支持表格識別！這款OCR開源神器發布！

計算機視覺Daily 昨天

相信大家在工作生活中經常會遇到表格識別的問題，比如導師說，把下面PDF文件裡面的表格取出來整理成Excel表。

(a) Total-Text (b) Total-Text (c) CTW1500 (d) CTW1500

Figure 8. Visual experimental results. The blue contours are boundary proposals, and the green contours are final detection boundaries.

Table 6. Experimental results on CTW-1500.

Methods	Ext	R	P	F	FPS
TextSnake [18]	Syn	**85.3**	67.9	75.6	-
CSE [17]	MLT	76.1	78.7	77.4	0.38
LOMO[40]	Syn	76.5	85.7	80.8	4.4
ATRR[35]	Sy-	80.2	80.1	80.1	-
SegLink++ [28]	Syn	79.8	82.8	81.3	-
TextField [37]	Syn	79.8	83.0	81.4	6.0
MSR[38]	Syn	79.0	84.1	81.5	4.3
PSENet-1s [33]	MLT	79.7	84.8	82.2	3.9
DB [12]	Syn	80.2	86.9	83.4	22.0
CRAFT [2]	Syn	81.1	86.0	83.5	-
TextDragon [5]	MLT+	82.8	84.5	83.6	-
PAN [34]	Syn	81.2	86.4	83.7	**39.8**
ContourNet [36]	-	84.1	83.7	83.9	4.5
DRRG [41]	MLT	83.02	85.93	84.45	-
TextPerception[23]	Syn	81.9	87.5	84.6	-
Ours	-	80.57	87.66	83.97	12.08
Ours	Syn	81.45	**87.81**	84.51	12.15
Ours	MLT	83.60	86.45	**85.00**	12.21

CTW1500. In testing, the threshold th_s is set to 0.8. Representative visible results are shown in Fig. 8 (c) and (d), which indicate our method precisely detects boundaries of long curved text with line-level. The quantitative results are listed in Tab. 6. Compared with the previous sate-of-the-art methods [12, 34, 36], our approach achieves

Table 7. Experimental results on MSRA-TD500.

Methods	R	P	F	FPS
SegLink [26]	70.0	86.0	77.0	8.9
PixelLink [4]	73.2	83.0	77.8	-
TextSnake [18]	73.9	83.2	78.3	1.1
TextField [37]	75.9	87.4	81.3	5.2
MSR[38]	76.7	87.4	81.7	-
FTSN [3]	77.1	87.6	82.0	-
LSE[30]	81.7	84.2	82.9	-
CRAFT [2]	78.2	88.2	82.9	8.6
MCN [16]	79	88	83	-
ATRR[35]	82.1	85.2	83.6	-
PAN [34]	83.8	84.4	84.1	30.2
DB[12]	79.2	**91.5**	84.9	**32.0**
DRRG [41]	82.30	88.05	85.08	-
Ours (SynText)	80.68	85.40	82.97	12.68
Ours (MLT-17)	**84.54**	86.62	**85.57**	12.31

