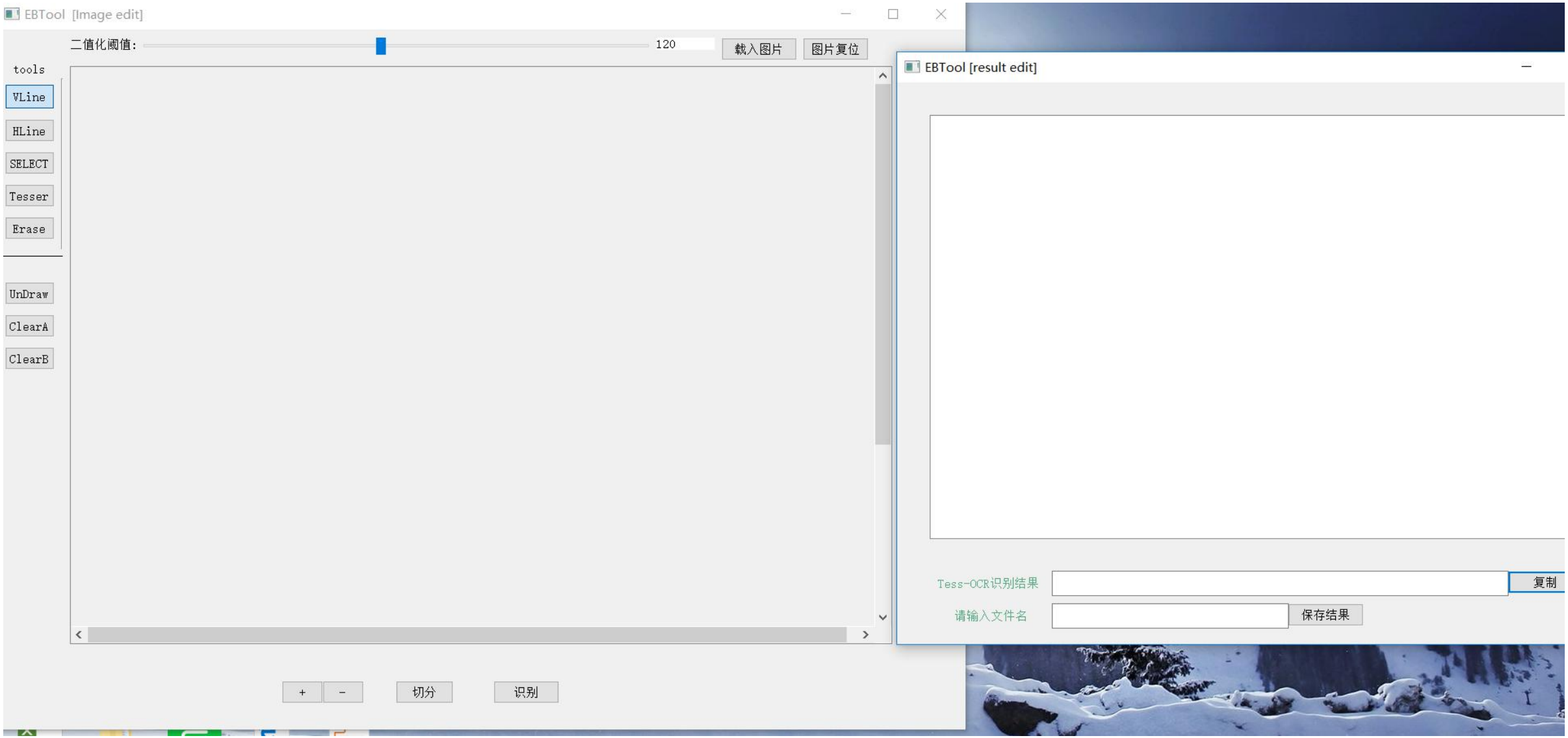
EBTool使用手册

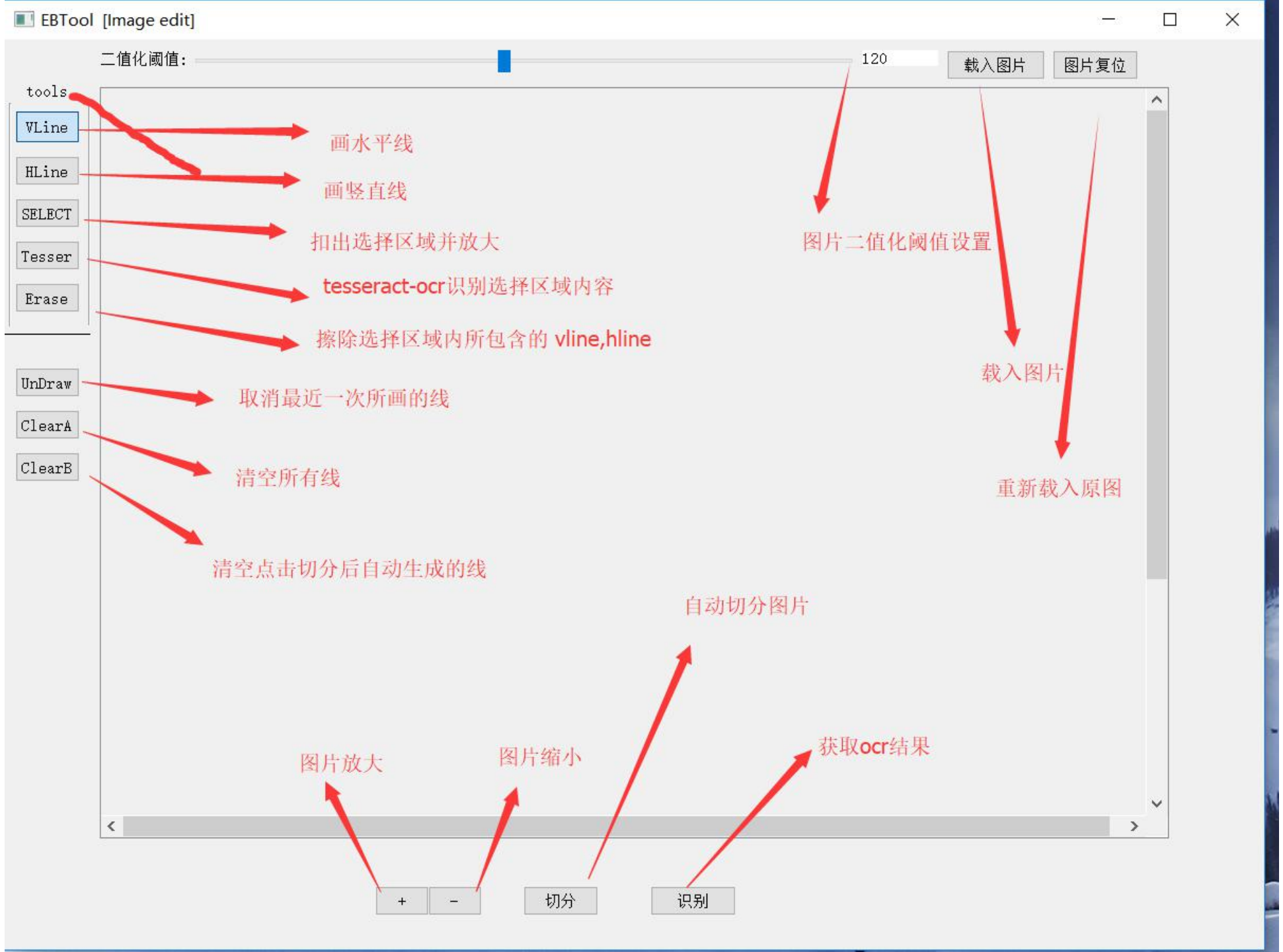
本工具采用的是字符分离再识别方向的OCR技术，至于其它更牛逼的OCR方法这里暂且不谈。

