npj | heritage science

Article

https://doi.org/10.1038/s40494-026-02391-0

InSwAV: involution enhanced feature clustering and swapped assignments for porcelain relic microscopic image classification

Check for updates

Yangyang Liu, Jiatong Liu, Xinda Liu ✉, Guohua Geng ✉ & Zhan Li

Accurate classification of porcelain relic fragments is essential for restoration. Traditional methods, which depend on visual traits like glaze color and patterns, perform poorly with visually similar fragments, especially when microscopic image samples are scarce. To solve these issues, we introduce InSwAV, which integrates involution-enhanced residual blocks for robust feature extraction and employs a swapped assignment mechanism. This mechanism aligns deep features with learnable cluster prototypes by enforcing consistency between differently augmented views of the same image. The model is optimized by minimizing a cross-entropy loss against cluster assignments, which reduces training time significantly. we constructed the Porcelain Relic Microscopic Images (PRMI) dataset of five classification, with data augmentation applied to enhance model robustness. Experimental results show that InSwAV achieves a classification accuracy of 96.2% on the porcelain relic microscopic image dataset, outperforming existing methods.

Over the past century of archeological excavations, numerous porcelain relics have been unearthed, serving not only as remnants of ancient manufacturing techniques but also as significant symbols of cultural evolution and technological advancement[1]. However, due to their inherent fragility and the cyclical processes of production, replacement, and degradation, archeological excavations frequently yield vast quantities of porcelain fragments. Traditional restoration faces significant challenges due to the vast quantity, diverse types, and wide geographical distribution of unearthed porcelain fragments, as illustrated in Fig. 1a. To address these challenges, we propose a comparative learning framework for the PRMI dataset based on involution and contrasting cluster assignments, which not only enhances restoration efficiency but also contributes substantially to cultural heritage preservation. Our laboratory has compiled and sorted out extensive porcelain fragments, as illustrated in Fig. 1b.

Prior to porcelain relic restoration, the accurate classification of fragment assemblages presents a technically demanding prerequisite task. Conventional manual classification approaches remain limited by their reliance on expert experience and low work efficiency. In contrast, computer-aided virtual restoration technology offers distinct advantages by minimizing human interference, operating without temporal or spatial constraints, and achieving a non-contact and reversible restoration process that effectively prevents the secondary damage to relics[2,3].

Due to the visual similarities among numerous specimens, reliable classification of porcelain fragments based solely on macroscopic appearance remains challenging. To address this, microscopic analysis of porcelain microstructure has emerged as a valuable alternative approach[4]. Microscopic analysis reveals the distinctive bubble structures within porcelain fragments, providing critical diagnostic features that enhance both the efficiency and accuracy of classification. Moreover, these microstructural characteristics provide a robust scientific foundation for the authentication of porcelain relics[5].

The formation of bubbles in porcelain is intricately linked to multiple stages of the manufacturing process, including material selection, crushing, slurry processing, clay refinement, shaping, cooling, glazing, and firing. Each of these stages exerts a direct influence on the development of bubble structures. Because most of the ancient porcelain production techniques were mastered by private workshops, even the porcelain produced by the same workshop had significant differences in the characteristics of bubbles[6]. These variations serve as critical diagnostic markers for classifying and authenticating porcelain relics.

In recent years, with the emergence of various microscopic instruments, people have begun to analyze and study microscopic images in various fields. Jin et al.[7] pioneered macroscopic ophthalmological examination of ancient porcelain, systematically categorizing distinct

School of Information Science and Technology, Northwest University, Xi'an, Shaanxi, China. ✉e-mail: liuxinda@nwu.edu.cn; ghgeng@nwu.edu.cn

Fig. 1 | Sort out a large number of porcelain fragments. a A large number of porcelain fragments excavated, **b** sorting porcelain fragments in the laboratory.

microstructural features that provide new dimensions for porcelain identification. Li et al.[8] developed a wavelet-based denoising approach combined with watershed segmentation, significantly improving bubble feature extraction from microscopic images. Zhu et al.[9] conducted a comprehensive analysis of bubble distribution patterns in Yaozhou porcelain across different eras, establishing scientific foundations for chronological classification while revealing technological influences on bubble formation. Liu[10] advanced this field through curvelet transform-based denoising and feature enhancement, achieving superior bubble boundary clarity via optimized watershed segmentation. While these traditional image processing methods offer certain improvements, they retain their sensitivity to noise and their inability to extract high-level semantic features.

Meanwhile, unsupervised and self-supervised learning paradigms have emerged as powerful alternatives, enabling effectively learning of representative features from unlabeled data. Notable among these are deep clustering techniques, such as Deep Embedding for Clustering[11] and Semantic Clustering by Adopting Nearest Neighbors[12], which iteratively refines feature representation and cluster assignments, demonstrating significant potential. Similarly, contrastive learning frameworks, including MoCo[13] and BYOL[14], have achieved performance comparable to supervised learning by constructing positive and negative sample pairs for instance discrimination. The success of these foundational methodologies extends beyond image classification; they have also been successfully adapted to more complex visual tasks such as 3D human pose estimation[15–18], highlighting their versatility and robustness in capturing intricate data patterns. Inspired by the success of methods that model complex relationships in data, such as graph-based approaches for 3D pose estimation[16,17], our framework is designed to learn meaningful semantic structures within the feature space. Instead of direct instance comparison, we employ online clustering and swapped assignment to guide the model in discovering consistent and robust data groupings, effectively learning a structured representation of the porcelain microscopic features. Inspired by these advancements, our work seeks to adapt and tailor these powerful approaches to the unique challenges inherent in porcelain relic microscopic image analysis.

Image classification techniques can mainly be divided into two categories, comprising traditional feature extraction methods and deep learning-based techniques. Traditional machine learning approaches[19], such as decision tree[20], k-nearest neighbor[21] and support vector machines[22], exhibit limitations in generalization performance, noise robustness, and complex image processing capability. while deep learning approaches[23] demonstrate superior performance in both efficiency and accuracy.

In computer vision, image classification techniques are primarily classified as supervised or unsupervised based on annotation requirements[24]. Given the considerable time and effort demanded by manual labeling, unsupervised learning approaches have emerged as viable alternatives that operate without labeled training data[25]. Unsupervised image classification techniques can be classified into methods based on traditional machine learning and deep learning. Traditional machine learning primarily relies on statistical methods, which are susceptible to overfitting[26]. Deep learning-based methods that integrate unsupervised principles with neural networks, utilizing contrastive learning to extract robust features from unlabeled data[27].

