# M$^5$HisDoc: A Large-scale Multi-style Chinese Historical Document Analysis Benchmark Supplementary Material

**Yongxin Shi, Chongyu Liu, Dezhi Peng, Cheng Jian, Jiarong Huang, Lianwen Jin***
South China University of Technology
yongxin_shi@foxmail.com, liuchongyu1996@gmail.com
pengdzscut@foxmail.com, eechengjian@mail.scut.edu.cn,
jiarong_huang@outlook.com, eelwjin@scut.edu.cn

## 1 Datasheets for M$^5$HisDoc

### 1.1 Motivation

**For what purpose was the dataset created?**
The purpose of creating M$^5$HisDoc dataset is to advance the development of historical document analysis in real-world scenarios. In the field of historical document analysis, some benchmarks [1, 2, 3] have been established and some methods [4, 5, 6] have reported promising performance. However, these methods fail to adequately address the challenges posed by real-world scenarios, including diverse page layouts, poor image quality, multiple font styles, and severe distortion, which are rarely considered in the existing benchmarks. To fill this gap, we introduce M$^5$HisDoc, a comprehensive and intricate benchmark for Chinese historical document analysis. M$^5$HisDoc encompasses a wide range of document styles, including diverse layouts, document types, calligraphy styles, backgrounds, and associated challenges. In contrast to existing benchmarks, M$^5$HisDoc offers a more thorough representation of the aforementioned issues. Consequently, we firmly believe that this complex dataset will significantly promote the development of historical document analysis in real-world scenarios.

**Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
The M$^5$HisDoc dataset is created by the Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology.

### 1.2 Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
The M$^5$HisDoc dataset comprises 8,000 images, accompanied by their respective annotation files. These images represent scanned or photographed historical documents and are stored in the Joint Photographic Experts Group (JPEG) format. The annotations, are stored in plain text (TXT) format, including bounding boxes for text lines/characters within the images, along with the corresponding text content. Furthermore, the texts are arranged in the correct reading order.

**How many instances are there in total (of each type, if appropriate)?**
The M$^5$HisDoc dataset comprises a collection of 8,000 images, with 403,824 text lines and 4,367,361 characters in 16,151 categories.

*Corresponding author.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
The M$^5$HisDoc dataset contains all possible instances.

**What data does each instance consist of?**
Each instance in the M$^5$HisDoc consists of an image along with corresponding annotations. These annotations include bounding boxes for text lines/characters within the images, along with the text content and the reading order between the texts.

**Is there a label or target associated with each instance?**
Yes. The label contains bounding boxes for text lines/characters, text content, and the reading order between the texts.

**Is any information missing from individual instances?**
NO. There is no missing information from individual instances in the M$^5$HisDoc dataset.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
There is no relationship between individual instances.

**Are there recommended data splits (e.g., training, development/validation, testing)?**
The two subsets of M$^5$HisDoc (M$^5$HisDoc-R and M$^5$HisDoc-H) are both divided into training, validation, and testing sets in a ratio of 2:1:1. The specific data splits can be found at `https://github.com/HCIILAB/M5HisDoc`.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**
The M$^5$HisDoc dataset is self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individuals' non-public communications)?**
The M$^5$HisDoc dataset comprises historical document images along with their corresponding manual annotations and does not include any confidential data.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**
The dataset does not include any data that could be considered offensive, insulting, threatening, or potentially causing anxiety.

### 1.3 Collection Process

**How was the data associated with each instance acquired?**
The collection process is described in Sec. 3.1 of the main paper. The images we obtained from the Internet are sourced from some open-copyright (under CC license) websites, such as Harvard-Yenching Library (`https://library.harvard.edu/`), National Archives of Japan (`https://www.digital.archives.go.jp/`).

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?**
The annotation process is described in Sec. 3.2 of the main paper. And we specifically developed a web-based platform for data annotation, as illustrated in Fig. 1. With this platform, annotators are able to label text boxes, text content, and correct their reading order.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
The sampling strategy employed in this study involved a manual selection process, where representative samples were carefully chosen from the larger set.

**Over what timeframe was the data collected?**
The data collection process spanned approximately 5 months.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
Our data collection process consists of three main sources. Firstly, we carefully select 300 images


Figure 1: The web-based platform designed for data annotation.

from the training set of MTHv2 [1] and 700 images from SCUT-CAB [7]. Secondly, we gather tens of thousands of scanned images from electronic ancient books available on the Internet and manually curate 2,799 historical document images taken from some representative books. Thirdly, we conduct realistic photo shoots to simulate photographing situations. By selecting four physical Chinese ancient books, we capture 201 images using a scanner, considering various scanning angles and lighting conditions.

### 1.4 Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
To replicate real-world conditions for historical document analysis applications, we incorporate image rotation, distortion, and resolution reduction into M$^5$HisDoc-R subset to form a new challenging subset named M$^5$HisDoc-H.

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**
Yes, the "raw" data for M$^5$HisDoc-H corresponds to M$^5$HisDoc-R.

### 1.5 Uses

**Has the dataset been used for any tasks already?**
No.

**What (other) tasks could the dataset be used for?**
The M$^5$HisDoc dataset can be used for various tasks of Chinese historical document analysis, including text line/character detection, recognition, and reading order prediction.

### 1.6 Distribution

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**
The M$^5$HisDoc dataset is available at `https://github.com/HCIILAB/M5HisDoc`.

