npj | heritage science

Article

https://doi.org/10.1038/s40494-025-02095-x

# Multi-modal ancient scripts recognition via deep learning with data homogenization and augmentation

Check for updates

Nan Wang[1,2,3] ✉, Weichen Wang[1], Bang Li[2], Han Zhang[1,3], Qingju Jiao[1,2,4] & Chaofan Liu[1]

Ancient scripts provide invaluable insights into ancient societies, and their effective recognition is crucial for cultural relic preservation, textual decipherment, and heritage. Current research primarily focuses on single mode ancient text data recognition such as processing rubbings or handwritten scripts independently, yet ancient scripts exhibit diverse forms across modalities. To address this, we propose a novel multi-modal recognition framework capable of processing hybrid inputs like rubbings of oracle bone inscriptions and handwritten scripts. Our method employs two additional modules, a cross-modal data homogenization module to unify heterogeneous data representations and a data augmentation module to enhance model robustness, then achieve the recognition with convolutional neural networks. Evaluated on oracle bone inscriptions and bronze inscriptions datasets, our approach outperforms baseline methods in recognition accuracy and generalization capability across modalities.

Character recognition is one of the foundational pillars of ancient scripts research and holds significant importance for exploring ancient civilizations. Ancient Greece, Rome, Egypt, and China each developed unique writing systems, many of which have faded or disappeared over centuries. Today, these texts are preserved in unearthed cultural relics, often fragmented and displaced—sometimes far from their original locations—due to natural decay or human activities such as trafficking.

Traditional methods of ancient script recognition rely heavily on accessing extensive repositories of information and the expertise of scholars. This process primarily depends on a researcher's accumulated experience and the corpus they have access to. When specialists study these inscriptions, they must invest considerable effort and care in organizing relevant materials, often engaging in high-threshold tasks such as reconstructing missing texts and conducting comprehensive literature reviews. As a result, traditional methods are highly complex, time-consuming, and require specialized workflows, which have increasingly faced limitations in recent years.

In order to change the current predicament, some researchers are starting to explore new methods with computer technology. The initial exploratory methods used traditional machine learning methods. In those methods, features of ancient scripts images are analyzed and transformed into corresponding structural encoding forms[1,2]. Then, classification algorithms such as support vector machine and K-nearest neighbor are used to classify the results to conduct the recognition process[3,4]. These traditional methods generally demonstrate robust recognition capabilities for these ancient inscriptions with uncomplicated line structures. However, challenges remain when processing characters with complex structures, particularly for multi-component characters of the ancient scripts, where recognition performance tends to be significantly weaker.

The advent of artificial intelligence (AI) and deep learning technologies has opened new avenues for researchers, enabling them to uncover and leverage intricate statistical patterns within vast datasets. A notable example is *Ithaca*[5], a groundbreaking tool that has reinforced confidence in this emerging research direction. By using Yolo[6], Resnet 50[7], Inception-v3[8] and other deep learning models, the research on ancient character recognition has obtained many valuable research results[9–11]. Research on ancient character recognition such as Oracle bone inscriptions, bronze inscriptions and ancient Egyptian texts, has gained more effective ideas[12–14]. Even some niche ancient texts have gained new research space, such as the Shui characters[15].

However, some practical problems faced by the research on ancient script recognition have constrained the development of related studies. Taking ancient Chinese scripts such as oracle bone inscriptions and bronze inscriptions as examples, existing research on character recognition faces the following challenges:

[1]School of Computer & Information Engineering, Anyang Normal University, Anyang, Henan, China. [2]Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education of China, Anyang Normal University, Anyang, Henan, China. [3]International Joint Research Laboratory for Perception Data Intelligent Processing of Henan, Anyang Normal University, Anyang, Henan, China. [4]Oracle Bone Inscriptions Application Big Data Development Innovation Laboratory, Anyang Normal University, Anyang, Henan, China. ✉e-mail: wn_ay@aynu.edu.cn

(1) Limited data scale and highly uneven distribution. Existing data on oracle bone script and bronze inscriptions primarily consist of rubbings and their replicas. The overall volume of data is relatively limited. For instance, in the case of oracle bone script, some characters appear thousands of times, whereas others occur only once or twice. In deep learning, particularly with large models, the limited data volume hampers the application of related technologies.
(2) Diverse forms of scripts. In addition to rubbings, handwritten forms are another primary medium for oracle bone inscriptions and bronze inscriptions. These handwritten forms are frequently encountered in academic papers and monographs. When retrieving information from diverse sources, both rubbings and handwritten forms are utilized. Therefore, the recognition system must accommodate hybrid type data.
(3) High resource demand. Deep learning, particularly large language models, requires substantial hardware resources. This poses significant constraints on the advancement of related research, especially given that the study of ancient Chinese scripts is a niche field. Therefore, exploring methods to achieve effective identification results with fewer resources is of critical importance.

These special issues of ancient scripts have led to existing research mostly focusing on single type character recognition. And the current situation, where the scale of ancient text data is generally small restricts the application of new methods such as large models in the field of ancient text recognition. For the cold research field of ancient character recognition, in order to better meet practical research needs, we propose an ancient scripts recognition model based on separate studies of oracle bone inscriptions and bronze inscriptions. The model comprises two main components: (a) A data preprocessing unit, which includes a cross-modal data homogenization module and a data augmentation module; (b) A recognition unit based on CNN model.

Essentially, the use of artificial intelligence technology to study ancient scripts recognition is aimed at providing useful methods and approaches for heritage preservation work such as character decoding and data organization. Our model can effectively address the problems of small data scale and hybrid inputs, and can use CNN models to improve the accuracy of ancient scripts recognition.

## Methods

### Datasets

The experimental data includes two types of ancient scripts: the oracle bone inscriptions and the bronze inscriptions. Either of these types ancient scripts images has two categories: rubbings and handwritings.

The oracle bone inscriptions dataset is collected from OBC306 and HWOBC on "Yin Qiwen Yuan Oracle Big Data Platform" (https://jgw.aynu.edu.cn/home/down/index.html). The HWOBC data set contains 3881 handwritten oracle bone inscriptions with 83,245 images. The OBC306 data set contains 306 oracle bone inscriptions with both handwriting and rubbing forms, as shown in Fig. 1. But as the uneven distribution of the data set some characters only has one or two images, which is too small to use for conducting experiments. Therefore, the original data set had been filtered and finally 165 oracle bone inscriptions were chosen for the experiments which have enough quantity and good image quality in both of the two data set as the experimental data.

Ultimately, the data set contains 12000 images, in which there are 8474 training images and 3526 testing images, as shown in Table 1.

The bronze inscriptions dataset is collected from "Jinwen scripts compilation". As shown in Fig. 2. This dataset is still in the processing and organization stage, Tencent and Key Laboratory of Oracle Bone Inscriptions Information Processing are currently responsible for organizing relevant data. But even so we selected 60 characters as the experimental data.

Ultimately, the data set contains 2551 images, including rubbings and handwritings, the detailed information is shown in Table 2.