In unsupervised deep learning classification, Wu et al.[28] first proposed InstDisc, a pioneering contrastive learning framework in computer vision. Their approach selects the computed image and its augmented version as positive samples, while treating all other images in the datasets as negative samples stored in a memory bank. However, this method demands substantial computational resources for large datasets. To address this limitation, Ye et al.[29] introduced an alternative negative sample selection strategy using mini-batches, where each image's augmented version served as the positive sample and other images in the batch as negative samples. This approach significantly reduced the negative sample quantity, leading to suboptimal performance. He et al.[13,30,31] subsequently developed a dynamic contrastive learning method employing a gradient-free queue to store numerous negative sample features, albeit with increased network complexity. Building upon these works, Chen et al.[32] proposed a simplified framework where features from the same encoder were processed through an MLP layer for contrastive training. More recently, Caron et al.[33] introduced SwAV, which utilizes cluster centers for comparison, achieving near-supervised classification accuracy while maintaining computational efficiency.

While recent foundation models like DINOv2[34] have set new benchmarks in self-supervised learning by leveraging large-scale curated data and distillation techniques, their computational demands and data requirements pose challenges for resource-constrained, domain-specific applications like cultural heritage analysis. In this work, we therefore focus on enhancing the data-efficient clustering-based paradigm of SwAV[33], tailoring it with novel architectural designs to address the unique challenges of porcelain relic microscopic image classification under limited samples.

Regarding image classification methods for Porcelain relic fragments. Wang et al.[35] developed an IGabor filter-based color-texture feature extraction model with SVM classification, achieving high accuracy in porcelain fragment categorization. Li et al.[36] proposed a lightweight capsule network incorporating ChannelTrans modules and MF-R loss functions to address feature extraction speed and class imbalance. In microscopic image analysis, Wang et al.[37] conducted a comprehensive comparative analysis of four deep learning architectures for archeological porcelain provenance classification, including VGG-16, Inception-v3, GoogLeNet and their ensemble model, while Liu et al.[38] introduced a multi-channel GCN architecture with multi-directional Gaussian filtering to handle noise and feature distribution challenges.

Porcelain relic microscopic images show special features that differ from regular image datasets. These data face three main problems. First, annotations rely on expert experience due to complex variations in texture, color, and morphology, along with potential interference from damage and stains. Second, sample sizes are typically limited, and category distributions are imbalanced due to excavation and preservation constraints. Third, taking microscope pictures is hard. To address these problems, we employ contrastive learning for microscopic image classification of porcelain relic

fragments. Compared with the supervised method, the unsupervised method is more suitable for small sample datasets.

For the porcelain relic microscopic images, Geng et al.[39] developed an enhanced SimCLR[32] framework incorporating multi-scale feature extraction of Res2Net, significantly improving local and global feature representation. Their method optimizes feature precision through contrastive learning that maximizes positive sample similarity while minimizing negative sample differences, achieving high-accuracy unsupervised classification. Zhou et al.[40] developed HySiam, a novel contrastive learning approach integrating dual-channel attention mechanisms. This hybrid attention system combines with convolutional networks to enhance semantic feature extraction while maintaining computational efficiency through reduced parameters, demonstrating improved network performance for porcelain analysis.

Although traditional convolution has achieved remarkable results in computer vision tasks, its inherent characteristics have gradually exposed limitations in the processing of microscopic images. The spatial invariance and channel specificity of the convolution kernel make it difficult to adapt to the diverse feature patterns of microscopic images of porcelain relic fragments. These constraints are particularly evident in two critical aspects: (1) The fixed-size convolution kernel cannot effectively capture the multi-scale characteristics of key features such as the distribution of glaze bubbles; (2) The parameter-sharing mechanism leads to computational redundancy when extracting distinctive textural features, including clay-glaze interfaces and crack propagation patterns.

To overcome these limitations, we present an innovative framework incorporating three key technical contributions. First, we introduce involution operations[41] to replace standard convolution, leveraging their dynamic kernel generation capability to enhance feature adaptation. Building upon this, we introduce InSwAV, a framework for porcelain relic microscopic image classification based on contrastive clustering and involution, which effectively addresses the core problems of imprecise feature extraction and low training efficiency. For model optimization, we adopt a dichotomy strategy with two components: (1) Quantitative evaluation of feature module stacking configurations to determine optimal architectural depth; (2) Systematic binary search within the [0, 300] range to identify ideal cluster center quantities. Experimental validation on the PRMI dataset demonstrates that our method significantly reduces computational complexity while maintaining classification accuracy. The proposed approach achieves a classification accuracy of 96.2%, representing improvements of 2.1%, 4.6%, and 12.8% over SwAV, SimCLR, and MoCo, respectively.

Overall, our contributions are as follows:

- We propose ResInv, a novel model combining involution operations with conventional convolution layers. This hybrid architecture effectively captures long-range spatial dependencies in porcelain relic microscopic images through position-sensitive modeling capability of involution. The proposed method overcomes the inherent limitations of conventional convolutional kernel dimensions, substantially mitigates inter-channel feature redundancy, and strategically employs large-scale convolutional kernels to expand the network's receptive field, thereby achieving refined extraction of microstructure characteristics.
- We perform feature clustering on the features extracted by ResInv through cross-prediction mechanisms, and we enforce consistency in cluster assignments across multi-view augmented samples. The optimization process employs cross-entropy loss minimization in an iterative manner. This indirect comparison mechanism not only avoids the consumption of computing resources caused by direct feature comparison in traditional methods, but also significantly improves the model training efficiency, achieving a balance between accuracy and speed.
- We constructed the PRMI dataset of five classifications and validated the proposed method on this dataset. This PRMI dataset will be publicly available, providing extensive support for both academic research and practical applications.

Methods

Overall framework

The InSwAV network architecture proposed in this paper is illustrated in Fig. 2, which consists of five key modules: (1) Data Augmentation: We generate eight augmented views for each input image through random cropping, followed by multiple transformations including Gaussian blur and color jitter to enhance feature diversity while preserving essential microstructure characteristics. (2) Feature Extractor: We propose ResInv, which combines Residual Block(Res Block) with involution operations, effectively capturing both local and global features of porcelain fragments while reducing channel redundancy. (3) Online Clustering: The extracted features are clustered onto trainable prototype vectors that are generated by random initialization. The resulting cluster assignment probabilities are normalized using the Sinkhorn-Knopp algorithm to prevent the model from lazily assigning all image features to the same or a few cluster centers. (4) Swapped Prediction: The cluster assignments of different augmented views of the same image are swapped and compared using a cross-entropy loss. (5) Cross-Entropy Loss: We employ a cross-entropy loss function to optimize feature distribution in the embedding space, minimizing distances between features from the same image while maximizing separation between different images.

Involution

Involution enhances model performance and computational efficiency by inverting the fundamental properties of conventional convolution operations. Figure 3 illustrates the comparative schematics between conventional convolution and involution operations. Involution reverses the spatial invariance and channel specificity characteristics of convolution to achieve spatial specificity and channel invariance. That is, differentiated convolution kernels are used at different positions within the same channel, while shared weights are employed across all channels at each position, thereby enhancing both spatial and channel-wise feature representation capabilities.

