Restoring Ancient Ideograph: A Multimodal Multitask Neural Network Approach

Siyu Duan[1,3], **Jun Wang**[1,3,4], **Qi Su**[2,3,4*]
[1] Department of Information Management, Peking University
[2]School of Foreign Languages, Peking University
[3]Center for Digital Humanities, Peking University
[4]Institute for Artificial Intelligence, Peking University
{duansiyu, junwang, sukia}@pku.edu.cn

Abstract

Cultural heritage serves as the enduring record of human thought and history. Despite significant efforts dedicated to the preservation of cultural relics, many ancient artefacts have been ravaged irreversibly by natural deterioration and human actions. Deep learning technology has emerged as a valuable tool for restoring various kinds of cultural heritages, including ancient text restoration. Previous research has approached ancient text restoration from either visual or textual perspectives, often overlooking the potential of synergizing multimodal information. This paper proposes a novel Multimodal Multitask Restoring Model (MMRM) to restore ancient texts, particularly emphasising the ideograph. This model combines context understanding with residual visual information from damaged ancient artefacts, enabling it to predict damaged characters and generate restored images simultaneously. We tested the MMRM model through experiments conducted on both simulated datasets and authentic ancient inscriptions. The results show that the proposed method gives insightful restoration suggestions in both simulation experiments and real-world scenarios. To the best of our knowledge, this work represents the pioneering application of multimodal deep learning in ancient text restoration, which will contribute to the understanding of ancient society and culture in digital humanities fields.

Keywords: text restoration, ancient ideograph, multimodal, digital humanities, cultural heritage

1. Introduction

Ancient cultural heritage stands as a record of human civilization, aiding in the understanding of human history and culture. Regrettably, many ancient artefacts have fallen prey to the ravages of time, natural deterioration, or deliberate human actions, expecting preservation and restoration. Deep learning technology has witnessed a series of remarkable advancements in the restoration of ancient cultural relics, including pottery (Farajzadeh and Hashemzadeh, 2021; Ostertag and Beurton-Aimar, 2020), architecture (Zou et al., 2021), murals (Wang et al., 2018; Zeng et al., 2020), etc. Among the myriad facets of cultural heritage, written language is the quintessential vessel of human thought, recording human history with symbols (Sommerschield et al., 2023). Restoring ancient texts aimed at proffering suggestions for the attribution of the fragmented scripts. Conventional methods for this task have leaned upon the knowledge of domain experts and the meticulous investigation of literature, which requires the mastery of philology and linguistics, rendering this undertaking a formidable and specialized task.

In this work, we applied the multimodal deep learning methodology to restore ancient texts, with a particular emphasis on the ideograph. Ideograms encapsulate semantics within visual symbols and endow each character with an intuitive visual correspondence. Consequently, restoring the ancient ideogram hinges on contextual information and visual cues. In this paper, we propose a novel Multimodal Multitask Restoring Model (MMRM) for ideograph restoration, synthesising cognizable context and the residual visual message of the damaged artefact to attribute damaged characters. It also employs a multitask learning paradigm to predict the damaged characters and generate restored images simultaneously.

We tested the MMRM model by experiments on both simulated data and authentic ancient inscriptions. The simulation experiments evince that the proposed MMRM model brings a substantive enhancement in ideograph restoration. In real-world scenarios, the model trained in the simulation experiment demonstrates its capacity to provide judicious recommendations for damaged characters. To the best of our knowledge, this work represents the pioneering application of multimodal deep learning methods to the restoration of ancient texts. By contributing to the building and improvement of ancient corpora, which is the basis of numerous digital humanities studies (Duan et al., 2023; Wang et al., 2024), this work will benefit historical, literary, and archaeological scholarship in the contemporary digital milieu.

*Corresponding Author: Qi Su, sukia@pku.edu.cn

LREC-COLING 2024, pages 14005–14015

2. Related Works

Ancient text restoration can be approached from two aspects: visual and textual. Visual-based restoration methods involved two computer vision techniques: handwritten text recognition and image inpainting. Ancient handwritten text recognition was extensively studied (Narang et al., 2020), however, this is generally for undamaged characters. Image-inpainting-based text restoration endeavours to regenerate the image of damaged texts. In character-level scenarios, the focus is the regeneration of deteriorated strokes. For example, Su et al. (2022) and Chen et al. (2022) use GAN to reconstruct ancient Chinese Han and Yi texts, respectively. Since the damage in this context often does not impede text recognition, the restoration purpose is to acquire higher-quality images. In document-level scenarios, the emphasis gravitates towards the amelioration of damaged areas, such as ink seepage, watermarks, blemishes, and blur (Wadhwani et al., 2021; Souibgui and Kessentini, 2022). These works aspire to procure higher-fidelity document images, not constricted to the meticulous reconstruction of individual words or characters.