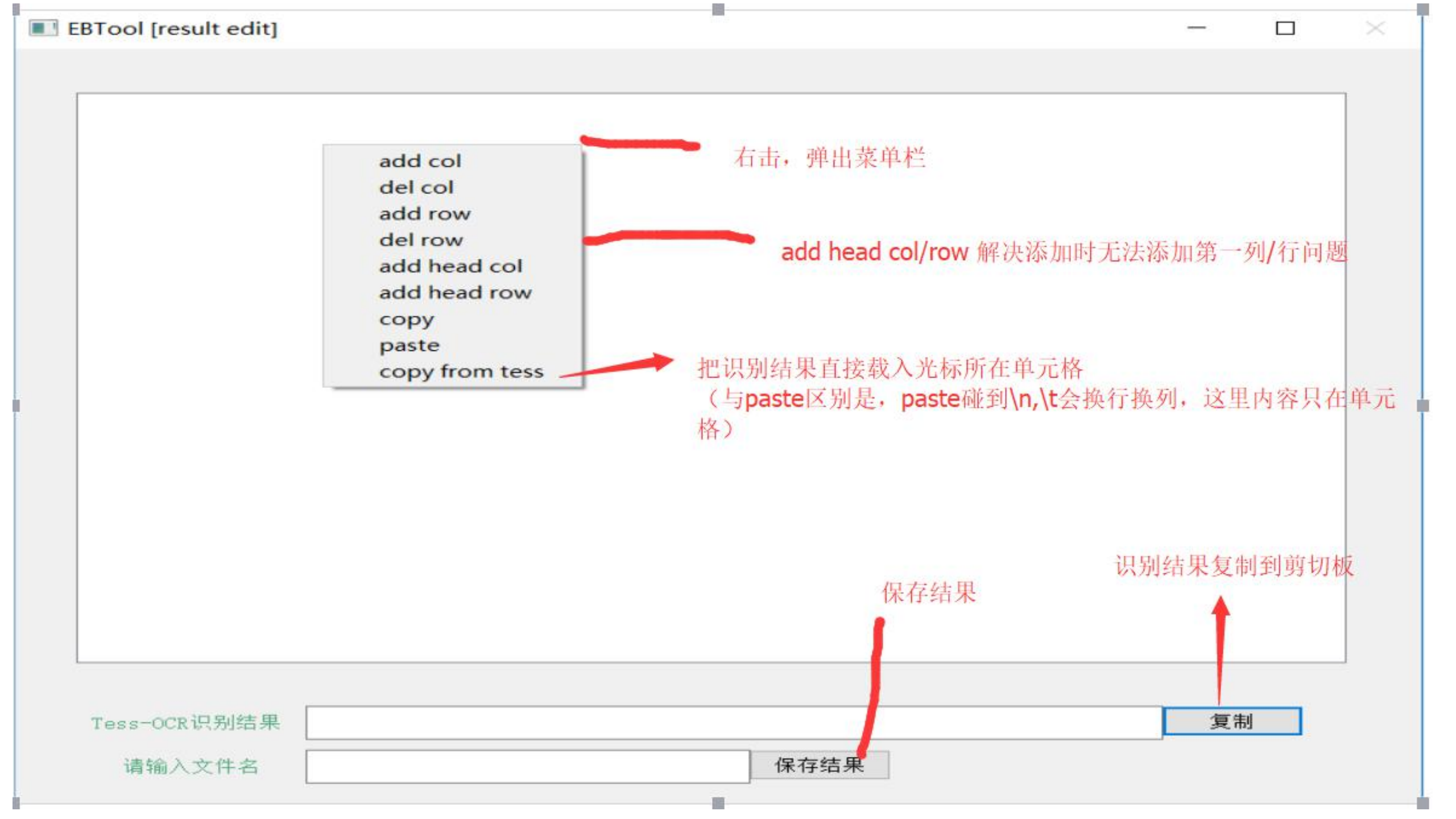
功能:
　辅助提取图片中的表格内容。其它方法无法提取的pdf，可以通过ghostscript转图片后再通过此工具提取。--技术有限、无法开发出复杂图片环境下全自动化高识别率的工具，只能辅助工具了。

主要方式:
　本着低成本原则，采用开源的tesseract对中文进行识别，自建模型对数字识别（需训练样本少，且准确率高）。
　基于OCR识别时复杂环境下准确率不能100%问题，以及表格非标的问题，提供了识别结果人工编辑校正功能。

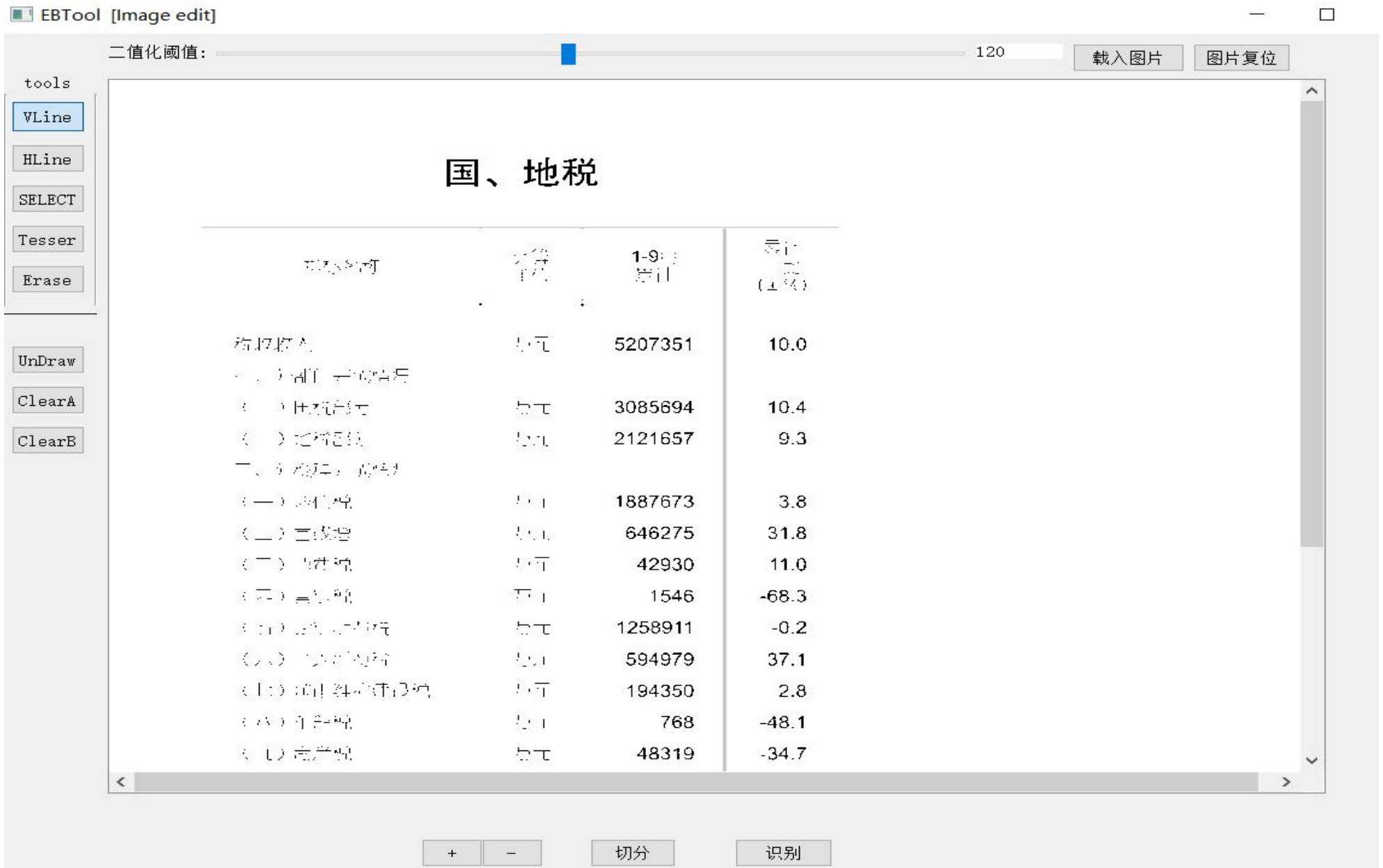
使用：


EBTool [Image edit]
二值化阈值:
120
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
+
-
切分
识别
EBTool [result edit]
Tess-OCR识别结果
复制
请输入文件名
保存结果