In the computation of involution, individual pixels serve as the fundamental processing units. Specifically, each pixel $\mathbf{X}_{i,j} \in \mathbb{R}^C$ uses different involution kernels $\mathcal{H}_{i,j,\cdot,\cdot,g} \in \mathbb{R}^{K\times K}$, $g = 1, 2, \ldots, G$. The output feature map of involution is derived by performing Multiply-Add operations on the input with such involution kernels, defined as

$$\mathbf{Y}_{i,j,k} = \sum_{(u,v)\in\Delta_K} \mathcal{H}_{i,j,u+\lfloor K/2\rfloor,v+\lfloor K/2\rfloor,\lfloor kG/C\rfloor}\mathbf{X}_{i+u,j+v,k}. \quad (1)$$

For an image X with a size of $H \times W \times C$, all channels are divided into groups G ($G = 1$ in this example for ease of demonstration). The shape of involution kernels $\mathbf{H}$ depends on that of the input feature map $\mathbf{X}$. The kernel generation function is symbolized as ϕ, and the function mapping at position (i, j) is abstracted as

$$\mathcal{H}_{i,j} = \phi(\mathbf{X}_{i,j}), \quad (2)$$

where the kernel generation function ϕ dynamically generates spatially adaptive convolution kernels at each position through a lightweight multi-layer perceptron(MLP):

$$\phi(\mathbf{X}_{i,j}) = \mathbf{W}_1\sigma(\mathbf{W}_0\mathbf{X}_{i,j}), \quad (3)$$

where σ implies the Batch Normalization and non-linear activation functions that interleave two linear projections:

$$\sigma(x) = \mathrm{ReLU}(\mathrm{BN}(x)). \quad (4)$$

In Formula(3), $\mathbf{W}_0 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_1 \in \mathbb{R}^{(K\times K\times G)\times\frac{C}{r}}$ represent two linear transformations that constitute the generation function of the involution kernel, the hyperparameter r is the channel compression ratio, and G is the number of channel groups. Each pixel is rearranged following computation by this function, which flattens the channel dimension into the spatial dimension to generate respective inner convolution kernels. Finally,

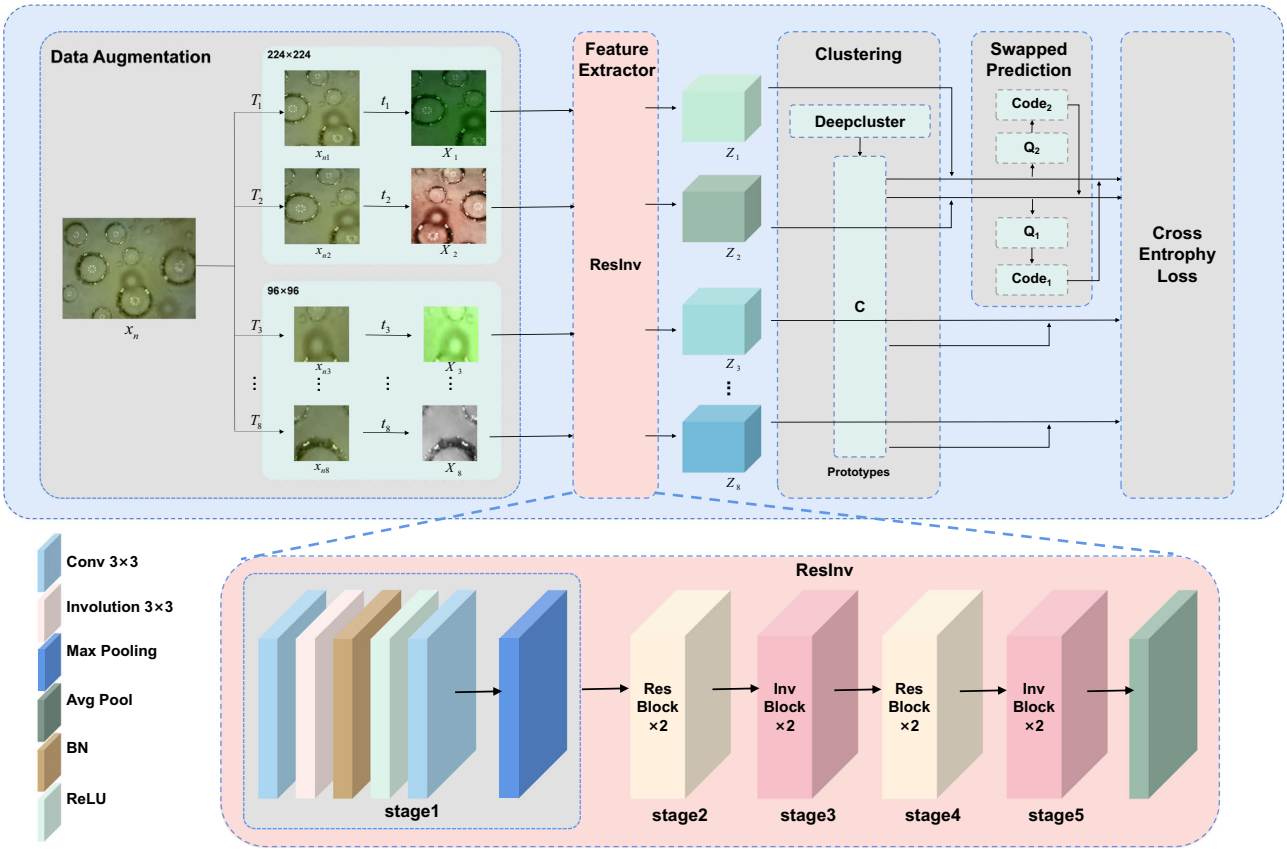
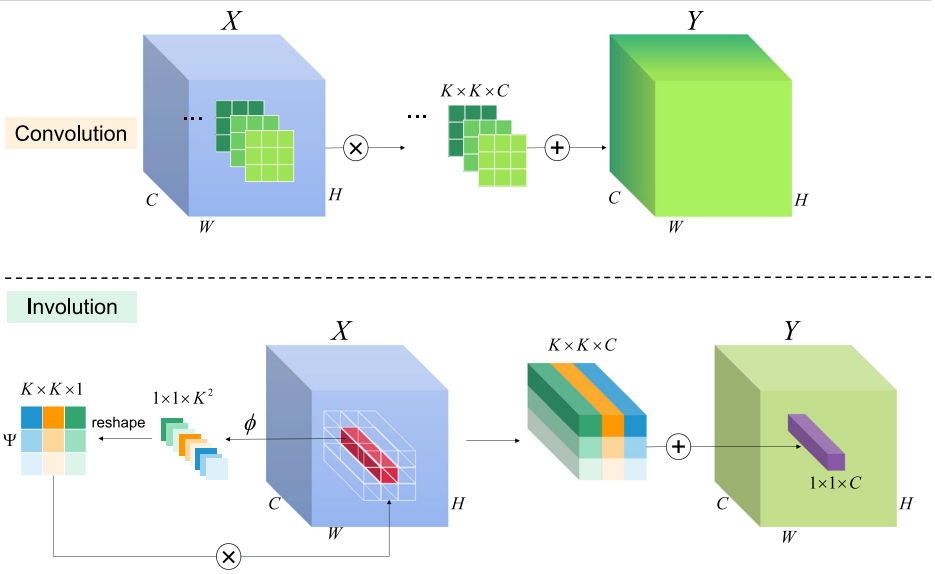


Fig. 2 | The network framework of our proposed InSwAV. The model comprises five parts:(1) Data augmentation generates eight views per image via cropping, blur, and jitter. (2) The ResInv extractor merges ResBlocks with involution for multi-scale features. (3) Online clustering uses Sinkhorn-Knopp normalization to assign features to prototypes. (4) Swapped prediction contrasts cluster assignments across different views. (5) A cross-entropy loss optimizes the embedding space, pulling together features from the same image while pushing others apart.