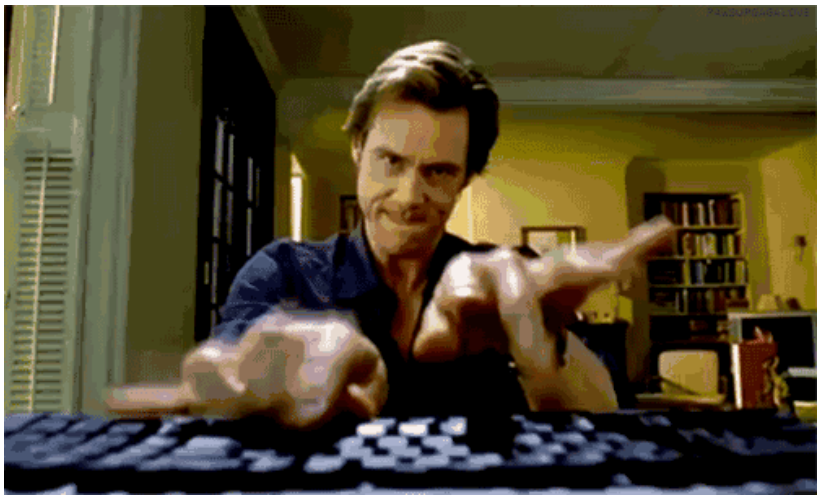
5. Conclusion

In this paper, we propose a novel adaptive boundary proposal network for arbitrary shape text detection, which adopt an boundary proposal model to generate coarse boundary proposals, and then adopt an adaptive boundary deformation model combined with GCN and RNN to per-

也可能會遇到，公司領導或者客戶發來一張截圖，需要裡面的表格取出來轉成Excel表。

品名	单位	入库		出库	
		日期	数量	日期	数量
钢管	米	2011-02-27	510.00		
钢管	米	2011-03-05	82.00		
钢管	米	2011-03-15	1087.50		
钢管	米	2011-03-19	486.00		
钢管小计			2165.50		
扣件	个	2011-02-27	450.00		
扣件	个	2011-03-05	20.00		
扣件	个	2011-03-15	820.00		
扣件小计			1290.00		
本期租金合计（元）		498.3			
应收金额合计（元）		597.71			

這種情況下你會怎麼做呢，新建一個Excel一個一個數據敲麼，辛辛苦苦半天趕出來，領導還會來一句，怎麼這麼慢，簡直鬱悶死.....



別著急，只要稍微會一點Python代碼，這個開源項目神器拯救你！

效果展示

版面分析+表格識别

代号	项目	结果	参考值	单位
ALT	谷丙转氨酶	25.6	0--40	U/L
TBIL	总胆红素	11.2	<20	umol/L
DBIL	直接胆红素	3.3	0--7	umol/L
IBIL	间接胆红素	7.9	1.5--15	umol/L
TP	总蛋白	58.9↓	60--80	g/L
ALB	白蛋白	35.1	33--55	g/L
GLO	球蛋白	23.8	20--30	g/L
A/G	白球比	1.5	1.5--2.5	
ALP	碱性磷酸酶	93	15--112	IU/L
GGT	谷氨酰转肽酶	14.3	<50	U/L
AST	谷草转氨酶	16.3	8--40	U/L
LDH	乳酸脱氢酶	167	114--240	U/L
ADA	腺苷脱氨酶	12.6	4--24	U/L

代号	项目	结果	参考值	单位
ALT	谷丙转氨酶	25.6	0--40	UA
TBIL	总胆红素	11.2	K20	umol/
DBIL	直接胆红素	3.3	0--7	umol/L
IBIL	间接胆红素	7.9	1.5-15	umol/L
TP	总蛋白	58.94	60--80	8A
ALB	白蛋白	35.1	33-55	81
GLO	球蛋白	23.8	20--30	8
A/G	白球比	1.5	1.5--2.5	
ALP	碱性磷酶	93	15--112	IUA
GGT	谷氨酰转肽酶	14.3	K50	UA
AST	谷草转氨酶	16.3	8--40	A
LDH	乳酸脱氢	167	114-240	UA
ADA	腺甘脱氨酶	12.	--24	U/L

如圖所示，針對一張完整的PDF圖片，這個開源項目可以對文檔圖片中的文本、表格、圖片、標題與列表區域進行分類。同時還可以利用表格識別技術完整地提取表格結構信息，使得表格圖片變為可編輯的Excel文件。

不僅僅是PDF文件轉excel，如果編程能力再強一些，結合版面分析技術，PDF轉Word都不在話下。

而且使用也是非常方便，在完成Python whl包安裝之後，簡單幾行代碼即可完成快速試用。

3 PP-Structure 快速开始

3.1 命令行使用（默认参数，极简）

```
paddleocr --image_dir=../doc/table/1.png --type=structure
```