EBTool [Image edit]
二值化阈值:
120
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
画水平线
画竖直线
扣出选择区域并放大
tesseract-ocr识别选择区域内容
擦除选择区域内所包含的 vline,hline
取消最近一次所画的线
清空所有线
清空点击切分后自动生成的线
图片二值化阈值设置
载入图片
重新载入原图
自动切分图片
图片放大
图片缩小
获取ocr结果
+
-
切分
识别


EBTool [result edit]
add col
del col
add row
del row
add head col
add head row
copy
paste
copy from tess
右击，弹出菜单栏
add head col/row 解决添加时无法添加第一列/行问题
把识别结果直接载入光标所在单元格
（与paste区别是，paste碰到\n,\t会换行换列，这里内容只在单元格）
识别结果复制到剪切板
保存结果
Tess-OCR识别结果
复制
请输入文件名
保存结果

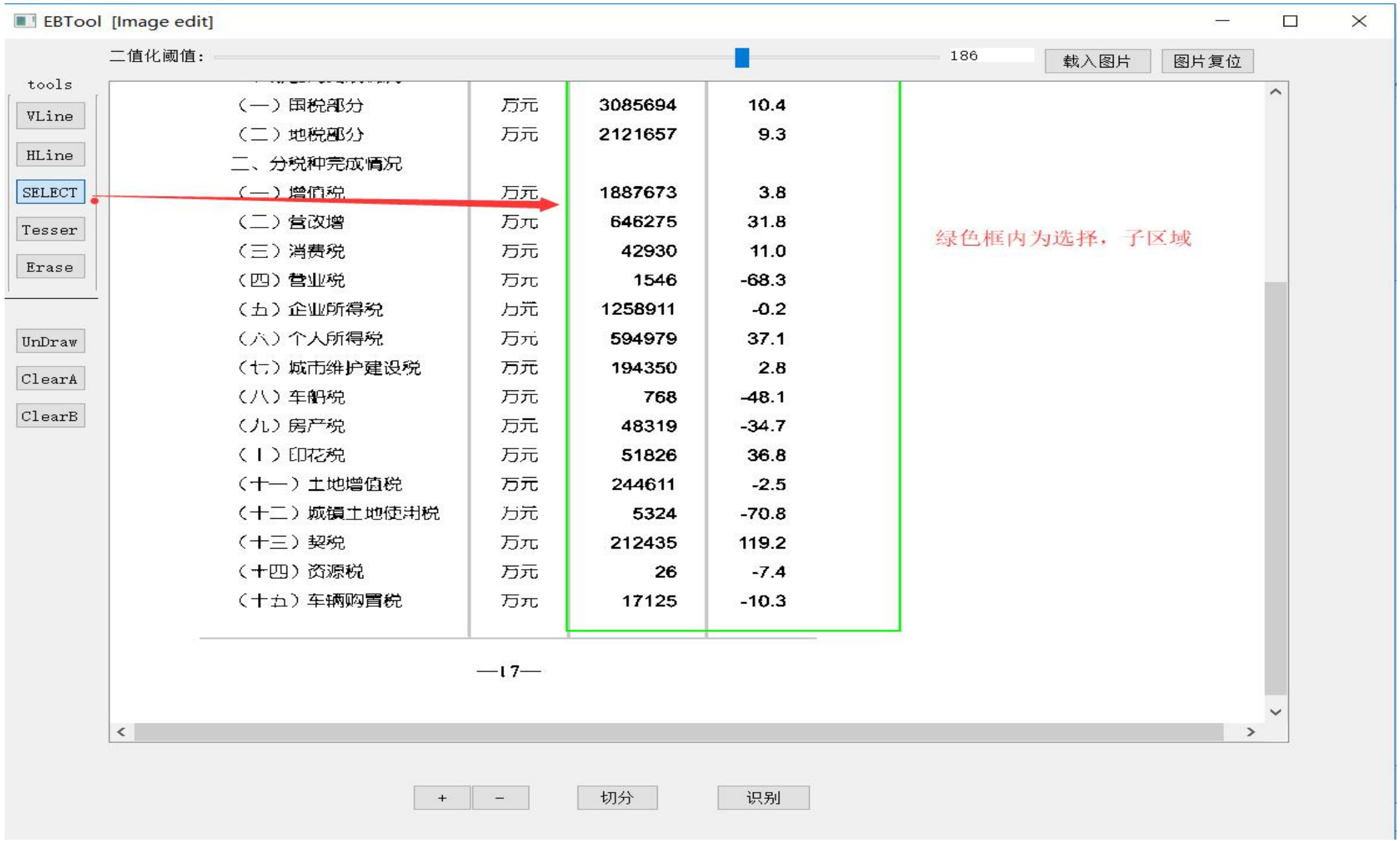
使用demo: 载入图片


EBTool [Image edit]
二值化阈值:
120
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
国、地税
1-9
10.0
10.4
9.3
3.8
31.8
11.0
-68.3
-0.2
37.1
2.8
-48.1
-34.7
+
-
切分
识别