Fig. 3 | The framework of convolution and involution. The above figure is a schematic diagram of a traditional convolution operation, and the following figure is a schematic diagram of involution operation.



the convolution kernels obtained at each pixel will be multiplied and added to its $K \times K$ neighborhood. After the calculation, $1 \times 1 \times C$ pixel in the output image Y is obtained.

Feature extractor

To more effectively capture long-range dependencies in image data, enhance semantic correlations among similar images, and increase the adaptability of convolution kernels, we propose a feature extraction feature extraction model called ResInv. We introduce involution to Res Block and propose Involution Block(Inv Block). As illustrated in Fig. 4, the Involution Block architecture sequentially consists of 1×1 convolution layer, 7×7 involution layer, and 1×1 convolution layer.

We combined Res Blocks with Inv Blocks in ResInv. Extensive experimental comparisons demonstrate that the optimal performance is

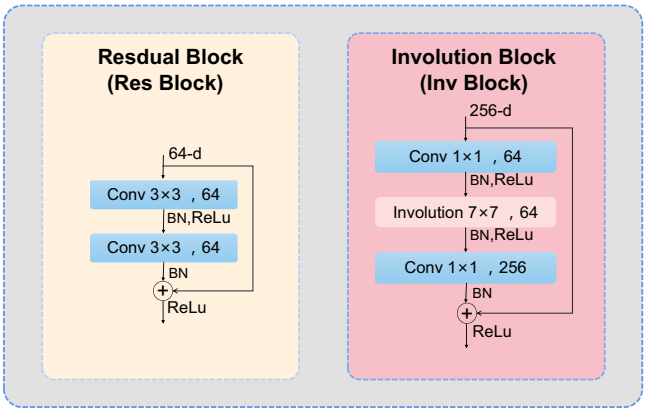


Fig. 4 | Residual Block and Involution Block. Res Block consists of two 3 × 3 convolutions layer, Inv Block consists of 1 × 1 convolution layer, 7 × 7 involution layer, and 1 × 1 convolution layer.

Table 1 | Architecture comparison of ResNet50 and ResInv

Stage	Output Size	ResNet50[47]	ResInv (Ours)
Conv1	112 × 112	7 × 7, stride=2	3 × 3, stride=2 Involution 3 × 3, stride=1
Conv2	56 × 56	3 × 3 max pool, stride=2 $\begin{bmatrix}1\times1, & 64\\ 3\times3, & 64\\ 1\times1, & 256\end{bmatrix}\times3$	3 × 3 max pool, stride=2 $\begin{bmatrix}3\times3, & 64\\ 3\times3, & 64\end{bmatrix}\times2$
Conv3	28 × 28	$\begin{bmatrix}1\times1,128\\ 3\times3,128\\ 1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\ \text{Involution}\,7\times7,128\\ 1\times1,512\end{bmatrix}\times2$
Conv4	14 × 14	$\begin{bmatrix}1\times1,256\\ 3\times3,256\\ 1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}3\times3,256\\ 3\times3,256\end{bmatrix}\times2$
Conv5	7 × 7	$\begin{bmatrix}1\times1,512\\ 3\times3,512\\ 1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\ \text{Involution}\,7\times7,512\\ 1\times1,2048\end{bmatrix}\times2$
	1 × 1	average pool	

achieved when Res Blocks and Inv Blocks are alternately combined in a ratio of 2: 2: 2: 2. The detailed architecture is shown in Table 1.

Clustering with swapped prediction

Unlike common contrastive learning methods that rely on direct feature comparison, our approach employs an online clustering strategy. The features extracted by ResInv are not compared pairwise. Instead, they are mapped to a set of K trainable prototype vectors, denoted as $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$. These prototype vectors are model parameters that are optimized simultaneously with the backbone network weights via gradient descent. Consequently, the network learns to generate features $\mathbf{z}$ that better align with the current prototypes $\mathbf{C}$, while the prototypes themselves are updated to more accurately represent the distribution centers of all feature vectors.

To ensure meaningful and balanced cluster assignments, the Sinkhorn-Knopp algorithm[42] is applied to the similarity scores between the entire mini-batch of features and the prototype vectors to generate soft normalized codes. For a given pair of features $\mathbf{z}_t$ and $\mathbf{z}_s$ produced from the same image, their corresponding codes $\mathbf{q}_t$ and $\mathbf{q}_s$ are then extracted from the resulting assignment matrix to compute the swapped prediction loss.

(1) Compute Assignment Probabilities: The similarity between each feature and all prototypes is calculated, producing a probability distribution over the prototypes using a softmax function of eq. (5).

$$\mathbf{p}_t^{(k)} = \frac{\exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\mathbf{z}_t^\top \mathbf{c}_{k'}\right)}, \mathbf{p}_s^{(k)} = \frac{\exp\left(\frac{1}{\tau}\mathbf{z}_s^\top \mathbf{c}_k\right)}{\sum_{k'} \exp\left(\frac{1}{\tau}\mathbf{z}_s^\top \mathbf{c}_{k'}\right)}. \tag{5}$$

where τ is the temperature parameter[43].

