传统与未来的结合：
基于滤波后配合AMPD与PageNet的古籍OCR

主讲人：Shanghai-C57团队 沈佳琦

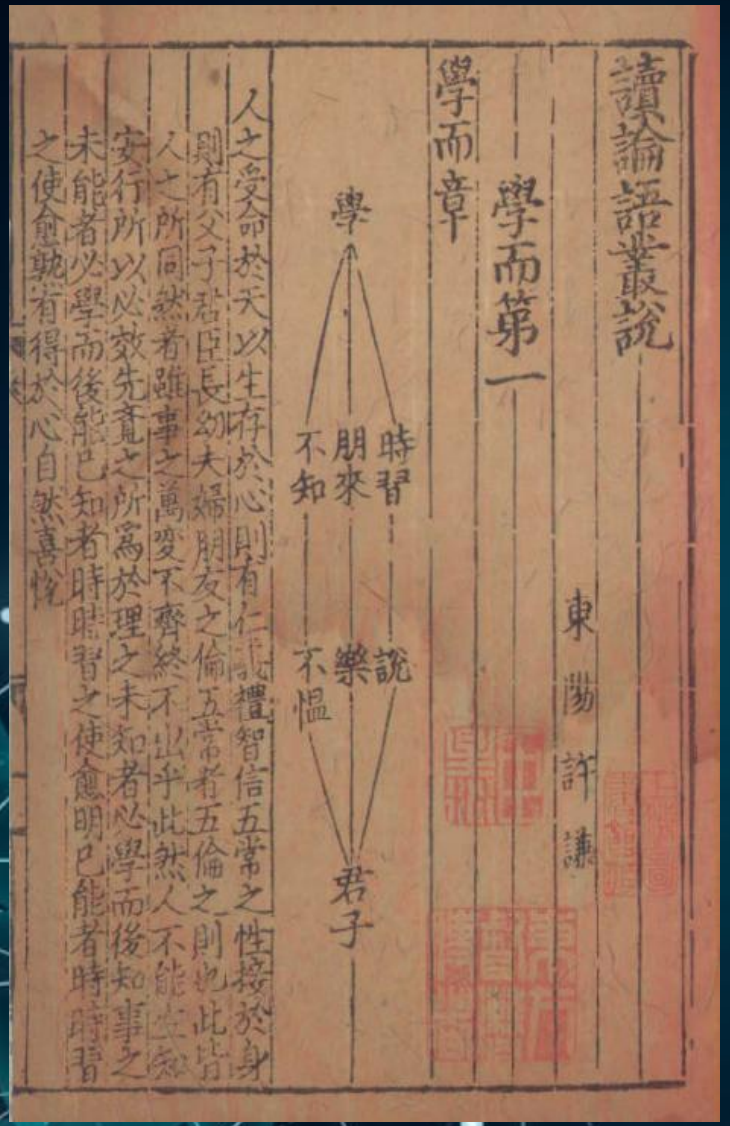
赛题介绍与分析

本组的赛道为古籍OCR，需要尽可能完整地提供图像信息，可包括但不限于正文文本、双行小字、朱批、文字坐标等古籍图像中所包含的信息。 在本次竞赛主办方提供的数据集中涵盖了多种可能会影响OCR识别准确性的数据变量，其中有：

1. 楷书与草书
2. 扫描后二值图像和照相后原图像
3. 有无边框修饰的图像
4. 字体的清晰程度

方案选择

由于竞赛数据集字体多变且部分排列不规则，传统汉字特征提取方法（四边码、粗外围、粗网格、笔画密度特征等）并不能十分有效地提取出能够进行大规模的多分类特征，且本次竞赛并没有提供数据集标签，所以在基于图像本身特征进行数据的预处理后使用预训练模型成为了最佳选择。



讀論語叢說

學而第一

東陽許謙

學而章

學 時習 說 朋來 樂 不知 不慍 君子

人之受命於天以生存於心則有仁義禮智信五常之性接於身則有父子君臣長幼夫婦朋友之倫五常者五倫之則也此皆人之所同然者雖事之萬變不齊終不出乎此然人不能安然安行所以必效先覺之所爲於理之未知者必學而後知事之未能者必學而後能已知者時時習之使愈明已能者時時習之使愈熟有得於心自然喜悅

