Digital Scholarship in the Humanities, 2024, **39**, 61–73
https://doi.org/10.1093/llc/fqad065
Advance access publication 24 November 2023
Full Paper

Image classification for historical documents: a study on Chinese local gazetteers

Jhe-An Chen[1], Jen-Chien Hou[1], Richard Tzong-Han Tsai [1,2,*], Hsiung-Ming Liao[1], Shih-Pei Chen [3], Ming-Ching Chang[1,4,*]

[1]Center for Geographic Information Science, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei 115201, Taiwan
[2]Computer Science and Information Engineering Department, National Central University, Taoyuan 320317, Taiwan
[3]Max Planck Institute for the History of Science, Berlin 14195, Germany
[4]Computer Science Department, University at Albany, State University of New York, Albany, NY 12222, USA

*Corresponding authors. Center for Geographic Information Science, Research Center for Humanities and Social Sciences, Academia Sinica, Taipei, 115201, Taiwan. E-mail: thtsai@g.ncu.edu.tw (R.T.H.T.); mchang2@albany.edu (M.C.C)

Abstract

We present a novel approach for automatically classifying illustrations from historical Chinese local gazetteers using modern deep learning techniques. Our goal is to facilitate the digital organization and study of a large quantity of digitized local gazetteers. We evaluate the performance of eight state-of-the-art deep neural networks on a dataset of 4,309 manually labeled and organized images of Chinese local gazetteer illustrations, grouped into three coarse categories and nine fine classes according to their contents. Our experiments show that DaViT achieved the highest classification accuracy of 93.9 per cent and F1-score of 90.6 per cent. Our results demonstrate the effectiveness of deep learning models in accurately recognizing and categorizing historical local gazetteer illustrations. We also developed a user-friendly web service to enable researchers easy access to the developed models. The potential for extending this method to other collections of scanned documents beyond Chinese local gazetteers makes a significant contribution to the study of visual materials in the arts and history in the digital humanities field. The dataset used in this study is publicly available and can be used for further research in the field.

Keywords: image classification; historical document; art; digital humanities; Chinese local gazetteers; Convoluational Neural Network; Vision Transformer; DaViT.

1. Introduction

This research explores the utilization of modern Deep Neural Networks (DNNs) for image classification of historical documents, a topic of increasing interest in the field of digital humanities. Chinese local gazetteers (difangzhi 地方志) are a primary source for studying Chinese local history, as they provide extensive coverage of local data including geography, infrastructure, natural conditions, politics, culture, and society. With over 8,000 extant titles dating from the 10th to the early 20th centuries, and covering almost all regions of historical China (Zhuang *et al.* 1985), a large number of digitized local gazetteers have become accessible through digitization and commercial database subscriptions (Chen *et al.* 2023).

Local gazetteers contain vast amounts of information that can be challenging to explore and analyze. To make this process easier, digital archives of these gazetteers are typically enriched with annotated metadata that includes bibliographic information, publication details, and edition information. Being mainly a texture genre, local gazetteers also contain rich graphical illustrations that can be difficult to explore or analyze without proper image recognition tools. In the past, the annotation of historical contents has usually been done manually, which is a labor-intensive and time-consuming task requiring expertise even with proper digital aids.[1] However, with the advent of automatic book image recognition (Antonacopoulos *et al.* 2013), text detection (Yang *et al.* 2018), and Optical Character Recognition (OCR; Du *et al.* 2020; Martínek, Lenc, and Král 2020), the computer-assisted annotation has become feasible. Despite these advances, effectively automating the annotation process and ensuring annotation consistency remains a major challenge for digital humanities researchers and computer scientists.

The goal of this research is to develop a classification system for illustrations in Chinese local gazetteers and other Chinese historical books using recent advancements in DNNs. Our study is based on a collection of 410 rare local gazetteers that are held at and digitally scanned by Harvard-Yenching Library.[2] Through a collaboration between the Library and the Max Planck Institute for the History of Science (MPIWG), pages among these gazetteers that contain illustrations were identified. Later, scholars who participated in the 'Visual Materials in Local Gazetteers Working Group' at MPIWG jointly developed a classification scheme for illustrations that would generally appear in local gazetteers.[3] The availability of digitally scanned Chinese local gazetteers and the manually assigned labels by domain experts has provided valuable data for the application of deep learning algorithms. We propose an image classification pipeline to classify local gazetteer illustrations into nine classes that are simplified from the seventeen classes developed by the Working Group: *text, scenic map, city map, administrative map, star map, human figure, photograph, building, and object* (as shown in Fig. 1). This classification system specifically focuses on identifying historical map types, as local gazetteers often depict geospatial environments. We demonstrate the effectiveness of our approach in Section 4, where our best-performing model achieves 93.9 per cent classification accuracy and 90.6 per cent F1-score. The developed classification system can automatically annotate local gazetteer illustrations, making the document query and retrieval process more efficient.

Our approach relies on a large amount of data and annotations to train DNNs, and the availability and accessibility of a proper dataset are crucial to its success. However, in the field of Chinese Studies, research institutions may be limited by image licensing agreements imposed by commercial vendors who own and sell databases of Chinese local gazetteers and other primary sources. These non-disclosure agreements can hinder collaboration with third-party research teams and the development of automatic annotation of licensed materials.

In order to comply with license restrictions and address copyright concerns (Besek 2003), we have developed our method using down-sampled, low-resolution images of original scans of local gazetteer pages, with a resolution of 128 × 128 pixels.

Our study is based on an open-access dataset of local gazetteers from Harvard-Yenching Library,[4] with the intention of providing support for researchers studying Chinese local gazetteers who may be limited by licensing restrictions associated with commercial databases. We have also created a user-friendly web service front-end for nontechnical researchers to easily access the developed image classification system. We believe that our developed system and framework can serve as a model for future research in digital arts and humanities, particularly in the study of cultural heritage materials.

The contribution of this article is summarized as follows:

1. We propose a novel image classification system utilizing state-of-the-art DNNs for the annotation, organization, and retrieval of Chinese local gazetteer illustrations. To the best of our knowledge, this is the first system of its kind. We evaluate the performance of eight DNN models extensively.
2. We introduce a web service that enables nontechnical researchers to easily classify historical document scans by uploading an image and accessing our model.
3. Our developed system has wide applicability beyond Chinese local gazetteer illustrations, especially in the field of digital humanities for the classification of cultural and heritage illustrations in historical documents.

2. Related works

2.1 Chinese local gazetteers

The Chinese local gazetteers are the main genre in Chinese history that consistently records local information about territorial units, such as topography, institutions, population, taxes, biographies, and literature (Dennis 2015). Since historical times, this genre has been offering scholars detailed and pervasive accounts of various aspects of local history that cannot be found in other sources.