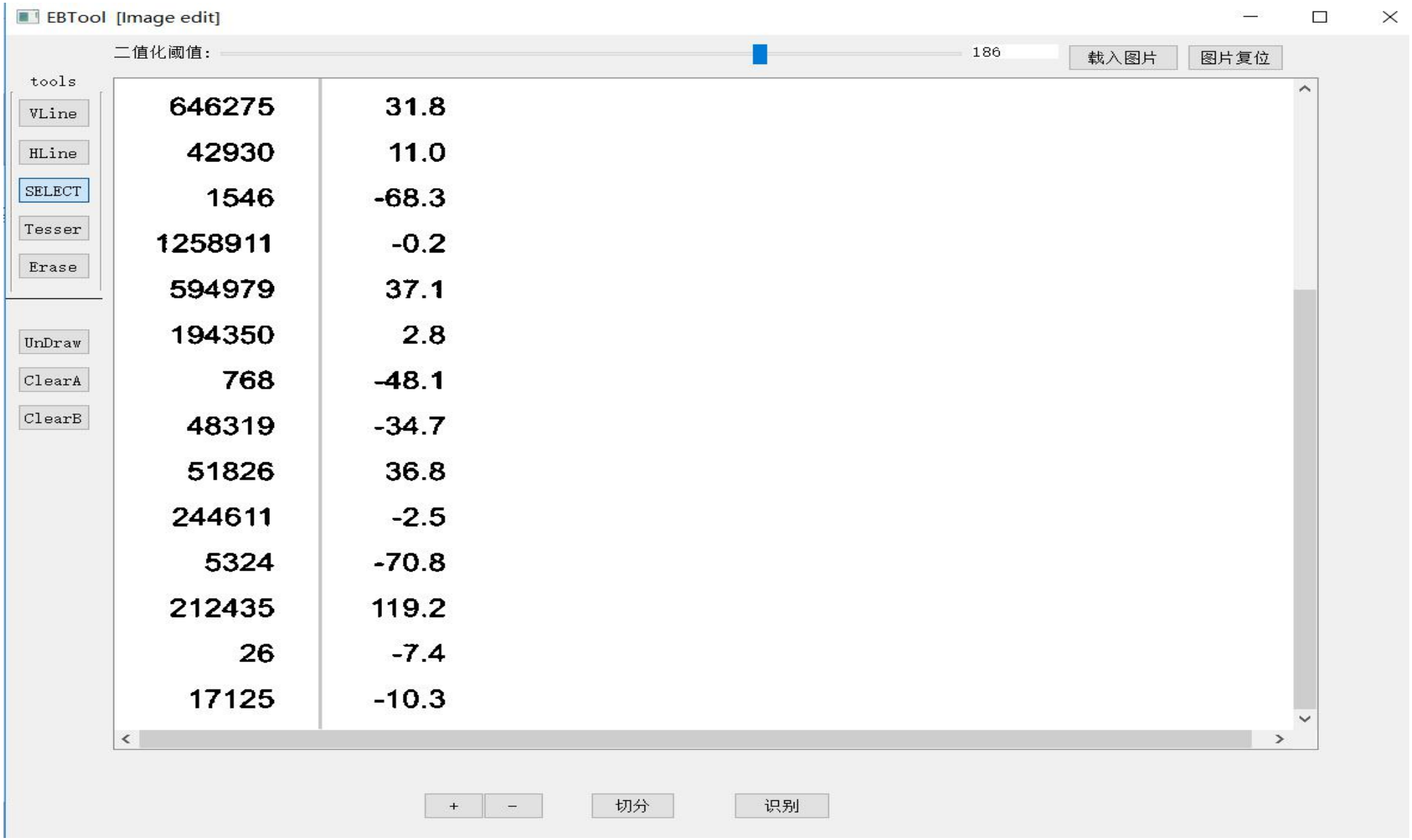
调整二值化阈值


EBTool [Image edit]
二值化阈值：
186
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
阈值调大，图片清晰了
国、地税
指标名称
计算单位
1-9月累计
累计同比(±%)
税收收入 万元 5207351 10.0
一、分部门完成情况
（一）国税部分 万元 3085694 10.4
（二）地税部分 万元 2121657 9.3
二、分税种完成情况
（一）增值税 万元 1887673 3.8
（二）营改增 万元 646275 31.8
（三）消费税 万元 42930 11.0
（四）营业税 万元 1546 -68.3
（五）企业所得税 万元 1258911 -0.2
（六）个人所得税 万元 594979 37.1
（七）城市维护建设税 万元 194350 2.8
（八）车船税 万元 768 -48.1
（九）房产税 万元 48319 -34.7
+
-
切分
识别