有边框的非二值图像

經內。於前四中。二三所攝。八意之內。正唯三七。無二五八。然又長行偈頌相望。有五對之例。謂有無。廣略。離合。先後。隱顯。至文。詳而知之。清涼云又長行偈頌有十例五對。謂有無。廣略。離合。先後。為八。九或超開。十或頌已重頌。故釋頌文。不可一例。上下準之。此頌九行。應頌諷頌。間雜相糅。連環起伏。展轉相生。大意破無明法執。令觀門明淨。備行謹入。一真法界也。二正頌標宗破執二比量正破經真性有為空。緣生故如幻。無為無起滅。不實如空華。國此文正破無明法執。執為無為。有實體者。皆迷真性。一真法界。本非對待。故此雙破。即是解結之所因也。此中四句。前二句破有為

無边框的二值图像

PageNet 介绍

PageNet 为IJCV 2022中《PageNet: Towards End-to-End Weakly Supervised Page-Level Handwritten Chinese Text Recognition》中所提供的模型。本模型的优点为

1. 采用character detection同时采用time line detection的信息由此可以在识别出字体的同时保证同行或同列中文字阅读顺序的正确。
2. 训练数据集庞大，有效包含了各种古文手写字体类型

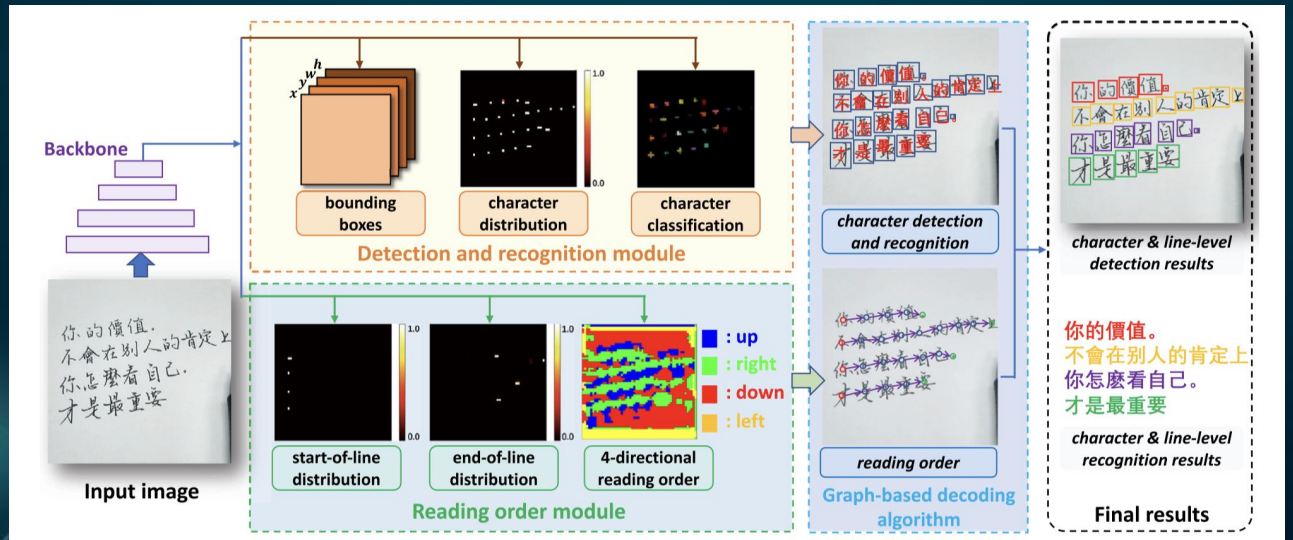


Fig. 1 Overall architecture of PageNet. Based on the features extracted by the backbone network, the detection and recognition module predicts the character detection and recognition results, while the reading order module predicts the reading order between characters. Combining these two predictions, the graph-based decoding algorithm outputs the final results containing detection and recognition results at both the character and line levels

直接使用PageNet所存在的问题

1. 尽管可以进行文本列内顺序的预测，但是PageNet对列整体顺序仍存在一定困难。与此同时，由于古文阅读顺序和现代阅读顺序的差异，输出的文本会顺序异常逻辑混乱。

