Optimizing HTR and Reading Order Strategies for Chinese Imperial Editions with Few-Shot Learning

Marie Bizais-Lillig, Chahan Vidal-Gorène, Boris Dupin

Optimizing HTR and Reading Order Strategies for Chinese Imperial Editions with Few-Shot Learning

Marie Bizais-Lillig[1,2][0000−0002−2426−2641], Chahan Vidal-Gorène[3,4][0000−0003−1567−6508], and Boris Dupin[3]

[1] UR-1340 GÉO and USIAS, Université de Strasbourg, France
[2] Huma-Num consortium DISTAM, CNRS, France
[3] Calfa, Paris, France, https://www.https://calfa.fr/
[4] École nationale des chartes, Université Paris, Sciences & Lettres, Paris, France

Abstract. In this study, we tackle key challenges in layout analysis, reading order, and text recognition of historical Chinese texts. As part of the CHI-KNOW-PO Corpus project, which aims to digitize and publish an online edition of 60,000 xylographed documents, we have developed and released a specialized small dataset to address this common issues in HTR of historical documents in Chinese. Our approach combines a CNN-based instance segmentation model with a local algorithmic model for reading order, achieving a mean precision of 95.0% and a recall of 93.0% in region detection, and a 97.81% accuracy in reading order. Text recognition is conducted using a CRNN model enhanced with GAN-augmented data, effectively addressing few-shot learning challenges with an average accuracy of 98.45%, demonstrating the effectiveness of a small and targeted dataset over a large-scale approach. This research not only advances the digitization and analytical processing of Chinese historical documents but also sets a new benchmark for subsequent digital humanities efforts.

Keywords: Chinese · HTR · Historical Documents · Layout Analysis · Reading Order · Dataset

1 Introduction

Most experiments in the field of text recognition applied to Chinese language focus on modern and contemporary texts onwards (beginning at the very end of the 19th century). Characters are not so diverse, texts are punctuated, and the layout generally differs, with many texts written on horizontal lines from left to write.[5] There have been experiments to try and extract text from ancient editions similar to the ones used in the present dataset. However, the datasets

[5] Layout may still be complex, when working on newspapers for instance. For this type of data, see the Heidelberg Centre for Transcultural Studies project[11] and dataset[10].

produced during the projects often stay private when they are not published under license. In Taipei, the Academia Sinica started developing models to extract text from images since the early 2000s. The rich collection[6] of texts neatly edited by this institution, which many sinologists use on an every-day basis, is, however, mostly accessible by subscribed membership, and its content is licensed. The Chinese Text database is another commonly used resource, freely accessible online, but whose content is also submitted to copyright limitations. Also, despite the progress made in the field of text recognition[7], available models are often so specialized that they don't seem very useful in other contexts, and training datasets are not so common.[8] In other words, the field of imperial Chinese studies is in need of an open dataset of texts and images, produced with the help of HTR technologies.

Within the scope of the CHI-KNOW-PO Corpus project, which aims to establish a comprehensive online and searchable corpus of 60,000 pages, we are experimenting with the development of models to address specific challenges associated with late imperial China editions. The project's goals include the definition of an efficient pipeline, the publication of a training dataset, and the availability a broader corpus of texts from the Chinese first millennium. The key challenges include: 1/ precisely analyzing and structuring page layouts to accurately order the full text, and 2/ recognizing a vast array of different glyphs. Despite years of research in this domain across both Eastern and Western contexts—as noted by recent studies from Fudan University (Shanghai)[31]—automatic extraction of ordered text based on images of Imperial China xylographed editions remains highly experimental and in need of established benchmarks like those the CHI-KNOW-PO Corpus project aims to provide.[10] To tackle these challenges effectively, we have created a specialized small dataset tailored to the project's needs, that can be used as a benchmark dataset for HTR purposes Chinese historical documents. We propose some results as a baseline for this dataset.

2 Sources: Corpus characteristics and dataset

The corpus was circumscribed according to several criteria, the main objective being to represent a literate library of the first millennium—excluding Buddhist

[6] The database is called Scripta sinica. Url: https://hanchi.ihp.sinica.edu tw/ihp/hanji.htm (Last accessed on May 5th 2024).

[7] See for instance the ICDAR 2017 Competition on Reading Chinese Text in the Wild[21] along with Donald Sturgeon's recent publication[22].

[8] Certainly, the Kanripo repository[9], which represents a huge amount of text mirroring black and which images of the *Siku quanshu* 四庫全書 is very interesting. Nevertheless, the quality of the images reduces the interest of such dataset.

[10] The Fudan team contributes to such benchmarks; see https://github.com/FudanVI/benchmarking-chinese-text-recognition (last accessed May 5th 2024). Unlike other initiatives, however, the CHI-KNOW-PO Corpus project adheres to the FAIR principles.

texts, which are collected within the framework of the collaborative project on Medieval China Buddhist texts directed by Christoph Anderl[11]. Exhaustivity once Buddhist texts had been put aside was still an impossible ideal. Most Classics—Confucian or Taoist in particular—commented on by scholars of the first millennium as well as historical works already available online have hence been excluded. We favored thematic coherence: our research project explores the phenomena of co-occurrences and repetitions or echoes between passages of texts which refer to plants. This prism allows us to include a wide variety of genres: knowledge texts such as lexicons and compendiums classified by categories (*leishu* 類書 in Chinese, often referred to under the misleading term encyclopedia), poetry, treatises on *materia medica*, and agricultural treatises in particular. The variety of genres is accompanied by a relative variety in form. Certain works are extracted from large collections compiled either on imperial order or privately by a bibliophile, while others constitute independent physical units. Texts also vary in form, length, and organization: the corpus includes lexicons, anthologies of poems, along with chaptered prose texts. Finally, some texts are commented on by one or more exegetes, while others are the work of a single author. Before illuminating more precisely the issues that the glosses represent in this project, let us describe the dataset and the material and formal characteristics of the corpus as a whole.

2.1 Qualitative description of targeted documents

The CHI-KNOW-PO Corpus project contains three types of texts, from which we extracted a sample in order to construct a dataset for model training adapted to its processing.

Fig. 1. Overview of the dataset, with typical page layout of a xylographed edition on the left and on the right. Main semantics zones and lines are displayed.

[11] https://www.database-of-medieval-chinese-texts.be/ (last accessed March 13 2024).