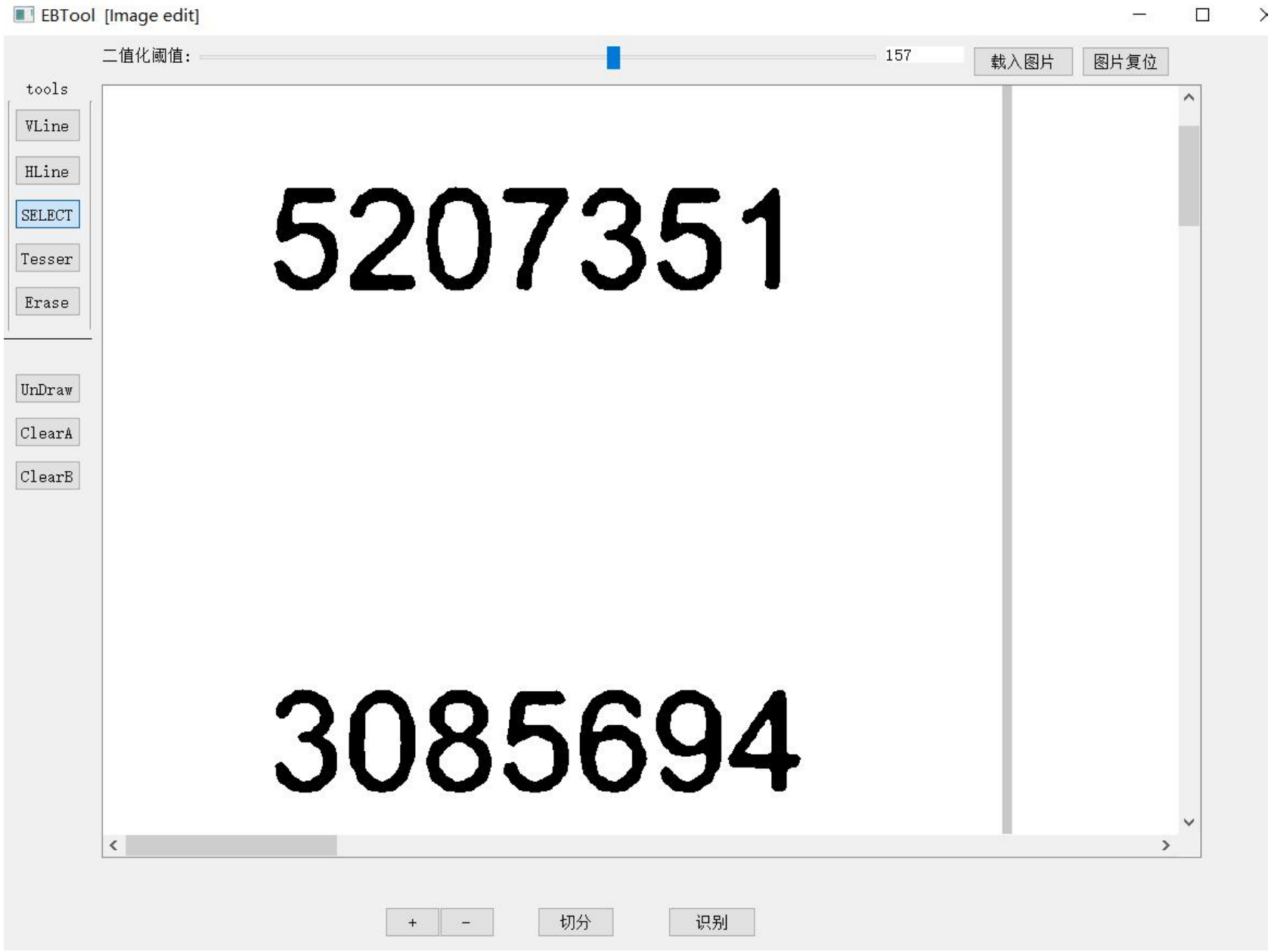
选择子区域：点击select后，光标移动至画图区，按下鼠标左键拖动画图。



跳转到选择的子区域

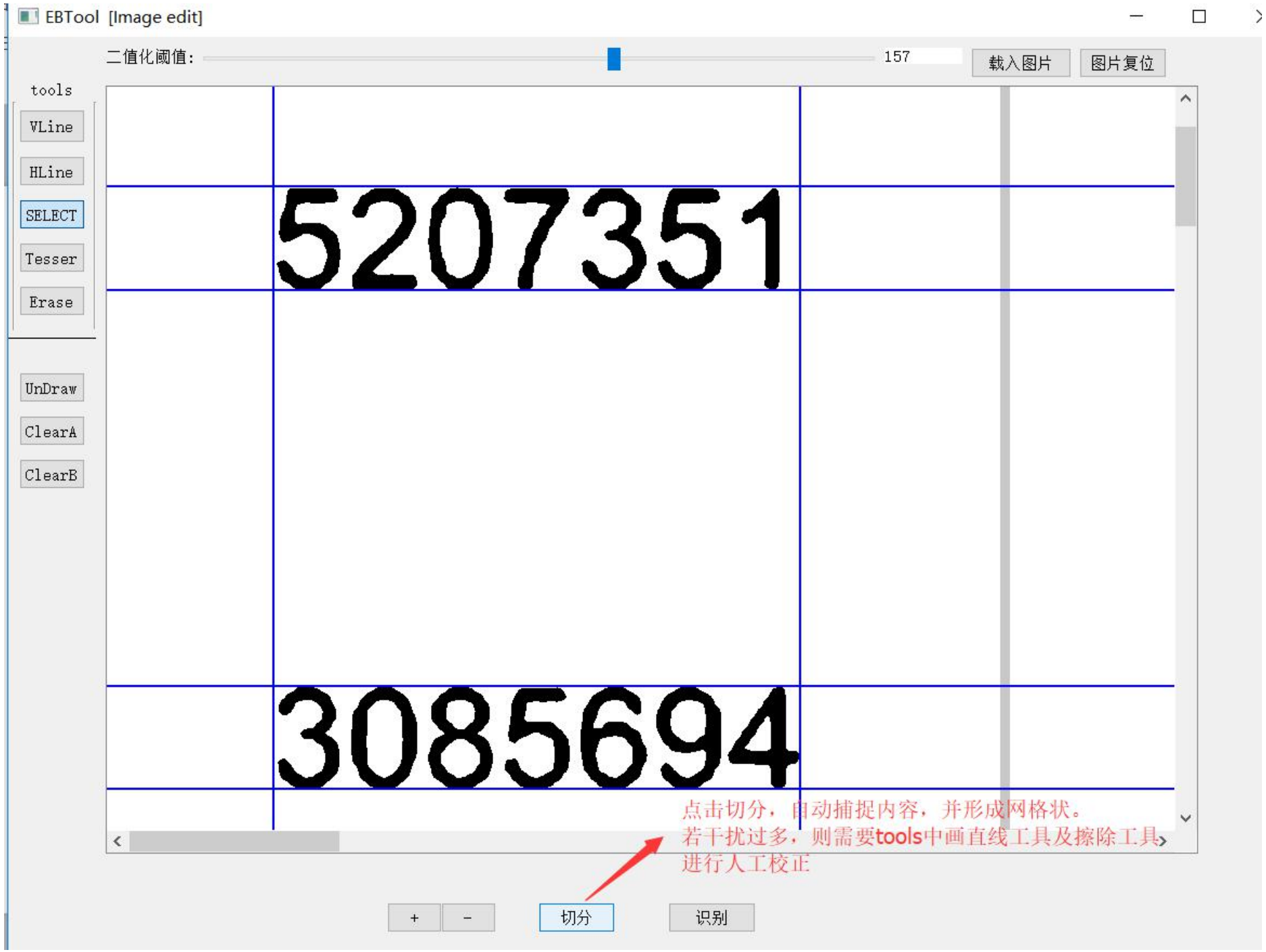

EBTool [Image edit]
二值化阈值：
186
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
646275 31.8
42930 11.0
1546 -68.3
1258911 -0.2
594979 37.1
194350 2.8
768 -48.1
48319 -34.7
51826 36.8
244611 -2.5
5324 -70.8
212435 119.2
26 -7.4
17125 -10.3
+
-
切分
识别

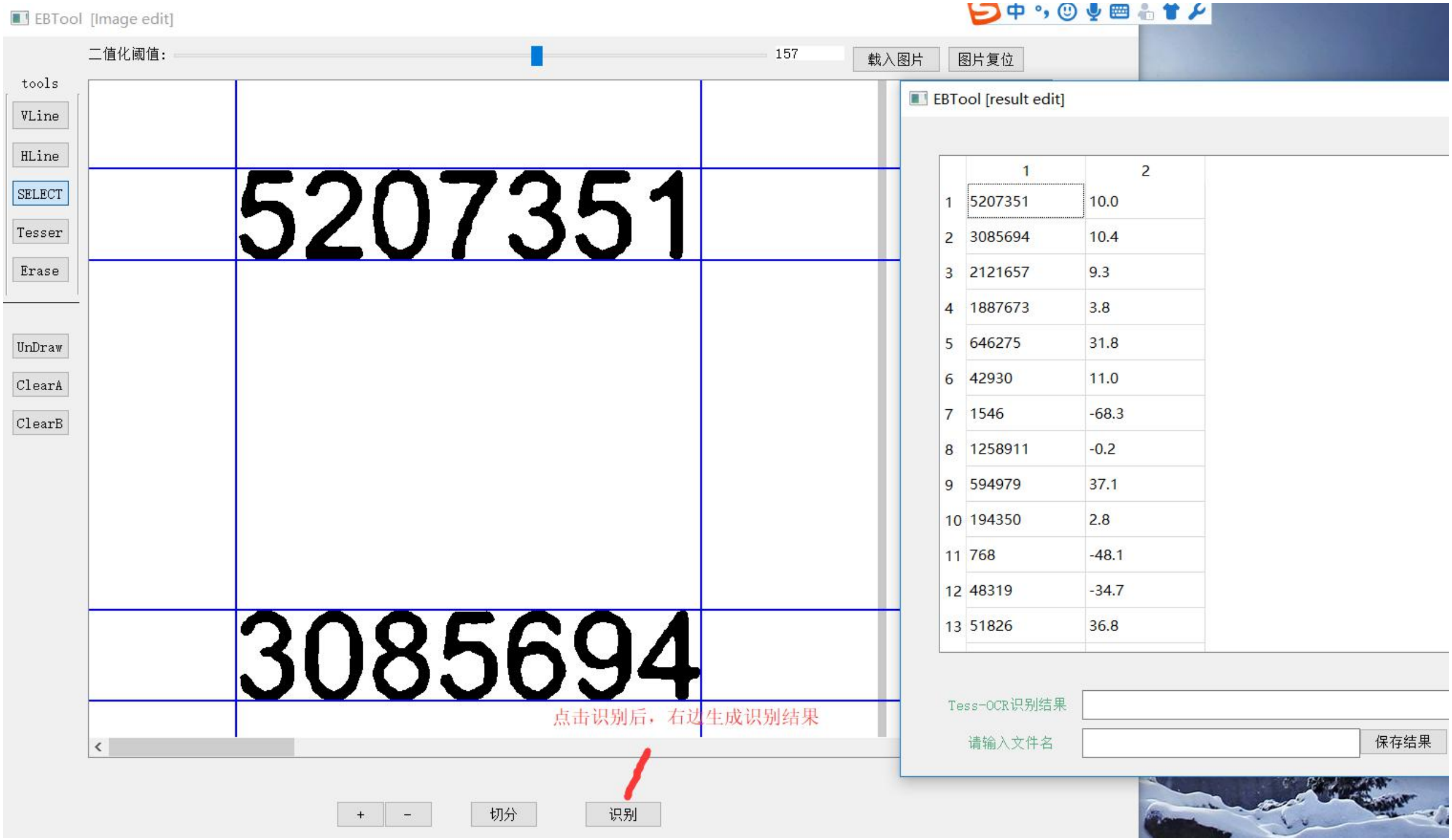
放大图片(+),调整阈值，使数字尽量不黏连。（一般 图片放越大越好，阈值160左右，字体看起来不带毛边且圆润即可)