Predicting blank spaces within sentences is a familiar task in natural language processing (Shen et al., 2020; Donahue et al., 2020), which has been applied in the preliminary attempted work of ancient text restoration (Assael et al., 2019; Fetaya et al., 2020). The pre-trained models also benefit to ancient text restoration since the Masked Language Model (MaskLM) task is widely employed (Devlin et al., 2019; Liu et al., 2019). For instance, Lazar et al. (2021) introduced multilingual pre-training for Ancient Akkadian text restoration. However, low-resource languages lack sufficient data for pre-training, thus the rudimentary RNN still needs attention, a salient instance being the Mycenaean Greek (Papavassileiou et al., 2023), where training samples are exceedingly scarce. Overall, these methods construct training data by masking out parts of characters in ancient texts, without utilizing visual signals.

Given the paucity of exploitable information within ancient texts, multitask learning fortified by supplementary guidance data was applied in text restoration tasks in some works. For example, the ancillary tasks predicting the region and age of text elicited a discernible enhancement in Ancient Greek restoration (Assael et al., 2022). However, annotating ancient texts with their provenance and temporal message constitutes a labour-intensive undertaking. Kang et al. (2021) employed the simultaneous pursuit of machine translation and text restoration tasks, while its enhancements in text restoration accuracy remained elusive. It is imperative to underscore that these multitask learning methods necessitate supplementary data annotations or parallel corpora. Besides, although both visual and contextual information play important roles in text restoration, the application of multimodal methodology is still a blank area.

3. Method

3.1. Task Define

This article proposes a multimodal ideograph restoration task that utilizes both visual and textual cues to predict damaged characters in ancient texts. The visual cue is the image of the damaged characters from cultural relics, and the textual cue is the recognizable context text. To restore the damaged ideogram, the model takes the undamaged context and the damaged image as input and predicts the damaged character.

3.2. Data

Ancient Chinese is a kind of widely used ideograph. It evolved from hieroglyphs and retained the characteristic of using visual elements to convey semantic meaning. Our experimental data consists of three parts: Classical Chinese corpus to simulate damaged context, images to simulate individual damaged characters and real inscription data to examine the model in a real-world scenario. [1] In this section, we will introduce how to simulate damaged text and images in our experiments.

3.2.1. Damaged Text

The ancient literature data was collected from the publicly available website 'xueheng (http://core.xueheng.net/)'. This Classical Chinese corpus includes the core classics in Chinese history, spanning over 2000 years with diverse topics. We segmented the texts into sentences, and when a sentence exceeded 50 characters, it was split into two. We got approximately 590,000 sentences for the simulation experiments. Its statistical information is shown in Table 1.

Train	Dev	Test	Max	Avg
575,398	10,000	10,000	50	14.4

Table 1: Statistics of textual dataset

In the simulation experiments, the characters in each sentence were randomly masked. It is worth noting that Classical Chinese was written in traditional Chinese characters, while in modern mainland China, simplified Chinese characters are

[1] Data and Code: https://github.com/CissyDuan/MMRM.

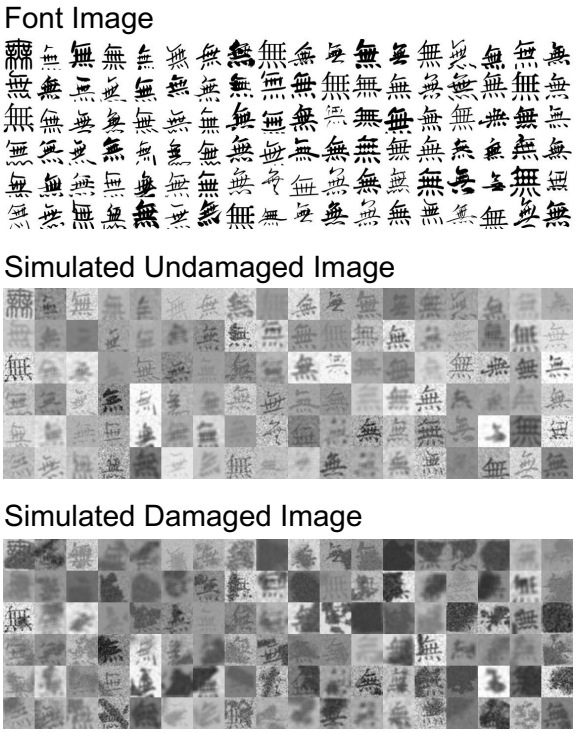


Figure 1: The process of simulating images of damaged characters. The first picture shows images of ideograms generated by 108 kinds of fonts. The second picture shows the simulated images of undamaged characters. The third picture shows the simulated images of damaged characters.

used. This has led to the fact that most common Chinese pre-trained models support simplified characters. Therefore, for each sentence, we prepared both simplified and traditional versions [2], one for inputting into the model and one for generating images of characters.