Anthologies The *Li Shan zhu Wenxuan* 李善注文選 (The Selection of Belletristic texts commented upon by Li Shan, abbreviated to Li Wenxuan, n°A-1) is a large anthology of 30 *juan* (rolls), organized by genre, and compiled by Xiao Tong 蕭統 (501-531), Prince Zhaoming 昭明太子, under the Southern dynasty of the Liang 梁 (502-557). It comprises more than 700 texts composed in many different genres between Antiquity and the early Medieval period. Each of these texts stood as a model for later compositions by literati. More than half of the anthology is made of *shi* 詩 poetry and *fu* 賦 rhymed prose, which were refined texts and whose vocabulary was very rich. To shed light on the meaning of the texts and their intertextual relations with other reference texts, these texts were commented upon less than a century after its compilation. The earliest commentary that survived was Li Shan's 李善 (630-689), which also became the most famous. Fragments of the early versions of the text can be found in the Dunhuang collections of the Bibliothèque nationale de France. Complete versions transmitted to us are relatively late. This one was printed in 1809 based on xylographic plates from the Song dynasty (960-1279). The work is composed of four cases of six leaflets each. It is 30 cm high and preserved at the National University Library in Strasbourg (BNU, call number: FP.12.4.0001)[12].

The *Liuchen zhu Wenxuan* 六臣注文選 (abbreviated Liuchen Wenxuan, n°A-2) corresponds to the same book. After Li Shan commented upon the *Wenxuan*, a group of scholars considered his commentary too elitist and difficult to understand. The collection that merges the commentaries produced by the five opponents to Li Shan with the ones by Li Shan is called the *Selection of Belletristic texts commented upon by the six ministers*. This text is interesting because of its pedagogical input and its richness. The edition used in the dataset is a rather late edition, printed in 1923. It belongs to the *Sibu congkan* reference collection. The work is composed of three cases of ten leflets each. Its is 20 cm high and is preserved at the BNU (FP.12.8.0001). It contains manual annotations in the margins.

The *Yutai xinyong* 玉臺新詠 (New Songs of the Jade Terrace, abbreviated Yutai, n°A-3) is a collection of 769 court poems from Antiquity and the Early Medieval period compiled by Xu Ling 徐陵 (507-583). The collection is famous for its tematic coherence, most poems dealing with love and women. This is a rather late edition, printed in 1879, precious because of the presence of commentaries. It consists of one case of ten 18 cm high leaflets. It is preserved at the Library of the Institut des hautes études chinoises at the Collège de France in Paris (BIHEC, call number: V XIV 69 (1-8)).

The *Quan Tang shi* 全唐詩 (Complete Poems of the Tang dynasty, abbreviated Tangshi, n°A-4) corresponds to the complete collection of Tang dynasty poems compiled during the Qing (1644-1911) dynasty. It contains some 50,000 poems. For each author, biographical elements are provided in the form of a commentary. This one is one of the earliest editions (possibly the first), printed

[12] The Libraries part of the CHI-KNOW-PO Corpus project are listed in the acknowledgement section.

in 1707. This work is composed of twelve boxes of ten leaflets each. It is 16.5 cm high and is preserved at the BIHEC (SB 4002 (1-12) (1-120)).

Scholarship The *Beitang shuchao* 北堂書鈔 (Written excerpts from the Northern Hall, abbreviated Beitang, S-1) is one of the earliest *leishu* composed in China. It is divided in categories, while each category is composed of a series of entries. Each entry includes 1/ a series of definitions (usually excerpted from ancient reference books), and 2/ a series of famous quotations (that can be reused by literati in their own writings). This edition, which is a reconstruction of the book, was printed in 1888. The structure of the book is not easily readible. This work is composed of four cases of fives leaflets each. It is 31 cm high and is preserved at the University Library des langues et des civilisations in Paris (BULAC, call number: BIULO CHI.1087).

The *Bowu zhi* 博物志 (Notes on things at large, abrreviated Bowu zhi, S-2) is a short book composed by Zhang Hua 張華 (232-300) during the Western Jin 西晉 (265-316). It is composed of ten chapters. Each chapters stores a series of tales on the world. The book, printed in 1875, consists in one leaflet. It is preserved at the BULAC (BIULO CHI.1140).

The *Chuxue ji* 初學記 (Records to begin studies, abbreviated Chuxue, S-3) is one of the most prestigious *leishu* of Chinese history. It was composed by a team on imperial order during the Tang 唐 dynasty (618-907). It defines categories to classify entries. Each entry is followed by different sections: definitions, citations from texts in different genres, so called "parallel matters". This 1587 edition is very clearly presented. It contains two cases of twelve leaflets each. It is 27.5 cm high and is a preserved at the BIHEC (SB 3701 (1-2)).

The *Erya yintu* 影宋鈔繪圖爾雅 (Illustrated Elegantiae, facsimile of a Song edition, abbreviated Erya, S-4). The *Erya* is a Confucian classic. It is a lexicon, ordered by category. Each entry is followed by a series of equivalents (not exactly synonyms). It is often considered to be essential to decipher the meaning of rare terms used in the Confucian Canon. This edition is characterized by the presence of phonological annotations and by its illustrations. The work was printed in 1801. It consists of three illustrated volumes. Each volume is 34 cm high. It is preserved at the BULAC (BIULO CHI.1938(1)-(3)).

The *Mao Shi caomu niaoshou chongyu shu* 毛詩草木鳥獸蟲魚疏 (Commentary on flora and fauna in the *Poems* according to Mao, abbreviated Maoshi shu, S-5) is a commentary established by Lu Ji 陸璣, who lived between 200 and 500. It is a commentary of the Confucian classic anthology of 300 poems known as the *Shijing* 詩經. The commentary doesn't include its base text. It is composed of entries that are named using one line of one poem. The selected lines all include elements of fauna or flora. This element is defined using multiple sources: it is a compilation of all information available on these elements (where they grow or live, what are their physical characteristics, how they are called in different places and times). This work is part of a larger collection, printed in 1857, and preserved at the BIHEC (V I 111 (1) 5).