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
All authors bear all responsibility for the M$^5$HisDoc dataset in case of violation of rights, etc. The

M$^5$HisDoc dataset should be used under Creative Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License for non-commercial research purposes.

### 1.7 Maintenance

**Who will be supporting/hosting/maintaining the dataset?**
The dataset will be maintained by the Deep Learning and Vision Computing Lab (DLVC-Lab) of South China University of Technology.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
Contact can be made via email at eelwjin@scut.edu.cn.

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
If substantial errors are raised by dataset users, we will update the dataset accordingly. The updated version of the dataset will be made available through the dataset release link.

**Will older versions of the dataset continue to be supported/hosted/maintained?**
Yes, with each update, the older versions will remain accessible through their original links.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**
If other researchers or individuals are interested in extending, augmenting, building on, or contributing to the dataset, they should contact us via email, clearly articulating their intentions and requesting our consent prior to any further actions.

## 2 Heuristic algorithm for reading order prediction

The heuristic algorithm employed in Sec. 4.5 of the main paper arranges text lines in a left-to-right, top-to-bottom order, primarily based on the conventional reading order observed in Chinese historical documents. A detailed description of the procedure is provided in Algorithm 1.

**Algorithm 1** Sort Bounding Boxes

**Input:** $boxes$: a list of bounding boxes to be sorted
**Output:** $sorted\_boxes$: the sorted list of bounding boxes

$sorted\_boxes \leftarrow copy(boxes)$
$length \leftarrow length(sorted\_boxes)$
**for** $i \leftarrow 1$ to $length - 1$ **do**
    **for** $j \leftarrow 1$ to $length - i$ **do**
        $box1 \leftarrow sorted\_boxes[j]$
        $box2 \leftarrow sorted\_boxes[j+1]$
        **if** $box1.center_x < box2.center_x$ **then**
            $sorted\_boxes[j] \leftrightarrow sorted\_boxes[j+1]$
        **else if** $box1.center_x = box2.center_x$ **and** $box1.center_y > box2.center_y$ **then**
            $sorted\_boxes[j] \leftrightarrow sorted\_boxes[j+1]$

## 3 Additional information of the dataset

### 3.1 Number of text lines/characters

Table 1: Comparison with existing Chinese historical document datasets. * indicates that we only consider the training set due to only the training set of ICDAR 2019 HDRC-CHINESE is available.

| Dataset | Images | Text lines | Characters |
|---|---|---|---|
| MTHv1 [8] | 1,500 | 40,656 | 521,375 |
| MTHv2 [1] | 3,199 | 105,578 | 1,081,663 |
| IC19 HDRC [2]* | 11,715 | 417,489 | 2,482,992 |
| CASIA-AHCDB [3] | - | - | 2,276,740 |
| M$^5$HisDoc (Ours) | 8,000 | 403,824 | 4,367,361 |

As shown in Table 1, M$^5$HisDoc contains 4,367,361 characters, and represents the largest collection in the field of historical document analysis.

### 3.2 Comparison between pre-annotation and manual re-annotation

As mentioned in the Sec. 3.2 in the main paper, we employ models trained on other datasets (MTHv2 [1] and CASIA-AHCDB [3]) for pre-annotation. To enhance the understanding of the extent to which the existing OCR model contributes to the annotation process, we conduct this comparison. That is, i.e., we use manually-labeled labels as GT to evaluate the performance of the models used for pre-annotation. We employ the metrics of character detection and recognition to measure the gap between pre-labeled and manually labeled data. The results are shown in Table 2 and Table 3, respectively. We can see that the models provide about 70% accuracy in bounding box and category labeling of characters.

Table 2: Performance of the character detector in pre-labeling on the manually labeled data.

| IoU thres | Precision↑ | Recall↑ | F1-score↑ |
|---|---|---|---|
| 0.5 | 96.38 | 85.65 | 90.70 |
| 0.6 | 94.03 | 83.56 | 88.49 |
| 0.7 | 80.76 | 71.77 | 76.00 |

Table 3: Performance of the character recognizer in pre-labeling on the manually labeled data.

| Top-1 acc↑ | Top-5 acc↑ |
|---|---|
| 76.50 | 84.78 |

## 4 Additional Experiments

### 4.1 Text line/character detection/recognition and reading order prediction

**Setting** The experimental settings in this section align with Sec. 4.1~4.5 of the main paper. We evaluate the performance of different models on the validation set of M$^5$HisDoc.

**Results and analysis** The experimental results are presented in Table 4, 5, 6, 7 and 8, respectively. It can be observed that the experimental findings on the validation set are consistent with those on the test set.

Table 4: Results of text line detection, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R validation set.