2. 在本次主办方提供数据集中经常会出现手写字体与页边框相重合情况。在这种情况下PageNet会将与边框重叠的文字错误识别为边框而导致忽略，不做输出。

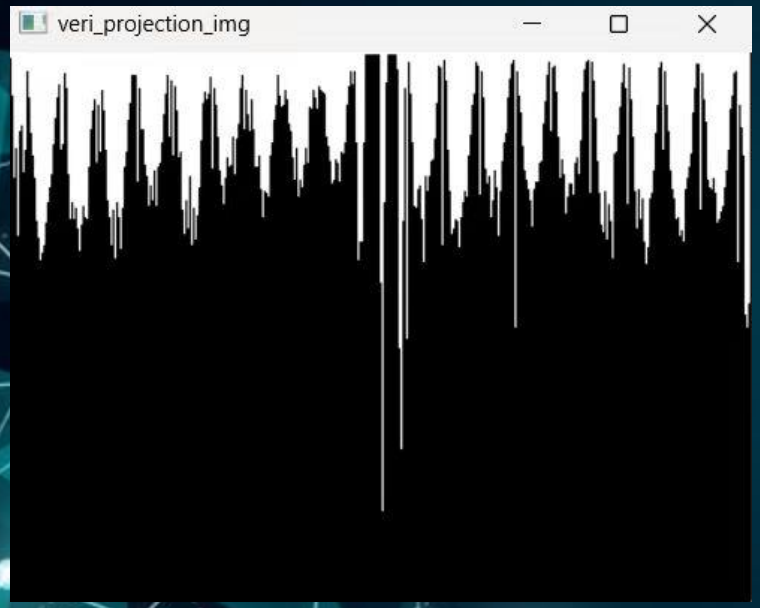
3. 处理低分辨率图像准确率低于处理高分辨率图像。

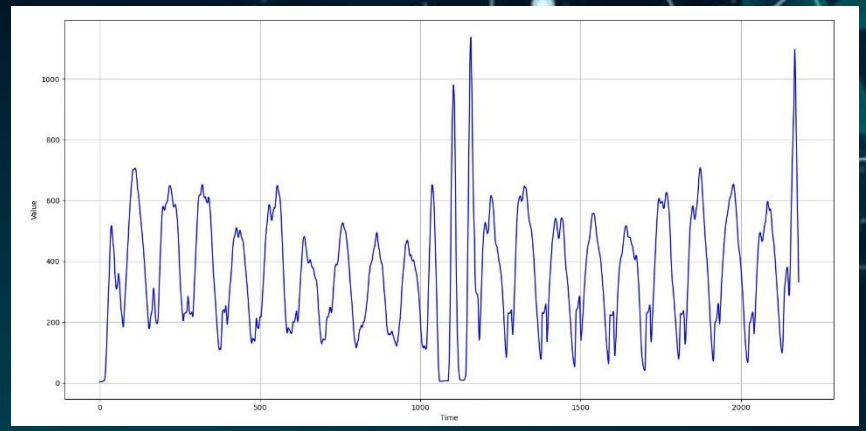
使用AMPD算法进行分列

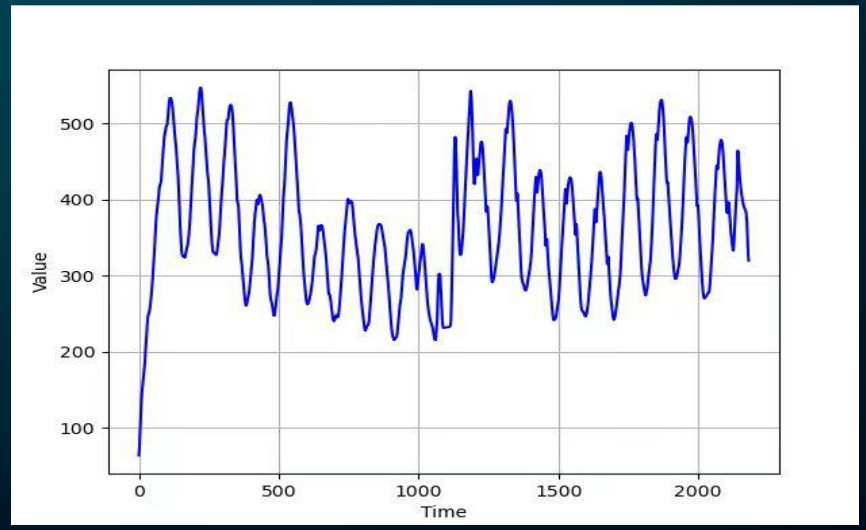
经过对于汉语手写古体数据集的分析我们可以看出，手写字体都会做出明显的“列分隔”。如果我们能够有效将每列识别出并规定列输出顺序就能保证输出符合现代汉语语序的逻辑性。与此同时，由于破坏了边框的连续性和完整性，与边框相重叠的文字也能够被成功识别。在分割图像后我们只对单列进行识别，可以尽可能地保留图像的高分辨率，从而做出更加准确的输出。