The *Yiwen leiju* 藝文類聚 (Compilation of texts and works ordered by category, abbreviated Yiwen, S-6) is, with the *Chuxue ji*, S-3, the other major *leishu* of early Chinese history. It was also composed during the Tang dynasty (618-907), by a team headed by the well-known scholar Ouyang Xun 歐陽詢 (557-641). This work was printed in 1879. It consists of eight boxes of five leaflets each. It is 20.5 cm high and is preserved at the BIHEC (CIII 5-7 (1-8)).

The *Zhi bu zu zhai cong shu* 知不足齋叢書 (Collection of the One who did not know enough, abbreviated Zhibuzu, S-7) corresponds to what is called *congshu* 叢書 (collection) in Chinese. It means that it is a compilation of a number of books—in this case, 207 books. The *Zhibuzu* collection was compiled by Bao Tingbo 鮑廷博 (1728-1814). It includes all sorts of books, mainly encyclopedias and other books of scholarship, from the first millennium. It is a very large and diverse collection, although our selection means to establish links between the books included in the *Zhibuzu* and other books present in the dataset (poems, texts on plants). This book was printed in 1822. The layout changes from one title to the other. It is composed of 120 leaflets, 30 cm high, and is preserved at the BIHEC (F X 2 (1-15) 1-120)).

Technical and Practical Knowledge The *Gujin shiwen leiju* 古今事文類聚 (Compilation of texts and activities from present and past ordered by categories, abbreviated shiwen leiju, T-1). This *leishu* is composed of a series of *leishu* from different periods (from the 12th century onwards). Although such a *leishu* resembles scholarship encyclopedias, this one focuses more clearly on practical issues. The work was printed in 1604. It was later bound in a Western way. It is composed of 26 volumes, 27.5 cm high. It is preserved at the BIHEC (SB 3705 (1-26)). It has suffered from humidity.

The *Qimin yaoshu* 齊民要術 (Essential techniques of the people of Qi, abbreviated Qimin yaoshu, T-2) is the earliest treatise on agriculture preserved in Chinese history. It is a short texte. This edition was printed in 1896. It is preserved at the BIHEC (V I 22 (1-4)).

The *Xinzhai shizhong* 心齋十種 (Tens types for the temperate heart, abbreviated Xinzhai, T-3) is a short collection of practical texts including treatises and encyclopedias. It was printed between 1785 and 1788. It consists of one box of four leaflets. It is 16.8 cm high and is preserved at the BIHEC (V I 53 (1)).

Although the selected medieval corpus (ca. 250-1000) covers relatively varied genres, it is generally studied based on imperial editions from the second millennium,[13] and is hence rather homogeneous in its physical form.

We are mostly dealing with xylographed accordion-bound books. This means that the medium is generally thin paper, which is written on only one side, and which is folded to form the equivalent of a sheet in a codex. The side 'trapped' in the fold remains inaccessible to the reader, its function being to reduce the noise created by the back of the sheet when reading the front, and vice versa.

[13] Most manuscripts disappeared in the transmission process.

Table 1. Dataset Distribution

Title	Pages	Lines	Characters
Li Wenxuan, A-1	27	771	16,697
Liuchen Wenxuan, A-2	29	889	16,271
Yutai, A-3	10	590	10,220
Tangshi, A-4	10	493	7,685
Beitang, S-1	35	1,508	27,388
Bowu zhi, S-2	23	302	7,874
Chuxue, S-3	20	1,267	24,411
Erya, S-4	40	2,577	20,121
Maoshi shu, S-5	10	397	8,357
Yiwen, S-6	11	357	7,409
Zhibuzu, S-7	49	1,804	29,524
shiwen leiju, T-1	20	1,055	15,629
Qimin yaoshu, T-2	20	885	18,307
Xinzhai T-3	23	864	14,143
TOTAL	327	13,759	224,036

Each page is composed of a regular number of columns, delineated with vertical and horizontal lines—in imitation of the string of bamboo strips on which texts were inscribed in Antiquity. The double-pages, which correspond to a leaf or a sheet, in a xylographed edition are printed from a single plate, and separated by a column of a particular shape (diamond or double fish, in salmon pink in figure 1) where metadata are inscribed. This area, which corresponds to the fold of the sheet, becomes difficult to read in a bound book: we only see half-characters which can correspond to the title of the work, the title of the part, and the number of the sheet (which sometimes starts again at the beginning of each chapter). In the project, despite the poor readability of this area, we decided to identify and transcribe it.

The text itself (in red in figure 1) is written in columns, from top to bottom, then from right to left. Commentaries are embedded within main text, in double columns and smaller font (in purple in figure 1).

Sometimes, addition text, either manuscript notes or printed metadata, appears in the margins (in yellow in figure 1).

2.2 Quantitative description of the dataset and guidelines

The dataset comprises 327 pages, totaling 13,759 transcribed lines (see Table 1). Annotations have been made using Calfa Vision[28], an online collaborative tool that incorporates active learning strategies. As annotations progress, the tool automatically generates and refines layout and text predictions, specifically flagging the most challenging pages for the model to facilitate targeted corrections.

The goal extended beyond simple transcription; it aimed to generate XML files structured according to TEI (Text Encoding Initiative) standards, incorpo-

rating semantic interest areas both at the region and baseline levels within those regions. The key types of elements annotated on the page are detailed in Table 2.

Semantic typing prioritizes MainText and Marginalia_MetaData regions, along with Text, Commentary, and Title lines (see Table 2). Despite the dataset's imbalance in character classes (discussed in part 4.2 below), it is representative of the targeted xylographic production theme.

Table 2. Region Types and corresponding baselines

Type	Zone	Count	Definition
Text Regions			
MainText		296	Main Zone of text
MainText TableOfContent		58	Table of Contents
Marginalia MetaData		492	MetaData informations
Marginalia Caption		49	Caption for an image
Marginalia PageNumber		106	Regions for arabic numbers
Marginalia		23	Additional notes by readers
Image		53	Regions for Illustrations
Image Stamp		27	Regions for Library Stamps
Image Seal		22	Regions for Seals
Text Lines			
Author_name	MainText*	220	
Commentary	MainText*	6633	
MarginaliaLine	Marginalia	89	
Page_Number	Marginalia MetaData	0	
Text	MainText*	4190	
Title	MainText*, Marginalia Metadata and Caption	2093	

In order to ensure the interoperability of the dataset with other common annotation tool or models, we have adopted an annotation by baseline from which polygons are generated. On the other hand, the baseline information is not used in the present experiments. The data is distributed in pageXML format, under Apache 2.0 license.