More than half of extant Chinese local gazetteers have been digitized and provided access, mainly through commercial databases, namely Erudition's Database of Chinese Local Records (with 6,000 titles),[5] East View's China Comprehensive Gazetteers (with 7,000 titles),[6] and Diaolong's Full-text Database of Chinese and Japanese Ancient Books (with 6,000 titles).[7] With a vast amount of digitized Chinese local gazetteers available, there is an obvious need for effective analytical tools to aid researchers in understanding and analyzing local gazetteers. An analytic tool named Local Gazetteers Research Tools (LoGaRT) was developed by Shih-Pei Chen *et al.* in 2015 to facilitate the collective analysis of a large amount of digital local gazetteers.[4] It has been shown in Chen *et al.* (2023) that LoGaRT has provided significant assistance in researching the local gazetteers for studying Chinese history from new angles that traditional close reading cannot cover.

Text-based digital analysis and retrieval methods (Peiyuan 1993; Liu *et al.* 2015a, 2015b; Li and Li

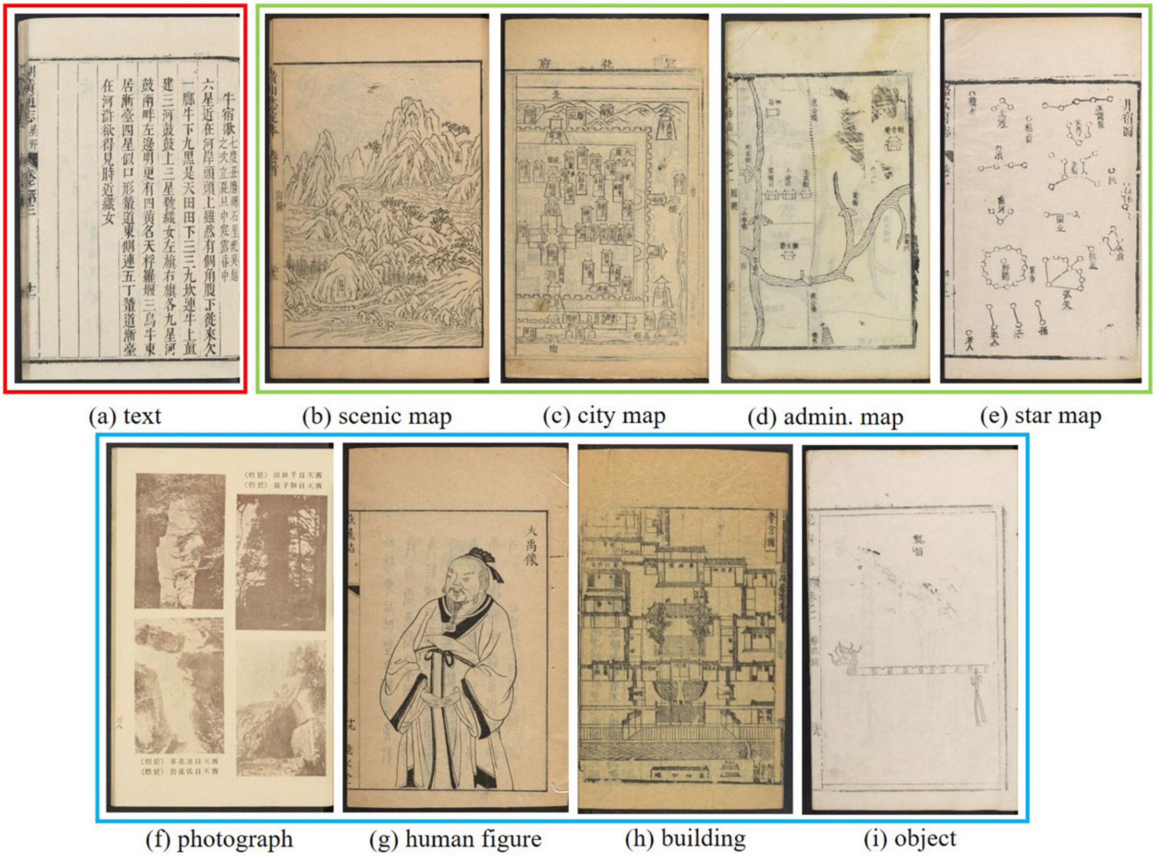


Figure 1. Examples of the nine classes and three categories in our Chinese local gazetteer dataset. The nine classes are: (**a**) text, (**b**) scenic map, (**c**) city map, (**d**) administrative map, (**e**) star map, (**f**) photograph, (**g**) human figure, (**h**) building, (**i**) object. The three categories are: text (**a**), target maps (**b, c, d, and e**), and non-target graphics (**f, g, h, and i**).

2022; Liu, Wang, and Bol 2022) have been applied to the study of full-text digitized Chinese local gazetteers. Liu, Wang, and Bol (2023) proposed a Bi-LSTM-CRF model to automatically extract biographical information from local gazetteers, which greatly improves the biographical information annotation efficiency and can be used as a supportive tool for humanities experts. Yuehua Li and Li and Li (2022) proposed an approach using named-entity recognition and clustering to identify and classify rice cultivars in local gazetteers. Lin *et al.* (2020) experienced simple neural networks with a small training dataset to build an automatic image tagging system for local gazetteer illustrations. Furthermore, Lin *et al.* (2020) developed a web Geographic Information System (GIS) platform for scholars to explore historical maps and local gazetteer illustrations in a simple, systematic way. This research focuses on image analysis, specifically the categorization of gazetteer illustrations such as maps, which is an area that has not been widely studied and lacks suitable methods as stated in Luo (2016).

2.2 DNN for image classification

Deep learning has revolutionized the field of image classification, particularly with the use of Convolutional Neural Networks (CNNs) for feature extraction and classification. CNNs have been shown to outperform traditional shallow machine learning models such as Support Vector Machines (SVMs; Hearst *et al.* 1998) and Random Forests (Breiman 2001).

2.2.1 *Early CNN prototypes*

One of the early prototypes of CNNs is LeNet (LeCun *et al.* 1998), which consists of simple convolution, pooling, and fully connected layers. However, the breakthrough of using deep CNNs for image classification came with the success of AlexNet (Krizhevsky, Sutskever, and Hinton 2017) in the ImageNet Large Scale Visual Recognition Challenge (Russakovsky *et al.* 2015) benchmark. AlexNet reduced the error rate from 26.2 per cent to 15.4 per cent and significantly outperformed other traditional machine learning methods.

2.2.2 *Deep CNN models*

Since the initial triumph of AlexNet, the field of deep CNNs has been invigorated by numerous proposals aiming to enhance the performance of these models, making them more accurate and efficient. The following are key examples of the evolution of these models.

VGGNet (Simonyan and Zisserman 2014) presented two novel variations namely VGG16 and VGG19. The innovation of VGGNet lies in deepening the network through the incorporation of additional convolutional layers. This approach improved the model's performance by amplifying its feature learning capability. Yet, the critique commonly associated with VGGNet revolves around its substantial computational resources demand, an issue future models sought to address. The introduction of the *inception structure* in InceptionNet (Szegedy *et al.* 2015) marked a significant advancement in the field. This design aimed to enhance efficiency by reducing the number of parameters and computational demand. It showcased that improvements in model performance could be attained not merely through network deepening but also through innovative restructuring of its architecture. ResNet (He *et al.* 2016) further brought a significant shift by proposing residual networks featuring skip connections, a solution designed to tackle the persistent issue of vanishing gradients in deep networks. ResNet has gained widespread adoption in various applications, its popularity being testament to its robust accuracy and efficiency.