3.2.2. Damaged Image

After obtaining the text with random blanks, we proceeded to generate simulated images for the damaged characters. In the simulation experiments, we transformed all images to grayscale mode and let the dark text be generated on a light background. This means that in scenarios with light-coloured text on a dark background, such as rubbings of engraved inscriptions, colour inversion needs to be applied in advance.

The simulation of the damaged image involves three steps: First, we generate binary images of the characters with random fonts. Next, various image processing techniques are applied to simulate the images of undamaged characters collected from cultural relics. Finally, by adding random masks to

[2]The conversion tool is OpenCC (https://github.com/BYVoid/OpenCC)

Figure 2: Two concrete cases of Additive Damage and Fading Damage. The left image is from a rubbing of the 'Cao Quan Stele'; the right image is from the paper book '*Sima Fa Ji Jie*'.

the images, we simulate the images of damaged characters. Some simulated samples are shown in the Figure 1.

Step 1: Generating Font Image

Various calligraphic styles should be considered to simulate images of ancient characters. We have collected 108 traditional Chinese fonts from the internet to simulate diverse calligraphic styles. Images for missing characters will be generated with random fonts.

Step 2: Simulating Undamaged Images

Each specific font image for a character is unique, which obviously cannot account for text images collected in real-world scenarios. To achieve this, additional image processing is applied, including the random applications of texture, brightness, contrast, translations, rotations, scaling, and Gaussian blur.

Step 3: Simulating Damaged Images

Two types of noises typically occur on the images of damaged texts: one that is close in colour to the text itself, and the other close in colour to the background. We refer to them as 'additive damage' and 'fading damage' respectively. We showed two concrete cases in Figure 2. In these two cases, additive damage is more predominant in the left engraved inscription rubbing, while fading damage is more noticeable in the right paper book. To simulate these damages, we randomly add large masks as well as a random number of spot-type masks. In practical applications, the types, areas, and quantities of masks can be adjusted based on the characteristics of the real conditions.

3.3. Model

We propose a multimodal multitask framework for ideograph restoration, employing four modules to encode and decode the text and image respectively. The model finally gives two types of outputs: the candidate character and the restored image. The model formulation is shown in Figure 3.

3.3.1. Modules

Context Encoder. The context encoder is used to extract contextual features of the damaged text. We employ a pre-trained $RoBERTa$ model to encode the masked context and extract the feature vectors of the masked positions. Before the multimodal training, we fine-tuned the $RoBERTa$ model on the textual dataset by predicting the missing characters based on context alone. In the subsequent multimodal training, these parameters are frozen. The masked position is denoted as i, the calculation of this module is as follows:

$$memory = RoBERTa(Context) \quad (1)$$

$$x_1 = memory[i] \quad (2)$$

Image Encoder. The image encoder is used to extract visual features from the damaged image. We apply a pre-trained $ResNet\ 50$ model and replace the final layer with a new linear layer that maps the features to the same dimension as the Context Encoder. Since the Context Encoder, which has undergone pre-training and fine-tuning, already has preliminary capabilities in predicting the missing character, this linear layer is initialized with all zeros to gradually increase the influence of visual features.

$$x_2 = Resnet50(Img) \quad (3)$$

Feature Fusion. We use additive fusion to combine contextual and visual features. The fused features encompass information from both the text and the image, which will be used to predict the damaged characters and restore the damaged images.

$$x = x_1 + x_2 \quad (4)$$

Since the pre-trained and finetuned textual features already possess preliminary capabilities for this task, the image features need to be learned anew. The use of additive fusion, coupled with the full zero initialization of the last linear layer in the image encoder, minimizes interference from image features on textual features during the early stages of training.

Text Decoder. The fused features will pass through an MLP layer to predict the missing character. This MLP layer is initialized with parameters from the $RoBERTa$ LM layer.

$$Y_{pred} = MLP(x) \quad (5)$$

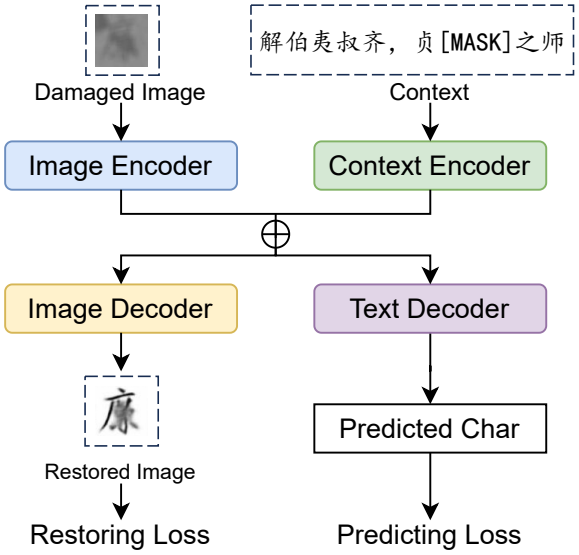


Figure 3: The formulation of Multimodal Multitask Restoration Model (MMRM).