3 Layout analysis and reading order of Chinese handwriting

3.1 Simple columns for text, double columns for commentaries

The order of the columns is complexified by a layout that mixes main text and exegetical apparatus. In the selected body of texts, the Confucian Classic called

Shijing 詩經 (Classic of Poems) best exemplifies this complexity. An official version of this anthology of 300 poems was established on imperial order and commented upon by a certain Kong Yingda 孔穎達 (574-648) and his team, who systematically cite three other commentaries: a preface which dates back to the 1st century BC, a system of glosses attributed to a certain Mao 毛, probably also from the 1st century BC, and a sub-commentary by the scholar Zheng Xuan 鄭玄 (127-200). Kong Yingda then adds his own prefaces and comments. Glosses and comments are inserted immediately after the passage or words to which they relate. They are recognizable thanks to their format: instead of occupying the entire width of a column, commentaries are written in double columns, in a font half the size of main text. The same layout applies in most editions, including that of the *Yutai xinyong* (figure 1), where we have drawn on the right page two red binding boxes to identify two strings of main text, and purple binding boxes for glosses. A blue binding box corresponds to the title of the next poem. In a large column on the page, each block corresponding to a single category is read before the one below. Also, within each block, baselines read from top to bottom and from right to left.

Several of the texts in our corpus are characterized by this textual millefeuille inside the columns. However, as the case of the *Classic of Poems* detailed above illustrates, the distinction between the different parts of the text is not trivial: each of the comments points to a certain part of the base text (the one that is just above), each one follows its own discursive modalities, and each dates from a period distinct from the others. The main text also dates to a specific period. In the case of the *Classic of Poems* anthology, each poem was composed between the 11th and 6th centuries BC, although the written version which has reached us was not been established before the 2nd century BC.

Distinction between types and correct sequence of baselines are hence critical.

3.2 Strategies to establish reading order

While layout analysis tasks, including complex ones, have been largely addressed in most common scenarios through targeted fine-tuning (excluding forms and newspapers) using either a basic U-net architecture for segmentation [8, 19, 1, 28] or, more recently, Transformer models [30, 32], the challenge of determining the reading order of text regions and lines persists, particularly in the context of historical documents. The reading order represents a semantic challenge that cannot be directly solved using only computer vision techniques. Yet, it remains a critical factor for ensuring the accuracy and relevance of OCR/HTR outputs.

A topological sorting [2] proves adequate for simple document layouts, such as those limited to two columns with marginal annotations, but it encounters limitations with more complex documents. To address this, more advanced approaches have been employed. These include Graph Segmentation without machine learning [29] and the use of a multi-layer perceptron with machine learning [20]. However, these methods primarily capture morphological features, such as the spatial relationships between text regions or lines, rather than semantic attributes. Vision Transformers offer a promising solution, as they integrate both

spatial and contextual information [30]. Although they have been tested predominantly on printed documents, training them requires a substantial data volume. This requirement often proves challenging for historical documents, particularly those with limited resources.

The Chinese reading order introduces additional complexity, typically following a hierarchical right-to-left and top-to-bottom pattern. While topological sorting based on centroid positions of text areas is straightforward in most scenarios (as demonstrated by the dataset in this study), applying the same logic to lines within text boxes proves problematic. This is illustrated in figure 2, which shows a mix of Text (in red) and Commentary (in purple). Global topological sorting here is significantly hindered by its reliance on the coordinates of the bounding boxes or polygons of the lines. An initial experiment using a simple global sort of line centroids achieved an accuracy of 82.31%. However, this accuracy dropped below 60% with variations in baseline inclination, page orientation, and the model's ability to predict complete baselines in practical scenarios.

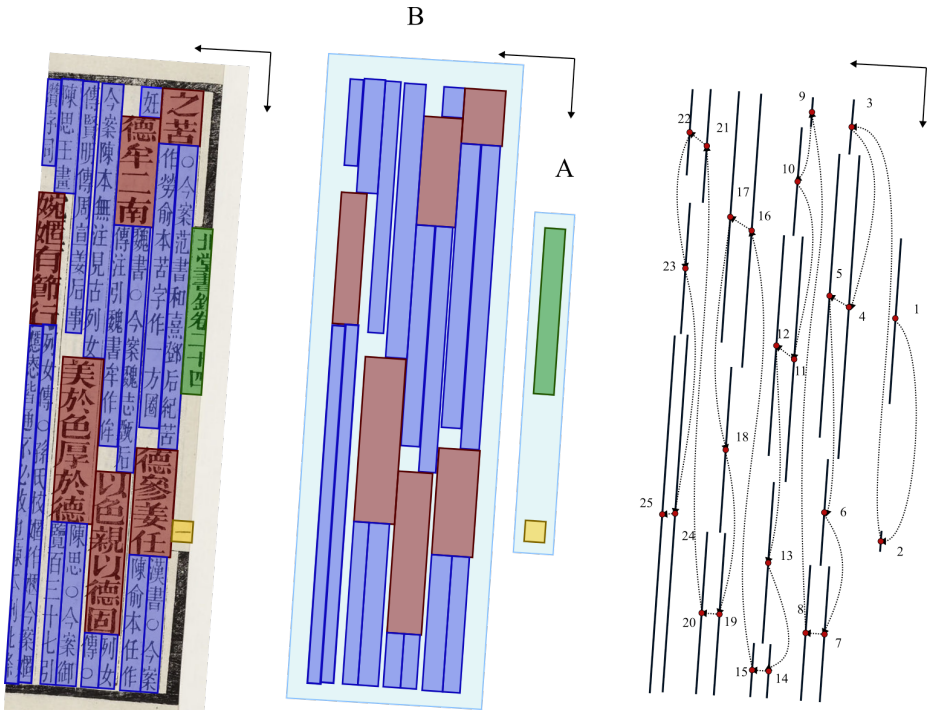

Fig. 2. Reading Order in Chinese Documents: Commentary line in purple, Text line in red, Title in green, and Page number in yellow. Traditional methods using box or baseline centroids often fail to accurately determine reading order in this case.

The top-bottom and right-left hierarchy often conflicts, as exemplified by the lines labeled 3-4-5, which may be incorrectly ordered as 4-3-5 or 3-5-4, and the lines 11-12, where the shorter line 12 has a higher centroid (see Figure 2).