3.2 Python脚本使用（自定义参数，灵活）

```
import os
import cv2
from paddleocr import PPStructure,draw_structure_result,save_structure_res

table_engine = PPStructure(show_log=True)

save_folder = './output/table'
img_path = '../doc/table/1.png'
img = cv2.imread(img_path)
result = table_engine(img)
save_structure_res(result, save_folder,os.path.basename(img_path).split('.')[0])
```

最終結果會輸出圖片文件夾，Excel表和文字識別結果，確實是非常方便。

傳送門：

https://github.com/PaddlePaddle/PaddleOCR/blob/release/2.2/ppstructure/README_ch.md

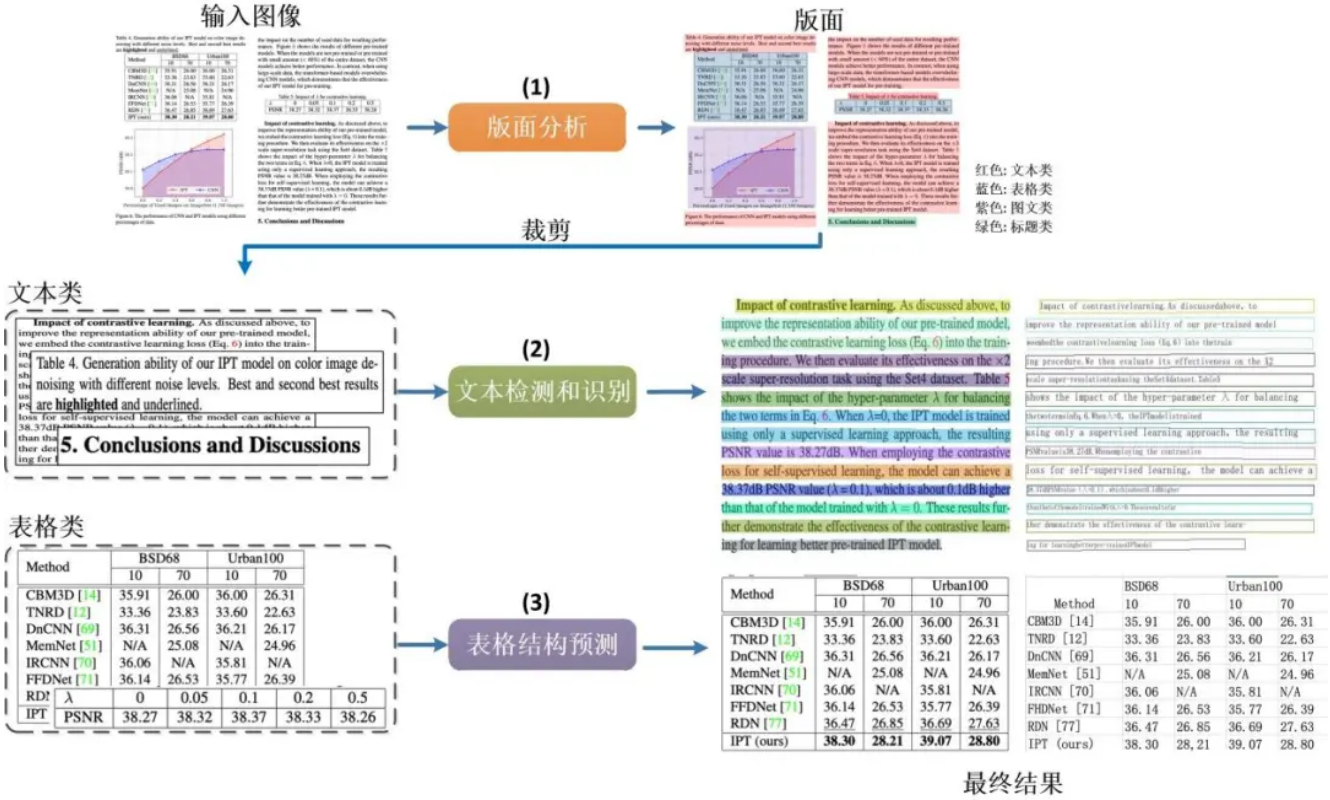
版面分析與表格識別核心技術概述

不管是版面分析還是表格識別，現有方案可大致分為基於圖像處理的傳統方法和基於深度學習的方法。

（1）传统方法：版面分析比较著名的是O'Gorman在1993年TPAMI中发表的算法Docstrum。通过自下而上的方法依次将图像中的黑白连通域划分为文字、文本行与文本块，从而得到版面布局。表格识别的传统方法通过腐蚀、膨胀等操作获得表格线、划分行列区域，然后将单元格与文本内容相结合重构为表格对象。但是传统算法主要问题在于，对于版面布局分析和表格结构的提取，图像处理的方法依赖各种阈值和参数的选择，对于不同场景下的文档图片难以保证泛化性。

（2）深度学习方法：除了直接使用检测模型来对版面内容进行分类以外，还融合了检测、分割、图神经网络、注意力机制等众多前沿技术能力。依赖算法工程师对于深度神经网络的精心设计，可以不再依赖阈值与参数，具有更好的泛化性。