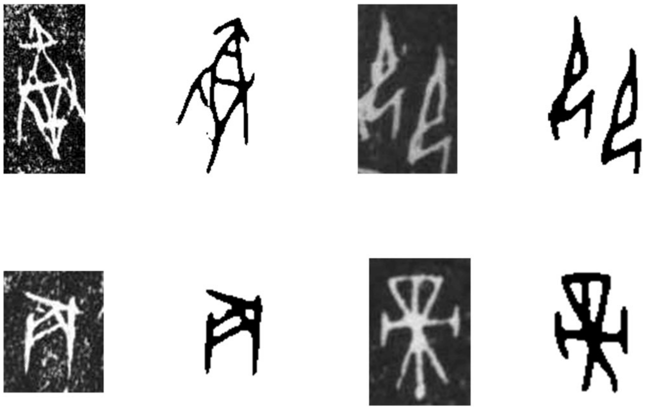


**Fig. 1 |** Examples of oracle bone inscriptions.

**Table 1 | The Oracle Bone Inscriptions Dataset**

| | Training data | Testing data | Total |
|---|---|---|---|
| Handwritings (HWOBC) | 3007 | 1254 | 4261 |
| Rubbings (OBC) | 5467 | 2272 | 7739 |
| Mixed data (MSO) | 8474 | 3526 | 12,000 |

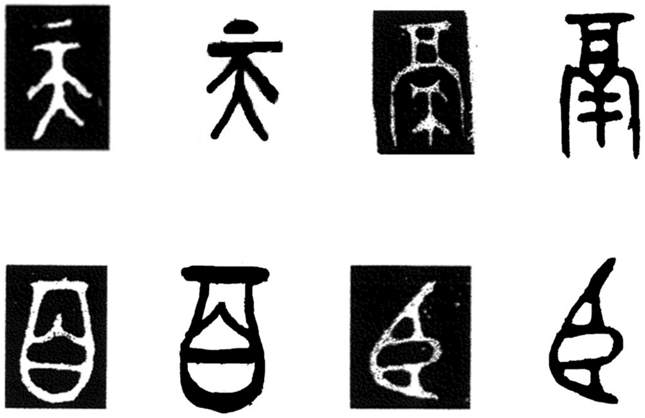


**Fig. 2 |** Examples of Bronze Inscriptions.

**Table 2 | The Bronze Inscriptions Dataset**

| | Training data | Testing data | Total |
|---|---|---|---|
| Bronze Inscription Handwriting | 897 | 452 | 1349 |
| Bronze Inscription Rubbings | 899 | 303 | 1202 |
| Bronze Inscription Mixed data | 1796 | 755 | 2551 |

### Proposed model

We proposed an idea which is homogenizing the input data and then applied the multi-modal ancient script recognition process. In the preliminary research, it was found that imitations have high accuracy in the recognition process, due to its simple image structure[16]. As for the rubbings, due to noise factors like wear and shield patterns in the image, the recognition accuracy of rubbings is relatively low. Therefore, a generative model that can transfer rubbings into handwritings to bypass the problem of rubbing recognition. Image generation like Image-to-Image, text-to-image, and image inpainting has been a popular research topic in many application areas over recent years[17–19]. Models such as U-net, diffusion and VAE have become important content in image generation research[20,21].

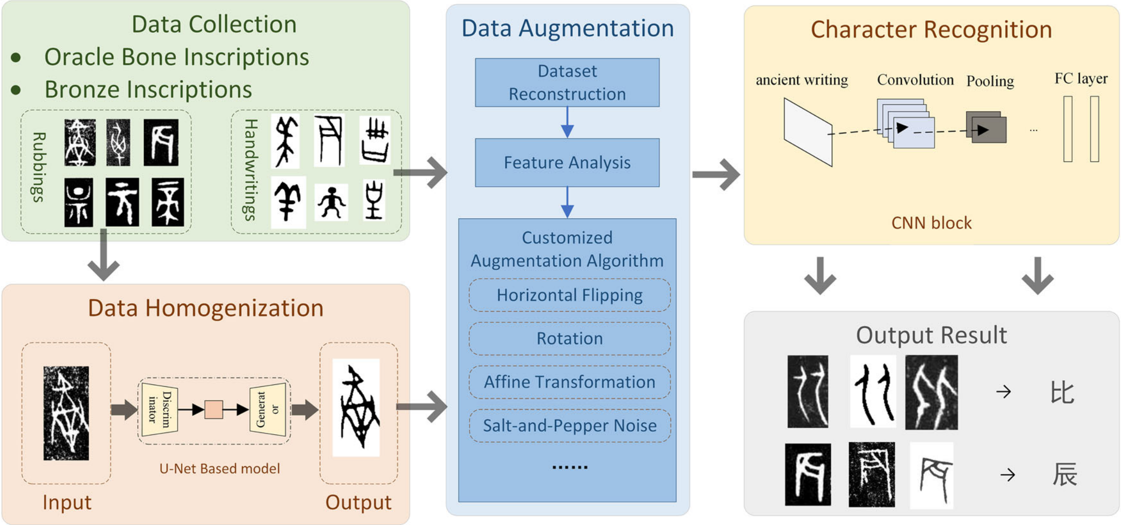
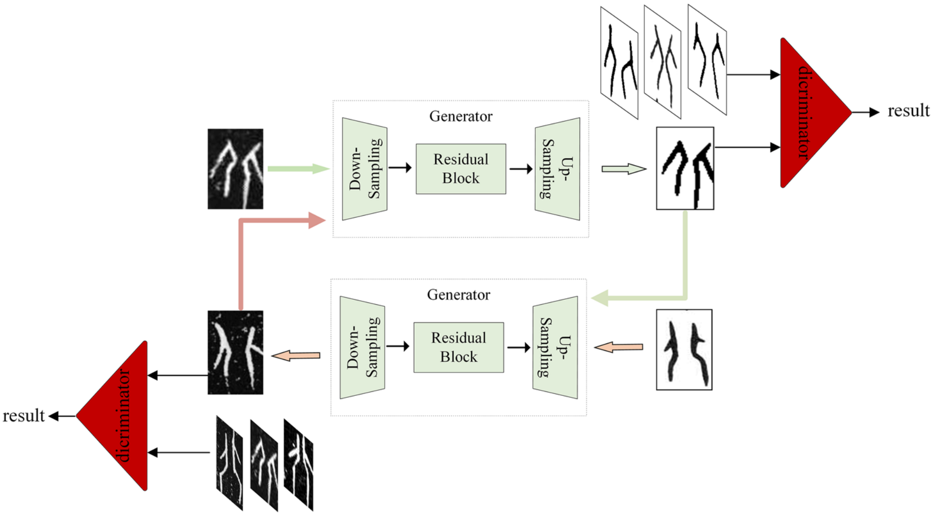


**Fig. 3** | The framework of The Proposed Method.

**Fig. 4** | The process of handwritten ancient script generation.



But for the rubbing image of ancient scripts, they are usually grayscale images with fewer features. Besides the limited information contained in the rubbings themselves, the content that can be used to describe the images is also not rich. Some characters only have obscure and difficult to understand classical Chinese explanations, and some characters have not even been deciphered. In addition, the available rubbings data for training is relatively small, and a more effective method needs to be selected based on actual performance.

Compared with diffusion and VAE models, we using a U-net based module to achieve the data homogenization which can converts rubbing data into handwriting data in a better way. In addition, in response to the scarcity and uneven distribution of ancient text data, in order to better enhance the generalization ability of recognition methods, we also proposed a method for data augmentation. The proposed method specifically targets pictographic characters such as oracle bone inscriptions and bronze inscriptions, and are more effective than the general random methods. The relevant work was reflected in the ablation experiments. The architecture of our model is shown in Fig. 3.

Rubbings are the forms of ancient scripts after being rubbed and imprinted, and their essence is also a kind of engraved or written text. From the perspective of image features, rubbings have similarities with medical images. Therefore, semantic segmentation of ancient script symbols on rubbings can be achieved through generation methods, enabling the conversion of ancient script rubbings to handwritten styles.