在此部分我们的基本思路采用了基于像素分布的分割方法。在对原图像进行二值化处理，进行了竖直向的膨胀与全图的腐蚀，之后对列像素进行统计得到分布曲线。在此分布曲线中峰值即为有文字区域，波谷即为分割边界。由于统计出的分布曲线存在噪声，我们对其进行了多次高频滤波。

滤波过程示意


veri_projection_img


Value
Time

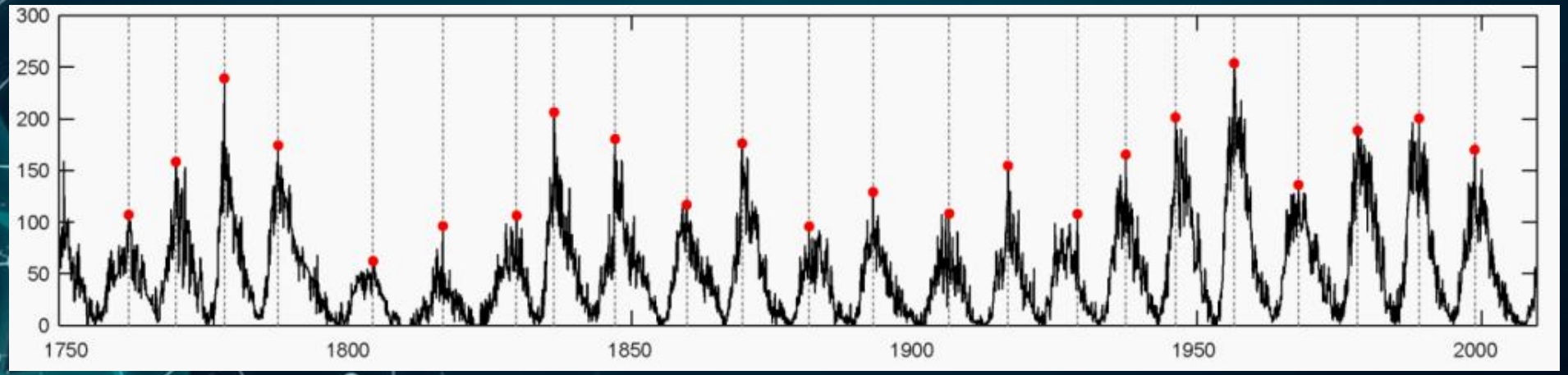

Value
Time
100
200
300
400
500
0
500
1000
1500
2000

AMPD算法介绍

AMPD，即自动多尺度峰值查找算法。
其优势是：

（1）算法本身对信号具有良好的自适应性，唯一的假设是信号是周期的或者准周期的。在古文识别中，列间距一般固定，能够提供良好的周期性辅助进行列分割；

（2）抗噪能力强，且对周期性的要求也不是很高。



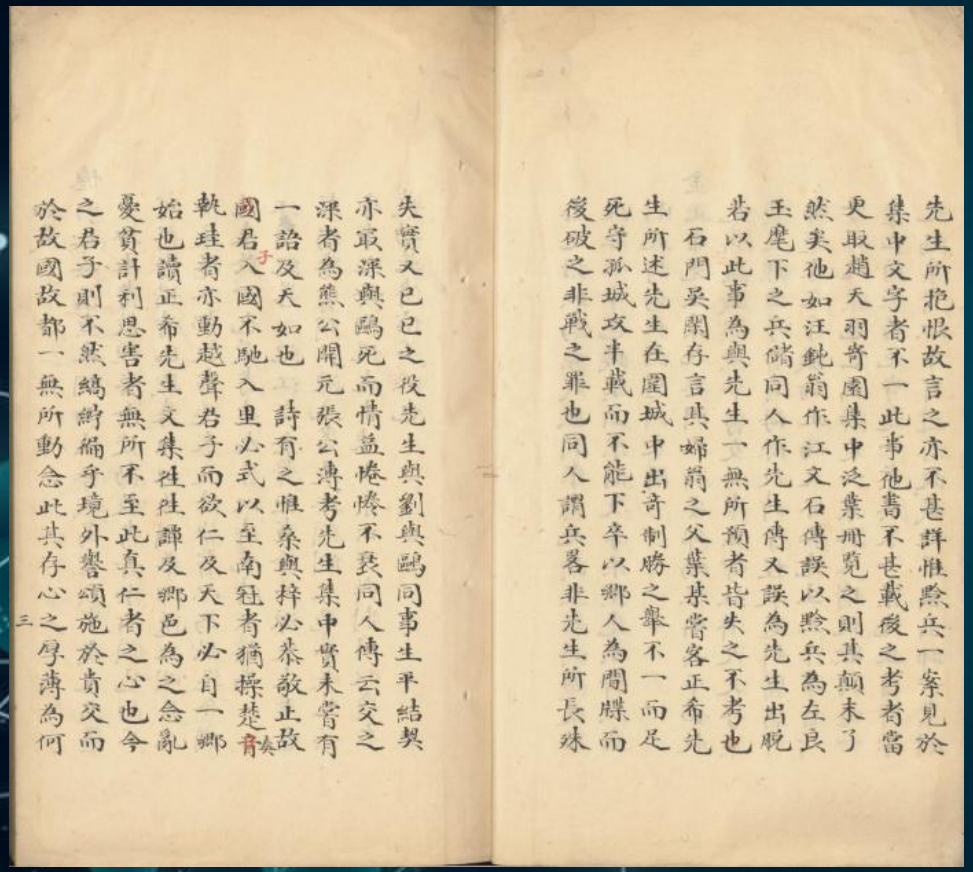
AMPD效果示意图

遂獨慮是經之傳出於朱子之門人苟一豪

既剖析而著明之矣先生受學之夕聞義之

本朝設科取士並紬衆說而專用古注疏蔡氏

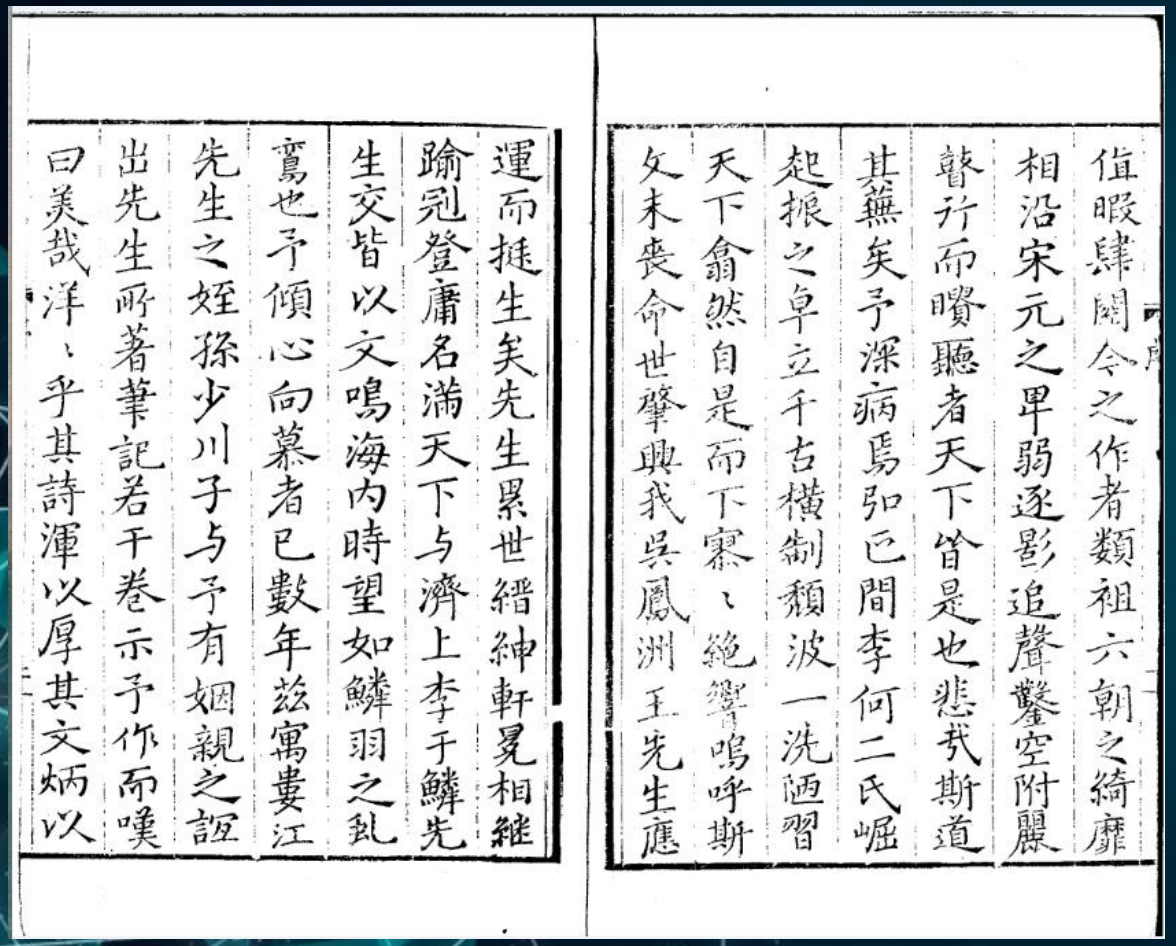
效果展示：