Image Decoder. The fused features will pass through multiple transposed convolution layers to generate the restored image.

$$Img_{res} = ConvT(x) \quad (6)$$

After these processes, the candidate characters and images are generated by the model.

3.3.2. Multitask Learning

Highly correlated multitask learning can leverage information sharing between different tasks to improve their respective performances. In the proposed model, we have it not only predict missing characters but also generate images of the restored characters. The multitask learning serves two purposes: one is to predict the missing characters, and the other is to make the restored image resemble the original font image. This is achieved by optimizing two loss functions. One is the restoring loss, which is the MSE loss between the restored image and the font image:

$$Loss_{res} = MSELoss(Img_{res}, Img_{font}) \quad (7)$$

One is the predicting loss, which is the cross-entropy loss between the predicted token and the actual character:

$$Loss_{pred} = CELoss(Y_{pred}, Label) \quad (8)$$

The final loss is the weighted sum of the restoring loss and the predicting loss:

$$Loss = \alpha * Loss_{res} + Loss_{pred} \quad (9)$$

The determination of α considers two factors: the text and image losses should be comparable, and the changes in text and image losses during the training should be comparable. Such a multitask learning design does not require the introduction of additional labelled data and is highly relevant to the missing character prediction task.

3.3.3. Curriculum Learning

Curriculum learning involves initially focusing on simpler cases during the training and progressively ramping up the difficulty to tackle more intricate situations (Bengio et al., 2009). In our task, images with smaller damaged areas are easier to identify. Consequently, we incrementally enlarge the damaged area with each training epoch. Assuming we want to simulate a damaged area with dimensions length l and width w, curriculum learning is carried out over k epochs. In the $j-th$ epoch, if $j < k$, both the l and w of this damaged area will be multiplied by j/k.

4. Experiment

4.1. Baselines

We employed single-modal methods as baselines, encompassing visual and textual modalities. In instances where the feature extraction models used in previous text restoration studies are outdated, we elevated them to frontier models and adapted them to ancient Chinese text restoration.

- **Img**. When predicting damaged text using only visual features, it becomes similar to the handwritten Chinese character recognition (HCCR) task (Zhang et al., 2017), which has many mature models that can be adapted to text restoration. In our experiments, we use $ResNet\ 50$ (He et al., 2016) to predict the character from the damaged image.

- **LM**. When predicting damaged text using only context features, like methods in Fetaya et al. (2020) and Lazar et al. (2021), it reduces to a MaskLM task. We use the pre-trained masked language model $RoBERTa$ (Liu et al., 2019) to predict the damaged character from context. This model, after pre-training on a large-scale dataset of ancient Chinese, possesses certain abilities in predicting masked characters.

- **LM ft**. Since pre-trained models serve multiple tasks, there may still be room for improvement in a single MaskLM task. We further fine-tuned the $RoBERTa$ model on our dataset.

4.2. Proposed Approach

As for multimodal models, we initially fuse visual and text features to assess the multimodal approaches (MRM), and then introduce multitask learning (MMRM) and curriculum learning (MMRM CL) to further enhance the model.

- **MRM** (Multimodal Restoring Model). We integrate both vision and text features to predict the damaged character in a multimodal manner.

- **MMRM** (Multimodal Multitask Restoring Model). In addition to predicting the missing text token, this model performs multitask learning, allowing the model to generate the restored image of the damaged character. Unlike previous multitask methods such as Assael et al. (2022) and Kang et al. (2021), it requires no extra labelling data.

- **MMRM CL** (Multimodal Multitask Restoring Model with Curriculum Learning). During the training phase, curriculum learning is introduced to gradually increase the degree of damage, enabling the model to gain the restoration ability in a progressive manner.

4.3. Metrics

A text restoration task that relies solely on context can be evaluated based on the damaged amount and string length (Assael et al., 2019). However, in simulated experiments involving visual signals, the masks undergo processes such as overlay and blur, making it hard to measure the degree of image damage in a standardized manner. Therefore, evaluation metrics in simulated experiments are to assess the performance of different model architectures under consistent simulated data. We chose three commonly used evaluation metrics in text restoration:

- **Accuracy**. Accuracy stands for the average accuracy when the model predicts damaged characters.

- **Hits**. Hits represent the probability of the correct character being in the $top\ k$ candidates. Here we set k to 5, 10, and 20.

- **MRR**. MRR stands for Mean Reciprocal Rank, which refers to the reciprocal of the rank of the correct character.

