scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A large-scale dataset for Chinese historical document recognition and analysis

Yongxin Shi[1], Dezhi Peng[1,2], Yuyi Zhang[1], Jiahuan Cao[1] & Lianwen Jin[1,3] ✉

The development of Chinese civilization has produced a vast collection of historical documents. Recognizing and analyzing these documents hold significant value for the research of ancient culture. Recently, researchers have tried to utilize deep-learning techniques to automate recognition and analysis. However, existing Chinese historical document datasets, which are heavily relied upon by deep-learning models, suffer from limited data scale, insufficient character category, and lack of book-level annotation. To fill this gap, we introduce HisDoc1B, a large-scale dataset for Chinese historical document recognition and analysis. The HisDoc1B comprises 40,281 books, over 3 million document images, and over 1 billion characters across 30,615 character categories. To the best of our knowledge, HisDoc1B is the largest dataset in the field, surpassing existing datasets by more than 200 times in scale. Additionally, it is the only dataset with book-level annotations and punctuation annotations. Furthermore, extensive experiments demonstrate the high quality and practical utility of the proposed HisDoc1B. We believe that HisDoc1B could provide valuable resources to boost the advancement of research in this domain.

Background & Summary

Historical documents are invaluable cultural heritages produced during the evolution of human civilization. Particularly, the long-standing Chinese civilization has left a vast collection of such documents, covering various fields such as history, art, and medicine. Therefore, the recognition and analysis of Chinese historical documents hold significant value for unveiling ancient Chinese culture, hence considered an important and urgent research topic. Advances in artificial intelligence (AI), particularly in deep learning, have facilitated the development of automatic recognition and analysis techniques[1,2] for these documents, substantially reducing the reliance on extensive manual-labor by human experts. Nonetheless, the efficacy of these deep-learning-based methods heavily depends on the availability of extensive annotated datasets for model training.

In this field, there have been several pioneering efforts in dataset construction. For instance, Xu *et al.*[3] established the CASIA-AHCDB, which contains over 2.2 million character images from Chinese historical documents. The MTHv1 proposed by Yang *et al.*[4] is the first dataset with page-level annotation, which consists of 1,500 historical document images with annotated texts and their reading order. However, it is limited to a single document type, i.e., Buddhist scriptures. Building upon MTHv1, Ma *et al.*[5] introduced MTHv2 by expanding the dataset to 3,199 images, but it still contains only Buddhist scriptures. Similar to MTHv1 and MTHv2, IC19 HDRC[6] provides page-level annotation and includes 12,850 historical document images, but it also covers only a single type of document, namely family genealogies. To address the limitations in style within the above datasets, Shi *et al.*[7] proposed M5HisDoc, which includes 8,000 historical document images and features multiple styles. However, there are still several key limitations that persist:

- Limited data scale: All existing datasets contain only a few thousand to tens of thousands of historical document images. This is very small compared to the massive amount of historical documents in the real world, which hinders the development of relevant methods.
- Insufficient character category coverage: The maximum number of character categories in the existing dataset is only 16,151, which hinders the models in addressing the challenges of category diversity in real-world scenarios.

[1]School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510641, China. [2]Huawei Cloud, 518129, Shenzhen, China. [3]SCUT-Zhuhai Institute of Modern Industrial Innovation, Zhuhai, 519175, China. ✉e-mail: eelwjin@scut.edu.cn

Dataset	#Books	#Document images	#Characters	#Character categories	Text punctuation
MTHv1[4]	—	1,500	521,370	4,058	×
MTHv2[5]	—	3,199	1,081,678	6,733	×
IC19 HDRC[6]	—	<u>11,715</u>	2,482,994	8,353	×
M5HisDoc[7]	—	8,000	<u>4,367,360</u>	<u>16,151</u>	×
CASIA-AHCDB[3]	—	—	2,276,740	10,350	×
HisDoc1B[8] (Ours)	**40,281**	**3,163,330** (270×)	**1,082,544,808** (248×)	**30,615** (1.9×)	✓

Table 1. Comparison of HisDoc1B with existing Chinese historical document datasets. The highest and second highest values within each column are denoted by **bold** and <u>underline</u>, respectively. HisDoc1B is more than 200 and 1.9 times greater than existing dataset in the count of image/character and character category, respectively. It is the only dataset that includes book-level and punctuation annotations.