(2) Generate Robust Targets with Sinkhorn-Knopp: Without normalization, the initial soft codes $\mathbf{p}_t$ and $\mathbf{p}_s$ would otherwise lead to a degenerate solution, characterized by most samples being assigned to a few clusters. To prevent this, we apply the Sinkhorn-Knopp algorithm[42] to the similarity matrix between a batch of features $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_B\}$ and prototypes $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K\}$. We denote this mapping or codes by $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_B\}$, and optimize $\mathbf{Q}$ to maximize the similarity between the features and the prototypes:

$$\max_{\mathbf{Q}\in\mathcal{Q}} \mathrm{Tr}(\mathbf{Q}^\top \mathbf{C}^\top \mathbf{Z}) + \varepsilon H(\mathbf{Q}), \tag{6}$$

where H is the entropy function, $H(\mathbf{Q}) = -\sum_{ij}\mathbf{Q}_{ij}\log\mathbf{Q}_{ij}$ and ε controls the smoothness of the mapping. The matrix $\mathbf{Q}$ is an allocation matrix obtained through optimization, whose function is to define an optimal representation of the results of the dot product of $\mathbf{Z}$ and $\mathbf{C}$, making the similarity between $\mathbf{Q}$ and the results of the dot product as large as possible. The constraints are as follows:

$$\mathcal{Q} = \left\{\mathbf{Q} \in \mathbb{R}^{K\times B} \,\middle|\, \mathbf{Q}\mathbf{1}_B = \tfrac{1}{K}\mathbf{1}_K,\ \mathbf{Q}^\top\mathbf{1}_K = \tfrac{1}{B}\mathbf{1}_B\right\}, \tag{7}$$

where 1 represents a B-dimensional vector consisting entirely of ones, and B denotes the batch size. Each row sums to $1/K$, ensuring uniform distribution across columns for each row. Each column sums to $1/B$, ensuring uniform distribution across rows for each column. We obtain a doubly random matrix $\mathbf{Q}^*$ after several iterations using the Sinkhorn-Knopp algorithm.

$$\mathbf{Q}^* = \mathrm{Diag}(\mathbf{u})\exp\left(\frac{\mathbf{C}^\top\mathbf{Z}}{\varepsilon}\right)\mathrm{Diag}(\mathbf{v}), \tag{8}$$

where $\mathbf{u}$ and $\mathbf{v}$ are both vectors with a sum of 1.

(3) Swapped Prediction: The key idea is to predict the code of one view from the representation of another view. The obtained codes $\mathbf{q}_t$ and $\mathbf{q}_s$ are used as targets to supervise the learning. Specifically, we swap the codes and compute a loss between the predicted probability of one view and the target code generated from the other view.

Cross-entropy loss

We select the corresponding sample columns $\mathbf{q}_t$ and $\mathbf{q}_s$ from the memory bank $\mathbf{Q}^*$. The symmetric contrastive loss function is then computed as follows:

$$L(\mathbf{z}_t, \mathbf{z}_s) = \ell(\mathbf{z}_t, \mathbf{q}_s) + \ell(\mathbf{z}_s, \mathbf{q}_t), \tag{9}$$

where $\ell(\mathbf{z}_t, \mathbf{q}_s)$ is cross-entropy loss:

$$\ell(\mathbf{z}_t, \mathbf{q}_s) = -\sum_{k=1}^{K} \mathbf{q}_s^{(k)} \log \mathbf{p}_t^{(k)}. \tag{10}$$

By substituting Equation (5) into Equation (10) and aggregating this loss across all images and their corresponding data augmentation pairs, we obtain the final loss function for the swapped prediction task:

$$-\frac{1}{N}\sum_{n=1}^{N}\sum_{s,t\sim T}\left[\frac{1}{\tau}\mathbf{z}_{nt}^\top\mathbf{C}\mathbf{q}_{ns} + \frac{1}{\tau}\mathbf{z}_{ns}^\top\mathbf{C}\mathbf{q}_{nt} - \log\sum_{k=1}^{K}\exp\left(\frac{\mathbf{z}_{nt}^\top\mathbf{c}_k}{\tau}\right) - \log\sum_{k=1}^{K}\exp\left(\frac{\mathbf{z}_{ns}^\top\mathbf{c}_k}{\tau}\right)\right]. \tag{11}$$

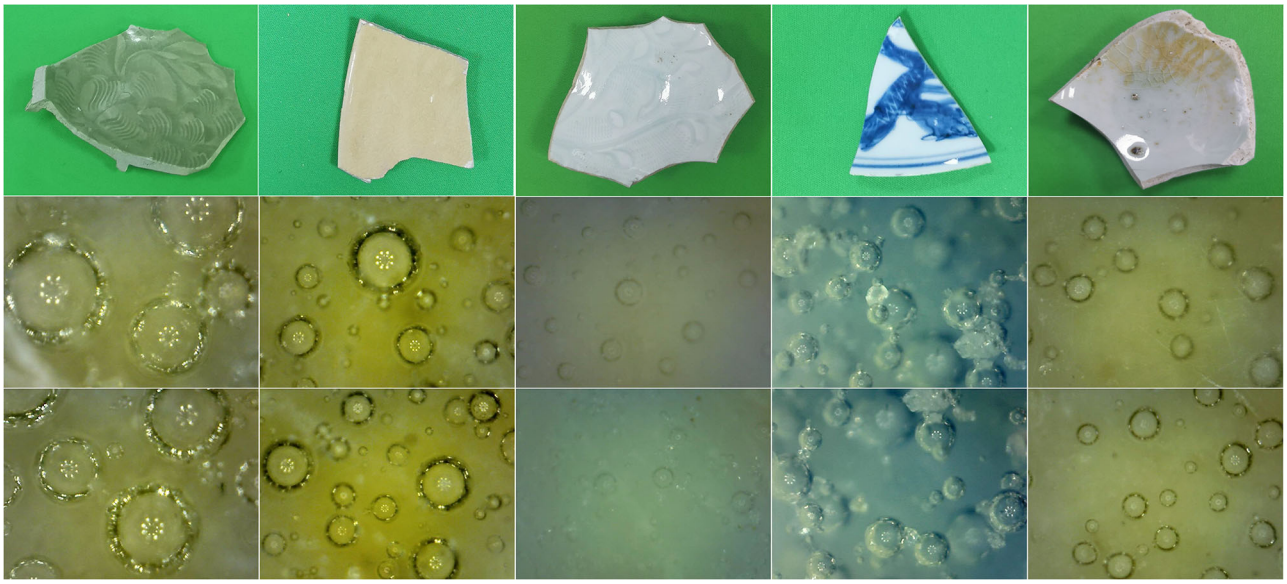


Fig. 5 | Porcelain relic images and Porcelain relic microscopic images. The figure shows five categories of porcelain fragments and their corresponding microscopic images.

Results

Dataset

Using a 3R Anyty microscope, we captured microscopic images at 600 × magnification of five types of porcelain relic fragments, constructing the Porcelain Relic Microscopic Image (PRMI) dataset. The PRMI dataset contains 7425 images in total, with 6282 images for training and 1143 images for testing. Representative samples of porcelain relics and their corresponding microscopic images are presented in Fig. 5. The PRMI dataset supports multiple research applications including intelligent classification of porcelain fragments, virtual restoration, authenticity identification, dynastic analysis, and kiln attribution studies, providing new opportunities for computational archeology research.