PP-Structure核心技术解读



- **版面分析技术**

PP-Structure的版面分析技术，主要是对图片形式的文档进行版面分析，将文档划分为文字、标题、表格、图片以及列表5类区域（与Layout-Parser联合使用）。其核心技术思路与Layout-Parser项目密切合作，参考了Layout-Parser的工程结构设计，配合PaddleDetection开源的高效检测算法PP-YOLO v2，在数据集TableBank和PubLayNet上mAP分别达到93.6和96.2， NVIDIA Tesla P40 耗时仅需66.6ms，且可以支持用户根据自己的数据自定义训练。

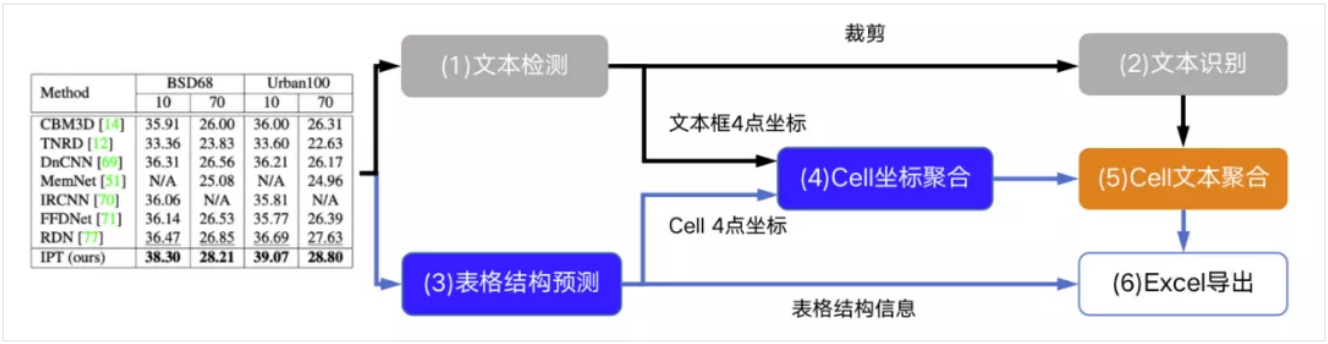
Layout-Parser是开源的基于深度学习的文档图像分析工具箱，可用于布局检测，字符识别和许多其他文档处理任务，包含大量丰富模型，支持自定义DL模型，支持多个文档布局检测数据集。

GitHub地址：

https://github.com/Layout-Parser/layout-parser

- **表格识别技术**

表格识别技术则主要使用基于注意力机制的图片描述模型RARE，整体流程如下图所示，对于其中的表格区域进行表格识别处理。



Method	BSD68		Urban100	
	10	70	10	70
CBM3D [14]	35.91	26.00	36.00	26.31
TNRD [12]	33.36	23.83	33.60	22.63
DnCNN [69]	36.31	26.56	36.21	26.17
MemNet [51]	N/A	25.08	N/A	24.96
IRCNN [70]	36.06	N/A	35.81	N/A
FFDNet [71]	36.14	26.53	35.77	26.39
RDN [77]	36.47	26.85	36.69	27.63
IPT (ours)	**38.30**	**28.21**	**39.07**	**28.80**