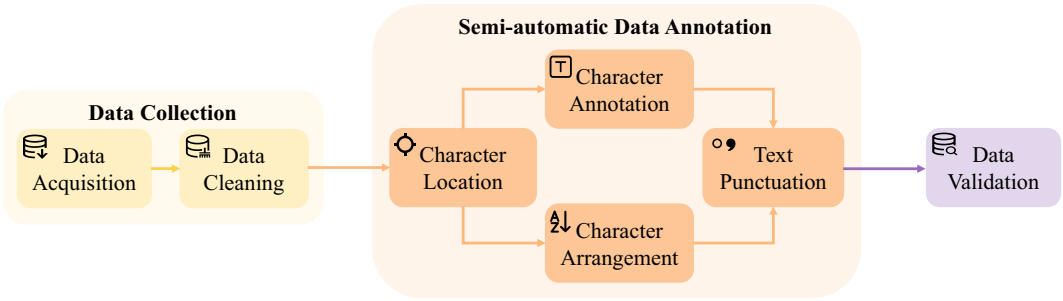


Fig. 1 The pipeline of building HisDoc1B dataset.

- Lack of book-level annotation: The annotations in the existing dataset are primarily focused on the page-level, without book-level annotations, making it impossible to conduct book-level research.

To address these limitations, we introduce HisDoc1B[8], a large-scale dataset of Chinese historical documents. For its construction, we design an effective semi-automatic annotation method, which contains four main processes: character location, character annotation, character arrangement, and text punctuation. Leveraging this method, we harness vast amounts of unlabeled web data to construct HisDoc1B. As detailed in Table 1, HisDoc1B includes annotations for over 40 thousand books, 3 million images, and 1 billion characters across 30,651 categories. HisDoc1B has the following features. (a) As **the largest** dataset in the field, it is over 200 times greater than the largest existing datasets. (b) It is **the most completely annotated** dataset of Chinese historical documents, offering unique annotations of book-level and punctuation. We believe this dataset will aid and inspire future research in the recognition and analysis of Chinese historical documents.

Methods

To construct a large-scale dataset of Chinese historical documents, we designed a systematic pipeline, as illustrated in Fig. 1. This pipeline contains three key steps: data collection, data annotation, and data validation. The data annotation process is based on a semi-automatic annotation method we proposed, as depicted in Fig. 2, which significantly reduces the burden of manual-labor. The semi-automatic data annotation method compromises four main steps: character localization, character annotation, character arrangement, and text punctuation. In this section, we will delve into the details of the data construction.

Data collection. *Data acquisition.* The main purpose of our study is to construct a large-scale dataset of Chinese historical documents. To achieve this goal, we need to acquire a sufficient amount of raw historical document data. Therefore, we selected a website with extensive historical document resources, GuoXueDaShi (https://www.guoxuedashi.net/guji/). This platform boasts an extensive collection of over one million books of historical documents, which were collected and organized by classical Chinese enthusiasts. The magnitude of data available on this website meets the requirements of our research. From this website, approximately 45,000 scanned books of historical documents were downloaded as source data, which are primarily in Portable Document Format (PDF) and DjVu formats.

Data cleaning. To guarantee the quality of the dataset, we engaged professional annotators to conduct a preliminary review of the downloaded books. The aim of the review is to filter out any content that deviated from our research standards, such as files with extensive watermarks and non-Chinese historical documents. This process resulted in the exclusion of about 10% of the data. Subsequently, we utilized automatic scripts to convert the scanned books into image formats (JPEG) and numbered them sequentially. After these two steps, we obtained over three million high-quality images of Chinese historical documents.

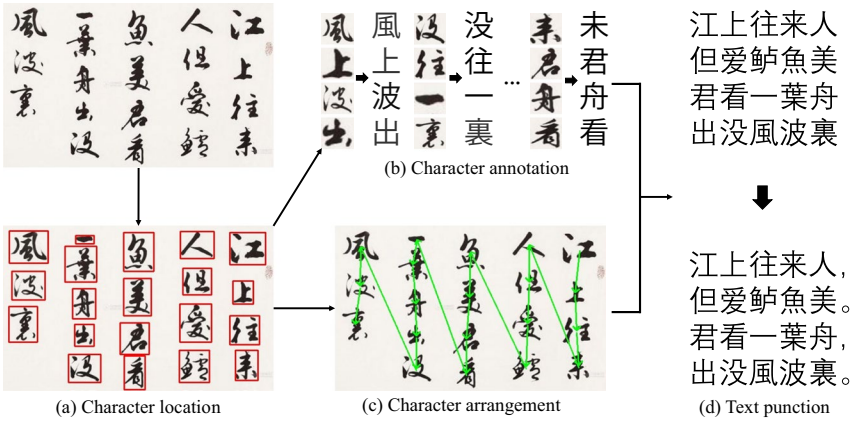


Fig. 2 The pipeline of data annotation.

Data annotation. *Character location.* In this step, our goal is to obtain the position of all the characters on the historical document images. Referring to previous works[4,7], we employ rectangular boxes that enclose the characters to indicate their position. Given the number of our images (over 3 million), it would be impractical to manually annotate all the rectangular boxes, as it would consume an enormous amount of human labor.