Data augmentation

To address the scarcity of microscopic image data of porcelain relics, this study employs a data augmentation strategy to expand the PRMI dataset while ensuring sample diversity. Using a random cropping strategy, we generate 8 images from each input image, including 2 images of 224 × 224 pixels and 6 images of 96 × 96 pixels. Each generated view undergoes transformations including random horizontal flipping, color jittering, and Gaussian blurring. The specific parameters for color jittering include adjustments to brightness, contrast, saturation, and hue with a strength factor of 0.8, while Gaussian blur is applied with a randomly selected kernel size. This comprehensive augmentation approach enhances feature diversity while preserving critical microstructural characteristics of the porcelain samples.

Experimental setting

We evaluate our method using classification accuracy as the primary metric, implementing the experimental framework in PyTorch with all models trained on NVIDIA A6000 GPUs. The optimization employs stochastic gradient descent (SGD) with an initial learning rate of 1×10^{-4}, batch size of 180, and 800 training epochs. The weight decay is set to 1×10^{-6}, momentum to 0.9, and warm-up learning lasts for 10 epochs. For hyper-parameter settings, the number of cluster centers is set to 20 for the PRMI dataset, and the temperature coefficient τ is 0.1. The Sinkhorn-Knopp algorithm runs for 3 iterations, and the smoothing control parameter is 0.05.

Comparison to SOTA methods

To validate the effectiveness of our proposed InSwAV method on the PRMI dataset, we conducted comprehensive comparisons with the methods of MoCo[13], SimCLR[32], and SwAV[33]. As demonstrated in Fig. 6, InSwAV exhibits superior performance stability and achieves significantly higher accuracy compared to these baseline methods. Specifically, our method attains a classification accuracy of 96.2%, outperforming MoCo by 12.8%, SimCLR by 4.6%, and SwAV by 2.1%, thereby establishing its superior capability for porcelain relic microscopic image analysis.

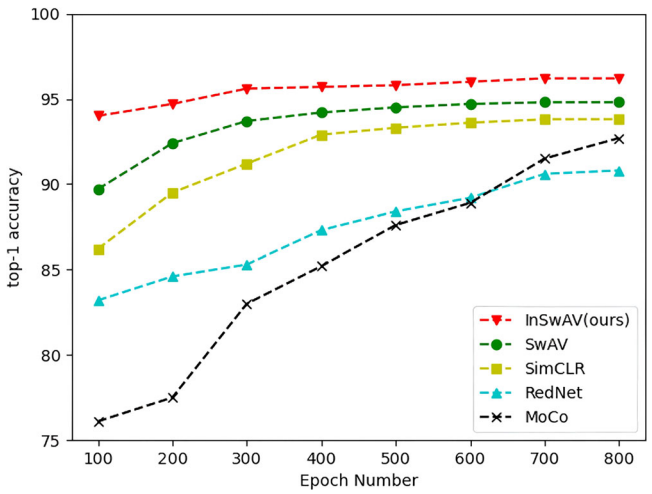


Fig. 6 | The accuracy-parameter envelope on PRMI dataset. Comparison of the proposed algorithm with state-of-the-art algorithms (SwAV, SimCLR, RedNet and MoCo) for porcelain relic microscopic image classification. The performance is evaluated across different epoch.

While MoCo initially demonstrates relatively low accuracy, its performance gradually improves with increasing training epochs, albeit with slower convergence speed. In contrast, our InSwAV method achieves stable accuracy after 300 training epochs, demonstrating significant advantages. After training for the full 800 epochs, InSwAV reaches 96.2% classification accuracy. The comparison demonstrates that while MoCo's momentum encoder and queue-based negative sampling are effective, they cannot match the efficiency of InSwAV. Notably, RedNet, as a supervised learning approach, fails to exceed the accuracy achieved by any of the four unsupervised contrastive learning methods in our study, including InSwAV. This clearly indicates that the PRMI dataset is particularly well-suited for contrastive learning methodologies, especially given its limited sample size. As

Table 2 | Comparison of training time and accuracy (300 epochs)

Method	Training time	Accuracy (%)
SimCLR[32]	18h 31min	91.2
MoCo[13]	18h 27min	83
RedNet[41]	10h 20min	85.1
SwAV[33]	18h 24min	93.7
InSwAV (Ours)	**4h 45min**	**95.8**

Table 3 | Architecture configurations and accuracy comparison

Layer1	Layer2	Layer3	Layer4	Accuracy (%)
Inv Block × 3	Inv Block × 3	Inv Block × 6	Inv Block × 9	20.9
Inv Block × 3	Inv Block × 4	Inv Block × 6	Inv Block × 3	25.9
Inv Block × 2	Inv Block × 2	Inv Block × 2	Inv Block × 2	47.1
Res Block × 2	Res Block × 2	Res Block × 2	Inv Block × 2	82.7
Inv Block × 2	Res Block × 2	Res Block × 2	Inv Block × 2	94.6
Res Block × 2	Inv Block × 2	Res Block × 2	Inv Block × 2	96.2
Res Block × 1	Inv Block × 2	Res Block × 2	Inv Block × 1	87.3

shown in Table 2, quantitative results further demonstrate that InSwAV substantially outperforms all four mainstream contrastive learning networks in both final accuracy and training efficiency metrics. These advantages make InSwAV particularly suitable for time-sensitive porcelain fragment classification scenarios, offering both faster convergence and reduced training time compared to existing approaches.

Architectural ablation study

To validate the effectiveness of the proposed components, we performed extensive ablation studies on the PRMI dataset. We varied the stacking ratio and arrangement of Residual Blocks (Res Blocks) and Involution Blocks (Inv Blocks) within the feature extractor. The results are presented in Table 3. We used an approximate binary search strategy to efficiently explore the parameter space that spans stacking ratios from 1: 2: 2: 1 to 3: 3: 6: 9, and experimented by changing from using the Involution Block in each layer initially to using only one layer. We found that directly using the hierarchical structure of ResNet50 or a deeper network structure would lead to overfitting, resulting in a very low precision in our experiments. Therefore, this chapter improves the network performance by appropriately simplifying the network. The results clearly demonstrate that simply using a deeper network or stacking only Involution block leads to overfitting and performance degradation. The optimal configuration, which alternates Res and Inv blocks in a ratio 2: 2: 2: 2, strikes a balance between capturing long-range dependencies and maintaining representational stability. This validates the core design of our hybrid ResInv backbone.

Parameter sensitivity analysis

We specifically analyzed the accuracy influenced by the number of cluster prototypes K, a crucial hyperparameter in our framework. We performed a binary search over the range [0, 300], and the results are summarized in Table 4. The results indicate that our method is not overly sensitive to K within a wide range from 20 to 50, where the performance plateaued at its peak. This robustness is a desirable property, suggesting that the method does not require an exact and hard-to-find parameter value. The choice of K = 20 represents an optimal trade-off between accuracy and computational efficiency.