先生所抱恨故言之亦不甚詳惟黔兵一案見於
集中文字者不一此事他書不甚載後之考者當
更取趙天羽寄園集中泛葉冊覽之則其顛末了
然矣他如汪鈍翁作江文石傳誤以黔兵為左良
玉麾下之兵儲同人作先生傳又誤為先生出脫
若以此事為與先生一無所預者皆失之不考也
石門吳閑存言其婦翁之父葉某皆客正希先
生所述先生在圍城中出奇制勝之舉不一而足
死守孤城攻半載而不能下卒以鄉人為魁牒而
後破之非戰之罪也同人謂兵畧非先生所長殊
失實又已已之役先生與劉與鷗同事生平結契
深者為熊公開元張公溥考先生集中實未嘗有
語及天如也詩有之惟桑與梓必恭敬止故
亦取深與鷗死而情益惓惓不衰同人傳云交之
深者為熊公開元張公溥考先生集中實未嘗有
一語及天如也詩有之惟桑與梓必恭敬止故
國君入國不馳入里以式以至南冠者猶操楚音
執珪者亦動越聲君子而欲仁及天下必自一鄉
始也讀正希先生文集往往譯及鄉邑為之念亂
憂貧計利思害者無所不至此真仁者之心也今
之君子則不然編紵徧乎境外譽頌施於貴交而
於故國故都一無所動念此其存心之厚薄為何

```
788262/D0000005.JPG
先生所抱恨故言之亦不甚詳惟黔兵一案見於  [1882, 1972]
集中文字者不一此事他書不甚載後之考者當  [1803, 1882]
更取趙天羽寄園集中泛葉冊覽之則其顛末了  [1718, 1803]