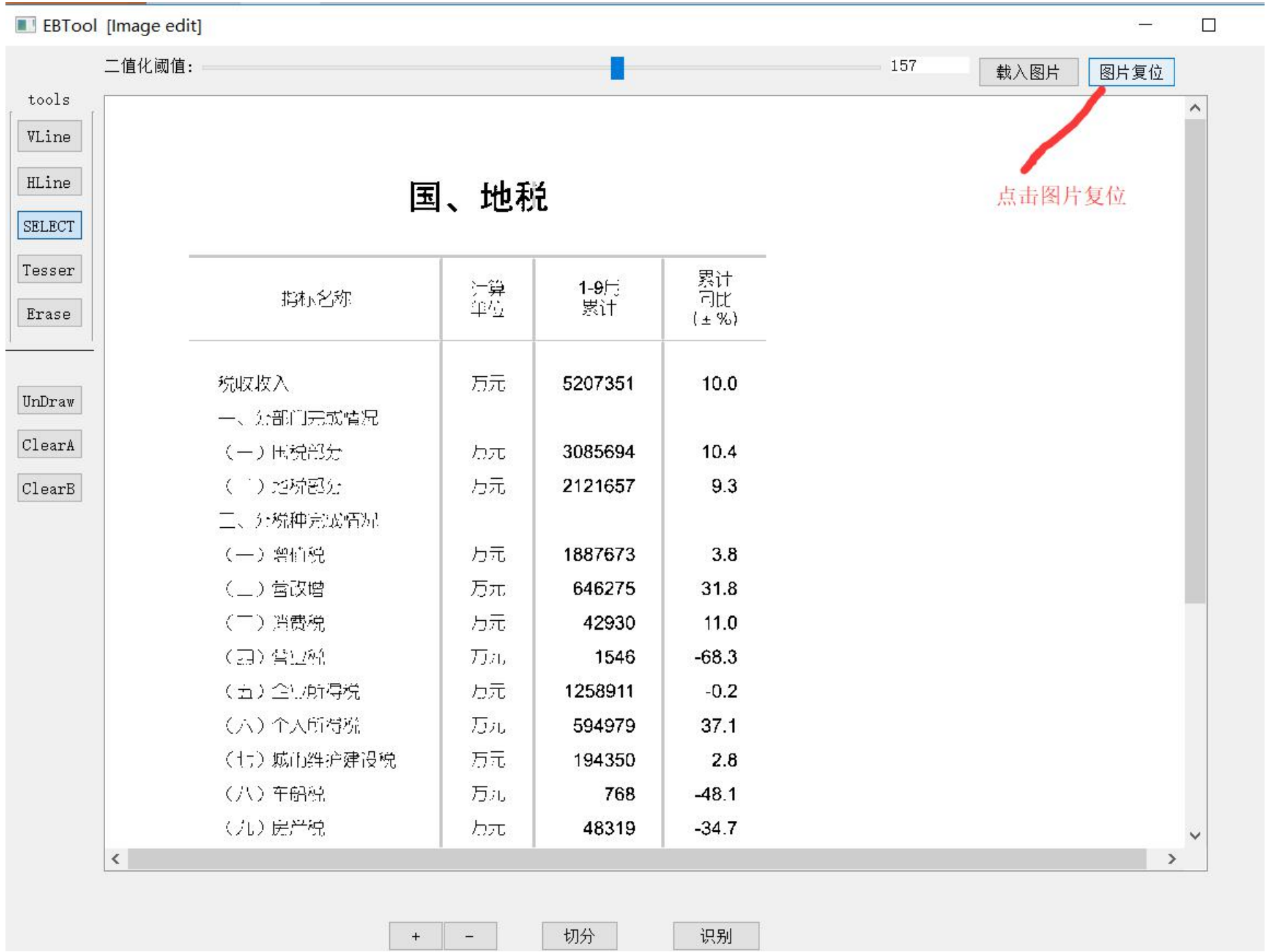


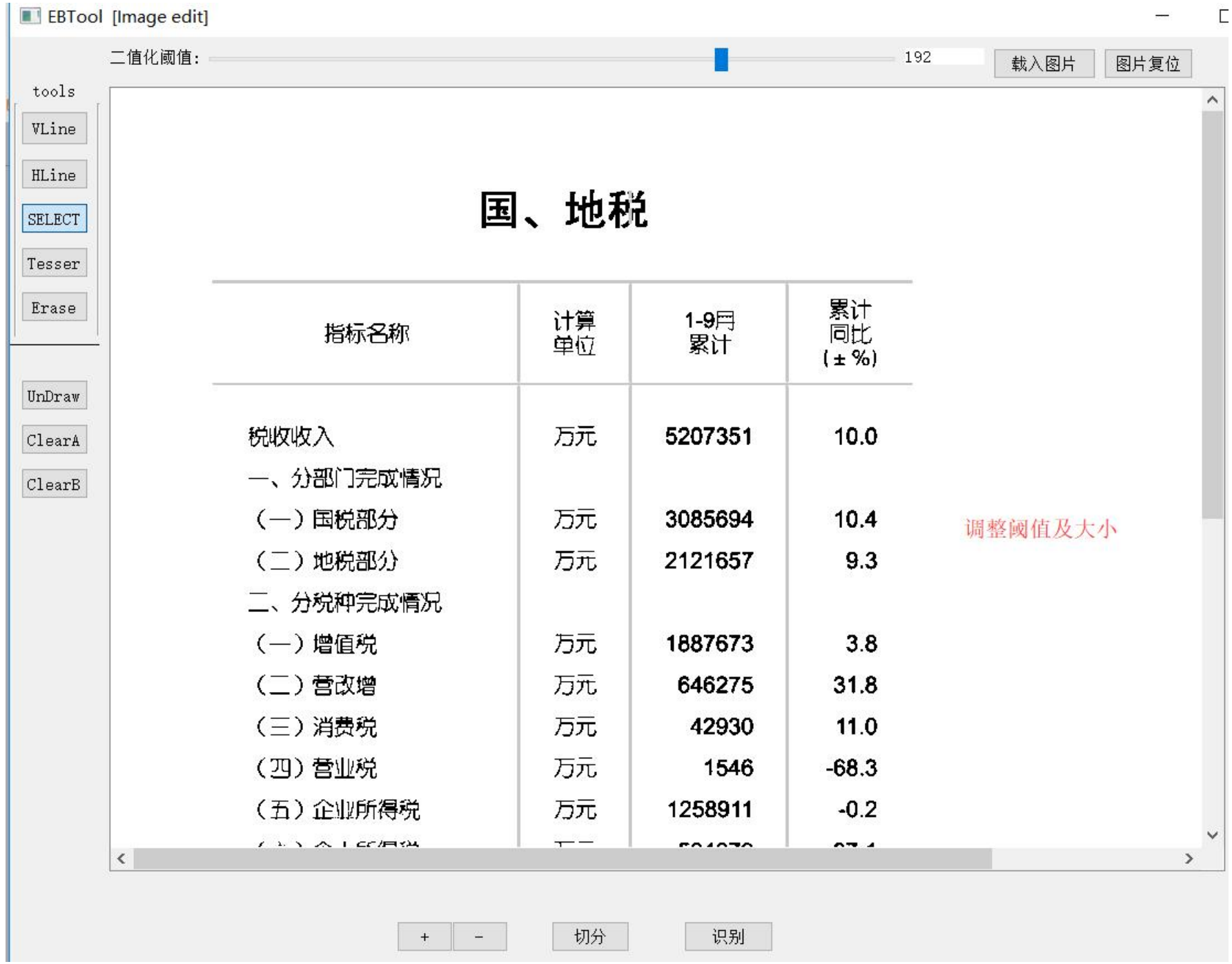
调整后字体特征明显，且不黏连

识别时单元格形式和直线(绿，蓝)网格类似（灰色直线忽略）

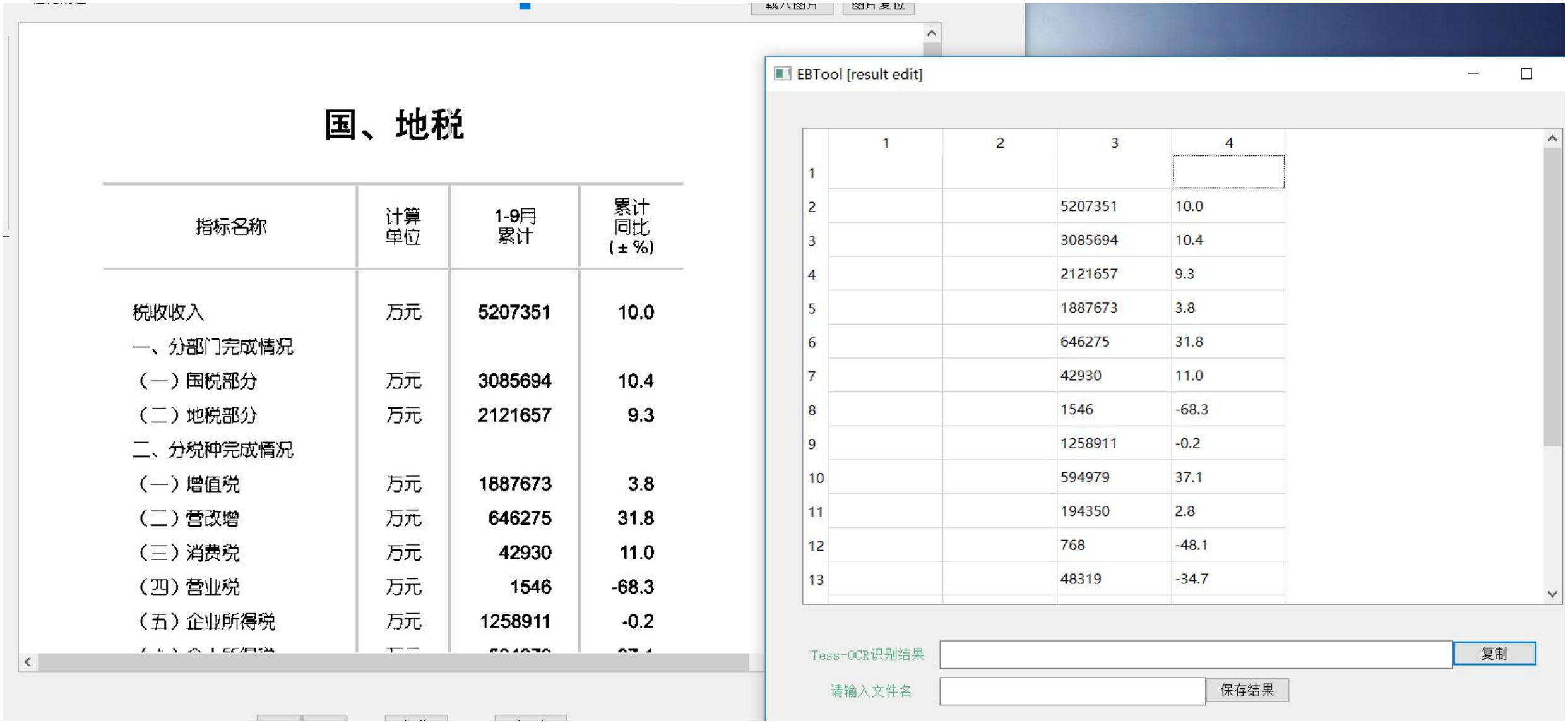


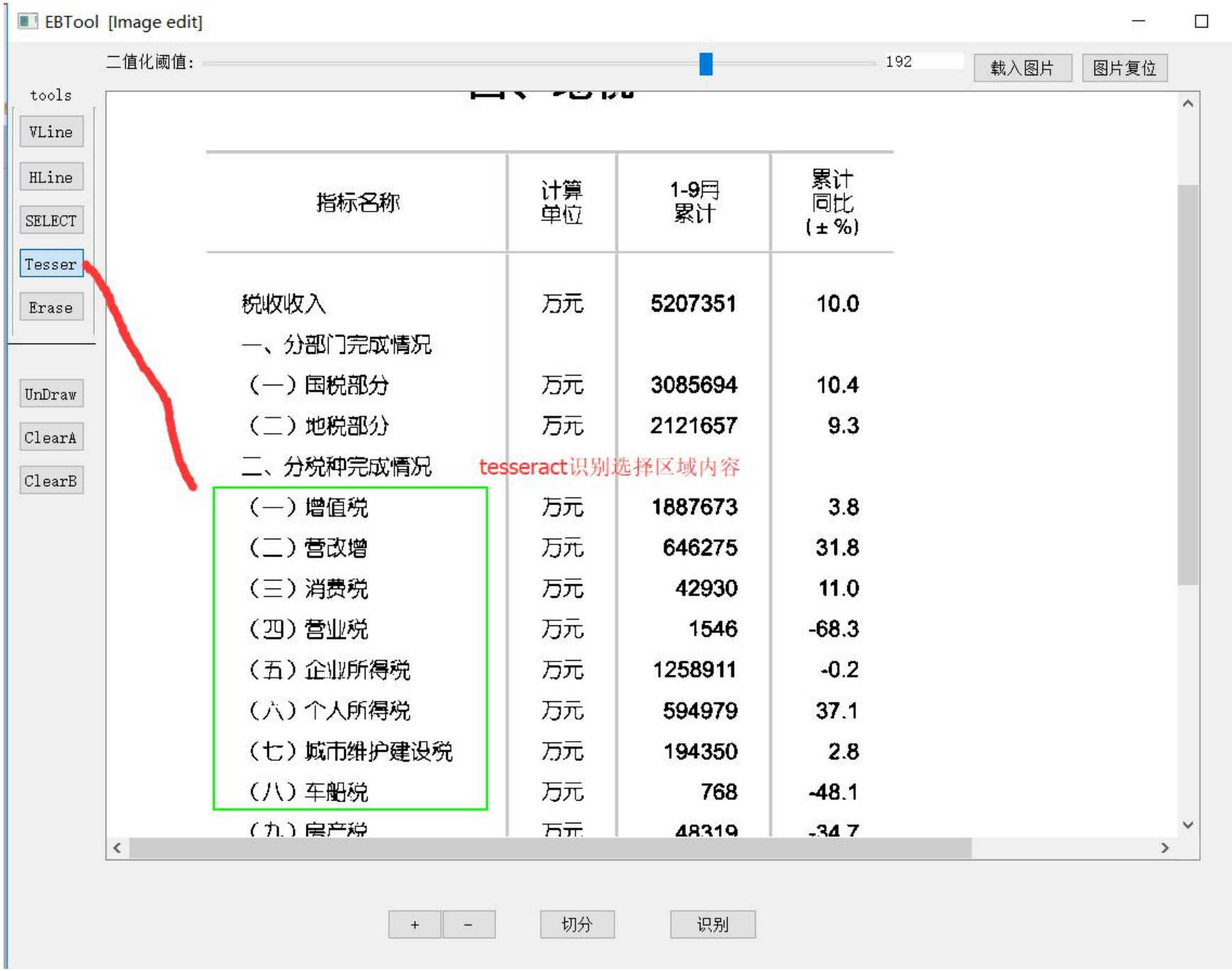

EBTool [Image edit]
二值化阈值:
157
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
5207351
3085694
点击识别后，右边生成识别结果
+
-
切分
识别
EBTool [result edit]
1
2
1 5207351 10.0
2 3085694 10.4
3 2121657 9.3
4 1887673 3.8
5 646275 31.8
6 42930 11.0
7 1546 -68.3
8 1258911 -0.2
9 594979 37.1
10 194350 2.8
11 768 -48.1
12 48319 -34.7
13 51826 36.8
Tess-OCR识别结果
请输入文件名
保存结果


EBTool [Image edit]
二值化阈值：
157
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
点击图片复位
国、地税