Therefore, we used a deep-learning based object detection algorithm as an automatic character position annotator, which treats all the characters as detection targets. To develop an accurate character-localization model, high-quality data for model training is required. To this end, we proposed a strategy to efficiently construct the training set: selecting one image sample from each book and manually annotating the character positions on these images. With this strategy, we only need to manually annotate 40,281 images, resulting in a 98.7% reduction in annotation compared to full annotation. This strategy is based on such a prior knowledge: The key to character location in Chinese historical documents lies in accurately distinguishing the foreground characters from the complex background. Furthermore, the stylistic attributes of the foreground characters and background are highly consistent with images from the same book. This insight suggests that a character location model capable of accurately localizing the characters in a sample image from a book can effectively handle all images in the entire book.

To efficiently annotate these 40,281 images, we adopted a hybrid approach that combines model-generated pseudo-annotations with manual refinement. Initially, the open-source datasets MTHv2[5] and M5HisDoc[7] were used to train the object detection model. For the choice of character localizers, we compared mainstream detectors such as Faster R-CNN[9], YOLOX[10] and YOLOv7[11]. Experimental results indicate that YOLOv7 achieves superior training and inference efficiency with comparable performance. To enhance efficiency, we chose YOLOv7 as the base model for the character localizer. Subsequently, the model was employed to generate preliminary annotations for the character positions of these images. Thereafter, we invited professional annotators to refine these annotations, including filling in omissions, removing redundancies, and correcting inaccuracies. The annotation refinement process was facilitated by Labelme (https://github.com/labelmeai/labelme), an open-source annotation tool. This hybrid annotating approach has further reduced human labor.

Through this method, we obtained the accurate character locations of 40,281 images. Combining this fine-tuned data with the open-source data as the training set, we trained the YOLOv7 model as a highly accurate character location annotator. Ultimately, the annotator was applied to our collection of over 3 million images, yielding accurate character location annotations. The character position obtained in this step totals over 1 billion.

Character annotation. The next crucial step is dedicated to the recognition of each character on the images, i.e., to represent each character with a computer-readable code. In the previous step, we obtained the location of over 1 billion characters, thus over 1 billion characters are required to be recognized. For such a large amount of data, manual annotation would be extremely labor-intensive and impractical.

Therefore, we adopted an automatic annotating method based on the deep-learning model, specifically employing a classifier model to automatically allocate the character images into their respective character categories. For character annotator, we evaluate both convolutional networks[12] and Vision Transformer (ViT)[13] architectures, with results showing that ViT achieves better performance. Therefore, to achieve higher-quality annotations, we selected ViT as the architecture for training the character annotator. The total number of categories set to 31,524, which encompasses the 27,533 characters in the national standard (GB 18030-2000) as well as other character categories in the open source dataset[3,5,7].

To develop a high-performance character annotator, we implemented the following three strategies:

- Self-supervised pre-training: Leveraging the precise annotations of character position obtained from the previous step, we cropped out character images from document images to form a self-supervised learning dataset. Utilizing this dataset, we pre-trained our model using the MAE[14] framework, a self-supervised learning method that has been proven to significantly enhance the model performance.

- Data synthesis: We employed MTHv2[5], CASIA-AHCDB[3], and M5HisDoc[7] as the foundation of the training set. However, due to the limited scale of these datasets, we integrated data synthetic techniques to augment our training set. From a font website (https://www.foundertype.com/), we selected 320 TrueType font (TTF) files that are similar in style to the characters on Chinese historical documents. The character images rendered using these font files were used to supplement the training set. Additionally, since the categories are not covered by the open source dataset and the TTF data, we used FontDiffuser[15], an advanced deep learning-based font generation approach for data completion. Specifically, utilizing the character images of existing categories within the TTF files as training data, we trained the FontDiffuser model to generate character images for each of the absent character categories. Furthermore, to enhance the authenticity of synthetic data, we randomly replaced the white background with actual historical document backgrounds.
- Category averaging sampling: In light of the long-tail distribution of character categories within open-source datasets, directly training models on them could lead to bias towards common categories and neglect of rare categories. To mitigate this, we adopted a class-balanced sampling strategy, randomly sampling the dataset during the training process to ensure that each category is given equal consideration, thereby improving the model's comprehensive recognition capabilities.

Through these strategies, we successfully developed an automatic character annotator. Subsequently, we fed all the cropped character images into this annotator, accomplishing the character annotation step with precision and efficiency.

Algorithm 1. Segmentation of Text Blocks.