Xception (Chollet 2017) and DenseNet (Huang *et al.* 2017) further adapted the key features of InceptionNet and ResNet to enhance model performance. These models signify an emerging trend where successful innovations are iteratively integrated and refined in subsequent models to achieve compounded performance improvements. Finally, EfficientNet (Tan and Le 2019), aiming for a more comprehensive enhancement, took a holistic view of model performance. This approach resulted in a series of models (EfficientNet-B0 to EfficientNet-B7) that balanced and optimized multiple performance metrics. The EfficientNet series underscores the ongoing evolution of deep CNN models, where both depth and breadth of perspective are considered crucial for continued advancements.

Through a critical examination of these models, it becomes evident that improvements in deep CNN architectures follow an iterative and cumulative pattern. Each advancement in this domain is built upon the successes and shortcomings of the preceding models, showcasing a continuous process of refinement and enhancement.

2.2.3 *Transformer-based models*

In the field of machine learning, transformer-based models have emerged as powerful contenders, demonstrating remarkable performance gains over conventional CNNs. This shift signals a movement toward architectures that can better handle complex relationships in the data. Here, we present some groundbreaking models that have propelled this advancement.

The Vision Transformer (ViT; Dosovitskiy *et al.* 2020), a trailblazer in this domain, takes a transformer-based approach toward image classification. The core innovation lies in its unique method of processing images as a sequence of patches rather than a matrix of pixels. This provides an entirely novel approach to interpreting visual information, capable of capturing complex, long-range dependencies between image features. Building on the success of ViT, the Swin Transformer (Liu *et al.* 2021) was proposed, introducing an additional layer of abstraction to the transformer architecture. Instead of processing the patches independently, Swin Transformer groups them into non-overlapping windows and applies a hierarchical transformation. This innovation further increases the model's ability to understand spatial relationships within the image data, enhancing its robustness in classification tasks. The Dual Attention ViT (DaViT; Ding *et al.* 2022) presents another compelling variation of the transformer model. DaViT integrates both spatial and channel attention mechanisms into the ViT architecture. This fusion of attention mechanisms empowers DaViT to effectively capture local and global contextual information, leading to an enhanced understanding of image content.

The progress in transformer-based models signifies a captivating trajectory within the realm of deep learning, highlighting architectures that possess the ability to offer sophisticated contextual interpretations of visual information. These advancements exemplify the dynamic nature of the field, characterized by its constant evolution through the critical synthesis and refinement of existing methodologies.

2.3 Addressing data imbalance and scarcity: strategies and techniques

The issues of data imbalance and scarcity are prevalent challenges in machine learning, with the potential to adversely affect the performance of deep learning models (Singh and Purohit 2015). These issues can create biased models and undermine their generalizability. A suite of strategies has been proposed to counter these problems, each with its own merits and applicability:

Data augmentation is a widely used technique that enhances the richness of the dataset by generating additional data samples derived from the original ones. This augmentation is often achieved by applying random transformations like rotation, flipping, and cropping to the existing data (Bloice, Roth, and Holzinger 2019; Buslaev *et al.* 2020). Through increasing data diversity and volume, data augmentation can help prevent overfitting, thereby improving the robustness of the model.

Data sampling is a common strategy to address data scarcity and imbalance by strategically selecting subsets of the original data for training the model. The selection can be carried out randomly, or it can employ more sophisticated techniques such as the Synthetic Minority Over-sampling Technique (Mohammed, Rawashdeh, and Abdullah 2020). This process helps in creating a balanced representation of all classes, improving the model's ability to generalize across diverse data points.

The Bootstrap Aggregating (or Bagging in short) is an ensemble technique that enhances model performance by reducing variance. It achieves this by training multiple models on different subsets of the original data and then averaging their predictions for the final decision (Breiman 1996). This diversifies the model perspective on data, leading to a more resilient and robust prediction.

The Boosting method works by training multiple models sequentially, each attempting to correct the errors made by the preceding model (Freund, Schapire, and Abe 1999; Seiffert *et al.* 2009). This method progressively focuses on hard-to-classify instances, enhancing the model's ability to handle complex data patterns.

When choosing the aforementioned strategies and techniques, it is essential to ensure they align with the specific requirements and limitations of the problem being addressed, as each approach entails distinct tradeoffs and assumptions. Therefore, gaining a comprehensive understanding of these nuances is pivotal in the development of a resilient and high-performing model.

2.4 Transfer learning

Transfer learning is a popular machine learning paradigm using a pre-trained model for a different task by fine-tuning on a smaller dataset. It offers significant advantages for the annotation of local gazetteer illustrations, which is a labor-intensive task due to the lack of large annotated datasets. The work conducted by Zhuang *et al.* (2020) primarily focused on the broader domain of machine learning and did not extensively explore the specific context of historical graphic datasets. As a result, its direct applicability to our study is limited, highlighting the need for further investigation. Instead of developing a new DNN, we leverage the efficiency of transfer learning. Notably, most high-performing image classification networks, such as those trained on ImageNet 1k (Deng *et al.* 2009) and ImageNet 21k (Ridnik *et al.* 2021), have been developed and assessed using photograph datasets. This is a crucial distinction as historical documents often contain drawings or illustrations that diverge significantly from photographic imagery.

Previous studies have ventured into the realm of transfer learning for tasks akin to ours. For instance, Granet *et al.* (2018) applied transfer learning to handwriting recognition in historical documents. They achieved promising results, but it is worth noting that the inherent dissimilarities between handwriting and illustration recognition might limit the direct transferability of their findings to our context. Roullet *et al.* (2021) utilized transfer learning to classify historical images in the collection of books named *Pitture e Mosaici* using several DNNs. Their findings, while encouraging, were constrained by the specific nature of their dataset. It remains an open question how well their approach would translate to the diverse and complex array of illustrations found in local gazetteers.

In summary, while the cited works provide a promising foundation for our study, they also highlight the need for a nuanced understanding of the potential and limitations of transfer learning in the context of historical graphic datasets. Through this study, we aim to contribute to this evolving discourse.

2.5 Image classification in digital humanities

While text analysis via OCR is a common focus in digital humanities, with several successful applications reported in the literature (Du *et al.* 2020; Martínek, Lenc, and Král 2020), visual materials hold equal significance. Hence, image analysis, particularly image classification, emerges as a crucial tool for scholars in this field.

There are many image analysis systems for digital humanity research applications, including visualization systems, visual search systems, object classification and recognition systems, and image restoration systems. In this research, our focus is centered on image classification. We aim to critically examine previous works in this area and identify potential gaps for further exploration.

Sarõ, Salah, and Akdag Salah (2019) employed VGGNet as a feature extractor kingma2014adam and trained SVMs and random decision forests with these features to classify the gender of portrait subjects. While their approach demonstrated the feasibility of using deep learning features in digital humanities, it primarily targeted a specific task, gender classification, leaving open questions about its applicability to other tasks or image types.