To address the challenge of sorting Chinese lines, two main approaches are prevalent in state of the art. The first approach employs a non-global, algorithmic sorting method, which establishes a topological hierarchy starting from pairs of characters and extending to larger groupings, achieving a final accuracy of 97.7%, even on pages that are wavy or misoriented[16]. This method, however, depends on glyph bounding boxes, which are not commonly included in annotations schemes and may be inaccurately defined by recognizers, as perfect alignment with glyphs is not essential for their recognition. The second approach utilizes machine learning, specifically a convolutional neural network (CNN) trained to discern line spacing and character sizes[17], coupled with a multi-layer perceptron (MLP)[19]. This method integrates both global (MLP) and local (CNN) sorting techniques, resulting in a Page Error Rate of 5%, where a page is deemed incorrect if any line is misordered.

3.3 Proposed method and results

The overall challenge of determining the reading order is partially mitigated in both scenarios through a local analysis of the relationships between objects. However, the unique characteristics of our dataset – marked by a high variety of document types and a significant mix of semantic line types (see Part 2.2) – complicate the reproducibility of these methods. Additionally, the lack of glyph-level annotations further limits their effectiveness. Consequently, training a multi-layer perceptron (MLP) for global sorting yields an accuracy of only 50%, which is effectively equivalent to random sorting.

In this paper, we propose incorporating the reading order task as an auxiliary component within the layout analysis pipeline (see Figure 3). Specifically, we apply a localized algorithmic topological sort that efficiently handles dual-column layouts:

- **Horizontal Overlap with Right Priority**: If two lines overlap vertically beyond a certain threshold, the line whose centroid is further to the right is considered to come first. This reflects a reading order from right to left within overlapping lines.
- **Vertical Priority without Overlap**: If two lines do not vertically overlap beyond the threshold, the line whose centroid is higher (lower y-coordinate) is considered to come first. This mimics the top-to-bottom reading order for non-overlapping lines.

We formalize the reading order determination through an overlap_ratio and a sorting function S that takes into account the bounding coordinates of two lines u and v. The function is defined as follows:

$$\text{overlap\_ratio}(u, v) = \frac{|\mathcal{I}_u \cap \mathcal{I}_v|}{|\mathcal{I}_u \cup \mathcal{I}_v|}$$

where:

- $\mathcal{I}_u = [y_{\min,u}, y_{\max,u}]$, and $\mathcal{I}_v = [y_{\min,v}, y_{\max,v}]$,
- $y_{\min,u}$ and $y_{\max,u}$ denote the minimum and maximum y-coordinates of line u's bounding-box (respectively for v).

We have finally, for a given δ threshold :

$$S(u,v) = \begin{cases} 1 & \text{if overlap\_ratio}(u,v) > \delta \text{ and } u_x > v_x \\ 1 & \text{if overlap\_ratio}(u,v) \leq \delta \text{ and } u_y < v_y \\ 0 & \text{otherwise} \end{cases}$$

where u_x, v_x, u_y, and v_y represent the x and y coordinates of the centroids of lines u and v.

The layout analysis process comprises three stages: First, the detection and semantic classification of text regions are performed using a CNN-anchor-based instance segmentation model, more relevant in this under-resourced case than state-of-the-art anchorless approach for Chinese[24, 25]. Instance segmentation ensures that closely positioned or similar regions do not merge, a well-know phenomenon with pixel-level semantic segmentation[15], and allows for accurate detection on curved or poorly oriented pages. Second, lines of text are detected and semantically classified, also through instance segmentation. Third, double columns of text are identified using a CNN-based object detection model. We are using a YOLOv8-s model for each step, with default hyperparameters and data augmentation[14].

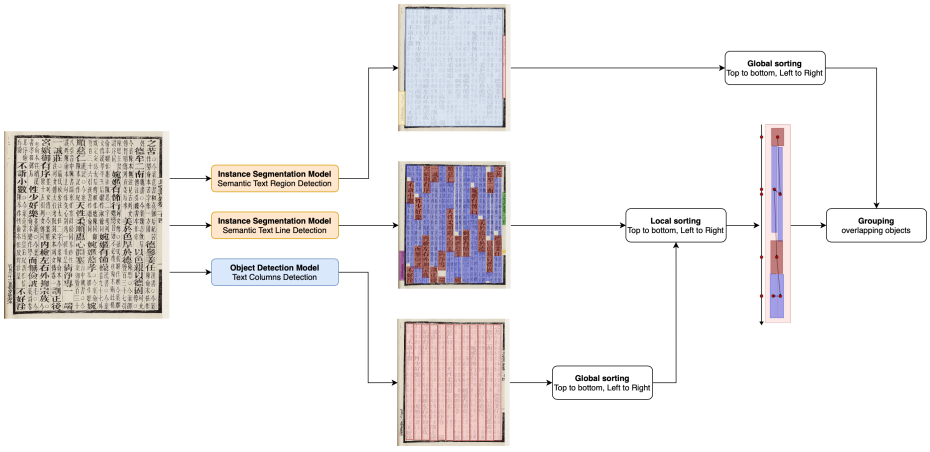

Fig. 3. Pipeline for layout analysis and global/local reading orders

After these three stages, a global sort is applied to both the regions and text columns. The lines within each column are then sorted according to the previously described algorithm. Finally, a global merge of all overlapping objects is

executed to generate the page's output (see Figure 3). While annotation tools commonly utilize baselines for quicker document markup[15, 28], our approach opts for direct detection and sorting of lines. This method proves more effective and expedient for managing the complexities of Chinese documents, which often feature dense text layouts where traditional baseline strategies may not be optimal.

Table 3 summarizes the detection results for each task. For text regions, the model achieves an overall precision of 0.95 and an overall recall of 0.93. This paper specifically targets the MainText, MainText TableOfContents, and Marginalia_Metadata regions. Detection and classification are nearly perfect for the first two categories, though there are some failures with MainText TableOfContents. Nevertheless, the distinction between MainText and MainText TableOfContents is ambiguous, primarily due to differences in line indentation within the region. When MainText TableOfContents is not identified, it is classified as MainText, which suffices for the purposes of this study.

The detection of Marginalia_Metadata presents more challenges due to its variable location along the right and left page borders, often complicated by page folds that crop the content. Despite these difficulties, the model still achieves both precision and recall rates around 0.9.