表格识别的难点主要在于表格结构的提取，以及将表格信息与OCR信息融合。整体流程可以分为上下两部分，其中上半部分（黑色支路）是普通的OCR过程，通过**（1）文本检测模块**对表格图片进行单行文字检测，获得坐标，然后通过**（2）文本识别模块**识别模型得到文字结果。

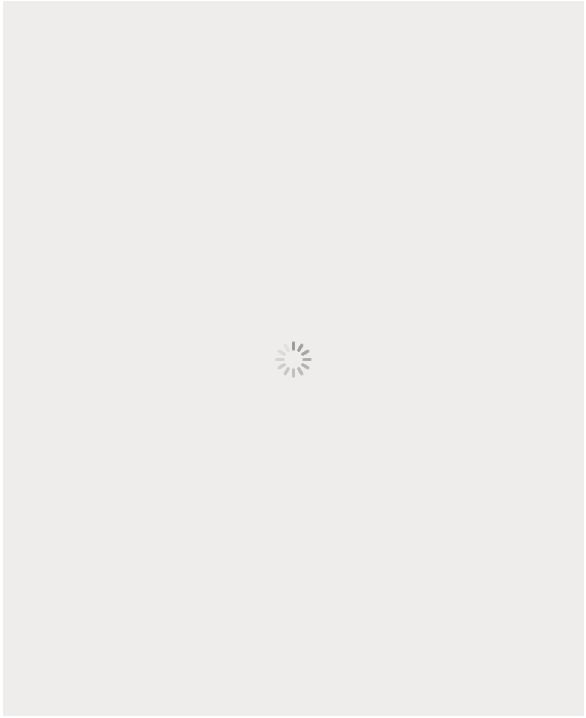
而在下半部分的在蓝色支路中，表格图片首先经过**（3）表格结构预测模块**，获得每个Excel单元格的四点坐标与表格结构信息。结合黑色支路文本检测获得的单行文字文本框4点坐标，共同输入**（4）Cell坐标聚合**模块，再通过**（5）Cell文本聚合模块**，将属于同一单元格的文本拼接在一起。最后结合表格结构信息，通过**（6）Excel导出模块**获得Excel形式的表格数据。

下面分别针每个模块分别展开介绍。

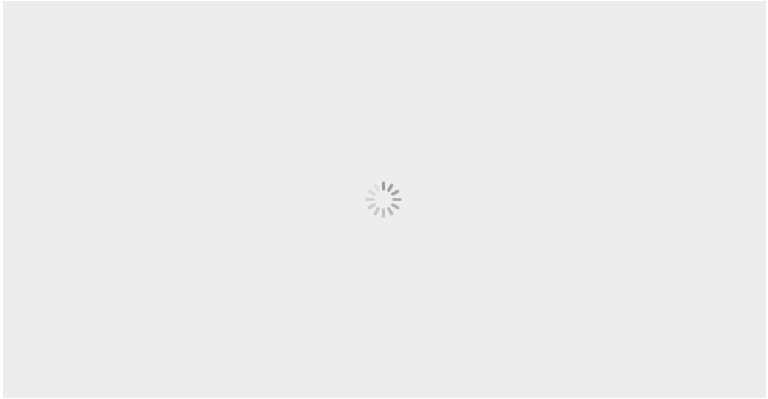
（1）文本检测模块 和**（2）文本识别模块**：

主要使用PP-OCR提供的检测和识别算法。

（3）表格结构预测模块，主要使用基于Attention的图片描述模型RARE，RARE模型可以实现：输入一张图片，通过带有注意力机制的网络输出一段文字，描述图片的内容，而针对于表格图片的图片描述网络，输入一张经过版面分析的表格图片，输出的是一串HTML字符（如下图所示）。表格的结构通过HTML的结构标记表示，其中的内容即为表格文本中的内容。通过进一步的HTML解析，可以获得每个文本的单元格四点坐标和表格结构信息。