| Type | Method | Venue | IoU thres | Precision↑ | Recall↑ | F1-score↑ | 1-NED↑ |
|---|---|---|---|---|---|---|---|
| Regression-based | Mask R-CNN [9] | CVPR'16 | 0.5 | 87.90/97.77 | 80.73/94.09 | 84.16/95.89 | 66.21/85.07 |
| | | | 0.6 | 87.40/97.52 | 80.28/93.85 | 83.69/95.65 | |
| | | | 0.7 | 86.24/96.75 | 79.20/93.11 | 82.57/94.90 | |
| | Cascade R-CNN [10] | CVPR'18 | 0.5 | 91.21/**98.27** | 82.82/94.22 | 86.81/**96.20** | 70.18/**85.71** |
| | | | 0.6 | 90.80/**98.00** | 82.44/**93.96** | 86.42/**95.94** | |
| | | | 0.7 | 89.90/**97.37** | 81.62/**93.36** | 85.56/**95.32** | |
| | OBD [11] | IJCV'21 | 0.5 | 94.56/97.05 | 82.44/91.60 | 88.08/94.25 | 72.45/82.92 |
| | | | 0.6 | 91.59/96.35 | 79.85/90.94 | 85.32/93.56 | |
| | | | 0.7 | 85.42/94.50 | 74.47/89.18 | 79.57/91.76 | |
| Segmentation-based | PSENet [12] | CVPR'19 | 0.5 | 78.92/93.93 | 87.62/**95.45** | 83.04/94.68 | 64.04/80.63 |
| | | | 0.6 | 74.93/90.10 | 83.20/91.56 | 78.85/90.83 | |
| | | | 0.7 | 64.69/78.39 | 71.82/79.66 | 68.07/79.02 | |
| | PAN [13] | ICCV'19 | 0.5 | 93.13/94.01 | 83.24/90.47 | 87.91/92.21 | 71.81/77.23 |
| | | | 0.6 | 89.15/87.60 | 79.68/84.29 | 84.15/85.91 | |
| | | | 0.7 | 77.01/68.72 | 68.83/66.13 | 72.69/67.40 | |
| | FCENet [14] | CVPR'21 | 0.5 | 87.31/90.86 | 85.59/87.92 | 86.44/89.37 | 67.76/72.58 |
| | | | 0.6 | 79.42/85.51 | 77.86/82.75 | 78.64/84.11 | |
| | | | 0.7 | 65.46/75.04 | 64.17/72.61 | 64.81/73.80 | |
| | DBNet++ [15] | T-PAMI'22 | 0.5 | 91.96/89.85 | **91.35**/87.97 | 91.65/88.90 | 75.36/72.03 |
| | | | 0.6 | 62.62/63.32 | 62.21/62.00 | 62.41/62.65 | |
| | | | 0.7 | 35.05/34.08 | 34.82/33.37 | 34.93/33.72 | |
| Connected component-based | TextSnake [16] | ECCV'18 | 0.5 | **95.54**/94.03 | 88.96/90.88 | **92.14**/92.43 | **77.64**/77.31 |
| | | | 0.6 | **93.75**/91.39 | **87.30**/88.33 | **90.41**/89.83 | |
| | | | 0.7 | **90.07**/87.85 | **83.87**/84.91 | **86.86**/86.35 | |

Table 5: Results of text line recognition, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R validation set. CR: correct rate, AR: accurate rate. * indicates the annotation of the character bounding box is used.

| Type | Method | Venue | CR↑ | AR↑ |
|---|---|---|---|---|
| CTC-based | CRNN [17] | T-PAMI'16 | 90.70/91.39 | 90.39/91.09 |
| | Ma et al. [1] | ICFHR'20 | **91.29**/92.07 | **90.93**/91.71 |
| | ZCTRN [18] | ICDAR'21 | 87.99/88.92 | 87.62/88.50 |
| Attention-based | ASTER [19] | T-PAMI'18 | 87.54/87.95 | 87.06/87.51 |
| | NRTR [20] | ICDAR'19 | 84.36/85.01 | 81.73/82.41 |
| | Robust Scanner [21] | ECCV'20 | 88.93/90.83 | 88.52/90.46 |
| Segmentation-based | Peng et al. [22] | TMM'22 | 88.80/89.19 | 88.57/88.99 |
| | Peng et al. [22]* | TMM'22 | 90.96/**92.63** | 90.71/**92.42** |

Table 6: Results of character detection, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R validation set. "Top-1 acc" is the metric of extracting the detected bounding box and feeding it into the recognizer. * indicates the accuracy of the missed and overchecked characters is set to 0, otherwise only the characters matching on GT are considered.

| Type | Method | Venue | IoU thres | Precision↑ | Recall↑ | F1-score↑ | Top-1 acc↑ | Top-1 acc↑* |
|---|---|---|---|---|---|---|---|---|
| Two-stage | Faster R-CNN [23] | NIPS'15 | 0.5 | 96.31/97.57 | **98.42**/98.44 | 97.36/98.01 | 94.40/94.73 | 89.54/91.02 |
| | | | 0.6 | 95.44/96.77 | **97.53**/97.63 | 96.48/97.20 | | |
| | | | 0.7 | 92.59/94.31 | **94.62**/95.16 | 93.59/94.73 | | |
| One-stage | YOLOv3 [24] | arXiv'18 | 0.5 | **98.99/98.98** | 95.54/96.30 | 97.23/97.62 | **94.80/95.07** | 89.70/90.65 |
| | | | 0.6 | **98.45/98.51** | 95.03/95.85 | **96.71**/97.16 | | |
| | | | 0.7 | **96.55/96.88** | 93.20/94.26 | **94.85/95.56** | | |
| | YOLOX [25] | CVPR'21 | 0.5 | 96.93/97.18 | 98.40/**99.18** | **97.66/98.17** | 94.52/94.85 | **90.20/91.45** |
| | | | 0.6 | 94.46/96.54 | 95.89/**98.53** | 95.17/**97.52** | | |
| | | | 0.7 | 85.44/94.57 | 86.74/**96.52** | 86.09/95.54 | | |

Table 7: Results of character recognition, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R validation set.