U-net model possesses characteristics suitable for small-sample scenarios and edge-sensitive scenarios, making it more suitable for the task of generating handwritten forms of different ancient script rubbings compared to models such as diffusion model. Therefore, in the Data homogenization module, U-net model will be used as a generator, employing the CycleGAN training process to convert ancient text data from different modalities into the handwritten form modality. The process is illustrated in Fig. 4.

As a consequence, the data homogenization module is based on the U-net and for better results the spatial attention and channel attention mechanism are introduced in the process. The generator is as shown in Fig. 5. For low-order features, a residual block with spatial attention is inserted before the third down sampling step. For the high-order features after down-sampling, each channel of a feature map is considered as a feature detector, channel attention is induced to the residual blocks.

During the down-sampling process, the Resblock can be described as follows:

$$F_{down}(x) = \sigma(W_2 \cdot \sigma(W_1 \cdot x)) \quad (1)$$

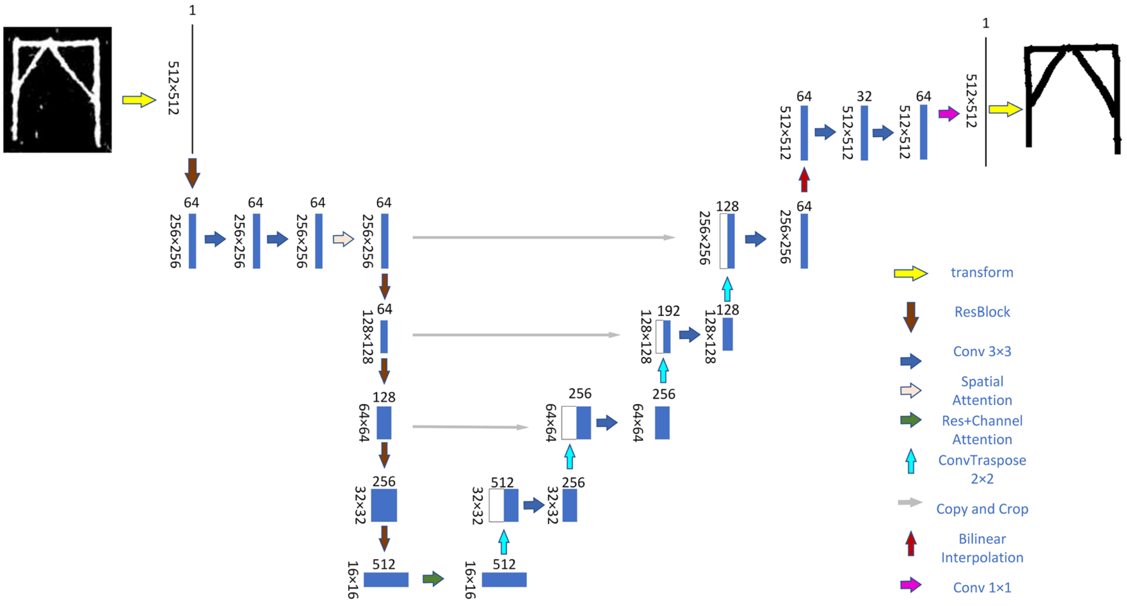


**Fig. 5** | Generator structure.

where $W_1$ and $W_2$ are the weights of two convolutions, the stride $S$ of the $W_1$ is set to 2. The output size is as shown in the following equation:

$$Hight_{out} = \left[\frac{Hight_{in}}{2}\right],\ Width_{out} = \left[\frac{Width_{in}}{2}\right] \tag{2}$$

The skip connection needs to synchronize down sampling, which is as follows:

$$W_s \cdot x \tag{3}$$

where S = 2, and the convolutional kernel size is 1 × 1. The final output is:

$$Y = F_{down}(x) + (W_s \cdot x) \tag{4}$$

The up sampling uses the transposed convolution with the output shown in Eq. 5:

$$N_{out} = (N_{in} - 1) \times S + F - 2P \tag{5}$$

*F* is the convolutional kernel with size 2 × 2, *S* is stride with value 2, *P* is the padding with value 0.

Based on the facsimile generation all rubbings can be converted to the handwriting form, which realizes the unified representation of each kind of oracle bone inscriptions data and bronze inscriptions data.

Based on the characteristics of ancient Chinese script such as oracle bone inscriptions and bronze inscriptions, we propose a targeted data augmentation plan, specifically, as follows:

Horizontal Flipping:

Giving a definition of a binary random variable α~Bernoulli(0.5), The flipped image $I_{flip}$ is given by:

$$I_{flip}(X, Y) = \begin{cases} I(x, W - y), \alpha = 1 \\ I(x, y), \alpha = 0 \end{cases} \tag{6}$$

where W is the image width.

Rotation:

Random rotation angle $\theta \sim \mathrm{U}(-20^\circ, 20^\circ)$, about center $(x_c, y_c)$, Transformed coordinates:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x - x_c \\ y - y_c \end{bmatrix} + \begin{bmatrix} x_c \\ y_c \end{bmatrix} \tag{7}$$

Affine Transformation:

Scaling factor s~U(0.6,1.2), and shear factor β~U(5,13)

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s & \beta \\ 0 & s \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{8}$$

Empty areas filled with: $I_{affine}(x', y') = 255$.

Salt-and-Pepper Noise:

Noise mask $M \in \{0,1\}^{H \times W}$, noise density, $\rho = 0.15$ (SNR = 0.85):

$$I_{noise}(x, y) = \begin{cases} 0,\ M(x, y) = 1(peper) \\ 255,\ M(x, y) = 0(salt) \\ I(x, y), otherwise \end{cases} \tag{9}$$

Application probability: *P*(apply)=0.3.

Gaussian Noise:

Noise $\eta$~N(0,0.05):

$$I_{gauss}(x, y) = clip(I(x, y) + \eta \cdot 255, 0, 255) \tag{10}$$

Application probability: *P*(apply)=0.3.

Brightness/Contrast Adjustment:

Random gain $\gamma$~U(0.8,1.2):

$$I_{bright}(x, y) = clip(\gamma \cdot I(x, y),\ 0, 255) \tag{11}$$

Grayscale Conversion:
Luminance transformation:

$$I_{grey}(x,y) = 0.299R(x,y) + 0.587G(x,y) + 0.114B(x,y) \tag{12}$$

Gaussian Blur:
Kernel $K$ of size 7×7 with $\sigma = 0.15$:

$$I_{blur} = I * K,\ K(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u^2+v^2}{2\sigma^2}} \tag{13}$$

**Table 3 | Experimental Environment**

| **Hardware Environment** | |
|---|---|
| Hardware Name | Specific Models |
| GPU | NVIDIA GeForce RTX 4090 |
| GPU memory | 24 G |
| CPU | Intel Core i9-14900KF |
| Computer memory | 64 G |
| **Software Environment** | |
| Software Name | Specific models |
| Operating system | Ubuntu 20.04.4 LTS |
| Deep learning framework | PyTorch 1.7.1 |
| develop environment | Python 3.12 + VScode |
| Graphics adapter | NVIDIA 566.14 |
| CUDA Version | V12.8 |

**Table 4 | Hyperparameters for Each Model**

| Model | Batch Size | Learning Rate | Weight Decay | K-fold | |
|---|---|---|---|---|---|
| | | | | Fold | Epoch |
| AlexNet | 32 | 0.0001 | 0.0001 | 1 | 81 |
| VGG19 | 32 | 0.00005 | 0.01 | 1 | 88 |
| ResNet50 | 32 | 0.00005 | 0.001 | 3 | 93 |
| ConvNext | 32 | 0.0005 | 0.05 | 2 | 89 |
| EfficientNet | 64 | 0.0001 | 0.05 | 1 | 98 |
| ShuffleNet | 32 | 0.001 | 0.0001 | 4 | 99 |
| ViT | 64 | 0.0001 | 0.01 | 5 | 81 |
| Swin Transformer | 64 | 0.0001 | 0.005 | 5 | 100 |

Application probability: $P(\text{apply})=0.3$

In the course of practice, after successfully performing the steps above an augmented experiment data set can be achieved. These randomly changed samples can reduce the model's dependence on certain attributes, thereby improving the model's generalization ability.