（4）Cell坐标聚合模块，主要用来解决如何将跨行单元格的文本重新拼接在一个单元格内的问题。它通过计算由文本检测算法获得的文本框坐标（红色框）与表格结构预测模块得到的Cell坐标（蓝色框）之间的IOU和顶点距离来进行单行到多行的聚合。使用IOU判断哪些红色框同属于一个蓝色框，使用顶点距离和IOU判断红色框的排列顺序。



（5）Cell文本聚合模块，根据已有的红色文本框顺序，按照从上到下从左到右顺序利用（4）Cell坐标聚合模块的结果将（2）文本识别结果和进行拼接，这样对于多行文本的单元格内容即可拼接成一个字符串。

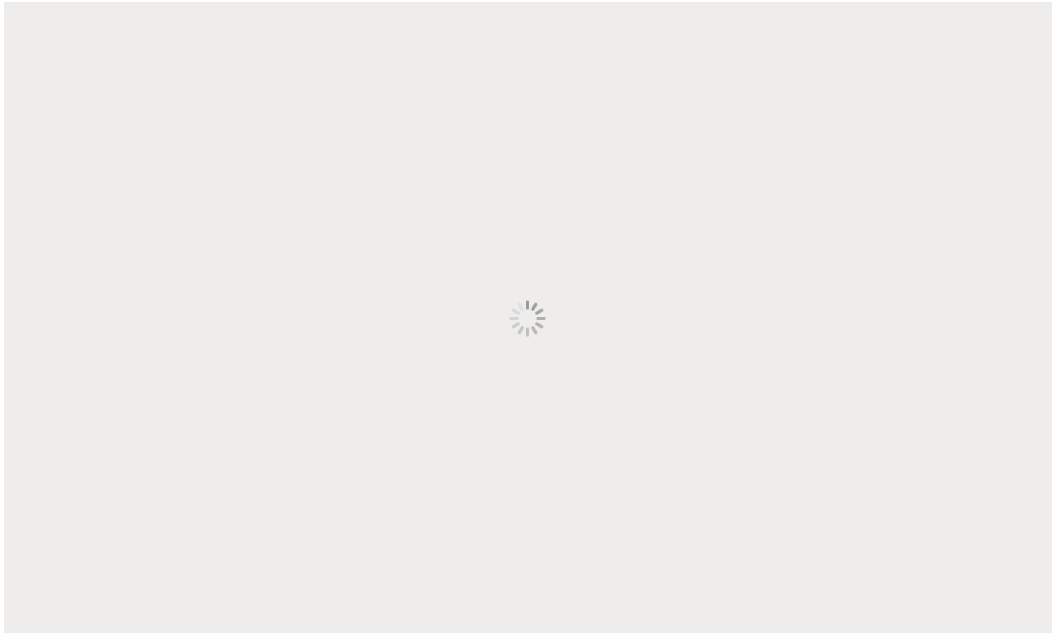
（6）Excel导出模块，将（3）表格结构预测结果html结果与（5）Cell文本聚合模块文本结果结合，最终导出为Excel输出。