Table 3. Layout analysis results for the three models of detection

	P	R	mAP
TextRegions detection (mask)			
Image	0.952	0.842	0.850
Image_Stamp	0.965	1.000	0.995
MainText	0.993	0.965	0.991
MainText_TableOfContents	0.955	1.000	0.995
Marginalia	0.910	1.000	0.995
Marginalia_Caption	1.000	0.953	0.995
Marginalia_Metadata	0.908	0.884	0.917
Marginalia_PageNumber	0.917	0.793	0.914
Column detection (bbox)			
Column	0.972	0.890	0.937
TextLines detection (mask)			
Author_name	0.419	0.714	0.722
Commentary	0.958	0.944	0.976
MarginaliaLine	0.774	0.977	0.959
Page_Number	0.748	0.676	0.72
Text	0.968	0.979	0.979
Title	0.896	0.855	0.929

Column detection exhibits varying precision and recall, suggesting an over-detection of columns compared to expectations (False Positives). The model was

trained exclusively on annotations from MainText and MainText TableOfContents columns (regions critical for maintaining correct reading order). Additionally, many pages at the end of chapters contain extensive blank spaces, which are not fully written. The model also tends to incorrectly tag columns delineated by lines, even though these are not present in the training annotations. While this leads to numerous false positives, they do not adversely affect the overall results, as they typically involve empty text columns or columns outside the target regions.

Line detection yields variable results, with Commentary and Text classes being accurately detected (mean precision of 0.963 and mean recall of 0.961). However, Author_Name and Title classes show greater variability due to their visual similarity to Text in MainText and MainText TableOfContents. These classes are differentiated only by the presence of indentations, which are not included in the annotations (bounding boxes start at the text, not the preceding space). For Title, the ambiguity increases in certain books where indented lines are classified as Text rather than Title, leading to notable misclassifications. Despite these challenges, the lines are in any case detected and sorted correctly, although their classification may be incorrect.

We evaluate reading order accuracy in two ways: Firstly, of the 412 lines assessed, 9 were incorrectly sorted, resulting in an accuracy of 97.81%. Secondly, from an editorial perspective, we consider a page faulty if at least one line is incorrectly sorted, yielding an accuracy of 93%. Current errors appear to be associated with very specific cases in how algorithm handles overlapping lines, indicating a need for further investigation.

4 Text recognition

Both classical and modern Chinese are written in characters (sinograms) used singly or in combination to form words that are not separated by spaces. The great dictionary of Chinese characters, *Hanyu da zidian*, listed over 60,000 sinograms in 2010[9]. The challenge this represents for machine recognition is further compounded, as we shall see, by the numerous graphic variants.

4.1 Words, characters and glyphs in ancient Chinese texts

It is historically attested that certain works were transcribed according to precise stylistic standards, although this constraint does not erase the diversity of hands, as Huang Mingli has shown in a recent article devoted to the court style *guange ti* 館閣體 in the great imperial collection entitled *Siku Quanshu* 四庫全書 (Complete Book of the Four Stores)[13]. In order to better characterize the graphic variety of the annotated corpus, we need to explain what the xylographic reproduction that is most common in imperial China is based on, clarify the notion of graphic variant, and return to the awareness of this variety evidenced in certain ancient writings. The reproduction of text by printing on paper actually

began in the 9th century CE, although this method is attested much earlier for the printing of images (notably Buddhist), charms and calendars[18, 26, 5].

Nevertheless, xylographic plates are derived from handwritten copies, so the results of this printing method do not lead to graphic standardization, as is the case with movable type printing introduced by Johann Gutenberg in 15th-century Europe. The consequences of the technical choice of xylographic printing include irregularity in the transcription of characters, a variety of hands within the same work, and a diversity of writing styles from one edition to the next.

The variety we just described is inherent to manuscript traditions and a priori expected in an HTR project. In the case described, the challenge is all the greater since Chinese language is by definition written using a very large number of characters—the classical Chinese language of the first millennium in particular. The problem becomes that of the annotator (as opposed to the computer scientist) when the question of character variants arises.

A variant (in Chinese *yi ti zi* 異體字) does not correspond to a regular graphic expression of a graphic unit, like an 'a' whose transcription varies according to the font or case of the character. The equivalent of such variety is writing styles (sometimes labelled calligraphic styles), and called *ti* 體 in Chinese. By contrast, a variant corresponds to a graphic expression of a script character that has ceased to be used, that is considered exceptional or even faulty, or that has been defined to enable the identification (by equivalence) of an otherwise taboo character[14] in an edition.

These variants stand as challenges for the transcriber, as they are not normally part of the unicode character set. The subject is made all the more thorny by the fact that a number of these variants exist in unicode. In other words, given that many rare characters (and therefore poorly known to annotators) are affected by these variant problems, it is difficult :

1. to identify the variant as such,
2. to transcribe it using the standard version of the character to which it corresponds.

The choice was made to standardize the transcription of the corpus, so as to avoid variant confusion as much as possible, and to limit, so to speak, the overall number of different characters.

4.2 Recognition results

Chinese HTR has predominantly advanced in online recognition and contemporary datasets, as seen in recent ICDAR publications[23]. CRNN+CTC models typically deliver good results on Chinese texts[12], although they demand large data volumes and struggle with generalization[3]. The primary challenge

[14] The personal name of the emperor was taboo during his lifetime, meaning that the characters that make up his personal cannot be written in any edition conceived during his lifetime and need to be replaced by substitute characters.

is the extensive number of classes (several thousands) that need to be recognized. Transformers present a viable alternative, offering excellent adaptability and handling of zero-shot learning scenarios[31], yet their effectiveness is limited on historical documents due to data scarcity. The SVTR-net (Single Visual Model) shows the best outcomes even for historical texts but requires training on several million characters[6, 17]. While transfer learning could be a relevant approach for our dataset, which features 5,581 unique characters – meaning 30.46% of which are represented only once, it is not currently feasible. Conversely, some characters are highly represented in our dataset, highlighting the imbalance. For instance, characters such as the following ones are particularly prevalent: 之 (2,239 samples), 也 (1,552), 曰 (1,549), 不 (942), 以 (895), 一 (889), 而 (769), 有 (765), 十 (742) and 為 (737)—as opposed to 佯, 棰, 帕, 伉, and 懾 that have only one sample.