## Results

### Baseline

Deep Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in image recognition tasks, owing to their inherent advantages, including local receptive fields, parameter sharing mechanisms, hierarchical feature learning, translation equivariance, and dimensionality reduction via pooling operations. Given these merits, we adopt seven representative deep learning architectures as baselines: AlexNet, VGG-19, ResNet-50, GoogLeNet, ShuffleNet, Vision Transformer (ViT), ConvNeXt and Swing Transformer. Extensive comparative experiments substantiate the superiority of our proposed approach.

### Experiment setup

To ensure the consistency and stability across all experiments, we used a dedicated deep learning device. The experiments are conducted on the same hardware environment and software environment. As shown in Table 3.

Due to the selection of multiple models in the experiments, in order to obtain more effective hyperparameters and ensure that each model can operate in its optimal state, we adopted the K-fold cross-validation method to estimate hyperparameters such as learning rate, batch size and weight decay, thereby guaranteeing the generalization ability of the model.

The experimental dataset of the ancient Chinese characters, encompassing oracle bone inscriptions and bronze inscriptions, is relatively small in terms of actual size. Therefore, K = 5 is selected for hyperparameter tuning using K-fold cross-validation and each fold has 100 epochs. Additionally, the training data is divided into a training set and a validation set at a ratio of 7:3, and the data within the sets are constructed through random sampling. Each model was trained for 100 epochs to ensure sufficient learning.

Finally, the hyperparameters of each model used in the experiment are shown in Table 4 (Taking the experiments on oracle bone inscriptions data as an example). The detailed experimental results will be discussed in the next section.

### Results

Data homogenization is one of the core ideas of the multi-modal ancient scripts recognition method proposed in this paper. It achieves the unification of ancient scripts data from different modalities through a U-net model with an attention mechanism added, thus providing a foundation for subsequent recognition work. To illustrate the effectiveness of this module, we compared it with diffusion and VAE methods. These two models are widely

**Table 5 | Comparison of The Impact of Different Data Homogenization Algorithms on The Final Recognition Results**

| Recognition Models | The recognition accuracy of data homogenization with different models | | | | | |
|---|---|---|---|---|---|---|
| | U-net | | Diffusion | | VAE | |
| | Oracle bone Inscriptions | Bronze Inscriptions | Oracle bone Inscriptions | Bronze Inscriptions | Oracle bone Inscriptions | Bronze Inscriptions |
| AlexNet | 0.861 | 0.763 | 0.699 | 0.552 | 0.774 | 0.756 |
| VGG19 | 0.904 | 0.727 | 0.736 | 0.568 | 0.802 | 0.689 |
| ResNet50 | 0.900 | 0.803 | 0.755 | 0.607 | 0.805 | 0.738 |
| ConvNext | 0.880 | 0.778 | 0.740 | 0.564 | 0.790 | 0.575 |
| EfficientNet | 0.863 | 0.744 | 0.732 | 0.561 | 0.791 | 0.742 |
| ShuffleNet | 0.902 | 0.754 | 0.743 | 0.537 | 0.805 | 0.735 |
| ViT | 0.774 | 0.441 | 0.629 | 0.312 | 0.623 | 0.371 |
| Swin Transformer | 0.858 | 0.718 | 0.745 | 0.501 | 0.744 | 0.710 |

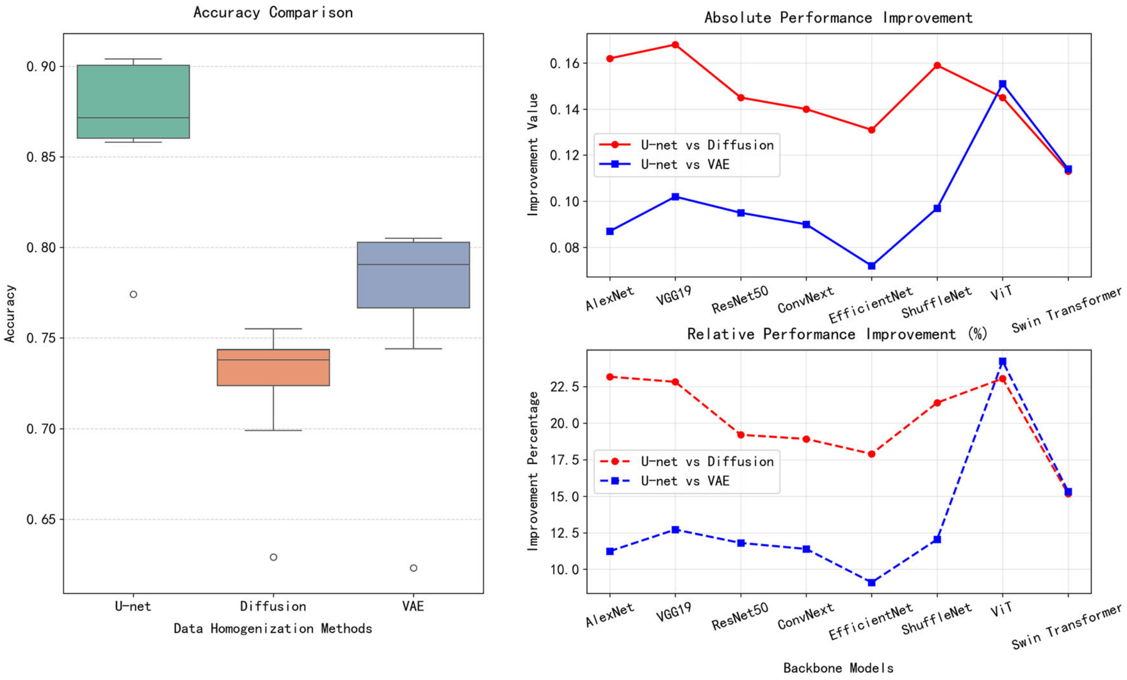


**Fig. 6** | Analysis of The Effectiveness of The U-net Algorithm in Processing Data Homogenization Tasks of Oracle Bone Inscriptions Rubbing Data.

used in image generation, but for ancient script rubbings, a special type of text image, the method proposed in this paper is more targeted.

We applied three image generation methods to the data homogenization process, and completed the recognition of ancient scripts after obtaining the corresponding generation results. Table 5 presents the recognition results of ancient scripts under the premise of different generation algorithms constituting the data homogenization module. The results indicate that the data processed based on U-net can achieve the best recognition results. This observation is reflected in both the oracle bone inscriptions dataset and the bronze inscriptions dataset.

From the view of result, it appears that the recognition accuracy of the bronze inscriptions data is relatively lower. This is due to the relatively small scale of bronze inscriptions data used in this study. Some models generally perform poorly on small datasets.

To analyze the strengths and weaknesses of various data homogenization methods from a more quantitative perspective, we conducted a comparative analysis from the perspectives of overall improvement in final recognition results and improvement across different models.