| Type | Method | Venue | Top-1 acc↑ | Top-5 acc↑ | Macro acc↑ |
|---|---|---|---|---|---|
| CNNs | ResNet50 [26] | CVPR'16 | 94.54/94.91 | 98.22/98.36 | 69.03/71.98 |
| | RegNet [27] | CVPR'20 | 94.71/94.99 | 98.27/98.35 | 68.91/71.29 |
| | ConvNeXt [28] | CVPR'22 | 94.67/95.01 | 98.27/98.36 | 70.01/**72.97** |
| Vision Transformer | ViT [29] | ICLR'21 | 94.24/94.63 | 98.07/98.21 | 64.91/68.10 |
| | SwinTransformer [30] | ICCV'21 | **94.77/95.03** | **98.32/98.42** | **70.34**/72.37 |

Table 8: Results of reading order prediction, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R validation set. ARD: Average Relative Distance.

| Type | Method | Venue | ARD↓ |
|---|---|---|---|
| Rule-based | Heuristic method (ours) | - | **4.53**/4.54 |
| | Augumented XY Cut [31] | CVPR'22 | 14.34/12.17 |
| Learning-based | LayoutReader [32](layout only) | EMNLP'21 | 7.05/**4.16** |

### 4.2 Reading order prediction on character-level

**Setting** We conduct experiments on M$^5$HisDoc for reading order prediction on character-level with both rule-based and learning-based methods. The rule-based method is Augmented XY Cut [31], we reverse the horizontal direction in this algorithm, as Augmented XY Cut is designed for modern documents, which are typically read from left to right. However, the reading order of Chinese historical documents is the opposite. For the learning-based method, we employ LayoutReader [32] for comparison. We utilize Average Relative Distance [32] (ARD) to evaluate the performance.

**Results and analysis** The experimental results are presented in Table 9. We can observe that the performance of the same method in character-level reading order prediction is significantly inferior compared to the text lines-level discussed in Sec. 4.5 of the main paper. This indicates that predicting the reading order at character-level presents a significant challenge, thus highlighting it as a promising research direction for future investigations.

Table 9: Results of reading order prediction on character-level, in the format of M$^5$HisDoc-H/M$^5$HisDoc-R test set. ARD: Average Relative Distance.

| Type | Method | Venue | ARD↓ |
|---|---|---|---|
| Rule-based | Augumented XY Cut [31] | CVPR'22 | 193.16/160.97 |
| Learning-based | LayoutReader [32](layout only) | EMNLP'21 | 51.72/37.97 |

### 4.3 Cross-validation between M$^5$HisDoc-R and M$^5$HisDoc-H

**Setting** We perform cross-validation between M$^5$HisDoc-R and M$^5$HisDoc-H using the Cascade R-CNN [10], CRNN [17], YOLOX [25], and ConvNeXt [28] models mentioned in Sec. 4.1~4.4 of the main paper.

Table 10: Cross-validation between the models trained on M$^5$HisDoc-R and M$^5$HisDoc-H. The evaluation metric for text line/character detection is F-score at 0.5 IoU, the metrics for text line/character recognition are AR and top-1 accuracy, respectively.

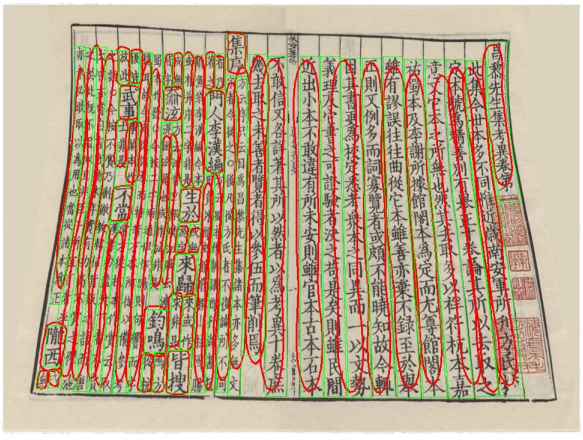
| Tasks | Text line det (F1-score) | Text line reg (AR) | Char det (F1-score) | Char reg (top-1 acc) |
|---|---|---|---|---|
| M$^5$HisDoc-R → M$^5$HisDoc-H | 81.89 | 87.48 | 79.24 | 93.39 |
| M$^5$HisDoc-H → M$^5$HisDoc-R | 95.39 | 90.78 | 98.35 | 94.94 |

**Results and analysis** The results are presented in Table 10, it can be observed that the models trained on M$^5$HisDoc-H outperforms the one trained on M$^5$HisDoc-R in cross-validation. These results highlight the significance of incorporating datasets such as M$^5$HisDoc-H in addressing the intricate challenges associated with historical document analysis, particularly in complex scenarios.

## 5 Computational Resources

All experiments in this study are conducted on four RTX 3090 GPUs.

## 6 Illustration of model output



(a) Text line detection

昌黎先生集考異卷第一
昌 # 先生集考異卷第

此集今世本多不同惟近歲南安軍所刊方氏校
此集今世本多不同惟近 # 南安軍所 # 者或考

定本號爲精善別有舉正十卷論其所以去取之
定本號爲精善別有舉正十卷論其所以去取之

意又它本之所無也然其去取多以祥符杭本嘉
意又宅本之所無也然其去取多以祥杭本嘉

祐蜀本及李謝所據館閣本爲定而尤尊館閣本
祐蜀本及李謝所據館閣本爲定而尤尊館閒本
...