以上所有内容均在PaddleOCR项目开源，目前star数量超过13.5k

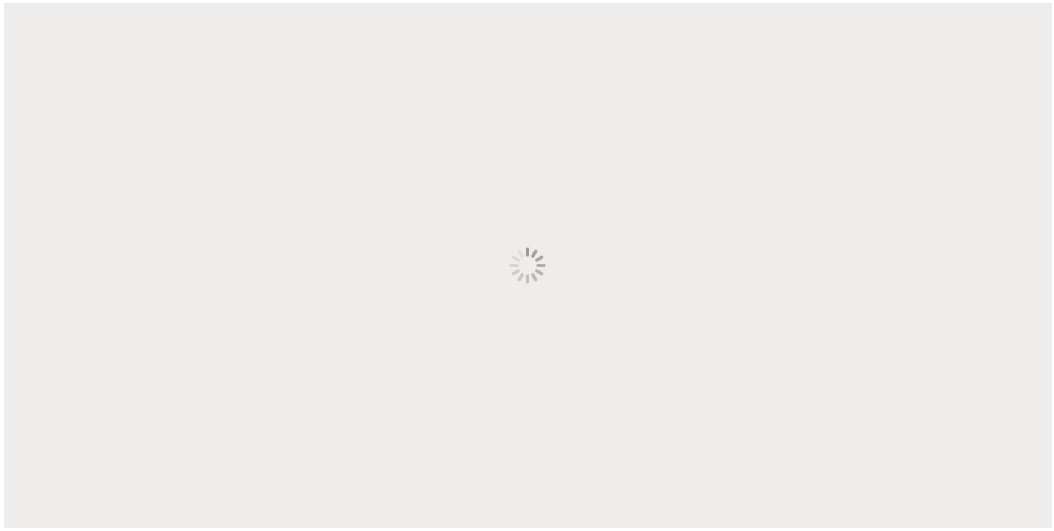
相关延伸阅读：
PaddleOCR历史表现回顾

- 2020年6月，8.6M超轻量模型发布，GitHub Trending 全球趋势榜日榜第一。
- 2020年8月，开源CVPR2020顶会算法，再上GitHub趋势榜单！
- 2020年10月，发布PP-OCR算法，开源3.5M超超轻量模型，再下Paperswithcode 趋势榜第一
- 2021年1月，发布Style-Text文本合成算法，PPOCRLabel数据标注工具，star数量突破10000+，截至目前已经达到11.5k，在《Github 2020数字洞察报告》中被评为中国GithubTop20活跃项目。