4.4. Settings

In missing character sampling, to cater to a broader range of characters rather than concentrating on high-frequency ones, we assign weighted probabilities to each character. We first calculated the

Model	Image	Text	LM ft	MT	CL	Acc	Hit 5	Hit 10	Hit 20	MRR
Img	+					66.00	79.43	83.33	86.55	72.18
LM		+				36.06	56.18	64.10	71.03	45.56
LM ft		+	+			44.75	66.07	73.23	79.48	54.57
MRM	+	+	+			86.74	94.16	95.61	96.87	90.09
MMRM	+	+	+	+		87.34	94.60	96.16	97.29	90.61
MMRM CL	+	+	+	+	+	**87.76**	**95.03**	**96.45**	**97.52**	**91.03**

Table 2: Results of simulation experiments. MT is for multitask learning, CL is for curriculum learning.

frequency of each character in the training dataset, the average frequency is f_{avg}. Assuming for character Ci, its frequency is represented as f_i, its sample weight w_i is calculated as follows:

$$w_i = \sqrt{\frac{1}{max(f_i, f_{avg})}} \tag{10}$$

This setting helps to mitigate the influence of high-frequency characters. We used two sampling methods in the experiment: one is to mask one character to compare its performance with baseline models; the other is to randomly mask 1-5 characters to verify the model in the multiple missing characters scenarios. More technical details for simulating damaged images can be found in the Appendix [3].

In model building, we used the pre-trained $ResNet\ 50$ [4] and pre-trained $RoBERTa$ [5]. The damaged image simulation is conducted on 64x64 images. In the image decoder, the number of transposed convolution layers is 5. The loss weight α was set to 100.

In training, the batch size is 256. The training lasted for 30 epochs, with curriculum learning applied in the first 10 epochs. The learning rate was set to 0.0001 and decayed to less than 1e-5. The optimizer is Adam (Kingma and Ba, 2014).

In metric calculation, all simulation results are the averages obtained after randomly sampling the damaged characters on the test set 30 times, ensuring statistical significance in the observed improvements.

4.5. Simulation Results

The results of simulation experiments are shown in Table 2. It can be observed that when using only context or image information, the prediction performance is not satisfactory. CNN-based models can achieve an accuracy of over 95% in the task of undamaged handwritten Chinese character recognition (HCCR) (Zhang et al., 2017; Li et al., 2020). However, in our simulation experiments, the $ResNet\ 50$ model, which performs well in the HCCR task, performed poorly in the damaged character recognition task ($Acc = 66.00\%$). Likewise, although the $RoBERTa$ model, trained and fine-tuned for the MaskLM task, exhibited some predictive capability ($Acc = 44.75\%$), it was insufficient for guiding real-world scenarios.

However, the introduction of the multimodal method (MRM) brought about a turning point, significantly improving prediction performance ($Acc = 86.74\%$). This underscores the necessity of jointly leveraging text and image information, which was overlooked in previous ancient text restoration research. The introduction of multitask learning (MMRM) further enhanced restoration accuracy and provided restored images ($Acc = 87.34\%$), and the curriculum learning approach (MMRM CL) also contributed a modest improvement to the model ($Acc = 87.76\%$). These experimental results on simulated data validate the superior performance of our proposed MMRM architecture in ideograph restoration.

In multiple missing characters scenarios, shown in Table 3, the results indicate that the model's performance slightly decreases but remains satisfactory. Moreover, in the single character damaged scenario, compared with specific training (Table 2, MMRM CL), there is only a very marginal decline in performance, suggesting that using one model is sufficient to address multiple types of characters missing.

Num	Acc	Hit 5	Hit 10	Hit 20	MRR
R	82.83	91.57	93.68	95.30	86.80
1	87.27	94.40	95.99	97.16	90.50
2	85.02	92.95	94.79	96.28	88.62
3	82.67	91.50	93.59	95.23	86.67
4	80.72	90.01	92.45	94.41	84.96
5	78.76	88.87	91.44	93.46	83.38

Table 3: Results for multiple missing characters. R stands for random missing 1-5 characters.

[3] Appendix: https://github.com/CissyDuan/MMRM.

[4] The pre-trained $ResNet\ 50$ was downloaded from https://download.pytorch.org/models/resnet50-19c8e357.pth

[5] The pre-trained $RoBERTa_{base}$ was downloaded from https://github.com/Ethan-yt/guwenbert, it has been pre-trained with Classical Chinese literature containing 1.7B characters.

Sentence	Proposal	Img	LM	LM ft
弗闡，則與之瑞節而以執之	瑞	端	同	旌
生死與道不相舍離，亦未曾即合	曾	增	尝	必
朝廷以克辰纂嚴，令與夙駕	纂	篡	戒	告

Figure 4: Cases from different models in simulation experiments

We present cases from different models in simulation experiments in Figure 4. Models that only use visual features will confuse text with similar shapes, and language models may suggest other characters that fit the context.