Guoxin *et al.* (2019) compared the performance of AlexNet, VGGNet, and a self-constructed CNN for classifying Dongba classical ancient books. Their work underscores the importance of choosing the right network architecture for the task at hand. However, their study was limited to a specific type of ancient book, and it remains unclear how well their findings translate to other types of historical documents.

Im, Kim, and Mandl (2022) applied AlexNet, ResNet, and EfficientNet to classify various printing technologies for digitized historical books. Their research contributes valuable insights into the application of different networks to the task of printing technology classification. Nonetheless, it is important to consider the specificity of their task, which may limit the generalizability of their results to other tasks within digital humanities.

In summary, while these works provide valuable insights into image classification applications in digital humanities, they also highlight the need for additional research to understand how these methods can be generalized to a wider range of tasks and types of historical documents. We aim to contribute to this ongoing discourse through our research.

3. Methodology

We present the methodologies employed to prepare the dataset and train the deep learning models for image classification in Chinese local gazetteers. We begin by introducing the dataset and annotation tools, followed by an overview of the organization of the dataset into two distinct label partitions. Subsequently, we provide a detailed account of our data preprocessing pipeline, encompassing image splitting, binarization, resizing, and border cropping utilizing object detection. Additionally, we elaborate on the data augmentation techniques implemented to address the issue of insufficient data. Furthermore, we explicate our approach for selecting and fine-tuning pre-trained networks, highlighting the rationale behind our choice of the DaViT (Ding *et al.* 2022) as the optimal fit for our research objectives. Lastly, we introduce a web service for document image classification that we have developed to assist scholars in their examination of local gazetteer illustrations.

3.1 Dataset and annotation tools

We used the high-resolution, open-access scans provided by Harvard-Yenching Library[2] and the manually assigned labels provided by the Visual Materials Working Group at MPIWG.[3] These digital images and annotations of local gazetteers are available through the Local Gazetteers Research Tools (LoGaRT),[4] a software developed by MPIWG for searching, analyzing, and collecting data from digitized Chinese local gazetteers.

3.2 Dataset organization and label partitioning

The dataset consists of 4,309 digital images of Chinese local gazetteers, with a train/test ratio of 3:1, resulting in 3,232 images in the train set and 1,077 images in the test set. We perform two types of dataset organization in producing two different label partitions, as follows:

Table 1. Statistics of our Chinese local gazetteer dataset

#	Class	Quantity	Percent
1	Text	427	9.9
2	Scenic map	2,000	46.4
3	City map	1,036	24.0
4	Administrative map	208	4.8
5	Star map	233	5.4
6	Photograph	166	3.9
7	Human figure	158	3.7
8	Building	45	1.0
9	Object	36	0.8
10	Text	427	9.9
11	Map	3,477	80.7
12	Non-map graphic	405	9.4

(Left) the nine classes and (Right) the three categories of the nine classes.

- Nine-Class Dataset: This dataset divides our gazetteer dataset into nine classes, as shown in Fig. 1. The classes are: *text, scenic map, city map, administrative map, star map, human figure, photograph, building, and object.* Table 1 (left) shows the sample quantity and statistics of the nine-class organization of the Chinese local gazetteer dataset.
- Three-Category Dataset: This dataset divides the same set into only three coarse categories, as shown in Fig. 1. The categories are: *text, map, and non-map graphic.* The map category includes scenic map, city map, administrative map, and star map. The non-map graphic category includes *human figure, photograph, building, and object.* Table 1 (right) shows the sample quantity and statistics of the three-category organization of the Chinese local gazetteer dataset.

These two types of dataset organization allow us to study how well the classification performs in different settings.

3.3 Data preprocessing pipeline

To prepare the data for deep learning algorithms, we perform the following preprocessing steps:

- Image Splitting and Relabeling: To avoid the problem of multi-label classification and ensure clarity for the model, we split one image into two and relabel the dataset. This adds approximately 2,000 images to our dataset and eliminates ambiguity for the model.

- Image Binarization and Resizing: In our dataset, the illustrations have varying background colors caused by the color differences of the papers used. To reduce color variability and improve the consistency of the dataset, we perform image binarization, converting color images to black-and-white images. Additionally, we resize images to a standard size to make them compatible with our models.
- Border Cropping using Object Detection: The illustrations in our dataset often contain white borders, which are considered irrelevant and can affect the performance of the classification models. To reduce differences in white border size, we use the YOLOv5 object detector (Jocher *et al.* 2020) to localize the illustrations and crop out the white borders. We manually annotate 220 images in our dataset with the illustration bounding boxes to train the YOLOv5 object detector and integrate it into our processing pipeline, as shown in Fig. 2.
- Data Augmentation for Mitigating Data Insufficiency: The data statistics in Table 1 show that the minority classes in our dataset face the challenge of data insufficiency, with scenic map and city map classes making up about 70 per cent of the dataset, while the other seven classes account for only 30 per cent of the dataset. To mitigate the data insufficiency issue and improve the classification performance, we apply image processing techniques such as horizontal flip, vertical flip, sharpen, and blur to generate additional training samples by up to 4 times using data augmentation (Bloice, Roth, and Holzinger 2019; Buslaev *et al.* 2020). With data augmentation in model training, the accuracy and F1-score of the classification model are increased by 3 per cent and 5 per cent, respectively. Additionally, the F1-scores for under-performing classes such as administrative map and building are increased by 7 per cent and 16 per cent, respectively.

3.4 Selection and fine-tuning of pre-trained networks

Once the dataset is well-prepared for deep learning algorithms, we perform supervised learning to train our classification models. We have observed that fine-tuning state-of-the-art pre-trained networks outperforms training our own DNN from scratch. However, due to the significant domain gap between the widely