Input: Image $\boldsymbol{I}$, Boxes of characters $\boldsymbol{B}$
Output: Coordinates of the horizontal segmentation line $\boldsymbol{L}$

```
I ← Binarization(I);
// Mask all the regions of character
for B[i] in B do
    I ← Mask character(I, B[i]);
// Horizontal projection
for i in range(I.height) do
    L[i] ← Sum(I[i, :]);
// Search for the segment lines
Thres ← max(L) × 0.3 ;
for L[i] in L do
    if L[i] < Thres then
        Delete L[i];
return L
```

Character arrangement. The objective of this stage is to arrange the characters within the images in the correct reading order. Analysis of Chinese historical documents reveals that the arrangement of characters follows certain patterns. For instance, as shown in Fig. 3(a), the characters are typically segregated into distinct blocks, and within each block, they are organized in a top-to-bottom and right-to-left sequence. Leveraging these patterns, we designed a character arrangement algorithm based on heuristic rules, which consists of three primary steps, as illustrated in Fig. 3:

- Segmentation of text blocks: In Chinese historical documents, the segmentation of text blocks is typically indicated by some black lines, as illustrated in Fig. 3(a). This feature provides a visual clue for the automatic segmentation of text blocks. Based on this feature, we employed image processing techniques to identify and extract these segment lines. Firstly, the image binarization technique from the OpenCV library was applied to the image. Secondly, a masking operation was conducted to set all pixels within the character regions to white. The purpose of this strategy is to reduce the potential interference of characters with the detection of segment lines, as the strokes of the characters might be pixel-wise similar to the segment lines and could thus affect the performance of the algorithm. Finally, the horizontal projection technique was engaged to ascertain the segment lines. This technique is implemented by calculating the total number of black pixels across each row of the image. By analyzing the distribution of the projection results, we can identify the locations of peaks, which are considered to be the segment lines between text blocks. The specific algorithmic process is demonstrated in Algorithm 1.
- Character arrangement within text blocks: Within each text block, characters are first arranged in the order from top to bottom, forming columns known as text lines. These text lines are then ordered from right to left. Notably, in Chinese historical documents, a special format sometimes occurs where two smaller text lines are below a larger text line, referred to as "double-column annotation". These smaller text lines typically serve to explain the larger text line above them. In the situation of "double-column annotation", the reading order should start with the larger text line, followed by the smaller text lines in a right-to-left order. Based on these

Algorithm 2. Sort Characters in Text Blocks.

```
Input: Boxes of characters B
Output: Sorted boxes of characters C
// Aggregate stage 1
G ← [] ;
for B[i] in B do
    if B[i] not in G then
        L ← [B[i]] ;
        for B[j] in B do
            x1, y1, x2, y2 ← B[j]; xu, yu, xd, yd ← L.bounding box.coordinates ;
            if (|x1-xu| < 10 or |x2-xd| < 10) and (|y1-yu| < 10 or |y2-yd| < 10) then
                L.append(B[j]);
        G.append(L);
// Aggregate stage 2
for G[i] in G do
    for G[j] in G do
        x1, y1, x2, y2 ← G[j].bounding box.coordinates; xu, yu, xd, yd ← G[j].bounding box.coordinates ;
        if (max((x1-xu)/(x2-x1), (x1-xd)/(xd-xu)) > 0.4) and ((y2-yu) < 300 or (yd-y1) < 300) then
            G[i].extend(G[j]);
            Delete G[j];
// Sort characters
for G[i] in G do
    Sort boxes in G[i] according to box[1] from low to high ;
    L1 ← []; L2 ← []; L3 ← [];
    xu, yu, xd, yd ← G[j].bounding box.coordinates; w ← xd - xu; x0 ← (xu + xd) / 2 ;
    for B[i] in G[i] do
        x1, y1, x2, y2 ← B[j]; x ← x2 - x1; ;
        if (x2-x1)/w > 0.6 or ((x2 - x1)/w > 0.45 and x1 < x0 and x2 > x0 and 0.4 < (x2-x0)/(x0-x1) < 2.5) then
            L1.append(B[i]);
        else
            if (x2 - x0) > (x0 - x1) then
                L2.append(B[i]);
            else
                L3.append(B[i]);
    C.extend(L1); C.extend(L2); C.extend(L3);
return C
```

reading order rules, we developed a heuristic algorithm, as depicted in Algorithm 2, which centralizes the idea of aggregating characters into horizontal regions and detecting the presence of "double-column annotation" within these regions. If such a situation is absent, the region is treated as an independent text line; otherwise, the characters within the "double-column annotation" will be further distinguished. After aggregating all the characters into separate text lines, the algorithm sorts the characters according to the top-to-bottom (within text line), left-to-right (between text lines) rules.