指标名称
计算单位
1-9月累计
累计同比(±%)
税收收入 万元 5207351 10.0
一、分部门完成情况
（一）国税部分 万元 3085694 10.4
（二）地税部分 万元 2121657 9.3
二、分税种完成情况
（一）增值税 万元 1887673 3.8
（二）营改增 万元 646275 31.8
（三）消费税 万元 42930 11.0
（四）营业税 万元 1546 -68.3
（五）企业所得税 万元 1258911 -0.2
（六）个人所得税 万元 594979 37.1
（七）城市维护建设税 万元 194350 2.8
（八）车船税 万元 768 -48.1
（九）房产税 万元 48319 -34.7
+
-
切分
识别


EBTool [Image edit]
二值化阈值:
192
载入图片
图片复位
tools
VLine
HLine
SELECT
Tesser
Erase
UnDraw
ClearA
ClearB
国、地税
指标名称
计算单位
1-9月累计
累计同比(±%)
税收收入 万元 5207351 10.0
一、分部门完成情况
（一）国税部分 万元 3085694 10.4
（二）地税部分 万元 2121657 9.3
二、分税种完成情况
（一）增值税 万元 1887673 3.8
（二）营改增 万元 646275 31.8
（三）消费税 万元 42930 11.0
（四）营业税 万元 1546 -68.3
（五）企业所得税 万元 1258911 -0.2
调整阈值及大小
+
-
切分
识别