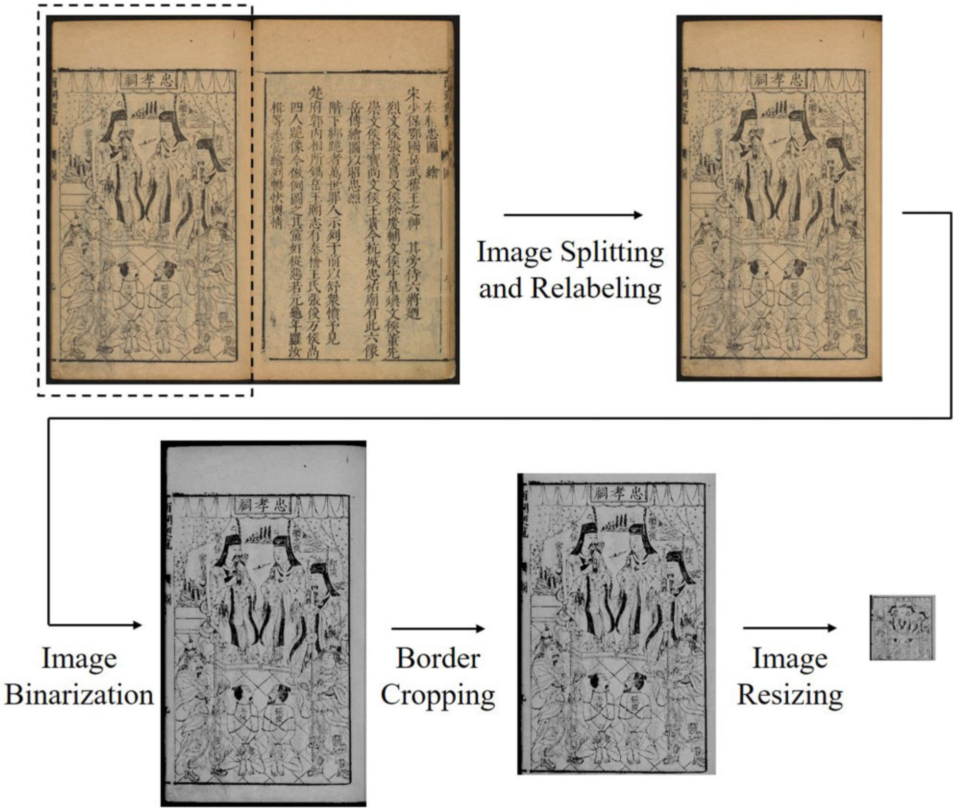


Figure 2. Data preprocessing pipeline.

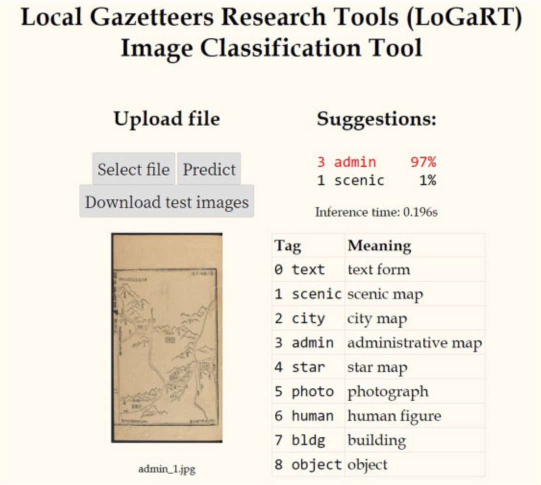
adopted ImageNet dataset (Deng *et al.* 2009; Ridnik *et al.* 2021) and our local gazetteer dataset, it is crucial to evaluate different pre-trained networks to find the best fit for our dataset. In our case, we have evaluated eight state-of-the-art DNNs and determined that DaViT is the best fit for our research, as it outperforms the other networks in most aspects (see Section 4.3).

3.5 Document image classification web service and its potential applications

To catalyze advancements in the field of digital humanities, we have developed a web service designed for document image classification (as detailed in Fig. 3).[8] This tool is intended to specifically aid scholars examining illustrations found within local gazetteers.

Upon using the service, users can upload a selected image and promptly receive the predicted classification by clicking the appropriate button. The interface is divided into two sections for ease of understanding: the left side presents the uploaded image along with its filename, while the right side elaborates on the classification results. The latter encompasses the probabilistic distribution across each class, detailed explanations for every label, and the time it takes to infer the classification of a single image.

Although we have described a high-level view of the potential benefits of the web service, it is crucial to examine the practicality and breadth of its applications. For example, a historian could utilize this service to quickly categorize an array of unclassified images from various gazetteers. Similarly, an archivist may find this tool useful for organizing a large collection of document images in a systematic manner. The adaptability of the service could also extend to materials from diverse sources, contingent on their similarity with the training dataset. However, the effectiveness of the service when applied to differing sources and types of materials is a subject requiring further exploration. Understanding the limitations and strengths of the service in various scenarios is crucial to its effective development and utility.



Local Gazetteers Research Tools (LoGaRT)
Image Classification Tool

Upload file

Suggestions:

```
3 admin    97%
1 scenic    1%
```

Inference time: 0.196s

Tag	Meaning
0 text	text form
1 scenic	scenic map
2 city	city map
3 admin	administrative map
4 star	star map
5 photo	photograph
6 human	human figure
7 bldg	building
8 object	object

admin_1.jpg

Figure 3. Screenshot of the developed web service.

In the spirit of open science, the code and the dataset pivotal to this research are publicly accessible at (https://github.com/AlanChen26/ChineseLocalGazetteers_ImageClassification/).[9] This move aims to stimulate further research and advancements in the field, fostering a collaborative community of scholars and developers.

4. Experimental results

In this section, we present the experimental results of our work, where we compare the performance of eight DNNs used in our Chinese local gazetteer-type classification model. First, we introduce the evaluation metrics used for comparing the classification performance.

4.1 Evaluation metrics

We use standard classification metrics to evaluate the performance of our model. These metrics provide a fair and accurate analysis of the classification results. The evaluation metrics used in our work are as follows:

- Overall accuracy: The overall accuracy of the model is the proportion of correctly predicted cases out of all cases. In the context of a multiple class prediction problem, True Positives (TP) denote correctly predicted classes where the ground truth is positive, while True Negatives (TN) represent correctly predicted classes where the ground truth is negative. False Positives (FP) indicate incorrectly predicted classes where the ground truth is positive, and False Negatives (FN) represent incorrectly predicted classes where the ground truth is negative.
- Precision: Precision is the proportion of real-positive cases out of all predicted positive cases:

$$\text{Precision} = \frac{TP}{TP + FP}.$$

- Recall: Recall is the percentage of correctly predicted positive cases out of all real positive cases:

$$\text{Precision} = \frac{TP}{TP + FN}.$$

- F1-score: The F1-score is a commonly adopted metric for evaluating performance in classification tasks. It combines precision and recall into a single metric, calculated as the harmonic mean of the two:

$$\mathrm{F1} = \frac{2 * Precision * Recall}{Precision + Recall}.$$

The precision, recall, and F1-score of the model are calculated as the mean of all classes by averaging the respective scores.

4.2 Experimental setup and training configurations

In our experiments, we use the Adam optimizer (Kingma and Ba 2014) and the cross-entropy loss function. The batch size is set to 32, and the learning rate is set to 0.001. We apply the early stopping method to prevent overfitting and the learning rate decay method to optimize performance. All models are trained on a Tesla V100 GPU.

4.3 Experimental results and performance evaluation of deep neural networks

We evaluate the performance of eight DNNs on both the nine-class and three-category datasets using a single test set consisting of 1,077 images. To ensure the test set is representative, we perform a stratified split on our dataset, maintaining the distribution of the original dataset in both the train set and the test set. Experiments are based on stratified 4-fold cross-validation to ensure the stability of the classification results.

Table 2. Model comparison results on the nine-class test set.

Model	Accuracy	Precision	Recall	F1-score
DaViT	0.9387	0.9413	0.8887	0.9059
Swin Transformer V2	0.9369	0.9509	0.8884	0.9133
ViT	0.9922	0.9257	0.8882	0.9011
DenseNet201	0.9171	0.8980	0.8628	0.8741
Xception	0.9116	0.9260	0.8424	0.8709
MobileNetV2	0.9090	0.9123	0.8426	0.8671
ResNet101V2	0.9011	0.9000	0.8131	0.8452
InceptionV3	0.8705	0.8704	0.7926	0.8195

Table 3. Model comparison results on the three-category test set.