It is important to note that these results are based on specific simulated data and are intended for comparing the performance of different models. This does not necessarily reflect the models' ability to perform text restoration in real-world scenarios.

4.6. Real-world Scenario

While we achieved commendable outcomes in the restoration of simulated data, it is still necessary to ascertain the model's continued efficacy within real-world scenarios. In this section, we conducted an empirical examination of the proposed model with a real historical artefact.

The 'Inscription of Sweet Spring in Jiucheng Palace (九成宮醴泉铭)' is a renowned calligraphic masterpiece during the Tang Dynasty (632). It was created by politician Wei Zheng and written by calligrapher Ouyang Xun and hailed as the 'peerless regular script' of China. Over the passage of time, this inscription has endured significant damage. Humanities scholars have reached generally accepted restoration conclusions for its damaged characters through careful investigation of literature. This significance and the availability of expert suggestions make it an ideal candidate for evaluating our proposed method.

The highest-quality extant rubbing of this inscription is the Li Qi edition, originating from the Song Dynasty (960 - 1279). We sourced a digital version of this edition from the internet and subjected it to restoration using the model developed in our simulation experiment. The full rubbing and photo of the inscription are shown in Figure 5. Subsequently, we compared the outcomes generated by our model with the results suggested by experts to gauge the model's effectiveness in real-world scenarios.

We cropped images of 38 damaged characters from the rubbings. These images are categorized into four levels based on the degree of damage and are shown in Figure 6.

- **I. Partially Damaged**: Some parts of the characters are damaged, but the overall shape is preserved. Experienced experts may be able to recognize the damaged characters from the remaining portions.

Figure 5: The full rubbing and photo of the 'Inscription of Sweet Spring in Jiucheng Palace'

- **II. Partially Preserved**: The characters in the images have a significant area of damage but still retain some key information, such as several strokes or radical components. While it narrows down the candidate's scope, variations in the damaged portion can point to different results.
- **III. Slightly Preserved**: Most area of the character is damaged, and it is impossible to determine the missing characters from the image alone. However, minimal details, such as a single stroke, are preserved.
- **IV. Totally Damaged**: The images provide no useful information and are completely damaged. In this case, the restoration can only rely on contextual information.

In the preprocessing stage, we converted the images into grayscale images. Since there were several signature seals on the images, we manually masked these seals using similar colours to the neighbouring area. Afterwards, we resized the images to the dimensions required by the model. We used the MMRM CL model (Table 2, MMRM CL) to restore these texts, the results are shown in Table 4. The MMRM mentioned in the subsequent text all refer to MMRM CL. Due to the limited amount of real data, the MRR metric is more important since it can alleviate the sparsity to a certain extent.

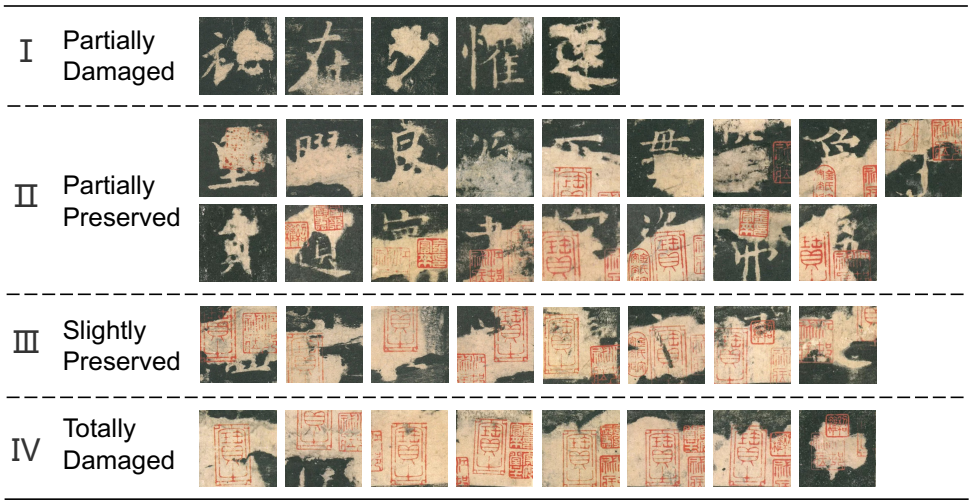


Figure 6: Four levels of damages from the 'Inscription of Sweet Spring in Jiucheng Palace'.