Existing models fail to perform adequately on our out-of-domain corpus, typically achieving no more than 40% weighted-accuracy[3]. This limitation is particularly critical in few-shot, single-shot, or zero-shot learning contexts where many characters may never be encountered during training, making existing models unsuitable for our needs.

To benchmark this dataset, we employ a word-based CRNN architecture, which provides greater contextual understanding for recognition[28]. Addressing the data scarcity issue, we implement qualitative data augmentation using CycleGAN[33], previously proven effective for Chinese text[4]. CycleGAN has been experimented to transfer handwriting styles to printed text for HTR data creation in Latin scripts[27], and authors have shown its limitations such as scale discrepancies and the absence of a recognizer to ensure real readability of the generated image, like the one used in ScrabbleGAN[7, 27]. The GAN approach is applied in a highly constrained setting, using cropped images of characters with a simple handwritten effect mapped onto printed characters. This reduces GAN instability, although no control experiments were conducted.

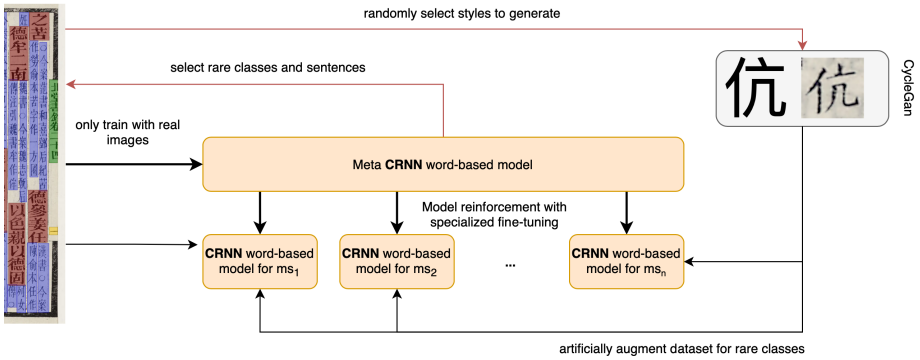

Fig. 4. Pipeline for training recognition model

For our purposes, we selectively generate single characters during training to enhance the representation of underrepresented classes. To maximize context in vocabulary, word sequences, and graphic variations, we implement a dual-learning approach. Initially, a model is trained using the entire dataset comprised solely of real data. This model then constitutes a base model for fine-tuning a specialized one tailored to the target manuscript. If needed, CycleGAN is utilized to augment the dataset with rare characters identified in the first training phase (see Figure 4). We ensure consistent data distribution across training, validation, and testing phases. Although this strategy results in some over-fitting, it is deemed acceptable for processing a specific target manuscript. Table 4 summarizes recognition results for each manuscript.

Table 4. Recognition results

Manuscript	**N°**	**Accuracy**
Li Wenxuan	A-1	99.38 (± 1.2)
Liuchen Wenxuan	A-2	98.84 (± 1.8)
Yutai	A-3	98.52 (± 1.2)
Tangshi	A-4	99.25 (± 1.8)
Beitang	S-1	98.76 (± 1.8)
Bowu zhi	S-2	99.18 (± 1.8)
Chuxue	S-3	97.57 (± 1.7)
Erya	S-4	96.57 (± 0.4)
Maoshi shu	S-5	98.42 (± 1.8)
Yiwen	S-6	98.72 (± 1.7)
Zhibuzu	S7	98.70 (± 1.8)
Shiwen leiju	T-1	97.47 (± 4.5)
Qimin yaoshu	T-2	99.35 (± 2.8)
Xinzhai	T-3	97.61 (± 3.2)

The results demonstrate that this strategy can yield high accuracy, even with small datasets; three manuscripts achieved around 97% accuracy.

If we look in more detail, we find the average confusion in character recognition reaches a peak for characters with 10-20 samples. However, for this specific case, the situation varies a lot depending on the manuscript taken into consideration. A further inquiry on these cases would be needed to explain and hopefully improve the results. On the other hand, characters with fewer than 10 samples show similar recognition rates to those with between 100 and 400 samples. Not that both situations are exactly similar: confusion rates appear more stable for characters with large samples in most manuscripts, when they tend to vary more in one-shot learning situations. The recognition accuracy for unknown characters, including those that were artificially generated, stands at 86.21%. Even though part of the explanation may lie in the fact that these characters are infrequent in both the training phase and during inference, these results show

that the strategy developed in this project is efficient when it comes to scarce characters.

To illustrate the real-world applicability of the models, we also provide adjusted accuracy figures (±1.8%) that account for predictions following uncontrolled layout detection. However, in real-world scenarios, the accuracy is notably reduced by issues such as Marginalia_MetaData texts being split in half, resulting in less than 60% accuracy. In the MainText category, the primary errors would stem from incomplete or noisy segmentation of lines, often caused by line polygons that cut through characters.

5 Conclusion

The dataset used in these experiments presents substantial challenges, aligning with the project's ambitious objectives. This dataset has been made publicly available, providing a base and benchmark for developing versatile HTR models for Chinese historical documents. While results for reading order need further enhancement, a local algorithmic approach proves adequate for the task when supplemented by additional semantic zone detection, reaching equivalent or better results than ML-based approaches but without training. Analysis of Chinese layouts is generally effective but lacks robustness in high-density text situations. Future plans include employing Transformer-based layout analysis on this dataset. Text recognition leveraged a GAN-supported CRNN, achieving an average accuracy of 98.45% (±1.9%). A refined analysis of the impact of character frequency in the dataset on character recognition rates shall contribute to a better understanding of the choices to be made at the stage of dataset selection. Finally, while the models are not adaptable to out-of-domain applications, they meet the specific editorial requirements of the project effectively. Given the great variety of the dataset, these models stand as great candidates to be tested and fine-tuned for ancient Chinese manuscript recognition.

Acknowledgments. The CHI-KNOW-PO Project was funded by the University of Strasbourg Institute for Advanced Studies (USIAS) and the COLLEX-Persée. It has been conducted in collaboration with three libraries in France, namely, the Bibliothèque universitaire des langues et civilisations (BULAC, Paris), the Bibliothèque nationale et universitaire de Strasbourg (BNU, Strasbourg), and the Bibliothèque de l'institut des hautes études chinoises at the Collège de France (BIHEC, Paris). The Calfa start-up was in charge of developing HTR models.