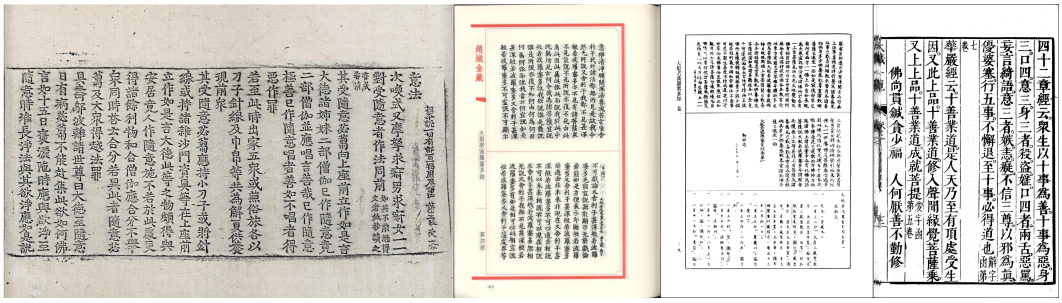
(b) Text line recognition

Figure 2: Illustration of model output. The red polygons and text represent the output of the model and the green ones represent GT.

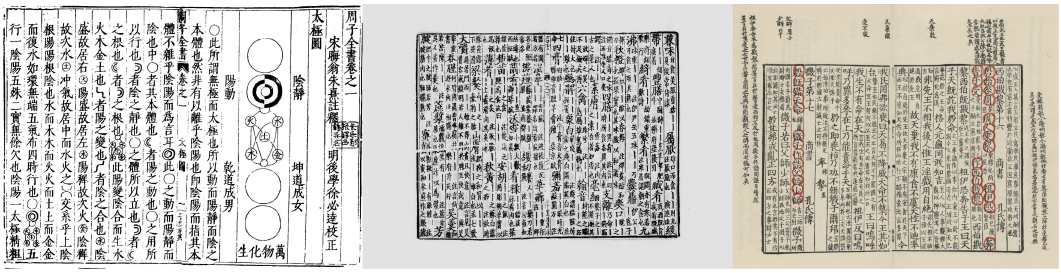
# 7 Additional visual illustrations of M$^5$HisDoc

## 7.1 Illustration of the data source

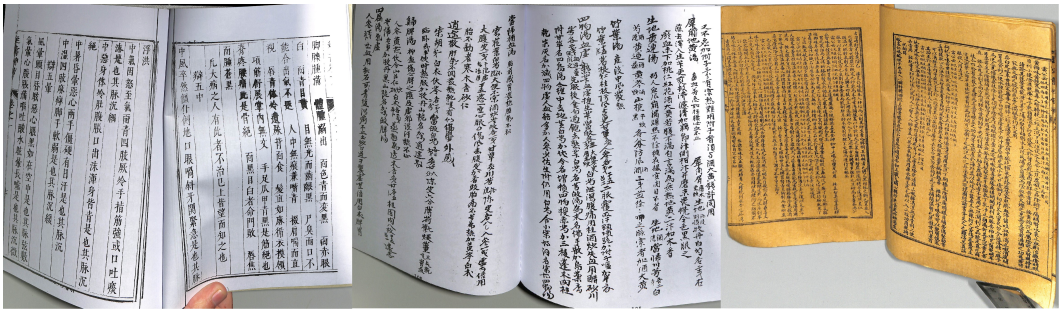
As mentioned in Sec. 3.1 in the main paper, our data source from MTHv2 (Fig. 3a), SCUT-CAB (Fig. 3b), pictures we took (Fig. 3c), and the Internet (Fig. 3d).



(a) From MTHv2



(b) From SCUT-CAB



(c) Photogenic

(d) From Internet

Figure 3: Illustration of the data source.

### 7.2 Illustration of various calligraphic styles

As described in Sec. 3.3 of the main paper, the M$^5$HisDoc contains various calligraphic styles, including regular, clerical, running, and cursive scripts (As shown in Fig. 4, respectively). These categories differ in Chinese calligraphy, each characterized by its unique writing style. For example, cursive script presents a more casual and flowing writing style, usually resulting in less recognizable characters, while regular script exhibits a more standardized and formal typeface.

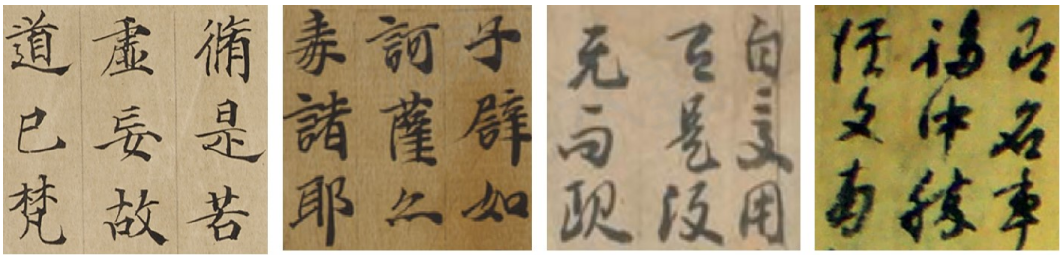


Figure 4: Illustration of various calligraphic styles.

### 7.3 Illustration of various layout types

"Layout type" refers to the various types of text arrangements and sizes within a document. This includes variations such as text divided into double blocks (left, right), text segregated into three parts (top, middle, and bottom), etc. Diverse layouts significantly influence text detection and reading order. As shown in Fig. 5, there are three different types of layout.