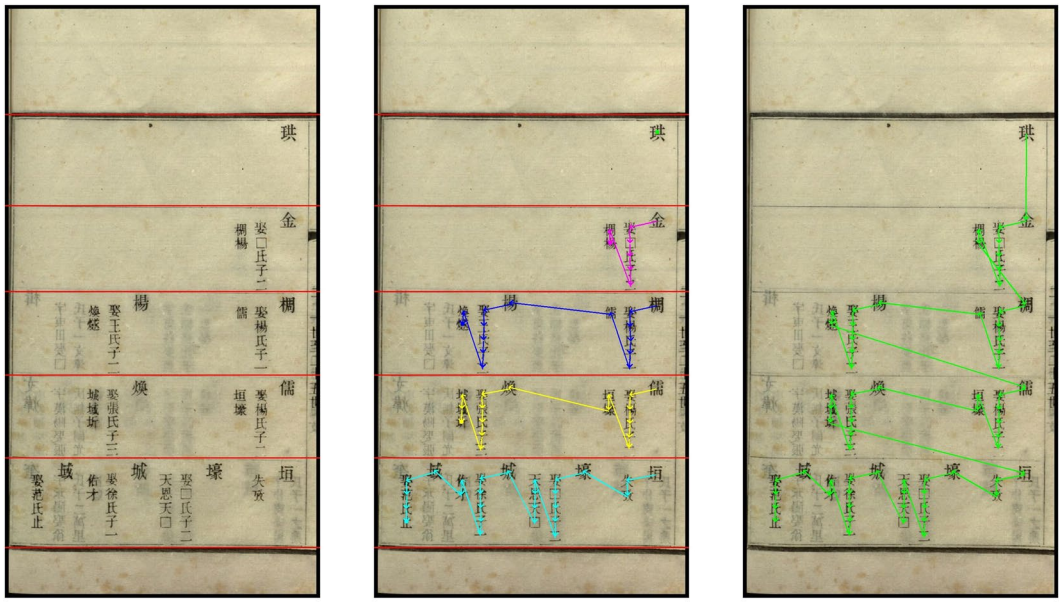
- Character arrangement between text blocks: The ordering between text blocks follows the basic principle of top-to-bottom arrangement. Therefore, we sort the text blocks based on the vertical coordinates (y-coordinates), from small to large.

By applying the aforementioned algorithms to all document images, the character arrangement step is completed efficiently.

Text punctuation. After successfully extracting individual text sequences from each image, it is crucial to punctuate these sequences to align with standard reading conventions. Given the extensive number of sequences to process, we employed an automatic punctuation method based on deep learning.

To develop an efficient automatic punctuation system, we utilized an ancient text corpus (https://github.com/garychowcmu/daizhigev20) as our training dataset. This corpus encompasses a rich collection of punctuated paragraphs sourced from historical documents. The model we employed was Transformer[16], which is widely recognized for its outstanding performance in handling sequential data.

The specific implementation steps are as follows: Initially, we preprocessed the selected ancient text corpus by removing all existing punctuation marks, yielding a set of unpunctuated text sequences for model input. Concurrently, the originally punctuated text served as the model's training target. Through this approach, the model can learn how to predict and add appropriate punctuation marks based on the text content.



(a) (b) (c)

Fig. 3 The visualization of character arrangement. (**a**) Segmentation of text blocks. (**b**) Sort characters within text blocks. (**c**) Sort characters between text blocks. Zoom in for a better view.

Character location ↑	Character annotation ↑	Character arrangement ↓	Text punction ↑
99.75	98.87	1.86	90.37

Table 2. Results of data validation.

Training set	MTHv2	M5HisDoc	HisDoc1B
MTHv2	**96.78**	80.81	92.11
M5HisDoc	95.34	**96.22**	98.70
HisDoc1B (Ours)	95.95	94.10	**99.45**

Table 3. Results of character detection.

To enhance the accuracy of automatic punctuation, we adopted a strategy of chunked input. Specifically, each unpunctuated text sequence was segmented into 30-character chunks, which were subsequently fed into the punctuation model incrementally. The punctuation outcomes of these chunks were then concatenated sequentially to reconstitute the fully punctuated text.

Data validation. To evaluate the accuracy and reliability of our semi-automatic annotation method, we conduct the following data validation. Firstly, we randomly selected 100 images from the annotated dataset. Secondly, we invited domain experts to independently and meticulously annotate these images. These annotations served as the ground truth for comparison with the results produced by our semi-automatic annotation system. Thirdly, we performed a thorough comparative analysis, the comparison results are presented in Table 2. The metrics of character location, character annotation, character arrangement, and text punction are F1-score in 0.7 IoU, top1-accuracy, ARDs, and F1-score[17], respectively. The results indicated that our semi-automatic annotation system performed well on the metrics, demonstrating the high quality of our dataset.

Data Records

The HisDoc1B[8] dataset consists of two main folders, one dedicated to storing historical documents in the form of e-books and the other containing their respective annotation files. The e-books are archived in PDF or DjVu formats. The annotation files are stored in JavaScript Object Notation (JSON) format, aligning with each corresponding e-book. The dataset employs a unique book identifier (ID) to pair e-books with their annotation files, ensuring precise alignment.

Each annotation file contains various annotation entries corresponding to the individual pages of the book. Each entry includes a detailed record of the following three key parts:

Training set	MTHv2	M5HisDoc	CASIA-AHCDB	HisDoc1B
MTHv2	75.27	34.59	55.87	26.89
M5HisDoc	82.27	66.49	79.35	52.28
CASIA-AHCDB	71.92	38.32	**95.21**	28.84
HisDoc1B (Ours)	**85.83**	**79.01**	92.13	**94.03**

Table 4. Results of character recognition.