然矣他如汪鈍翁作江文石傳誤以黔兵為左良  [1640, 1718]
玉麾下之兵儲同人作先生傳又誤為先生出脫  [1469, 1640]
若以此事為與生一無所預者皆失之不考也    [1469, 1640]
石門吳閑存言其婦翁之父葉某皆客正希先    [1393, 1469]
生所述先生在圍城中出奇制勝之舉不一而足  [1311, 1393]
死守孤城攻半載而不能下卒以鄉人為魁牒而  [1228, 1311]
後破之非戰之罪也同人謂兵畧非先生所長殊  [1144, 1228]
失實又已已之役先生與劉與鷗同事生平結契  [556, 880]
深者為熊公開元張公溥考先生集中實未嘗有  [556, 880]
語及天如也詩有之惟桑與梓必恭敬止故  [556, 880]
亦取深與鷗死而情益惓惓不衰同人傳云交之  [556, 880]
執珪者亦動越聲君子而欲仁及天下必自一鄉  [389, 556]
國君入國不馳入里以式以至南冠者猶操楚音  [389, 556]
始也讀正希先生文集往往譯及鄉邑為之念亂  [307, 389]
憂貧計利思害者無所不至此真仁者之心也今  [230, 307]
之君子則不然編紵徧乎境外譽頌施於貴交而  [64, 230]
於故國故都一無所動念此其存心之厚薄為何  [64, 230]
```