Figure 6 illustrates the results on the oracle bone dataset. The left part of Fig. 6 displays the overall distribution of the accuracy rates of different recognition models based on the data obtained from three image generation algorithms. According to the boxplot, the overall recognition accuracy of the model using U-net generated data is between 0.85 and 0.90, which is generally higher compared to the recognition results based on data generated by the other two algorithms.

The right part of Fig. 6 shows the recognition comparison results of different models using data from three algorithms. The upper graph depicts the difference in recognition accuracy obtained based on the U-net method compared to the results obtained based on the other two methods, while the lower graph shows the corresponding improvement ratio. From the two comparison methods, it can be seen that using U-net for data homogenization is more effective when dealing with ancient text rubbings.

Compared with the experimental results on oracle bone inscriptions data, a similar situation also occurred in the experiments with bronze inscriptions dataset as shown in Fig. 7. The results indicated that for the data homogenization our method based on U-net can outperform methods based on diffusion or VAE.

Based on the aforementioned experimental results, we have decided to complete the task of data homogenization using the U-net model. This approach is more suitable for the processing of ancient script rubbings and can provide a better data foundation for subsequent work.

In this study, we systematically implemented both the data homogenization and augmentation modules across all evaluated models, followed by comprehensive performance analysis.

Table 6 presents the comparative results on the mixed dataset, including four key metrics: Top-1 accuracy, F1 score, AUC, and Top-5 accuracy. The table contrasts two configurations: “Base” representing the baseline approach where models are trained solely on the original heterogeneous dataset containing both rubbing and handwritten images without any preprocessing; and “CA” our proposed framework incorporating the synergistic combination of data homogenization and augmentation components.

The experimental results shown in Table 6 indicate that on the hybrid dataset, the baseline models initially demonstrated only marginal performance. However, with the integration of our proposed data homogenization and augmentation modules, the results exhibited substantial improvement. The aforementioned situation occurred in every selected comparison metric. To better illustrate some detailed processes, taking the Top-1 accuracy as an example, Fig. 8 shows more information.

The blue solid line represents the accuracy of the proposed methods (CA Acc), the red solid line represents the corresponding loss (CA Loss). The blue dashed line represents the accuracy of the base method (Base Acc), and the red dashed line represents the corresponding loss (Base Loss). The eight sub-figures correspond to the methods listed in Table 6. From the results, it can be seen that the method proposed in this paper not only achieves better recognition accuracy but also converges faster in the process, with a smoother loss curve and a faster and more reasonable convergence rate. In contrast, these basic models have lower accuracy and greater fluctuation in loss when dealing with multimodal oracle bone inscriptions data. Especially for models like ViT, due to factors such as their own model structure, their accuracy is even worse when directly processing oracle bone inscriptions data.

In an effort to provide a better assessment of our model, we measured the performances of the proposed method with the bronze inscriptions data.

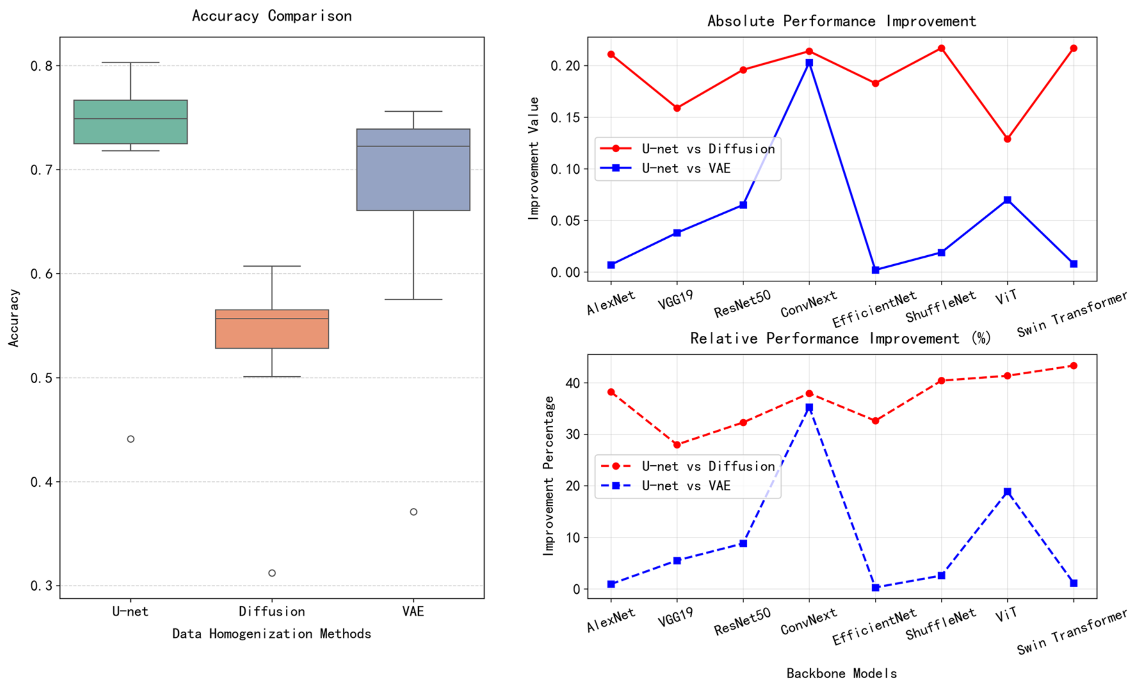


**Fig. 7** | Analysis of the effectiveness of the U-net algorithm in processing data homogenization tasks of bronze inscriptions rubbing data.

### Table 6 | Performance Comparison on Oracle bone inscriptions Dataset

| Model | Top-1 Accuracy | | F1 Score | | AUC | | Top-5 Accuracy | |
|---|---|---|---|---|---|---|---|---|
| | Base | CA | Base | CA | Base | CA | Base | CA |
| AlexNet | 0.657 | 0.861 | 0.649 | 0.860 | 0.987 | 0.997 | 0.876 | 0.966 |
| VGG19 | 0.643 | 0.904 | 0.640 | 0.904 | 0.975 | 0.999 | 0.853 | 0.978 |
| ResNet50 | 0.671 | 0.900 | 0.669 | 0.900 | 0.985 | 0.999 | 0.863 | 0.980 |
| ConvNext | 0.644 | 0.880 | 0.649 | 0.879 | 0.981 | 0.998 | 0.840 | 0.973 |
| EfficientNet | 0.693 | 0.863 | 0.690 | 0.862 | 0.988 | 0.998 | 0.870 | 0.965 |
| ShuffleNet | 0.538 | 0.902 | 0.538 | 0.901 | 0.968 | 0.999 | 0.756 | 0.978 |
| ViT | 0.288 | 0.774 | 0.283 | 0.776 | 0.874 | 0.990 | 0.519 | 0.901 |
| SwinTransformer | 0.421 | 0.858 | 0.414 | 0.856 | 0.945 | 0.999 | 0.662 | 0.973 |

Table 7 shows the performances of the baseline methods and the proposed method.

The results shown that for the Top-1 accuracy, F1 score, AUC and Top-5 accuracy indices, with the proposed data homogenization module and the augmentation module the recognition results had been improved, even the data scale of bronze inscriptions is relatively small.

Take the Top-1 accuracy as an example, compared with the experimental results of oracle bone inscriptions data, the bronze inscriptions recognition process better demonstrates the effectiveness of the method proposed in this paper, as shown in Fig. 9. The curve (red) of the Base method is more chaotic, which is due to the small scale of bronze inscriptions experimental data. Some models fail to exhibit their expected performance when recognizing bronze inscriptions. However, the method proposed in this paper (blue curve) yields better results in terms of convergence speed, changes in Loss value, and its final value.