Level	Number	Count	Accuracy	Hit 5	Hit 10	Hit 20	MRR
I	5	5	100.00	100.00	100.00	100.00	100.00
II	17	11	64.70	82.35	82.35	94.11	71.17
III	8	5	62.50	87.50	87.50	100.00	70.83
IV	8	0	0	25.00	37.50	50.00	10.06
Total	38	21	**55.26**	**73.68**	76.31	**86.84**	**62.28**
LM	38	11	28.94	68.42	**84.21**	86.84	46.57
LM ft	38	13	34.21	73.68	78.95	84.21	50.32

Table 4: The results of real-world scenario experiments

As the results indicate, when compared to the language model that solely relies on textual data, our proposed method has demonstrated substantial improvements in accuracy and MRR within real-world scenarios. This underscores the effectiveness of our damaged image simulation and the MMRM framework.

Specifically, the MMRM model exhibits the capability to provide reasonable candidates for damage levels I to III ($MRR > 70.00$). In the case of damage level III, where only minimal strokes are preserved, the model can still provide reliable recommendations ($MRR = 70.83$). However, its effectiveness diminishes when confronted with damage level IV ($MRR = 10.06$), falling short of achieving the average MRR achieved by the language model ($MRR = 50.32$). Consequently, in totally damaged cases, employing a language model for restoration assistance may be a more viable choice, while MMRM remains a suitable option when residual visual information is still present.

To further analyze the relationship between the degree of damage and model performance, we simulated damaged areas with different sizes separately. We added square-shaped damage of different sizes, using the side length ranging from 0 to 1.0 times the image size. The results are shown in Figure 7. It can be observed that when the side length of the damaged square is larger than 0.9 times the image's side length, the multimodal model is weaker than the fine-tuned language model. This indicates that when the damaged area is too large, it can no longer bring effective information but introduce extra noise.

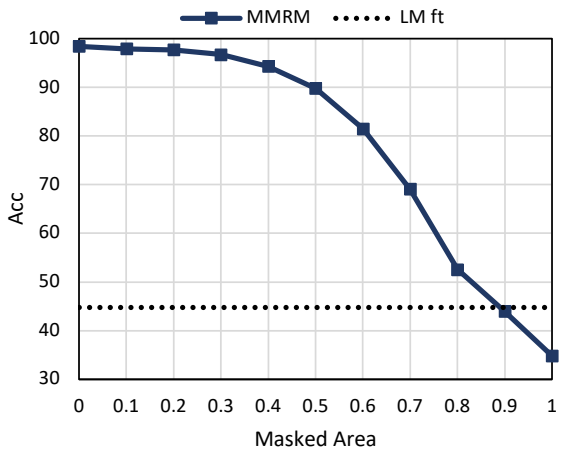
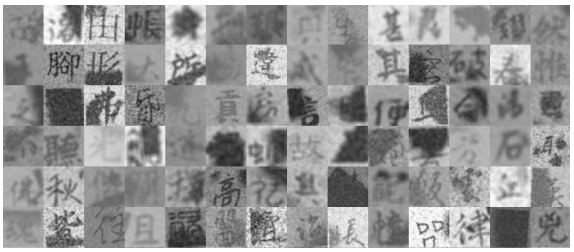


Figure 7: Relation between masked area and restoring accuracy

Damaged Image



Restored Image

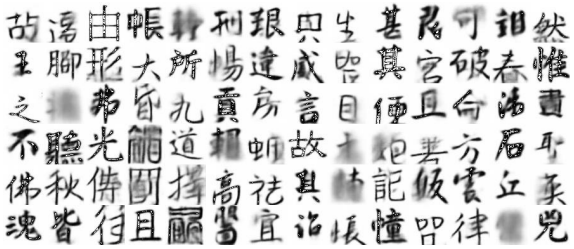


Figure 8: Restored images by MMRM in simulation experiment

Additionally, we show the restored images from the simulation experiment in Figure 8. Some severely damaged images can not achieve reasonable restoring results. In this case, it is a wiser choice to directly use the language model to assist text restoration.

5. Limitations

As the experimental results show, the proposed method still has limitations: it cannot give valuable results when facing large damaged areas; When multiple characters are damaged in the context, the model's performance inevitably drops. What's more, some ancient languages are endangered and do not have enough digital text and font resources to simulate the data, and the meanings of some characters may have not yet been deciphered, which poses further obstacles to their restoration.

In addition, there are challenges in generalizing to other ancient languages. Limited by our knowledge base, the experiments in this paper are conducted on Classical Chinese only. Scholars proficient in other ancient languages can try to adapt this method to other ancient texts.

6. Conclusion and Future Work

In this article, we propose a multimodal multitask model for ancient ideograph restoration, gaining a marked enhancement in simulation experiments and providing reasonable suggestions in real-world applications. This work represents a novel attempt that leverages the strengths of previous attempts in NLP and CV fields for restoring ancient textual artefacts, extending to situations bearing greater fidelity to real-world scenarios through multimodal methodologies.