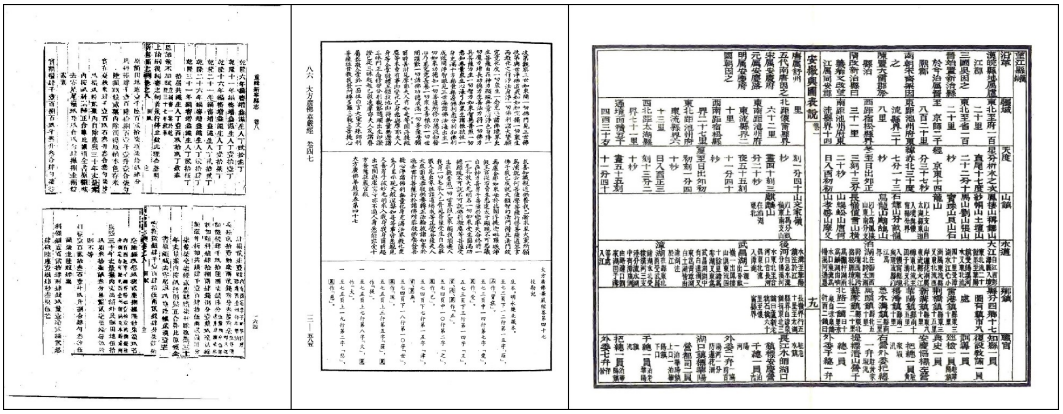


Figure 5: Illustration of various layout types.

### 7.4 Various features of M$^5$HisDoc

Fig. 6, 7, 8, 9, 10, and 11 demonstrate the features blurry texts, complex handwritten texts, variations in font sizes, complex arrangement of text content, dense texts, and distorition of M$^5$HisDoc, respectively.

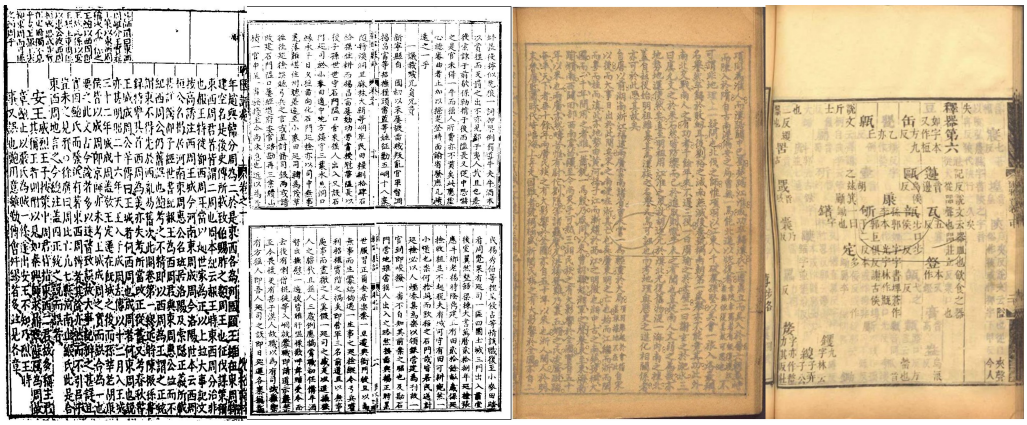


Figure 6: Blurry texts.

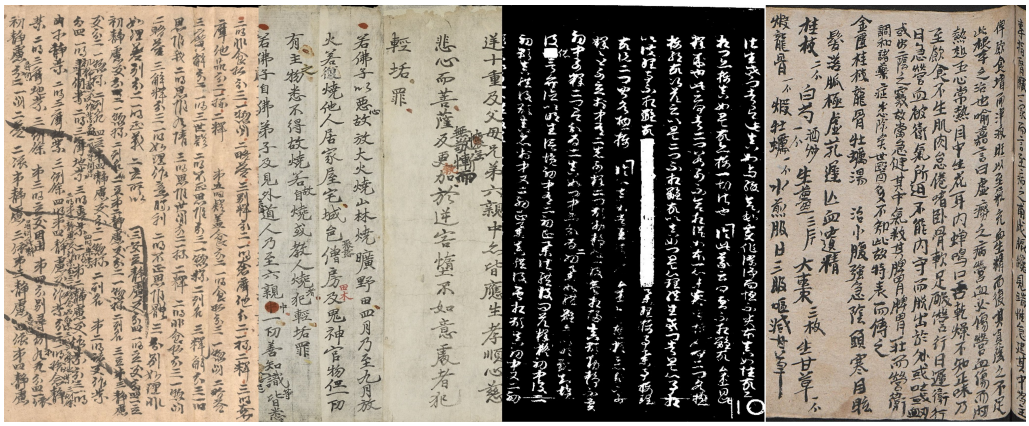


Figure 7: Complex handwritten texts.