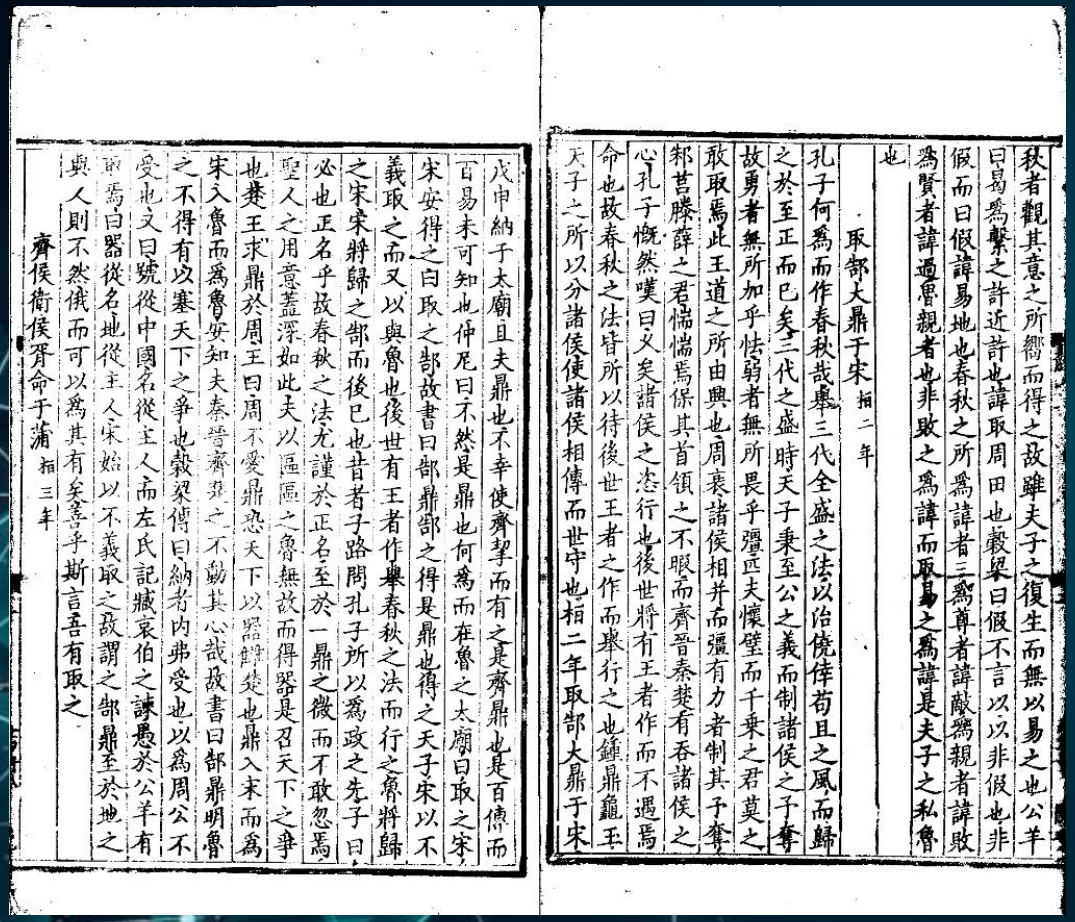
效果展示：



值暇肆閱今之作者類祖六朝之綺靡
相沿宋元之卑弱逐曷追聲鑿空附麗
聲分而賢聽名天下合是也悲夫斯道
其無矣予深病焉弘色間李何二氏崛
起振之阜立千古橫制頹波一洗陋習
天下翕然自是而下寖色響鳴呼斯
文未喪命世肇興我吳鳳洲王先生應
運而挺生矣先生累世縉紳軒冕相繼
踰冠登庸名滿天下与濟上李于鱗先
生交皆以文鳴海内時望如鱗羽之鈍
寬也予傾心向慕者已數年茲寓婁江
先生之姪孫少川子与予有姻親之誼
出先生所著筆記若干卷示予作而嘆
曰美哉洋乎其詩渾以厚其文炳以

```
786201/A0003A.TIF
值暇肆閱今之作者類祖六朝之綺靡   [2071, 2224]
相沿宋元之卑弱逐曷追聲鑿空附麗   [1922, 2071]
聲分而賢聽名天下合是也悲夫斯道   [1753, 1922]
其無矣予深病焉弘色間李何二氏崛   [1613, 1753]
起振之阜立千古橫制頹波一洗陋習   [1467, 1613]
天下翕然自是而下寖色響鳴呼斯     [1329, 1467]
久未喪命世肇興我受爲洲王光生應   [920, 1329]
運而挺生矣先生累世縉紳軒冕相繼   [920, 1329]
踰別登庸名滿天下与濟上李于鱗先   [779, 920]
生交皆以文鳴海内時望如鱗羽之鈍   [612, 779]
寬也矛傾心向慕者已數年茲寓婁江   [476, 612]
先生之姪孫少川子与子有如親之設   [306, 476]
出先生所著筆記若干卷示予作而嘆   [171, 306]
曰美哉洋乎其詩渾以厚其文炳以     [41, 171]
```

