102601 (2024).
17. Achlioptas, P., Diamanti, O., Mitliagkas, I. & Guibas, L. Learning representations and generative models for 3d point clouds. In *Proc. International Conference on Machine Learning* (40–49) (PMLR, 2018).
18. Yang, Y., Feng, C., Shen, Y. & Tian, D. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (206–215) (IEEE, 2018).
19. Wang, H., Liu, Q., Yue, X., Lasenby, J. & Kusner, M. J. Unsupervised point cloud pre-training via occlusion completion. In *Proc. IEEE/CVF International Conference on Computer Vision.* (9782–9792) (IEEE, 2021).
20. Rao, Y., Lu, J. & Zhou, J. PointGLR: Unsupervised structural representation learning of 3D point clouds. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 2193–2207 (2022).
21. Xie, S., Gu, J., Guo, D., Qi, C. R, Guibas, L. & Litany. O. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In In *Proc. Part III 16 Computer Vision–ECCV 2020: 16th European Conference* (574–591) (Springer, Glasgow, 2020).
22. Afham, M., Dissanayake, I., Dissanayake, D., Dharmasiri, A., Thilakarathna, K. & Rodrigo R. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* (9902–9912) (IEEE, 2022).
23. Sanghi, A. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Proc. Part XXIX 16 Computer Vision–ECCV 2020: 16th European Conference* (626–642) (Springer, Glasgow, 2020).
24. Huang, S., Xie, Y., Zhu, S.-C. & Zhu, Y. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proc. IEEE/CVF International Conference on Computer Vision* (6535–6545) (IEEE, 2021).
25. Wang, Y., Zhou, P., Geng, G., An, L. & Zhou, M. Enhancing point cloud registration with transformer: cultural heritage protection of the Terracotta Warriors. *Herit. Sci.* **12**, 314 (2024).
26. Xu, X., Song, D., Geng, G., Zhou, M., Liu, J. & Li, K. et al. CPDC-MFNet: conditional point diffusion completion network with Muti-scale Feedback Refine for 3D Terracotta Warriors. *Sci. Rep.* **14**, 8307 (2024).
27. Su, H., Maji, S., Kalogerakis, E. & Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. IEEE International Conference on Computer Vision* (945–953) (IEEE, 2015).
28. Wei, X., Yu, R. & Sun, J. View-GCN: View-based graph convolutional network for 3D shape analysis. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1850–1859 (IEEE, 2020).
29. Maturana, D. & Scherer, S. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 922–928 (IEEE, 2015).
30. Zhao, H., Jiang, L., Fu, C.-W. & Jia, J. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5565–5573 (IEEE, 2019).
31. Li, Y., Bu, R., Sun, M., Wu, W., Di, X. & Chen, B. Pointcnn: Convolution on x-transformed points. *Adv. Neural Inf Process. Syst.* https://doi.org/10.48550/arXiv.1801.07791 (2018).
32. Xu, Y., Fan, T., Xu, M., Zeng, L. & Qiao Y. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proc. European Conference on Computer Vision* 87–102 (ECCV, 2018).
33. Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R. & Hu, S.-M. Pct: Point cloud transformer. *Comput. Vis. Media* **7**, 187–199 (2021).
34. Zhao, H., Jiang, L., Jia, J., Torr, P. H, & Koltun, V. Point transformer. *In Proc IEEE/CVF International Conference on Computer Vision* 16259–16168 (IEEE, 2021).
35. Yu, X., Tang, L., Rao, Y., Huang, T., Zhou, J. & Lu, J. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 19313–19322 (IEEE, 2022).
36. Pang, Y., Wang, W., Tay, F. E, Liu, W., Tian, Y. & Yuan L. Masked autoencoders for point cloud self-supervised learning. In *Proc. European Conference on Computer Vision* 604–621 (Springer, 2022).
37. Zhang, Z., Girdhar, R., Joulin, A. & Misra, I. Self-supervised pretraining of 3d features on any point-cloud. In *Proc. IEEE/CVF International Conference on Computer Vision* (10252–10263) (IEEE, 2021).