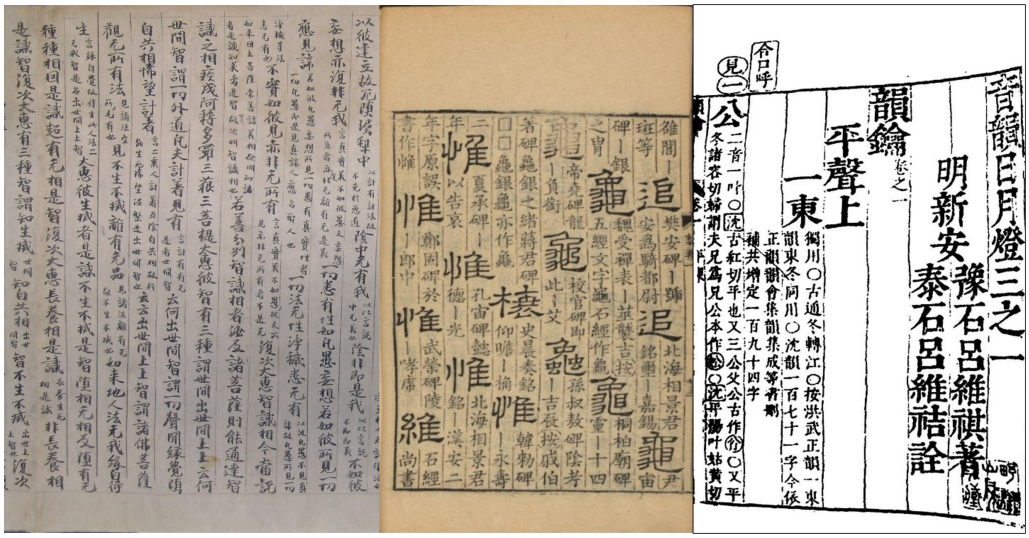



Figure 8: Variations in font sizes.

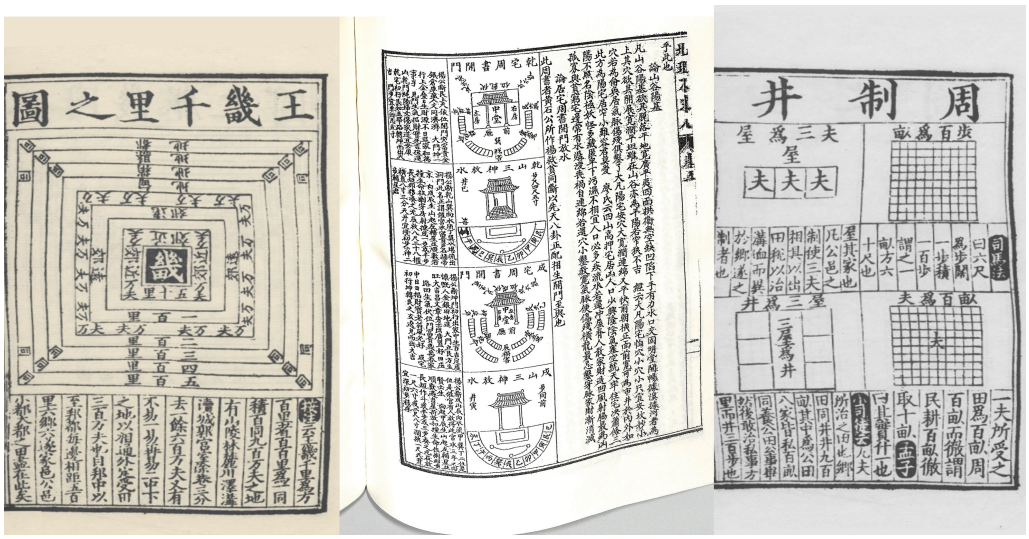



Figure 9: Complex arrangement of text content.

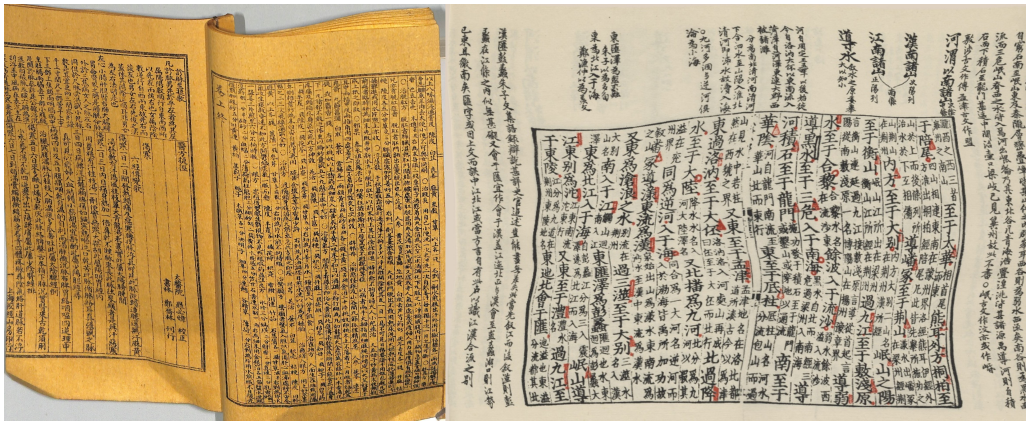


Figure 10: Dense texts.