效果展示：



804782-8/D0000017.TIF
者觀其意之所而得之故雖夫子之復生而無以易之也羊 [1836, 1900]
曰曷爲槃之許近許也諱取周田也穀梁曰假不言以以非假也非 [1772, 1836]
假而曰假諱易地也春秋之所爲者三爲尊者爲親者諒敗 [1708, 1772]
爲賢者諱過魯親者也非敗之爲譯而取易之爲諱是夫子之私魯 [1640, 1708]
取部大鼎子宋獨二年 [1519, 1640]
孔子何爲而作春秋哉舉三代全盛之法以治饒倖苟且之風而歸 [1452, 1519]
之於至正而已矣三代之盛時天子乘至公之義而制諸侯之子奪 [1388, 1452]
故勇者無所加乎性弱者無所乎彊匹夫懷璧而千乘之君莫之 [1326, 1388]
敢取焉此王道之所由興也周衰諸侯相并而彊有力者制其子奪 [1263, 1326]
邪苴勝薛之君惴惴焉保其首領之不暇而齊晉秦楚有吞諸侯之 [1198, 1263]
孔子慨然嘆曰久矣諸侯之恣行後世將有王者作而不遇焉 [1132, 1198]
命故春秋之法皆所以待後世王者之作而舉行之也鐘鼎龜玉 [1068, 1132]
天子之所以分諸侯使諸侯相傳而世守也相二年取部大鼎于失 [963, 1068]
戊申納于太廟且夫鼎也不幸使齊挈而有之是齊鼎也是百傳而 [866, 963]
百易未可知也仲尼曰不然是鼎也何爲而在魯之太廟曰取之宋 [801, 866]
宋安得之曰取之部故書曰部鼎部之得是鼎也得之天子宋以不 [734, 801]
義取之而又以與魯也後世有王者作舉春秋之法而行之魯將歸 [670, 734]
之宋宋將歸之而後已也昔者子路問孔子所以爲政之先子曰 [607, 670]
必也正名乎故春秋之法尤謹於正名至於一鼎之微而不敢忽焉 [544, 607]
聖人之用意蓋深如此夫以區區之魯無故而得器是召天下之爭 [480, 544]
也楚王求鼎於周王曰周不愛鼎恐天下以器餘楚也鼎入末而爲 [414, 480]
[414, 480]
宋入魯而爲魯安知夫秦晉齊之不動其心哉故書曰部鼎明魯 [352, 414]
之不得有以塞天下之爭也穀梁傳曰納者內弗受也以爲周公不 [287, 352]
受也又曰號從中國名從主人而左氏記臧哀伯之諫愚於公羊有 [222, 287]
印焉自器從名地從主人宋始以不義取之故謂之部鼎至於地之 [158, 222]
與人則不然佛而可以爲其有矣善乎斯言吾有取之 [93, 158]
齊侯衛侯胥命于蒲相三年 [41, 93]

中途遇到的困难：

1. 技术路线选则：刚开始认为在无标签情况下使用传统特征提取算法效果可能更优，基础技术路线是爬取楷书和行书手写字体特征并采用多种传统特征组合，但是在做完实现后发现由于分类数过多，传统特征的容错率分辨率整体上偏小，目标汉字基本只能进入前50首选，效果不好。

2. 空白判定：条带分割后会出现分割出边缘或中间空白的情况，我们设定了一个像素占比阈值来对其进行判定来决定是否输入网络。

本次比赛的收获：

1. 本次比赛让我能够比较深入地分析汉字古文典籍的模式特点。之前有做过英文表格OCR之类的项目，单显然汉字古文典籍的模式和英文模式有很大不同，汉字古文的形体、笔画和结构都有很大的差异和复杂性，给OCR技术带来了很高的要求和难度。本次比赛给了我一个机会去仔细地思考古文的独特模式并根据其模式进行技术路线设计。

2. 本次比赛中我也学习了汉字古文的历史、特点和变化，感受了汉字古文的美感和韵味，也了解了汉字古文的价值和意义。我认识到，汉字古文是中华文化的重要组成部分，是中华民族的瑰宝和骄傲，也是人类文明的贡献和遗产。我不仅提高了自己的OCR技术水平，也增进了自己的汉字古文文化素养。最后我希望这次比赛能够为汉字古文的保护、传承和发展做出一些贡献，也能够激发更多人对汉字古文的兴趣和热爱。

感谢聆听