There are many potential directions for future research: How to retrieve and utilize information from external databases to enhance text restoration? How to employ deep learning methods to aid in recognizing low-resource ancient ideographs, such as the ancient Chinese Oracle Bone Inscriptions from 3000 years ago? How to design an interactive tool for ancient text restoration, serving humanities scholars who lack the necessary programming skills? We look forward to the application of this work in both academic and industrial contexts.

7. Acknowledgments

This research is supported by the NSFC project "the Construction of the Knowledge Graph for the History of Chinese Confucianism" (Grant No. 72010107003).

Yannis Assael, Thea Sommerschield, and Jonathan Prag. 2019. Restoring ancient text using deep learning: a case study on Greek epigraphy. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6368–6375, Hong Kong, China. Association for Computational Linguistics.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Shanxiong Chen, Ye Yang, Xuxin Liu, and Shiyu Zhu. 2022. Dual discriminator gan: Restoring ancient yi characters. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training

of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Siyu Duan, Jun Wang, Hao Yang, and Qi Su. 2023. Disentangling the cultural evolution of ancient china: a digital humanities perspective. *Humanities and Social Sciences Communications*, 10(1):1–15.

Nacer Farajzadeh and Mahdi Hashemzadeh. 2021. A deep neural network based framework for restoring the damaged persian pottery via digital inpainting. *Journal of Computational Science*, 56:101486.

Ethan Fetaya, Yonatan Lifshitz, Elad Aaron, and Shai Gordin. 2020. Restoration of fragmentary babylonian texts using recurrent neural networks. *Proceedings of the National Academy of Sciences*, 117(37):22743–22751.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kyeongpil Kang, Kyohoon Jin, Soyoung Yang, Soojin Jang, Jaegul Choo, and Youngbin Kim. 2021. Restoring and mining the records of the Joseon dynasty via neural language modeling and machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4031–4042, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koren Lazar, Benny Saret, Asaf Yehudai, Wayne Horowitz, Nathan Wasserman, and Gabriel Stanovsky. 2021. Filling the gaps in Ancient Akkadian texts: A masked language modelling approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4682–4691, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyuan Li, Qi Wu, Yi Xiao, Min Jin, and Huaxiang Lu. 2020. Deep matching network for handwritten chinese character recognition. *Pattern Recognition*, 107:107471.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sonika Rani Narang, Manish Kumar Jindal, and Munish Kumar. 2020. Ancient text recognition: a review. *Artificial Intelligence Review*, 53:5517–5558.

Cecilia Ostertag and Marie Beurton-Aimar. 2020. Matching ostraca fragments using a siamese neural network. *Pattern Recognition Letters*, 131:336–340.

Katerina Papavassileiou, Dimitrios I. Kosmopoulos, and Gareth Owens. 2023. A generative model for the mycenaean linear b script and its application in infilling text from ancient tablets. *J. Comput. Cult. Herit.*, 16(3).

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, 49(3):703–747.

Mohamed Ali Souibgui and Yousri Kessentini. 2022. De-gan: A conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1180–1191.

Benpeng Su, Xuxing Liu, Weize Gao, Ye Yang, and Shanxiong Chen. 2022. A restoration method using dual generate adversarial networks for chinese ancient characters. *Visual Informatics*, 6(1):26–34.

Mayank Wadhwani, Debapriya Kundu, Deepayan Chakraborty, and Bhabatosh Chanda. 2021. *Text Extraction and Restoration of Old Handwritten Documents*, pages 109–132. Springer International Publishing, Cham.

Han-Lei Wang, Ping-Hsuan Han, Yu-Mu Chen, Kuan-Wen Chen, XinYi Lin, Ming-Sui Lee, and Yi-Ping Hung. 2018. Dunhuang mural restoration using deep learning. In *SIGGRAPH Asia 2018 Technical Briefs*, SA '18, New York, NY, USA. Association for Computing Machinery.

Jun Wang, Siyu Duan, Binghao Fu, Liangcai Gao, and Qi Su. 2024. Evol project: a comprehensive online platform for quantitative analysis of ancient literature. *Humanities and Social Sciences Communications*, 11(1):1–13.

Yuan Zeng, Yi Gong, and Xiangrui Zeng. 2020. Controllable digital restoration of ancient paintings using convolutional neural network and nearest neighbor. *Pattern Recognition Letters*, 133:158–164.

Xu-Yao Zhang, Yoshua Bengio, and Cheng-Lin Liu. 2017. Online and offline handwritten chinese character recognition: A comprehensive study and new benchmark. *Pattern Recognition*, 61:348–360.

Zheng Zou, Peng Zhao, and Xuefeng Zhao. 2021. Virtual restoration of the colored paintings on weathered beams in the forbidden city using multiple deep learning algorithms. *Advanced Engineering Informatics*, 50:101421.