Model	Accuracy	Precision	Recall	F1-score
DaViT	0.9898	0.9883	0.9711	0.9793
Swin Transformer V2	0.9886	0.9841	0.9707	0.9771
ViT	0.9838	0.9708	0.9651	0.9674
DenseNet201	0.9842	0.9804	0.9558	0.9674
Xception	0.9835	0.9758	0.9584	0.9667
MobileNetV2	0.9775	0.9617	0.9434	0.9546
ResNet101V2	0.9766	0.9594	0.9460	0.9524
InceptionV3	0.9740	0.9623	0.9327	0.9460

The eight DNNs we compare are: DaViT, Swin Transformer, ViT, DenseNet, Xception, MobileNet, ResNet, and InceptionNet. Tables 2 and 3 show the quantitative comparison results.

DaViT is the best-performing model in both the nine-class dataset (93.9% accuracy and 90.6% F1-score) and the three-category dataset (98.9% accuracy and 97.9% F1-score). Swin Transformer V2 also delivers impressive results, surpassing DaViT in precision and F1-score on the nine-class dataset. Nevertheless, since accuracy is the primary metric, Swin Transformer V2 emerges as the runner-up. ViT also performs well, while DenseNet201, Xception, MobileNetV2, and ResNet101V2 produce acceptable results. However, InceptionV3 performed less satisfactorily. The average inference time of our classification model for a single image is approximately 0.01 s.

Most classification models perform better on the three-category dataset than on the nine-class dataset. This is likely due to the distinct visual styles of *text, maps, and non-map graphics*, which makes it easier for models to classify them correctly in the three-category case. On the other hand, different types of maps may have similar visual styles, making it more challenging for models to categorize them correctly in the nine-class case.

Transformer-based models achieve better performance than CNN-based models in our research. This is because CNN-based models are based on local convolution operations, while transformer-based models use attention mechanisms to process the image as a sequence of patches, which allows them to learn the relationships between patches and capture global contextual information.

5. Discussions

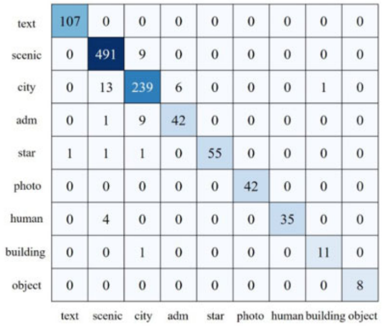
In this section, we discuss the classification results of our best-performing model, DaViT, on the nine-class dataset. We examine the performance of specific classes and analyze failure cases, seeking to identify the causes and characteristics of these errors. We also present the confusion matrix and provide examples of misclassifications to illustrate the challenges of the classification task.

5.1 Analysis of failure cases

5.1.1 *Performance of text class*

Our method demonstrates exceptional performance in classifying text document images with the precision of 99 per cent, recall of 100 per cent, and F1-score of 99.5 per cent as shown in Fig. 4 (left). This is expected as text-form images are visually distinct from graphic-form images, making it easy for the classification

Class	Precision	Recall	F1-score
Text	0.9907	1.0000	0.9953
Scenic map	0.9627	0.9820	0.9723
City map	0.9228	0.9228	0.9228
Administrative map	0.8750	0.8077	0.8400
Star map	1.0000	0.9453	0.9735
Photograph	1.0000	1.0000	1.0000
Human figure	1.0000	0.8974	0.9459
Building	0.9167	0.9167	0.9167
Object	1.0000	1.0000	1.0000



	text	scenic	city	adm	star	photo	human	building	object
text	107	0	0	0	0	0	0	0	0
scenic	0	491	9	0	0	0	0	0	0
city	0	13	239	6	0	0	0	1	0
adm	0	1	9	42	0	0	0	0	0
star	1	1	1	0	55	0	0	0	0
photo	0	0	0	0	0	42	0	0	0
human	0	4	0	0	0	0	35	0	0
building	0	0	1	0	0	0	0	11	0
object	0	0	0	0	0	0	0	0	8

Figure 4. The left shows the best classification results of the nine classes using DaViT. The right shows the classification confusion matrix.

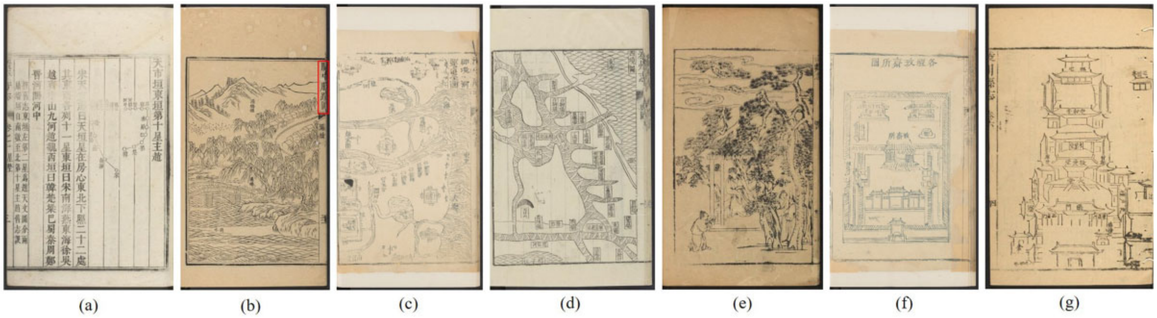


Figure 5. Ambiguous and error cases. (**a**) An example of a star map that is predicted as text. (**b**) An example of an administrative map that is predicted as a scenic map. (**c**) An example of an administrative map that is predicted as a city map. (**d**) An example of a city map that is predicted as an administrative map. (**e**) An example of a human figure that is predicted as a scenic map. (**f**) An example of a building that is predicted as a city map. (**g**) An example of a city map that is predicted as a building.

model to make accurate predictions. However, the confusion matrix in Fig. 4 (right) illustrates that a star map is misclassified as text. With further examination, Fig. 5 (a) shows that this misclassification may not necessarily be an error as there is text surrounding the star map. Our classification system can detect visual similarities and aid professionals in analyzing local gazetteer illustrations. In many cases, it can even identify and correct human annotation errors.

5.1.2 *Performance of administrative map class*

The administrative map class exhibits relatively lower performance with the precision of 87.5 per cent, recall of 80.7 per cent, and F1-score of 84 per cent as seen in Fig. 4 (left). The confusion matrix in Fig. 4 (right) illustrates that administrative maps are misclassified as city maps and scenic maps. Figure 5b shows an example of an administrative map classified as a scenic map. However, this misclassification may not necessarily be an error as the administrative map contains Chinese words (勝蹟圖) which are associated with a scenic map. Figure 5c and d shows administrative maps and city maps can have similar visual appearances. This is why we organized the nine classes into three categories, as maps in the same category are visually similar which can cause ambiguity for the classification model.