To sum up, for the multi-modal ancient scripts, the proposed method performed notably better than the baseline methods.

## Ablation study

To rigorously validate the efficacy of our proposed method, we conducted systematic ablation studies. The key contribution of this work lies in the integration of the data homogenization module and augmentation module with the recognition unit. Through controlled experiments, we quantitatively assessed the individual impact of each module on recognition performance.

In the first ablation experiment, we analyzed the impact of data homogenization. We isolated the effect of data homogenization by exclusively removing the data augmentation module while retaining other components. Subsequently, we evaluated the recognition performance using only the data homogenization module across both oracle bone inscriptions and bronze inscriptions datasets. For quantitative comparison, Top-1 and Top-5 accuracy were selected as evaluation metrics. The detailed results of this configuration are presented in Table 8.

The incorporation of the data homogenization module enhanced the recognition capability of baseline models, though the improvement margin remained limited. When evaluated against the complete model, architectures utilizing only the homogenization block exhibited inferior Top-1 and Top-5 accuracy across all benchmarks.

Among these methods, AlexNet, VGG19, ResNet50, ConvNext, and EfficientNet have all achieved further improvements on top of their original good results. And models like ShuffleNet, ViT, and SwinTransformer, which perform poorly even when directly processing raw data, have also seen significant improvements through data homogenization process. The complete model configuration consistently achieved significant accuracy improvements, underscoring the synergistic effect of integrated modules.

The same situation also occurred on experimental results with the bronze inscriptions data. Table 9 shows the Top-1 accuracy and Top-5 accuracy for all the methods to be compared.

From an overall perspective, the results shown in Table 9 are similar to these shown in Table 8. The data homogenization module could improve the performance of each baseline model but the improvement is limited. Compared with the complete model, the gap in accuracy still existed. This situation is attributed to the small scale of the bronze inscriptions dataset used in the experiment.

In the second ablation experiment, we analyzed the impact of data augmentation and performed comparative experiments analyzing its isolated contribution. Table 10 presents the Top-1 and Top-5 accuracy metrics for three configurations on the Oracle Bone Inscriptions dataset: Firstly, baseline models without data augmentation, secondly, models with only the data augmentation module, and thirdly, the complete integrated model.

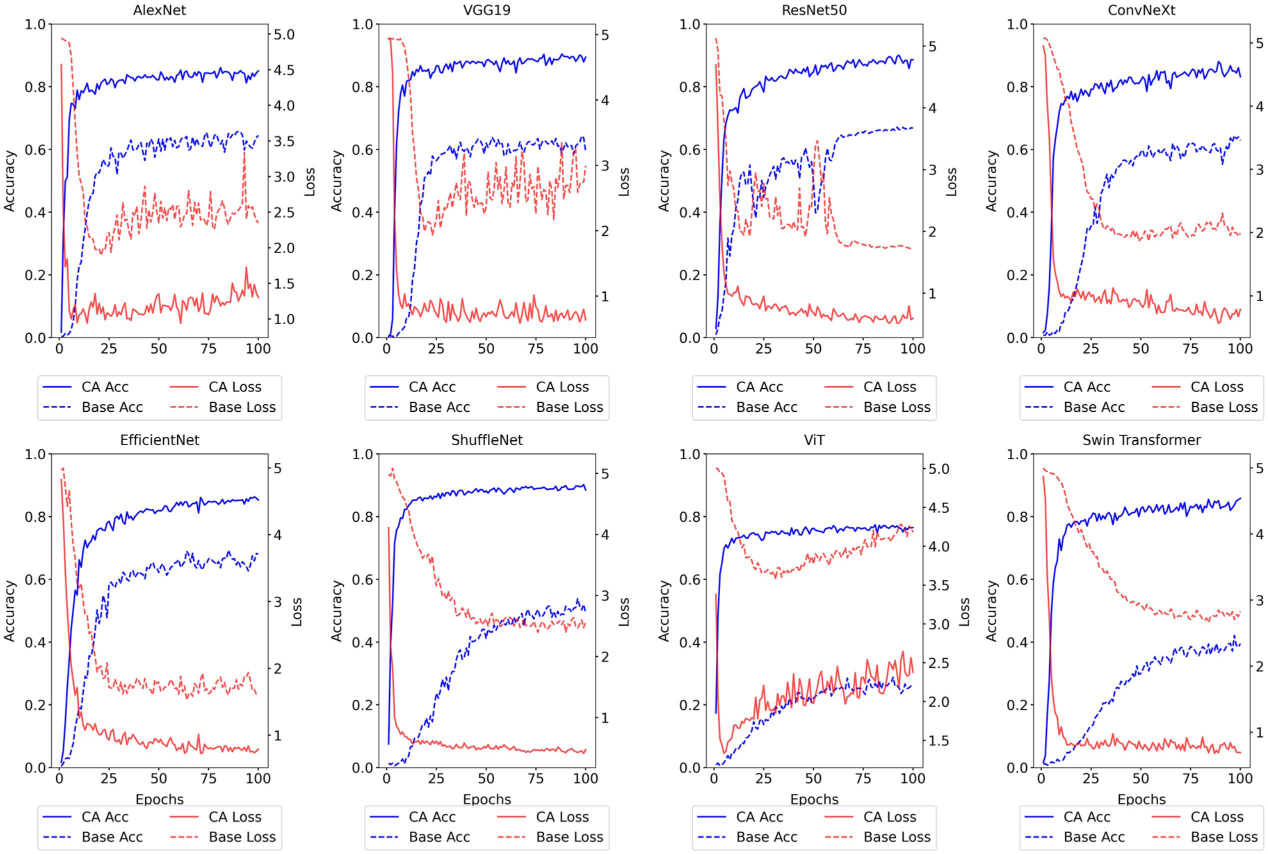


Fig. 8 | Accuracy curves of various models on the oracle bone inscriptions dataset: with and without the proposed CA strategy.

## Table 7 | Performance comparison on bronze inscription dataset

| Model | Top-1 Acc | | F1 Score | | AUC | | Top-5 Acc | |
|---|---|---|---|---|---|---|---|---|
| | Base | CA | Base | CA | Base | CA | Base | CA |
| AlexNet | 0.631 | 0.763 | 0.610 | 0.778 | 0.962 | 0.991 | 0.842 | 0.942 |
| VGG19 | 0.588 | 0.727 | 0.568 | 0.721 | 0.947 | 0.983 | 0.831 | 0.911 |
| ResNet50 | 0.615 | 0.803 | 0.615 | 0.797 | 0.965 | 0.994 | 0.850 | 0.960 |
| ConvNext | 0.573 | 0.778 | 0.555 | 0.772 | 0.949 | 0.991 | 0.792 | 0.942 |
| EfficientNet | 0.623 | 0.744 | 0.613 | 0.739 | 0.962 | 0.983 | 0.850 | 0.921 |
| ShuffleNet | 0.531 | 0.754 | 0.514 | 0.746 | 0.923 | 0.987 | 0.762 | 0.931 |
| ViT | 0.285 | 0.441 | 0.267 | 0.429 | 0.856 | 0.930 | 0.569 | 0.742 |
| SwinTransformer | 0.427 | 0.718 | 0.418 | 0.711 | 0.931 | 0.985 | 0.742 | 0.911 |

The results in Table 10 indicate that the baseline models exhibit the worst recognition performance. The recognition results improved after data processing through the data augmentation module compared to the baseline, while the method proposed in this paper achieved the best recognition results. The experimental results demonstrate that the data augmentation module has a positive impact on the recognition of ancient scripts. Similarly, related experiments were also conducted on the bronze inscription dataset to evaluate the effect of data augmentation.