- 2021年4月，开源AAAI顶会论文PGNet端到端识别算法，Star突破13k
- 2021年8月，开源版面分析与表格识别算法

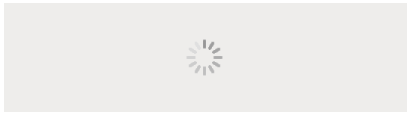


文本检测识别效果：



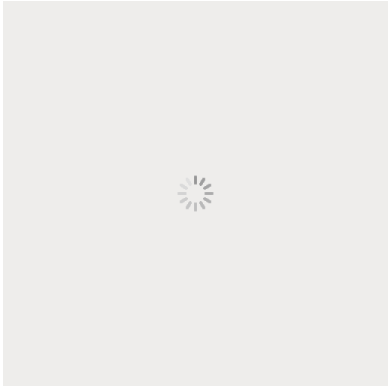
这个最强OCR项目，你值得拥有：

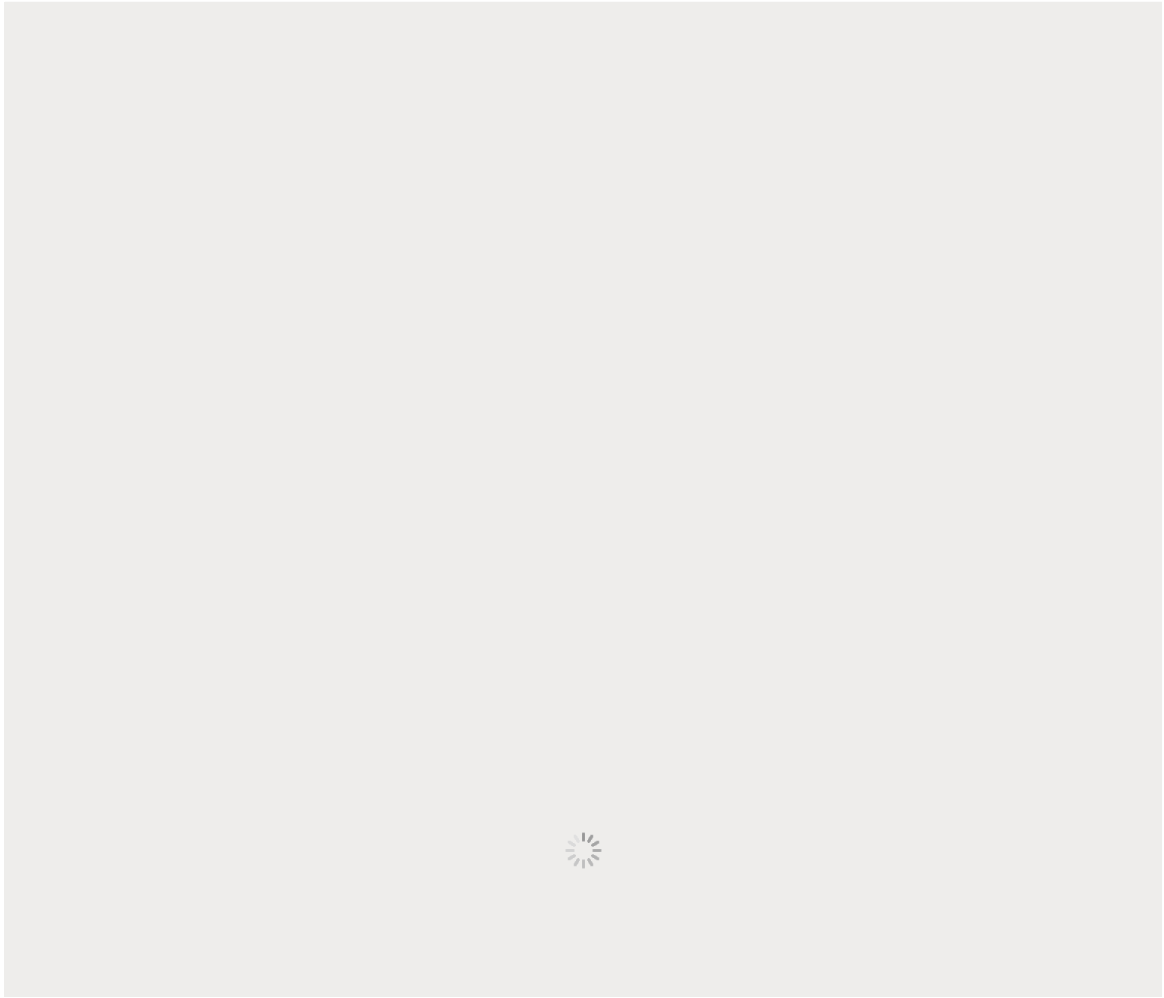
https://github.com/PaddlePaddle/PaddleOCR

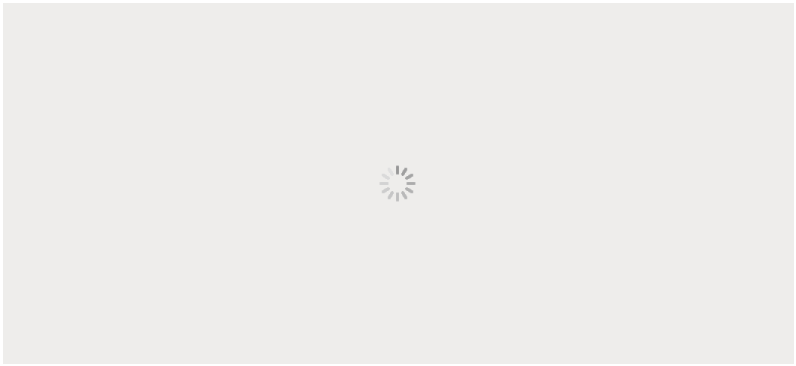


8月12日（週四）20:15-21:30

掃描二維碼報名，立即加入交流群







如果您想詳細了解更多飛槳的相關內容，請參閱以下文檔。

·PaddleOCR項目地址·

GitHub:
https://github.com/PaddlePaddle/PaddleOCR

Gitee:
https://gitee.com/paddlepaddle/PaddleOCR

.

https://www.paddlepaddle.org.cn/

閱讀原文