38. Chang, A. X, Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., et al. Shapenet: An information-rich 3d model repository. Preprint at *arXiv:151203012*. https://doi.org/10.48550/arXiv.1512.03012 (2015).
39. Uy, M. A., Pham, Q.-H., Hua, B.-S., Nguyen, T. & Yeung, S.-K. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proc. IEEE/CVF International Conference on Computer Vision* 1588–1597 (IEEE, 2019).
40. Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L. Tang, X., et al. 3d shapenets: A deep representation for volumetric shapes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 1912–1920 (IEEE, 2015).
41. Yi, L., Kim, V. G., Ceylan, D., Shen, I.-C., Yan, M. & Su, H. et al. A scalable active framework for region annotation in 3d shape collections. *ACM Trans. Graph. (ToG)* **35**, 1–12 (2016).
42. Wu, J., Zhang, C., Xue, T., Freeman, B., & Tenenbaum, J. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Adv. Neural. Inf. Process Syst.* https://doi.org/10.48550/arXiv.1610.07584 (2016).
43. Li, J., Chen, B. M. & Lee, G. H. So-net: Self-organizing network for point cloud analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 9397–9406 (IEEE, 2018).
44. Gadelha, M., Wang, R. & Maji, S. Multiresolution tree networks for 3d point cloud processing. In *Proc. European Conference on Computer Vision* 103–118 (ECCV, 2018).
45. Zhao, Y., Birdal, T., Deng, H., & Tombari, F. 3D point capsule networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 1009–1018 (2019).
46. Han, Z., Wang, X., Liu, Y.-S. & Zwicker, M. Multi-angle point cloud-VAE: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 10441–10450 (IEEE, 2019).
47. Sauder, J. & Sievers, B. Self-supervised deep learning on point clouds by reconstructing space. *Adv. Neur. Inf. Process. Syst*. https://doi.org/10.48550/arXiv.1901.08396 (2019).
48. Tran, B., Hua, B.-S., Tran, A. T. & Hoai, M. Self-supervised learning with multi-view rendering for 3d point cloud analysis. In *Proc. Asian Conference on Computer Vision* 3086–3103 (2022).
49. Hu, Y., Zhou, W., Geng, G., Li, K., Hao, X. & Cao, X. Unsupervised segmentation for terracotta warrior with seed-region-growing CNN (SRG-Net). In *Proc 5th International Conference on Computer Science and Application Engineering* 1–6 (2021).
50. Hu, Y., Geng, G., Li, K., Guo, B. & Zhou, P. Self-Supervised Segmentation for Terracotta Warrior Point Cloud (EGG-Net). *IEEE Access* **10**, 12374–12384 (2022).

Acknowledgements

We would like to acknowledge Dr. Chuang Niu from Rensselaer Polytechnic Institute for providing technical guidance and expertise that greatly assisted our research. We thank Emperor Qinshihuang's Mausoleum Site Museum

for providing the Terracotta Warriors data. This work was supported in part by the Key Research and Development Program of Shaanxi Province (2019GY-215, 2021ZDLSF06-04, 2024SF-YBXM-681).

Author contributions

Conceptualization, Xin Cao and Xinxin Han; methodology, Xin Cao and Xinxin Han; software, Wenlong Tang; validation, Kang Li; formal analysis, Yong Ren; investigation, Xin Cao; resources, Ping Zhou; data curation, Ping Zhou; writing—original draft preparation, Xin Cao; writing—review and editing, Xin Cao; supervision, Linzi Su; project administration, Linzi Su; funding acquisition, Kang Li. All authors have read and agreed to the published version of the manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

Written informed consent has been obtained from the School of Information Science and Technology of Northwest University, and all authors for this article, and consent has been obtained for the data used.

Additional information

Correspondence and requests for materials should be addressed to Kang Li or Linzhi Su.