Data availability Webpage of the project: https://www.collexpersee.eu/projet/chi-know-po-corpus/ ; Gitlab of the project: https://gitlab.huma-num.fr/chi-know-po ; Dataset: https://github.com/calfa-co/chi-know-po

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Boillet, M., Kermorvant, C., Paquet, T.: Multiple document datasets pre-training improves text line detection with deep neural networks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 2134–2141. IEEE (2021)
2. Breuel, T.M.: High performance document layout analysis. In: Proceedings of the Symposium on Document Image Understanding Technology. vol. 5 (2003)
3. Brisson, C., Constant, F., Bui, M.: Chinese Historical documents Automatic Transcription (CHAT) models (2023). https://doi.org/10.5281/zenodo.8383732
4. Chang, B., Zhang, Q., Pan, S., Meng, L.: Generating handwritten chinese characters using cyclegan. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 199–207. IEEE (2018)
5. Drège, J.P.: Le livre manuscrit et les débuts de la xylographie. In: Le livre et l'imprimerie en Extrême-Orient et en Asie du Sud. Société des bibliophiles de Guyenne (1986)
6. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.G.: Svtr: Scene text recognition with a single visual model. Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence (IJCAI-31) (2022), arXiv preprint arXiv:2205.00159
7. Fogel, S., Averbuch-Elor, H., Cohen, S., Mazor, S., Litman, R.: Scrabblegan: Semi-supervised varying length handwritten text generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
8. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019)
9. Han yu da zi dian bian ji wei yuan hui 漢語大字典編輯委員會 (The Editorial Committee of the Large dictionary of Chinese characters): 'Hanyu da zidian' 漢語大字典 (Large dictionary of Chinese characters). Sichuan cishu chubanshe (2010)
10. Henke, K., Arnold, M.: Jing bao ground truth -text block crops and annotations (2023). https://doi.org/10.11588/data/PVYWKB
11. Henke, K., Arnold, M.: Language model assisted ocr classification for republican chinese newspaper text. Journal of Digital Archives and Digital Humanities **11**, 1–19 (2023)
12. Hu, S., Wang, Q., Huang, K., Wen, M., Coenen, F.: Retrieval-based language model adaptation for handwritten chinese text recognition. IJDAR **26**(2), 109–119 (2023)
13. Huang, M.L. .: 'siku quanshu' tenglu shufa fengmao fenlei chutan - yi wenyuange ben wei zhu 《四庫全書》謄錄書法風貌分類初探——以文淵閣本為主 (manuscript calligraphy styles of siku quanshu based on wenyuan pavilion version). Zhongguo xueshu niankan 中國學術年刊 **40**(2), 27–57 (2018)
14. Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLO (jan 2023), https://github.com/ultralytics/ultralytics
15. Kiessling, B., Tissot, R., Stokes, P., Ezra, D.S.B.: escriptorium: An open source platform for historical document analysis. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19. IEEE (2019)
16. Lee, A., Yu, H., Min, G.: An algorithm of line segmentation and reading order sorting based on adjacent character detection: A post-processing of ocr for digitization of chinese historical texts. Journal of Cultural Heritage **67**, 80–91 (2024)
17. Ma, H.Y., Huang, H.H., Liu, C.L.: Reading between the lines: Image-based order detection in ocr for chinese historical documents. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 23808–23810 (2024)

18. Pelliot, P.: Les débuts de l'imprimerie en Chine. Imprimerie Nationale - Adrien Maisonneuve (1953)
19. Quirós, L.: Multi-task handwritten document layout analysis. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18). p. 1057–1063 (2018), arXiv preprint arXiv:1806.08852
20. Quirós, L., Vidal, E.: Reading order detection on handwritten documents. Neural Computing and Applications **34**(12), 9593–9611 (2022)
21. Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., Bai, X.: Icdar2017 competition on reading chinese text in the wild (rctw-17). In: 2017 14th iapr international conference on document analysis and recognition (ICDAR). vol. 1, pp. 1429–1434. IEEE (2017)
22. Sturgeon, D.: Large-scale optical character recognition of pre-modern chinese texts. International Journal of Buddhist Thought & Culture **28**, 11–44 (2018)
23. Su, T., Zhang, T., Guan, D.: Corpus-based hit-mw database for offline recognition of general-purpose chinese handwritten text. International Journal of Document Analysis and Recognition (IJDAR) **10**, 27–38 (2007)
24. Tang, C.W., Liu, C.L., Chiu, P.S.: Hrcenternet: An anchorless approach to chinese character segmentation in historical documents. In: 2020 IEEE International Conference on Big Data (Big Data). pp. 1924–1930. IEEE (2020)
25. Tang, C.W., Liu, C.L., Chiu, P.S.: Hrregionnet: Chinese character segmentation in historical documents with regional awareness. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16. pp. 3–17. Springer (2021)
26. Twitchett, D.C.: Printing and publishing in medieval China. Frederic C. Beil (1983)
27. Vidal-Gorène, C., Camps, J.B., Clérice, T.: Synthetic lines from historical manuscripts: An experiment using gan and style transfer. In: International Conference on Image Analysis and Processing. pp. 477–488. Springer (2023)
28. Vidal-Gorène, C., Dupin, B., Decours-Perez, A., Riccioli, T.: A modular and automated annotation platform for handwritings: Evaluation on under-resourced languages. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16. pp. 507–522. Springer (2021)
29. Wang, R., Fujii, Y., Bissacco, A.: Text reading order in uncontrolled conditions by sparse graph segmentation. In: International Conference on Document Analysis and Recognition. pp. 3–21. Springer (2023)
30. Wang, Z., Xu, Y., Cui, L., Shang, J., Wei, F.: Layoutreader: Pre-training of text and layout for reading order detection. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. p. 4735 - 4744 (2021). https://doi.org/10.18653/v1/2021.emnlp-main.389, https://aclanthology.org/2021.emnlp-main.389, arXiv preprint arXiv:2108.11591
31. Yu, H., Wang, X., Li, B., Xue, X.: Chinese text recognition with a pre-trained clip-like model through image-ids aligning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11943–11952 (2023)
32. Zhang, N., Cheng, H., Chen, J., Jiang, Z., Huang, J., Xue, Y., Jin, L.: M2doc: A multi-modal fusion approach for document layout analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7233–7241 (2024)
33. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV) (2017)