5.1.3 *Human figures misclassified as scenic maps*

The confusion matrix in Fig. 4 (right) reveals that four human figures are misclassified as scenic maps. Figure 5e shows that the human figures only occupy a small portion of the images and the presence of trees may cause the classification model to misclassify them as scenic maps. This illustrates the multivalent nature of visuality, where one image can contain elements of multiple classes.

5.1.4 *Performance of building class*

The building class also exhibits relatively lower performance with the precision, recall, and F1-score of 91.7 per cent as seen in Fig. 4 (left). The confusion matrix in Fig. 4 (right) illustrates that a building is misclassified as a city map and a city map is misclassified as a building. Figure 5f and g demonstrates that buildings can be visually similar to city maps. However, visual similarity does not always reflect conceptual organization. Therefore, although our developed classification tool can provide satisfactory accuracy, it can only assist professionals in organizing local gazetteer documents and cannot replace their expertise.

6. Limitations

In this section, we discuss the limitations of our study.

6.1 Limited recognition of classes

Our image classification models (like most other DNNs) are data-driven, so they can only recognize the nine classes of graphics they are trained on. This means that other local gazetteer illustrations such as diagrams and ritual layouts that are not included in our train set due to data scarcity may be misclassified into the nine classes or not recognized. Therefore, the recognition ability of our models is limited to the specific classes for which they were trained.

6.2 Possibility of errors

Although the developed models achieve 87–94 per cent accuracy and 82–91 per cent F1-scores in classifying the Chinese local gazetteer illustrations, errors can still occur. The classification performance may vary depending on the quality of the image and the complexity of the graphics. Thus, the classification model should be used as an assistive tool rather than a replacement for experts. The expertise of human analysts is still necessary to interpret the results and make informed decisions.

7. Conclusion

This study resulted in the development of a document image classification system specifically designed to identify and categorize illustrations from Chinese local gazetteers. By enabling the automatic annotation and organization of these collections, our system has made significant contributions to the field of Chinese historical research. It provides an avenue for researchers to navigate visual materials in digital archives without exclusive reliance on text-based search. Furthermore, our methods can facilitate the identification of similar documents in large digital collections and assist in the analysis of visual trends.

The method we developed employs a combination of data preprocessing, data augmentation, and supervised learning techniques. This approach has achieved accuracy rates ranging from 87 to 94 per cent, and F1-scores between 82 and 91 per cent. The methods we have developed are adaptable and extendable to the analysis and research of historical documents beyond Chinese local gazetteers.

To make our classification system accessible to researchers, we have also developed a user-friendly web service. This allows nontechnical researchers to categorize Chinese local gazetteer illustrations via a straightforward interface.

Future work stemming from this study includes the development of an API, which will provide researchers with easy access to our classification system. This will enable the verification of manually labeled data and the automatic annotation of unlabeled data. Additionally, we plan to incorporate OCR technology (Mori, Suen, and Yamamoto 1992) for the development of a text detection and recognition module. This addition would enable automatic analysis of significant textual information such as illustration captions, graphic text, and toponyms on maps. By integrating Natural Language Processing tools (Manning and Schutze 1999; Piotrowski 2012), we can further enhance automatic analysis and retrieval of these documents. We foresee vision-language models making a significant contribution to future research trends in digital humanities and history studies, playing a pivotal role in advancing these fields.

In conclusion, our classification system and web service will greatly aid researchers in the field of Chinese historical studies by providing efficient, accurate access to local gazetteer illustrations. The dataset used in this study is publicly available and can support further research in this area.

Acknowledgements

The authors warmly thank Harvard Yenching Library for digitizing their rare holdings and offering them to the world through open access, which allows this research to take place. We also thank scholars from the “Tu (圖) in Local Gazetteers” Working Group at the MPIWG, in particular Kenneth J. Hammond, Anne Gerritsen, Shellen Wu, Jiajing Zhang, and Daniel Knorr, for developing the classification scheme for gazetteer illustrations and for manually labeling them, part of which constitutes to the training data of this research. We thank Calvin Yeh, IT Architect of LoGaRT at MPIWG, for pulling out the training data from LoGaRT to support this study.

Author contributions

Jhe-An Chen (Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing), Jen-Chien Hou (Visualization, Writing—review and editing), Richard Tzong-Han Tsai (Conceptualization, Project administration, Resources, Supervision, Writing—review and editing), Ming-Ching

Chang (Conceptualization, Project administration, Supervision, Writing—review and editing), Hsiung-Ming Liao (Conceptualization, Project administration, Supervision), and Shih-Pei Chen (Conceptualization, Data curation, Writing—review and editing)

Notes

1. An example of image annotation tools for historical contents is the Ten Thousand Rooms Project at Yale University. It can be accessed via https://tenthousandrooms.yale.edu/ (accessed June 30, 2023).
2. The rare local gazetteers used in this research can be found in the Chinese Rare Book Collection at Harvard University via https://library.harvard.edu/collections/chinese-rare-books.
3. More information about the Visual Materials (Tu 圖) in Local Gazetteers Working Group can be found at https://www.mpiwg-berlin.mpg.de/research/projects/tu-tu-local-gazetteers. Related works see Chen *et al.* (2020) and Lin *et al.* (2020).
4. More information about Local Gazetteers Research Tools (LoGaRT) can be accessed via https://www.mpiwg-berlin.mpg.de/research/projects/logart-local-gazetteers-research-tools. Related works see Chen *et al.* (2020)
5. Erudition's Database of Chinese Local Records can be accessed via http://er07.com/home/pro_ 87.html
6. East View's China Comprehensive Gazetteers can be accessed via https://www.eastview.com/resources/e-collections/china-comprehensive-gazetteers/
7. Diaolong's Full-text Database of Chinese and Japanese Ancient Books can be accessed via http://hunteq.com/ancientc/ancientkm
8. The web service we designed for document image classification can be accessed via http://logart-ict.map.net.tw/
9. The code and the dataset pivotal to this research can be accessed via https://github.com/AlanChen26/ChineseLocalGazetteers_ImageClassification/tree/main

References

Antonacopoulos, A. *et al.* (2013) Icdar 2013 competition on historical book recognition (hbr 2013). *2013 12th International Conference on Document Analysis and Recognition*, pp. 1459–63. Washington, DC, USA: IEEE.

Besek, J. M. (2003) *Copyright Issues Relevant to the Creation of a Digital Archive*. Berlin, Germany: Walter de Gruyter GmbH and Co. KG.

Bloice, M. D., Roth, P. M., and Holzinger, A. (2019) 'Biomedical Image Augmentation Using Augmentor', *Bioinformatics*, **35**: 4522–4.

Breiman, L. (1996) 'Bagging Predictors', *Machine Learning*, **24**: 123–40.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, **45**: 5–32.

Buslaev, A. *et al.* (2020) 'Albumentations: Fast and Flexible Image Augmentations', *Information*, **11**: 125.