Visualization

To better understand the feature extraction capabilities of our method, we conducted comparative visualization analysis by generating channel heat maps for both our proposed ResInv architecture and ResNet50, as shown in Fig. 7. The results demonstrate that ResNet50 predominantly focuses on background information when processing the PRMI data, failing to effectively identify and extract the most discriminative features. In contrast, our ResInv architecture achieves the effective extraction of bubble features in the PRMI dataset through an efficient design of just 8 convolutional blocks. This streamlined architecture not only maintains computational simplicity but also significantly outperforms ResNet50 in capturing clear bubble contour information, demonstrating superior feature localization capabilities despite its relatively shallow network depth. The comparative visualization clearly illustrates the enhanced ability of ResInv to focus on relevant microstructural characteristics while effectively suppressing irrelevant background noise, a critical advantage for accurate analysis of cultural heritage materials.

Table 4 | Clustering centers setting and accuracy

Prototypes setting	Accuracy (%)
300	89.3
200	89.3
150	94.7
70	95.8
50	96.2
20	96.2
15	96.0
10	95.8

Discussion

In this work, we introduce InSwAV, an image classification network for porcelain relics based on contrastive clustering and involution. We introduce ResInv, a novel feature extraction model that combines involution with convolution to accurately identify the most distinguishable bubble features in the PRMI dataset. The features extracted by ResInv are clustered, and cluster assignment consistency across different augmented views is enforced via a cross-prediction mechanism, with iterative optimization performed using a cross-entropy loss function. Additionally, we implement a new random-cropping data augmentation strategy that effectively captures both global and local features without significantly increasing computational overhead, thereby improving classification accuracy. Experimental results demonstrate that InSwAV achieves significant accuracy improvements on the PRMI dataset.

This study highlights the potential of deep learning in analyzing porcelain relic microscopic images for digital cultural heritage restoration. While the current work establishes a strong baseline accuracy for unsupervised classification, we acknowledge that small-sample classification remains a fundamental challenge in artifact digitization. In future work, we plan to employ K-fold cross-validation and conduct a more extensive hyperparameter search to further validate model robustness. We will also evaluate the proposed method on public benchmarks such as ImageNet to better assess its broader applicability and generalization capability. Meanwhile, we plan to evaluate the transferability of the features learned by InSwAV in supervised settings on larger datasets and to benchmark against state-of-the-art supervised architectures.

We also envision further enhancing the practical deployment potential of the model by integrating advanced efficiency-aware techniques such as neural network pruning[44], tensor decomposition[45], and compact module design[46]. These approaches could be synergistically combined with the inherent efficiency of involution parameters to enable faster inference and lower memory footprint, facilitating real-time analysis on resource-constrained devices in field environments.We also intend to establish a comprehensive PRMI database through systematic collection and expert annotation, develop multimodal data sets that integrate microstructural, material composition, and 3D data, and explore cross-domain adaptation techniques to mitigate data scarcity. Furthermore, future iterations of the

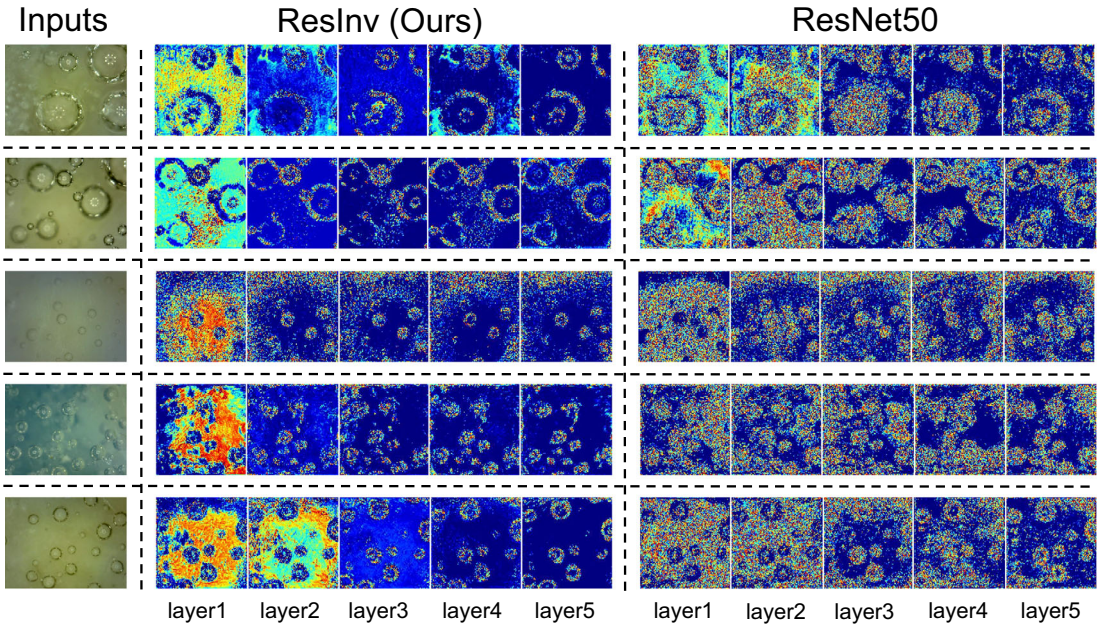


Fig. 7 | Heat maps of ResInv and ResNet50 model on the PRMI dataset. Comparative Visualization of Feature Activation Maps across Five Network Layers for Porcelain Micrographs using ResNet-50 and ResInv.

framework will explore architectural extensions to enable probabilistic output, facilitating a more comprehensive evaluation including ROC and precision-recall analysis.

Data availability

The PRMI dataset is hosted on GitHub to enable access and use by the research community. Dataset link: https://github.com/sunocean0716/InSwAV.

Received: 16 July 2025; Accepted: 13 February 2026;
Published online: 24 February 2026