Training set	Before finetune	Finetuned
DaiZhiGe	31.78	38.77
HisDoc1B (Ours)	**34.05**	**40.76**

Table 5. Results of incremental pre-training of language model.

- Character position: an array of rectangular boxes, sequenced in the reading order. Each rectangular box indicates the position of a character within the document image. The coordinates are represented in the format of 'x1, y1, x2, y2', where 'x1, y1' specifies the coordinates of the top-left corner of the rectangle, 'x2, y2' denotes the bottom-right corner.
- Character content: a sequence of Unicode-encoded symbols, with the order corresponding to the character positions.
- Punctuated text: the complete text sequence that includes punctuation marks.

HisDoc1B focuses mainly on commonly used font types in Chinese historical documents. However, some special fonts, such as Oracle[18,19] and Inscriptions on bronze, are missing from our dataset. This is a direction for future work.

Technical Validation

In order to access the utility of the proposed HisDoc1B dataset for Chinese historical document recognition and analysis, we conducted three technical validation tasks: character detection, character recognition, and incremental pre-training of the language model.

Character detection. The aim of this validation task is to validate the application value of the character location annotations within our HisDoc1B. To this end, we devised and executed a cross-validation experiment for character detection: Employing a consistent model architecture, we conducted training across diverse datasets of Chinese historical documents and subsequently assessed their performance on each dataset. The comparative datasets included MTHv2[5] and M5HisDoc[7]. Given the extensive data scale of HisDoc1B, we implemented a sampling strategy, randomly selecting two images per book to construct the training and test sets, one for each subset, to improve the efficiency of the experiment. We chose the YOLOv7[11] model and trained it for 100 epochs using the above training set, respectively. Subsequently, we validated the model performance on the test set of each dataset. The validation metric is the F1-score at 0.7 Intersection over the Union (IoU) threshold. The results are shown in Table 3. Based on the results, we observed that the model trained on the HisDoc1B dataset demonstrates the most robust generalization performance on the non-homologous dataset. This is attributed to the diversity of styles and large-scale of our dataset, which enhances the model's generalization capabilities. It also reflects the accuracy of the annotations. Conversely, models trained on other datasets perform poorly on the test set of HisDoc1B, mainly due to the extensive range of character categories in our dataset, which introduces new challenges to the field. The model trained on M5HisDoc exhibits the second-best generalization, which is attributed to its inclusion of multiple styles.

Character recognition. To evaluate the practical utility of the character annotations within the HisDoc1B dataset, we designed and conducted a character recognition cross-validation experiment similar to the previous experiment. The datasets we compare with HisDoc1B include MTHv2[5], CASIA-AHCDB[3], and M5HisDoc[7]. For HisDoc1B, we randomly sampled 10 images per category to construct the test set, the remaining images constituted the training set. In order to speed up the training, for categories with more than 300 samples in each training set, we downsampled the samples to 300. Considering the character categories vary across datasets, the zero-shot recognition model Hiercode[20] was employed, to give the model the ability to generalize beyond the training categories. The model was trained for 90 epochs using the above training sets respectively. Subsequently, we validated the model performance on the test set of each dataset. The validation metric is the macro accuracy[21]. The results are shown in Table 4. Based on the results, we observed that the model trained on the HisDoc1B dataset exhibits the best generalization performance on the non-homologous dataset. This is attributed to the diversity of styles and large-scale of our dataset, which enhances the model's generalization capabilities. It also reflects the accuracy of the annotations. Conversely, models trained on other datasets perform poorly on the test set of HisDoc1B, primarily due to the diverse range of character categories within our dataset, introducing new challenges to the field. The model trained on M5HisDoc exhibits the second-best generalization, which is attributed to its inclusion of multiple styles.

Method	Precision↑	Recall↑	F1-score↑
Faster R-CNN[9]	**99.30**	98.18	98.74
YOLOX[10]	99.09	98.87	98.98
YOLOv7[11]	99.07	**98.91**	**98.99**

Table 6. Results of baseline experiments on character detection.

Method	Top-1 acc↑	Macro acc↑
ResNet50[12]	95.06	91.86
ViT[13]	95.13	91.89
HDE[24]	95.42	93.76
RIE[25]	95.30	93.64
HierCode[20]	**95.89**	**94.03**

Table 7. Results of baseline experiments on character recognition.

Backbone	Parameters	FLOPs	Trained epochs		
			30	60	90
ResNet18	11.2M	1.8G	91.28	92.24	92.66
ResNet50	25.6M	3.8G	91.94	93.64	93.76
ResNet101	44.6M	7.6G	93.30	94.02	94.31

Table 8. Results of exploration experiments.