Chen, S. P. *et al.* (2020) 'Local Gazetteers Research Tools: Overview and Research Application', *Journal of Chinese History* 中國歷史學刊, **4**: 544–58.

Chen, S.P. *et al.* (2023) 'Treating a Genre as a Database: A Digital Research Methodology for Studying Chinese Local Gazetteers', *International Journal of Digital Humanities*, **4**: 171–93.

Chollet, F. (2017) Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–8. Honolulu, HI, USA: IEEE.

Deng, J. *et al.* (2009) ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Miama, Florida, USA: IEEE.

Dennis, J. R. (2015) *Writing, Publishing, and Reading Local Gazetteers in Imperial China, 1100–1700*. Leiden, Netherlands: BRILL.

Ding, M. *et al.* (2022) Davit: Dual attention vision transformers. *Computer Vision–ECCV 2022: 17th European Conference, Proceedings, Part XXIV*, October 23–27, 2022, , pp. 74–92. Israel: Tel Aviv.

Dosovitskiy, A. *et al.* (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Du, Y. *et al.* (2020) PP-OCR: A practical ultra lightweight ocr system. arXiv:2009.09941.

Freund, Y., Schapire, R., and Abe, N. (1999) 'A Short Introduction to Boosting', *Journal-Japanese Society for Artificial Intelligence*, **14**: 1612.

Granet, A. *et al.* (2018) Transfer learning for handwriting recognition on historical documents. *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Prague, Czech Republic: Springer.

Guoxin, W. *et al.* (2019) Dongba classical ancient books image classification method based on ReN-Softplus convolution residual neural network. *2019 14th IEEE International Conference on Electronic Measurement and Instruments (ICEMI)*, pp. 398–404. Changsha, China: IEEE.

He, K. *et al.* (2016) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–8. Las Vegas, NV, USA: IEEE.

Hearst, M. A. *et al.* (1998) 'Support Vector Machines', *IEEE Intelligent Systems and their Applications*, **13**: 18–28.

Huang, G. *et al.* (2017) Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA*, pp. 4700– 8.

Im, C., Kim, Y., and Mandl, T. (2022) 'Deep Learning for Historical Books: Classification of Printing Technology for Digitized Images', *Multimedia Tools and Applications*, **81**: 5867–88.

Jocher, G. *et al.* (2020) Yolov5. Code repository. https://github.com/ultralytics/yolov5.

Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017) 'ImageNet Classification with Deep Convolutional Neural Networks', *Communications of the ACM*, **60**: 84–90.

LeCun, Y. *et al.* (1998) 'Gradient-based Learning Applied to Document Recognition', *Proceedings of the IEEE*, **86**: 2278–324.

Li, Y. and Li, H. (2022) 'Exploring the Rice Cultivars in Large-scale Chinese Local Gazetteers: A Computational Approach', *Plants*, **11**: 3403.

Lin, N. Y. *et al.* (2020) 'Displaying Spatial Epistemologies on Web GIS: Using Visual Materials from the Chinese Local Gazetteers as an Example', *International Journal of Humanities and Arts Computing*, **14**: 81–97.

Liu, C. L. *et al.* (2015a) Mining local gazetteers of literary chinese with crf and pattern based methods for biographical information in chinese history. *2015 IEEE International Conference on Big Data (Big Data)*, pp. 1629–38. Santa Clara, CA, USA: IEEE.

Liu, C. L. *et al.* (2015b) Toward algorithmic discovery of biographical information in local gazetteers of ancient China. *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 87–95. Shanghai, China: Shanghai Jiao Tong University.

Liu, Z. *et al.* (2021) Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–22.

Liu, Z., Wang, H., and Bol, P. K. (2023) 'Automatic Biographical Information Extraction from Local Gazetteers with Bi-lstm-crf Model and Bert', *International Journal of Digital Humanities*, **4**: 195–212.

Luo, Q. (2016) Research on the illustrations of chinese local gazetteers: Overview, evaluation, and potential approach for future study. *Proceedings the 2nd International Conference on Social Sciences*. http://ase-scoop.org/papers/IWAHS-2016/5.Luo_IWAHS. pdf.

Manning, C. and Schutze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Martínek, J., Lenc, L., and Král, P. (2020) 'Building an Efficient OCR System for Historical Documents with Little Training Data', *Neural Computing and Applications*, **32**: 17209–27.

Mohammed, R., Rawashdeh, J., and Abdullah, M. (2020) Machine learning with oversampling and undersampling techniques: Overview study and experimental results. *2020 11th International Conference on Information and Communication Systems (ICICS)*, pp. 243–8. Copenhagen, Denmark: Springer.

Mori, S., Suen, C. Y., and Yamamoto, K. (1992) 'Historical Review of OCR Research and Development', *Proceedings of the IEEE*, **80**: 1029–58.

Peiyuan, Z. (1993) 'Extraction of Climate Information from Chinese Historical Writings', *Late Imperial China*, **14**: 96–106.

Piotrowski, M. (2012) 'Natural Language Processing for Historical Texts', *Synthesis Lectures on Human Language Technologies*, **5**: 1–157.

Ridnik, T. *et al.* (2021) Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972.

Roullet, C. *et al.* (2021) Transfer learning methods for extracting, classifying and searching large collections of historical images and their captions. *International Conference on Pattern Recognition*, pp. 185–99. Taichung, Taiwan: Springer.

Russakovsky, O. *et al.* (2015) 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision*, **115**: 211–52.

Sarõ, C., Salah, A. A., and Akdag Salah, A. A. (2019) 'Automatic Detection and Visualization of Garment Color in Western Portrait Paintings', *Digital Scholarship in the Humanities*, **34**: i156–71.

Seiffert, C. *et al.* (2009) 'Rusboost: A Hybrid Approach to Alleviating Class Imbalance', *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, **40**: 185–97.

Simonyan, K., and Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Singh, A. and Purohit, A. (2015) 'A Survey on Methods for Solving Data Imbalance Problem for Classification', *International Journal of Computer Applications*, **127**: 37–41.

Szegedy, C. *et al.* (2015) Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9. Piscataway, NJ: IEEE.

Tan, M. and Le, Q. (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning, Long Beach, CA, USA*, pp. 6105–14.

Yang, H. *et al.* (2018) 'Dense and Tight Detection of Chinese Characters in Historical Documents: Datasets and a Recognition Guided Detector', *IEEE Access*, **6**: 30174–83.

Zhuang, F. *et al.* (2020) 'A Comprehensive Survey on Transfer Learning', *Proceedings of the IEEE*, **109**: 43–76.

Zhuang, W. *et al.* (1985) Zhongguo di Fang Zhi Lian he mu lu/zhongguo ke Xue Yuan Beijing Tian Wen Tai Zhu Bian; [Zong Bian Zhuang Weifeng, Zhu Shijia, Feng Baolin; Bian ji wang Shuping ... *et al.*; Bian Shen wu Fengpei, Zhang Xiumin, Yang Dianxun] (Di 1 ban.). Zhonghua shu ju: Xin hua Shu Dian Beijing fa Xing Suo fa Xing Beijing.