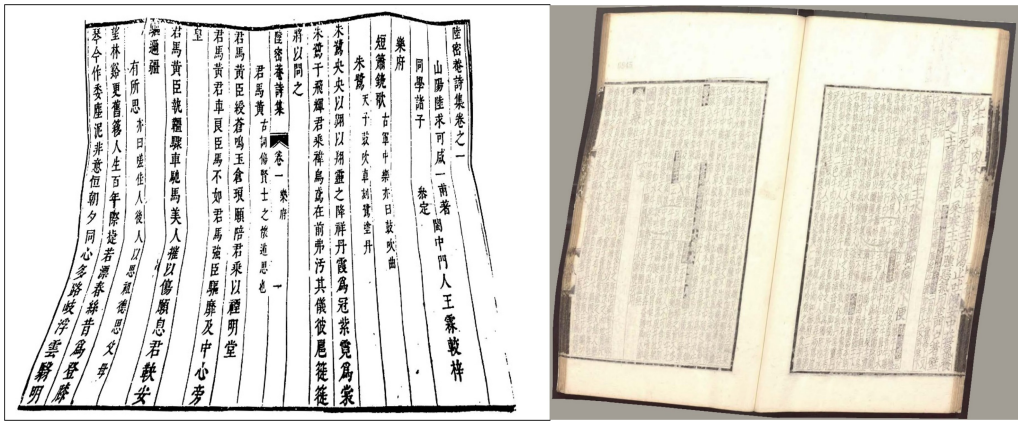


Figure 11: Distorition.

## References


[1] Weihong Ma, Hesuo Zhang, Lianwen Jin, Sihang Wu, Jiapeng Wang, and Yongpan Wang. Joint layout analysis, character detection and recognition for historical document digitization. In *Proc. ICFHR*, pages 31–36. IEEE, 2020.

[2] Rajkumar Saini, Derek Dobson, Jon Morrey, Marcus Liwicki, and Foteini Simistira Liwicki. ICDAR 2019 historical document reading challenge on large structured Chinese family records. In *Proc. ICDAR*, pages 1499–1504. IEEE, 2019.

[3] Yue Xu, Fei Yin, Da-Han Wang, Xu-Yao Zhang, Zhaoxiang Zhang, and Cheng-Lin Liu. CASIA-AHCDB: A large-scale Chinese ancient handwritten characters database. In *Proc. ICDAR*, pages 793–798. IEEE, 2019.

[4] Dezhi Peng, Lianwen Jin, Yuliang Liu, Canjie Luo, and Songxuan Lai. PageNet: Towards end-to-end weakly supervised page-level handwritten Chinese text recognition. *Int. J. Comput. Vis.*, 130(11):2623–2645, 2022.

[5] Oldřich Kodym and Michal Hradiš. Page layout analysis system for unconstrained historic documents. In *Proc. ICDAR*, pages 492–506. Springer, 2021.

[6] Olfa Mechi, Maroua Mehri, Rolf Ingold, and Najoua Essoukri Ben Amara. A two-step framework for text line segmentation in historical Arabic and Latin document images. *Int. J. Doc. Anal. Recognit.*, 24(3):197–218, 2021.

[7] Hiuyi Cheng, Cheng Jian, Sihang Wu, and Lianwen Jin. SCUT-CAB: A new benchmark dataset of ancient Chinese books with complex layouts for document layout analysis. In *Proc. ICFHR*, pages 436–451. Springer, 2022.

[8] Hailin Yang, Lianwen Jin, Weiguo Huang, Zhaoyang Yang, Songxuan Lai, and Jifeng Sun. Dense and tight detection of chinese characters in historical documents: datasets and a recognition guided detector. *IEEE Access*, 6:30174–30183, 2018.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proc. ICCV*, pages 2961–2969, 2017.

[10] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *Proc. CVPR*, pages 6154–6162, 2018.

[11] Yuliang Liu, Tong He, Hao Chen, Xinyu Wang, Canjie Luo, Shuaitao Zhang, Chunhua Shen, and Lianwen Jin. Exploring the capacity of an orderless box discretization network for multi-orientation scene text detection. *Int. J. Comput. Vis.*, 129:1972–1992, 2021.

[12] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proc. CVPR*, pages 9336–9345, 2019.

[13] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proc. ICCV*, pages 8440–8449, 2019.

[14] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang. Fourier contour embedding for arbitrary-shaped text detection. In *Proc. CVPR*, pages 3123–3131, 2021.

[15] Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):919–931, 2022.

[16] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proc. ECCV*, pages 20–36, 2018.

[17] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2016.

[18] Yuhao Huang, Lianwen Jin, and Dezhi Peng. Zero-shot Chinese text recognition via matching class embedding. In *Proc. ICDAR*, pages 127–141. Springer, 2021.

[19] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2035–2048, 2018.

[20] Fenfen Sheng, Zhineng Chen, and Bo Xu. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition. In *Proc. ICDAR*, pages 781–786. IEEE, 2019.

[21] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *Proc. ECCV*, pages 135–151. Springer, 2020.

[22] Dezhi Peng, Lianwen Jin, Weihong Ma, Canyu Xie, Hesuo Zhang, Shenggao Zhu, and Jing Li. Recognition of handwritten Chinese text by segmentation: a segment-annotation-free approach. *IEEE Trans. Multimedia*, 2022.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Proc. NIPS*, 28, 2015.

[24] Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[25] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[27] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proc. CVPR*, pages 10428–10436, 2020.

[28] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proc. CVPR*, pages 11976–11986, 2022.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, pages 1–22, 2021.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *Proc. ICCV*, pages 10012–10022, 2021.

[31] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. XYLayoutLM: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proc. CVPR*, pages 4583–4592, 2022.

[32] Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. LayoutReader: Pre-training of text and layout for reading order detection. In *Proc. EMNLP*, pages 4735–4744. Association for Computational Linguistics, 2021.