Incremental pre-training of language model. To validate the utility and quality of the book-level text provided in the HisDoc1B dataset, we conducted an experiment of large language model incremental pre-training. For comparative analysis, the DaiZhiGe (https://github.com/garychowcmu/daizhigev20) data was utilized, an open-source corpus dataset of Chinese ancient texts. The language model selected for this experiment is Qwen1.5-1.5B[22], which was trained for 1 epoch using the above training set, respectively. Subsequently, we employed ACLUE[23] as the dataset for evaluating the performance of pre-trained language models, it is a comprehensive benchmark of Ancient Chinese understanding. The training set of ACLUE was utilized for instruction fine-tuning of the pre-trained model, respectively. The scores on the test set of ACLUE are compared after the fine-tuning is completed. The results are detailed in Table 5, showing that the large language model pre-trained on the HisDoc1B dataset outperforms the model pre-trained on the DaiZhiGe corpus, both before and after fine-tuning. This superior performance can be attributed to the large-scale and diversity of the HisDoc1B dataset, which provides the model with rich knowledge of ancient Chinese culture and linguistic priors. Despite the DaiZhiGe corpus being larger than HisDoc1B, it contains some noise that negatively affects model performance. In contrast, the data quality in HisDoc1B is maintained at a high standard through our meticulous data construction pipeline. These findings not only confirm the accuracy and quality of annotations in the HisDoc1B dataset but also demonstrate its effectiveness in advancing language modeling for the understanding of ancient Chinese texts.

Baseline experiments. The baseline experiment for character detection follows the same data split as in Section Character detection. We conduct baseline experiments using a typical two-stage method, Faster R-CNN[9], as well as two one-stage methods, YOLOX[10] and YOLOv7[11]. The models are trained for 12 epochs with a batch size of 8. We use stochastic gradient descent (SGD) for optimization, with an initial learning rate of 0.001, which decays by a factor of 0.1 at the 8th and 11th epochs. Input images are resized and padded to a resolution of 1024×1024 while maintaining the aspect ratio. The evaluation metrics include precision, recall, and F1 score at a 0.7 IoU threshold. The results are shown in Table 6. We can find that various methods demonstrate comparable performance.

The baseline experiment for character recognition uses the same data split as in Section Character recognition. We conduct baseline experiments on two types of methods: (1) conventional image classifiers, including ResNet50[12] and ViT[13], and (2) zero-shot character recognition models, including HDE[24], RIE[25], and HierCode[20]. The models are optimized using AdamW[26] with a base learning rate of 1e-3, which decays to 1e-6 following a cosine annealing schedule. During the first 5 epochs, the learning rate linearly warms up from 1e-4 to 1e-3. We train the models for 90 epochs with a batch size of 1,024. All models use RandAugment[27] for data augmentation. Input images are resized and padded to a resolution of 96×96 while maintaining the aspect ratio. Evaluation metrics include top-1 accuracy and macro accuracy[21]. Based on the results in Table 7, we can gain the following findings: (1) ViT slightly outperforms ResNet50 among conventional image classifiers, and (2) zero-shot character recognition models outperform conventional classifiers due to variations in character categories between training and test sets.

Exploration of the challenges associated with the size of the dataset. The large-scale of our dataset introduces a wide variety of styles and character categories, presenting challenges discussed in Sections Character detection and Character recognition. To further explore the challenges posed by the size associated

with training resources, we conducted the following experiments. Following the setup in Section Character recognition, we trained the HierCode[20] model with backbones of varying parameter sizes and training durations. The results are shown in Table 8, where a clear performance improvement is observed with increased model parameters and training time. This demonstrates that the scale of our dataset necessitates using models with a larger number of parameters and longer training durations to achieve optimal results, which poses a significant challenge to training resources.

Usage Notes

The HisDoc1B dataset consists of two main folders. One folder contains the e-books of historical documents in the format of PDF or DjVu files. The other folder contains the corresponding annotation files for each book in JSON format. We have provided a Python script that facilitates the conversion of e-books into image format (JPEG), reads annotation data from the JSON files, and organizes them within the appropriate data structures.

Code availability

The inference codes to generate the dataset will be made available in the following GitHub repository (https://github.com/SCUT-DLVCLab/HisDoc1B).