Table 11 shows the results of different models on bronze inscriptions Dataset. The results shown in the tables indicated that data augmentation module could also improve the performance of each baseline model. But compared with the complete model, the results of models with data augmentation were relatively low.

To summarize, all the above ablation experiments indicate that the two block we added in the recognition process can indeed improve the accuracy of ancient character recognition. Each single block can have a certain effect, and with the complete model we can effectively improve the performance of baseline models.

## Discussion

This paper presents a multi-modal ancient script recognition method designed to recognize diverse script modalities, including oracle bone inscriptions, bronze inscription rubbing samples, and handwritten forms. The proposed framework comprises two core functional units. The first unit incorporates data homogenization and data augmentation modules for normalizing and enriching the input data. The second module implements the specific recognition task through a CNN-based deep learning architecture.

The experimental results show that compared to the baseline model, the method proposed in this paper achieves better performance on both oracle bone inscriptions and bronze inscriptions datasets. Especially for some smaller datasets, this method is also applicable.

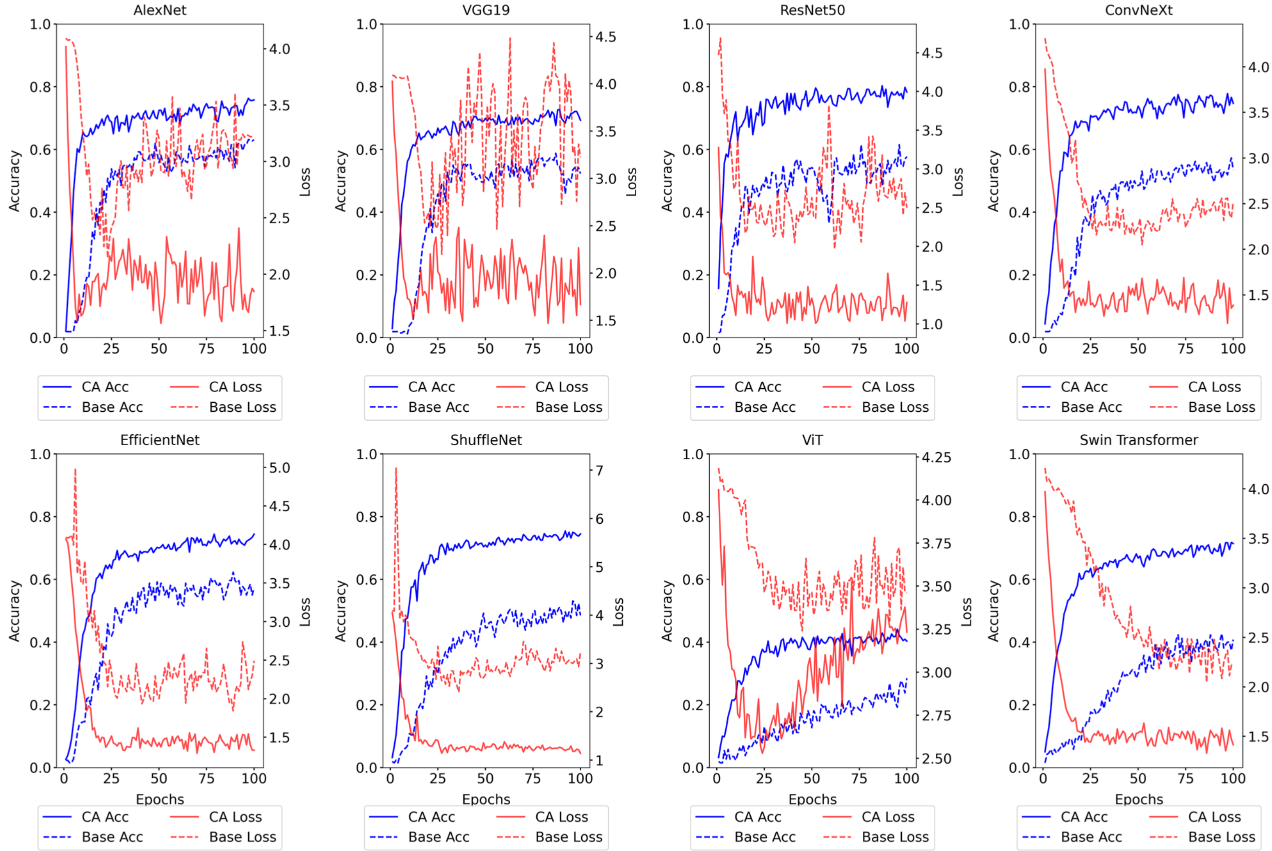


**Fig. 9 |** Accuracy curves of various models on the bronze inscription dataset: with and without the proposed CA strategy.

### Table 8 | Baseline models vs proposed method with only data homogenization module (marked as DH only) and CA accuracy comparison on oracle bone inscriptions dataset

| Model | Top-1 Acc | | | Top-5 Acc | | |
|---|---|---|---|---|---|---|
| | Base | DH only | CA | Base | DH only | CA |
| AlexNet | 0.657 | 0.797 | 0.861 | 0.876 | 0.940 | 0.966 |
| VGG19 | 0.643 | 0.735 | 0.904 | 0.853 | 0.904 | 0.978 |
| ResNet50 | 0.671 | 0.862 | 0.900 | 0.863 | 0.969 | 0.980 |
| ConvNext | 0.644 | 0.730 | 0.880 | 0.840 | 0.915 | 0.973 |
| EfficientNet | 0.693 | 0.775 | 0.863 | 0.870 | 0.917 | 0.965 |
| ShuffleNet | 0.538 | 0.713 | 0.902 | 0.756 | 0.901 | 0.978 |
| ViT | 0.288 | 0.386 | 0.774 | 0.519 | 0.691 | 0.901 |
| SwinTransformer | 0.421 | 0.562 | 0.858 | 0.662 | 0.806 | 0.973 |

### Table 9 | Baseline models vs proposed method with only data homogenization block (marked as DH only) and CA accuracy comparison on bronze inscriptions dataset

| Model | Top-1 Acc | | | Top-5 Acc | | |
|---|---|---|---|---|---|---|
| | Base | DH only | CA | Base | DH only | CA |
| AlexNet | 0.631 | 0.669 | 0.763 | 0.842 | 0.877 | 0.942 |
| VGG19 | 0.588 | 0.665 | 0.727 | 0.831 | 0.858 | 0.911 |
| ResNet50 | 0.615 | 0.619 | 0.803 | 0.850 | 0.850 | 0.960 |
| ConvNext | 0.573 | 0.638 | 0.778 | 0.792 | 0.850 | 0.942 |
| EfficientNet | 0.623 | 0.654 | 0.744 | 0.850 | 0.858 | 0.921 |
| ShuffleNet | 0.531 | 0.558 | 0.754 | 0.762 | 0.785 | 0.931 |
| ViT | 0.285 | 0.292 | 0.441 | 0.569 | 0.581 | 0.742 |
| SwinTransformer | 0.427 | 0.577 | 0.718 | 0.742 | 0.842 | 0.911 |