References

1. Duan, J., Dai, Y. & Xu, T. Research on the development and evolution of ancient chinese ceramic kilns. *Identification and Appreciation to Cultural Relics*. 120–123 (2022).
2. Wang, P., Geng, G. & Zhang, Y. Fragment splicing method based on surface texture characteristic. *Laser Optoelectron. Progr.* **55**, 274–282 (2018).
3. Zhang, Y., Geng, G., Wei, X., Zhang, J. & Zhou, M. Reassembly of fractured fragments based on skeleton graphs matching. *Acta Autom. Sin.* **43**, 622–633 (2017).
4. Qu, Y., Xu, J., Xi, X., Huang, C. & Yang, J. Microstructure characteristics of blue-and-white porcelain from the folk kiln of Ming and Qing Dynasties. *Ceram. Int.* **40**, 8783–8790 (2014).
5. Yu, H., Chen, M., Zhou, J., Dong, X. & Cheng, J. Comparison of glaze composition between the southern song guan kiln and longquan ge kiln. *J. Ceram.* **42**, 1072–1079 (2021).
6. Zhang, J. On bubbles in ancient ceramics. *Collection World*. 118–119 (2012).
7. Jin, X. Exploration of the validity of microscopic traces in the identification of ancient ceramics. *Technology Wind*. 145–147 (2022).
8. Li, Q., Wang, J., Sheng, H., He, C. & Wang, J. Ceramic microscopic image processing based on wavelet packet analysis. In *Proc. of Signal Processing Branch of China Electronics Society*, 321–324 (2005).
9. Zhu, S. A preliminary study on bubble measurement of ancient ceramic microscience–a case study of yue kiln celadon. *China Cultural Heritage Scientific Research*. 38–41 (2008).
10. Liu, G. Ceramic microscopic image processing based on fast curvelet transform. *China Ceramic Industry*. 17–21 (2008).
11. Xie, J., Girshick, R. & Farhadi, A. Unsupervised deep embedding for clustering analysis. In *Proc. of the 33rd International Conference on Machine Learning*, vol. 1, 740–749 (2016).
12. Van Gansbeke, W. et al. Scan: Learning to classify images without labels. In *Proc. of the European Conference on Computer Vision*, 268–285 (2020).
13. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).
14. Grill, J.-B. et al. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*., vol. 33, 21271–21284 (2020).
15. Hong, C., Yu, J., Zhang, J., Jin, X. & Lee, K.-H. Multimodal face-pose estimation with multitask manifold deep learning. *IEEE Trans. Ind. Inform.* **15**, 3952–3961 (2019).
16. Xie, Y., Hong, C., Zhuang, W., Liu, L. & Li, J. Hogformer: High-order graph convolution transformer for 3d human pose estimation. *Int. J. Mach. Learn. Cybern.* **16**, 599–610 (2024).
17. Hong, C., Chen, L., Liang, Y. & Zeng, Z. Stacked capsule graph autoencoders for geometry-aware 3d head pose estimation. *Comput. Vis. Image Underst.* **208–209**, 103224 (2019).
18. Hong, C., Yu, J., Tao, D. & Wang, M. Image-based 3d human pose recovery by multi-view locality sensitive sparse retrieval. *IEEE Trans. Ind. Electron.* **62**, 3742–3751 (2015).
19. Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. II—Recent Progress. In Levy, D. N. L. (ed.) *Computer Games I*, 366–400 (Springer, New York, NY, 1988).
20. Quinlan, J. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).
21. Fix, E. & Hodges, J. L. Discriminatory analysis: Nonparametric discrimination: Small sample performance. project 21-49-004, report 4 (1952).
22. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
23. Yann, L., Yoshua, B. & Geoffrey, H. Deep learning. *Nature* **521**, 436–444 (2015).

24. Lu, D. & Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **28**, 823–870 (2007).
25. Shifat-E-Rabbi, M., Yin, X., Fitzgerald, C. E. & Rohde, G. K. Cell Image Classification: A Comparative Overview. *J. Int. Soc. Anal. Cytol.* **97**, 347–362 (2020).
26. Hu, M. & Yang, S. Overview of image mining research. In *Proc. of Signal Processing Branch of China Electronics Society*, 1431–1433 (2010).
27. Chang, H. et al. Unsupervised domain adaptation based on cluster matching and Fisher criterion for image classification. *Comput. Electr. Eng.* **91**, 107041 (2021).
28. Wu, Z., Xiong, Y., Yu, S. & Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. arXiv: 1805.01978 (2018).
29. Ye, M., Zhang, X., Yuen, P. C. & Chang, S.-F. Unsupervised Embedding Learning via Invariant and Spreading Instance Feature. arXiv: 1904.03436 (2019).
30. Gao, S., Tsang, I. W.-H. & Ma, Y. Learning Category-Specific Dictionary and Shared Dictionary for Fine-Grained Image Categorization. *IEEE Trans. Image Process.* **23**, 623–634 (2014).
31. Wang, X., Wang, F. & Niu, Y. A convolutional neural network combining discriminative dictionary learning and sequence tracking for left ventricular detection. *Sensors* **21**, 3693 (2021).
32. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. arXiv: 2002.05709 (2020).
33. Caron, M. et al. Unsupervised learning of visual Features by contrasting cluster assignments. arXiv: 2006.09882 (2021).
34. Oquab, M. et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193* (2023).
35. Geng, G., Wang, K. & Li, K. Image classification based on color and texture features and its application. *J. Northwest Univ. (Nat. Sci. Ed.).* **40**, 53–56 (2010).
36. Li, R. et al. LBCapsNet: A lightweight balanced capsule framework for image classification of porcelain fragments. *Herit. Sci.* **12**, 1–14 (2024).
37. Wang, Q., Xiao, X. & Liu, Z. Using microscopic imaging and ensemble deep learning to classify the provenance of archaeological ceramics. *Sci. Rep.* **14**, 32024 (2024).
38. Liu, X., Zhen, J., Liu, Y., Geng, G. & Shui, W. Multiple channel GCN with multiple directional Gaussian for porcelain microscopic image classification. *npj. Herit. Sci.* **13**, 254 (2025).
39. Geng, G. et al. Ceramic microscopic image classification based on the combination of contrastive learning and multi-scale methods. *J. Northwest Univ. (Nat. Sci. Ed.)* **51**, 734–741 (2021).
40. Zhou, M., Liu, J., Feng, L., Lu, Z. & Liu, Y. Microscopic images of ceramic relics fragments classification method based on dual channel attention mechanism and comparative learning. In *proc. of 3rd International Conference on Optics and Machine Vision*, vol. 13179 (SPIE, Nanchang, China, 2024).
41. Li, D. et al. Involution: Inverting the Inherence of Convolution for Visual Recognition. In *proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 12316–12325 (2021).
42. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*. vol. 26, (2013).
43. Wu, Z., Xiong, Y., Yu, S. X. & Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. of the IEEE conference on computer vision and pattern recognition*, 3733–3742 (2018).
44. Pham, V. T., Zniyed, Y. & Nguyen, T. P. Singular values-driven automated filter pruning. *Neural Netw.* **192**, 107857 (2025).
45. Pham, V. T., Zniyed, Y. & Nguyen, T. P. Enhanced network compression through tensor decompositions and pruning. *Neural Netw. Learn. Syst. IEEE Trans.* **36**, 4358–4370 (2025).
46. Pham, V. T., Zniyed, Y. & Nguyen, T. P. Coupled tensor decomposition for compact network representation. *IEEE Transactions on Neural Networks and Learning Systems* 1–15 (2025).
47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778 (2016).

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (62271393,62576274), by Archeological Talent Promotion Program of China(2024-267), by Shaanxi Province Technology Innovation Guidance Special Project(Fund)(2024QY-SZX-11), by Natural Science Basic Research Program of Shaanxi (2025JC-QYXQ-039), and by Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2024C02).

Author contributions

Y.L.: Writing review and editing, data collection. J.L.: Conceptualization, writing original draft. X.L.: Writing review and editing, data collection. G.G.: Supervision, Writing review, and editing. Z.L.: Writing review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Xinda Liu or Guohua Geng.