Received: 16 July 2024; Accepted: 17 January 2025;
Published online: 29 January 2025

References

1. Peng, D., Jin, L., Liu, Y., Luo, C. & Lai, S. PageNet: Towards end-to-end weakly supervised page-level handwritten Chinese text recognition. *International Journal of Computer Vision* **130**, 2623–2645, https://doi.org/10.1007/s11263-022-01654-0 (2022).
2. Liu, C. *et al.* A robust and efficient algorithm for Chinese historical document analysis and recognition. *National Science Review* **10**, nwad115, https://doi.org/10.1093/nsr/nwad115 (2023).
3. Xu, Y. *et al.* CASIA-AHCDB: A large-scale Chinese ancient handwritten characters database. In *International Conference on Document Analysis and Recognition (ICDAR)*, 793–798, https://doi.org/10.1109/ICDAR.2019.00132 (2019).
4. Yang, H. *et al.* Dense and tight detection of Chinese characters in historical documents: Datasets and a recognition guided detector. *IEEE Access* **6**, 30174–30183, https://doi.org/10.1109/ACCESS.2018.2840218 (2018).
5. Ma, W. *et al.* Joint layout analysis, character detection and recognition for historical document digitization. In *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 31–36, https://doi.org/10.1109/ICFHR2020.2020.00017 (2020).
6. Saini, R., Dobson, D., Morrey, J., Liwicki, M. & Liwicki, F. S. ICDAR 2019 historical document reading challenge on large structured Chinese family records. In *International Conference on Document Analysis and Recognition (ICDAR)*, 1499–1504, https://doi.org/10.1109/ICDAR.2019.00241 (2019).
7. Shi, Y. *et al.* M5HisDoc: A large-scale multi-style Chinese historical document analysis benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, vol. 36 (2023).
8. Shi, Y., Peng, D., Zhang, Y., Cao, J. & Jin, L. A large-scale dataset for Chinese historical document recognition and analysis. *figshare* https://doi.org/10.6084/m9.figshare.26272375 (2024).
9. Ren, S., He, K., Ross, G. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)* (2015).
10. Ge, Z., Liu, S., Wang, F., Li, Z. & Sun, J. YOLOX: Exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430* https://doi.org/10.48550/arXiv.2107.08430 (2021).
11. Wang, C.-Y., Bochkovskiy, A. & Liao, H.-Y. M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7464–7475, https://doi.org/10.1109/CVPR52729.2023.00721 (2023).
12. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).
13. Dosovitskiy, A. *et al.* An image is worth 16 × 16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)* (2020).
14. He, K. *et al.* Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009, https://doi.org/10.1109/CVPR52688.2022.01553 (2022).
15. Yang, Z. *et al.* FontDiffuser: One-shot font generation via denoising diffusion with multi-scale content aggregation and style contrastive learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, vol. 38, 6603–6611, https://doi.org/10.1609/aaai.v38i7.28482 (2024).
16. Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30 (2017).
17. Zhang, Z., Liu, J., Chi, L. & Chen, X. Word-level BERT-CNN-RNN model for Chinese punctuation restoration. In *International Conference on Computer and Communications (ICCC)*, 1629–1633, https://doi.org/10.1109/ICCC51575.2020.9344889 (2020).
18. Wang, M. & Deng, W. A dataset of oracle characters for benchmarking machine learning algorithms. *Scientific Data* **11**, 87, https://doi.org/10.1038/s41597-024-02933-w (2024).
19. Wang, P. *et al.* An open dataset for oracle bone character recognition and decipherment. *Scientific Data* **11**, 976, https://doi.org/10.1038/s41597-024-03807-x (2024).
20. Zhang, Y. *et al.* HierCode: A lightweight hierarchical codebook for zero-shot Chinese text recognition. *Pattern Recognition* **158**, 110963, https://doi.org/10.1016/j.patcog.2024.110963 (2025).
21. Zhang, Y., Kang, B., Hooi, B., Yan, S. & Feng, J. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* https://doi.org/10.1109/TPAMI.2023.3268118 (2023).
22. Bai, J. *et al.* Qwen technical report. *arXiv preprint arXiv:2309.16609* https://doi.org/10.48550/arXiv.2309.16609 (2023).
23. Zhang, Y. & Li, H. Can large language model comprehend ancient Chinese? A preliminary test on ACLUE. In Proceedings of the Ancient Language Processing Workshop (ACLW) (2023).
24. Cao, Z., Lu, J., Cui, S. & Zhang, C. Zero-shot handwritten Chinese character recognition with hierarchical decomposition embedding. *Pattern Recognition* **107**, 107488, https://doi.org/10.1016/j.patcog.2020.107488 (2020).
25. Luo, G.-F. *et al.* Self-information of radicals: A new clue for zero-shot Chinese character recognition. *Pattern Recognition* **140**, 109598, https://doi.org/10.1016/j.patcog.2023.109598 (2023).
26. Ilya, L. & Frank, H. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR) (2019).
27. Cubuk, E. D., Zoph, B., Shlens, J. & Le, Q. V. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 702–703, https://doi.org/10.1109/CVPRW50498.2020.00359 (2020).

Acknowledgements
This research is supported in part by National Natural Science Foundation of China (Grant No.: 62441604, 62476093).

Author contributions
Yongxin Shi built the dataset, conceived the technical validation experiment(s), and write this paper. Yuyi Zhang and Jiahuan Cao conducted the technical validation experiment(s). Dezhi Peng reviewed and revised the manuscript. Lianwen Jin reviewed and revised the manuscript and supervised this study.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.J.

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

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

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025