### Table 10 | Baseline models vs proposed method with only data augmentation module (marked as DA only) and CA accuracy comparison on oracle bone inscriptions dataset

| Model | Top-1 Acc | | | Top-5 Acc | | |
|---|---|---|---|---|---|---|
| | Base | DA only | CA | Base | DA only | CA |
| AlexNet | 0.657 | 0.772 | 0.861 | 0.876 | 0.934 | 0.966 |
| VGG19 | 0.643 | 0.863 | 0.904 | 0.853 | 0.967 | 0.978 |
| ResNet50 | 0.671 | 0.905 | 0.900 | 0.863 | 0.977 | 0.980 |
| ConvNext | 0.644 | 0.748 | 0.880 | 0.840 | 0.905 | 0.973 |
| EfficientNet | 0.693 | 0.832 | 0.863 | 0.870 | 0.950 | 0.965 |
| ShuffleNet | 0.538 | 0.741 | 0.902 | 0.756 | 0.894 | 0.978 |
| ViT | 0.288 | 0.432 | 0.774 | 0.519 | 0.642 | 0.901 |
| SwinTransformer | 0.421 | 0.660 | 0.858 | 0.662 | 0.844 | 0.973 |

### Table 11 | Baseline models vs proposed method with only data augmentation block (marked as DA only) and CA accuracy comparison on bronze inscriptions dataset

| Model | Top-1 Acc | | | Top-5 Acc | | |
|---|---|---|---|---|---|---|
| | Base | DA only | CA | Base | DA only | CA |
| AlexNet | 0.631 | 0.723 | 0.763 | 0.842 | 0.921 | 0.942 |
| VGG19 | 0.588 | 0.672 | 0.727 | 0.831 | 0.889 | 0.911 |
| ResNet50 | 0.615 | 0.758 | 0.803 | 0.850 | 0.929 | 0.960 |
| ConvNext | 0.573 | 0.705 | 0.778 | 0.792 | 0.903 | 0.942 |
| EfficientNet | 0.623 | 0.743 | 0.744 | 0.850 | 0.912 | 0.921 |
| ShuffleNet | 0.531 | 0.703 | 0.754 | 0.762 | 0.900 | 0.931 |
| ViT | 0.285 | 0.362 | 0.441 | 0.569 | 0.682 | 0.742 |
| SwinTransformer | 0.427 | 0.606 | 0.718 | 0.742 | 0.866 | 0.911 |

To further demonstrate the effectiveness of the model, we also conducted detailed ablation experiments. The results showed that the two core modules in this method, data homogenization and data augmentation, both contribute to the improvement of the overall recognition performance, while the complete method achieves the best recognition results.

This study's primary contribution lies in establishing a standardized preprocessing pipeline for multimodal ancient scripts. The proposed two-module mechanism provides both theoretical foundations and practical methodologies for archaeological text analysis, with experimentally validated efficacy.

## Data availability

The original data used in this paper, HWOBC and OBC306, can be obtained from the website: http://jgw.aynu.edu.cn/DownPage; the complete data can be obtained from the website: https://github.com/Augety88/oracle-jinwen-code-data.

## Code availability

The underlying code for this study is available in https://github.com/Augety88/oracle-jinwen-code-data.

Received: 11 June 2025; Accepted: 3 October 2025;
Published online: 15 October 2025

## References

1. Gu, S. T. Identification of oracle-bone script fonts based on fractal geometry. *J. Chin. Inf. Process*. **32** (2018).
2. Qu, H. Y., Liu, J. Z. & Wu, J. Oracle-bone inscriptions recognition based on topological features. *Comput. Sci. Appl.* **9**, 1111–1117 (2019).
3. Liu, Y. G. & Liu, G. Y. Oracle bone inscription recognition based on SVM. *J. Anyang Norm. Univ.* **2**, 54–56 (2017).
4. Zhao, R. Q., Wang, H. Q., Wang, K., Wang, Z. & Liu, W. T. Recognition of bronze inscriptions image based on mixed features of histogram of oriented gradient and gray level co-occurrence matrix. *Laser Optoelectron. Prog.* **57**, 90–96 (2020).
5. Assael, Y. et al. Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022).
6. Fujikawa, Y. et al. Recognition of oracle bone inscriptions by using two deep learning models. *Int. J. Digit. Humanit.* **5**, 65–79 (2023).
7. Qiao, Y. G. & Xing, L. Z. Applying deep learning algorithms for automatic recognition and transcription of texts in oracle bones and golden texts. *Appl. Math. Nonlinear Sci.* **9**, 1–16 (2023).
8. Guo, Z. Y., Zhou, Z. H., Liu, B. S., Li, L. Q. & Jiao, Q. J. An improved neural network model based on inception-v3 for oracle bone inscription character recognition[J]. *Sci. Programm.* 7490363 (2022).
9. Liu, M. T., Liu, G. Y., Liu, Y. G. & Jiao, Q. J. Oracle bone inscriptions recognition based on deep convolutional neural network. *J. image Graph.* **8**, 114–119 (2020).
10. Meng, L., Kamitoku, N. & Yamazaki, K. Recognition of oracle bone inscriptions using deep learning based on data augmentation. *2018 Metrology for Archaeology and Cultural Heritage*. 33–38 (IEEE, 2018).
11. Mai, C., Penava, P. & Buettner, R. Oracle bone inscription character recognition based on a novel convolutional neural network architecture. *J.]. IEEE Access.* **12**, 197021–197034 (2024).
12. Wu, X. Q., Wang, Z. Y. & Ren P. CNN-based Bronze Inscriptions Character Recognition. *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering*. 514-519 (IEEE, 2022).
13. Xu, Y., Zhang, X. Y., Zhang, Z. X. & Liu, C. L. Large-scale continual learning for ancient Chinese character recognition. *Pattern Recognit.* **150**, 110283 (2024).
14. Barucci, A. et al. A deep learning approach to ancient Egyptian hieroglyphs classification. *IEEE Access* **9**, 123438–123447 (2021).
15. Zhao, H. S., Chu, H. Z., Zhang, Y. Y. & Yu, J. Improvement of ancient Shui character recognition model based on convolutional neural network. *IEEE Access* **8**, 33080–33087 (2020).
16. Wang, N., Wang, C. J. & Jiao, Q. J. Research on Handwritten Oracle Bone Inscriptions Recognition Based on EasyDL. *Electron. Technol. Softw. Eng.* **3**, 184–187 (2023).
17. Parmar, G. et al. Zero-shot image-to-image translation. ACM SIGGRAPH 2023 conference proceedings. 1-11 (ACM, 2023).
18. Li, Y. H. et al. Gligen: Open-set grounded text-to-image generation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 22511-22521 (IEEE, 2023).
19. Zhang, X. B., Zhai, D. H., Li, T. R., Zhou, Y. X. & Lin, Y. Image inpainting based on deep learning: a review. *Inf. Fusion* **90**, 74–94 (2023).
20. Kaneko, H., Yoshizu, Y., Ishibashi, R. & Meng, L. An attempt at zero-shot ancient documents restoration based on diffusion models. 2023 International Conference on Advanced Mechatronic Systems (ICAMechS). 1-6 (IEEE, 2023).
21. Chen, B. Z., Liu, Y. S., Zhang, Z., Lu, G. M. & Kong, A. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* **8**, 55–68 (2023).

## Acknowledgements

This study was funded by Natural Science Foundation of Henan Province, grant number 242300420680, Henan Province Science and Technology Research Project, grant number 222102320036.

## Author contributions

N.W.: Conceptualization. N.W. and B.L.: Methodology. W.W: Software. W.W. and N.W.: Validation. N.W.: Writing original draft preparation. H.Z.: Writing review and editing. Q.J. and C.L.: Data preparation. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to Nan Wang.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.