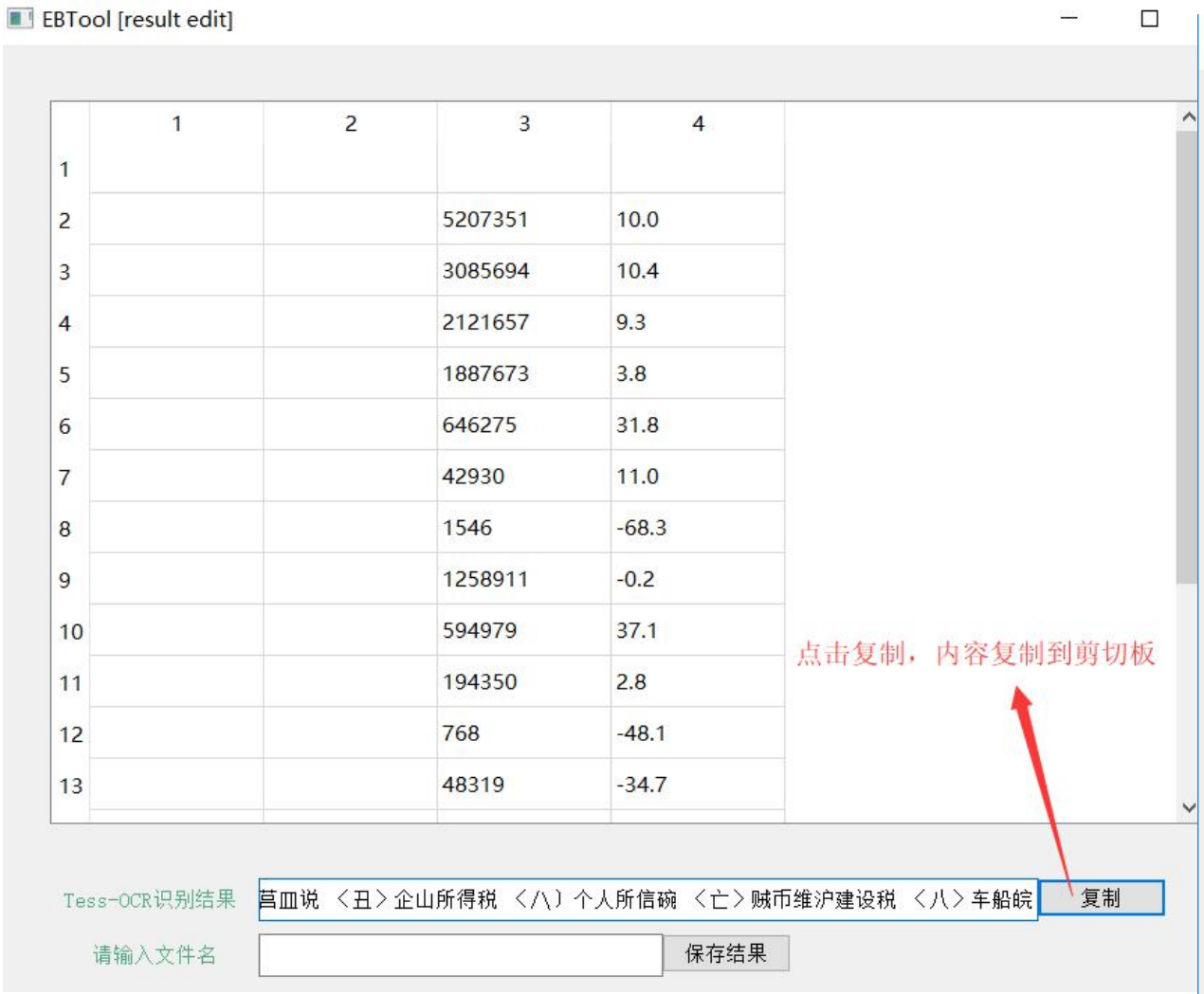
编辑结果编辑工具，添加横列

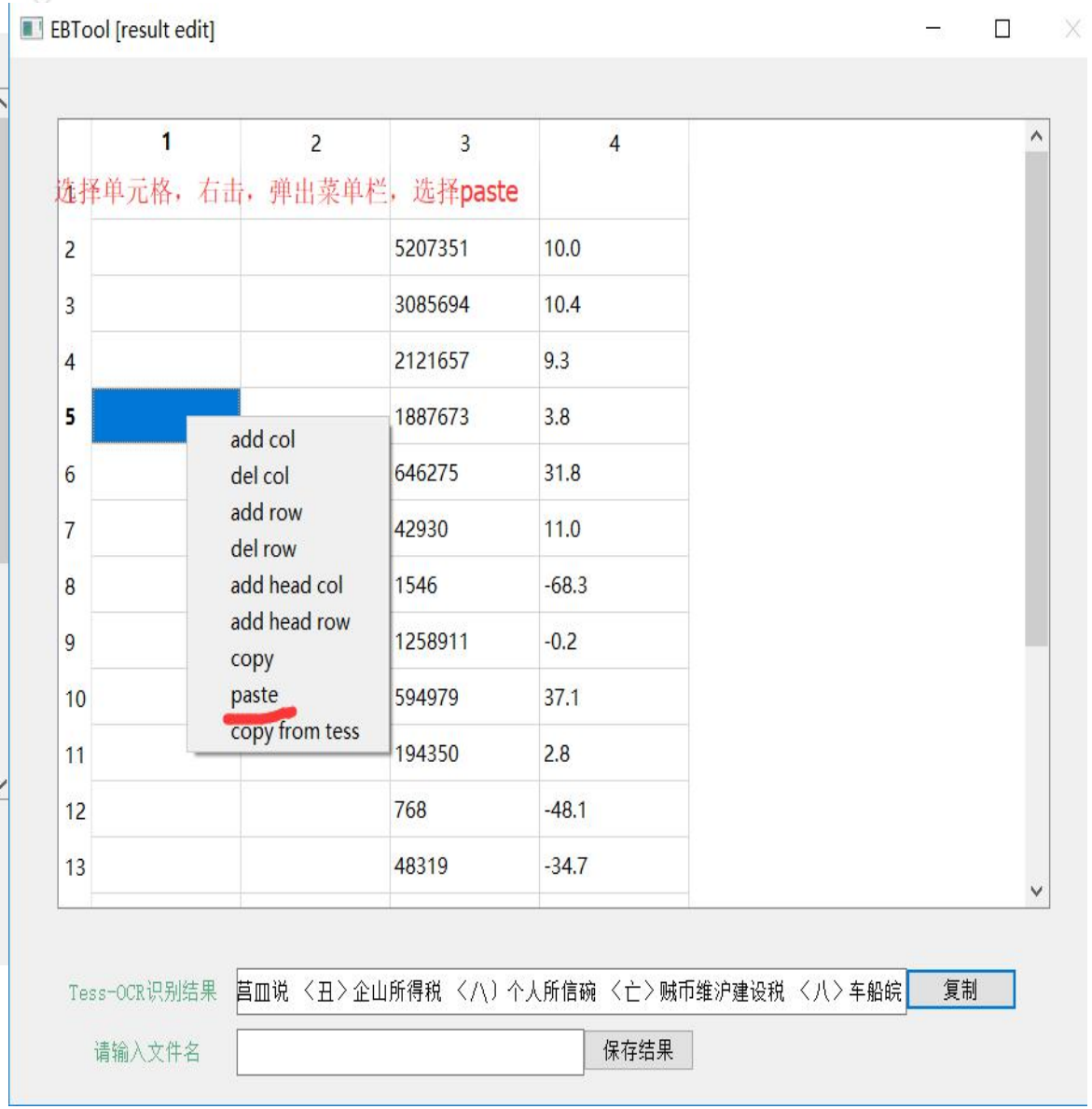




tesseract接口识别较慢，可能有一定的延迟


二、分税种完成情况
（一）增值税 万元 1887673 3.8
（二）营改增 万元 646275 31.8
（三）消费税 万元 42930 11.0
（四）营业税 万元 1546 -68.3
（五）企业所得税 万元 1258911 -0.2
（八）个人所得税 万元 594979 37.1
（七）城市维护建设税 万元 194350 2.8
（八）车船税 万元 768 -48.1
（九）房产税 万元 48319 -34.7
（十）印花税 万元 51826 36.8
4 2121657 9.3
5 1887673 3.8
6 646275 31.8
7 42930 11.0
8 1546 -68.3
9 1258911 -0.2
10 594979 37.1
11 194350 2.8
12 768 -48.1
13 48319 -34.7
识别结果
Tess-OCR识别结果
莒皿说 〈丑〉企山所得税 〈八〉个人所信碗 〈亡〉贼币维沪建设税 〈八〉车船皖
请输入文件名
保存结果
+
-
切分
识别

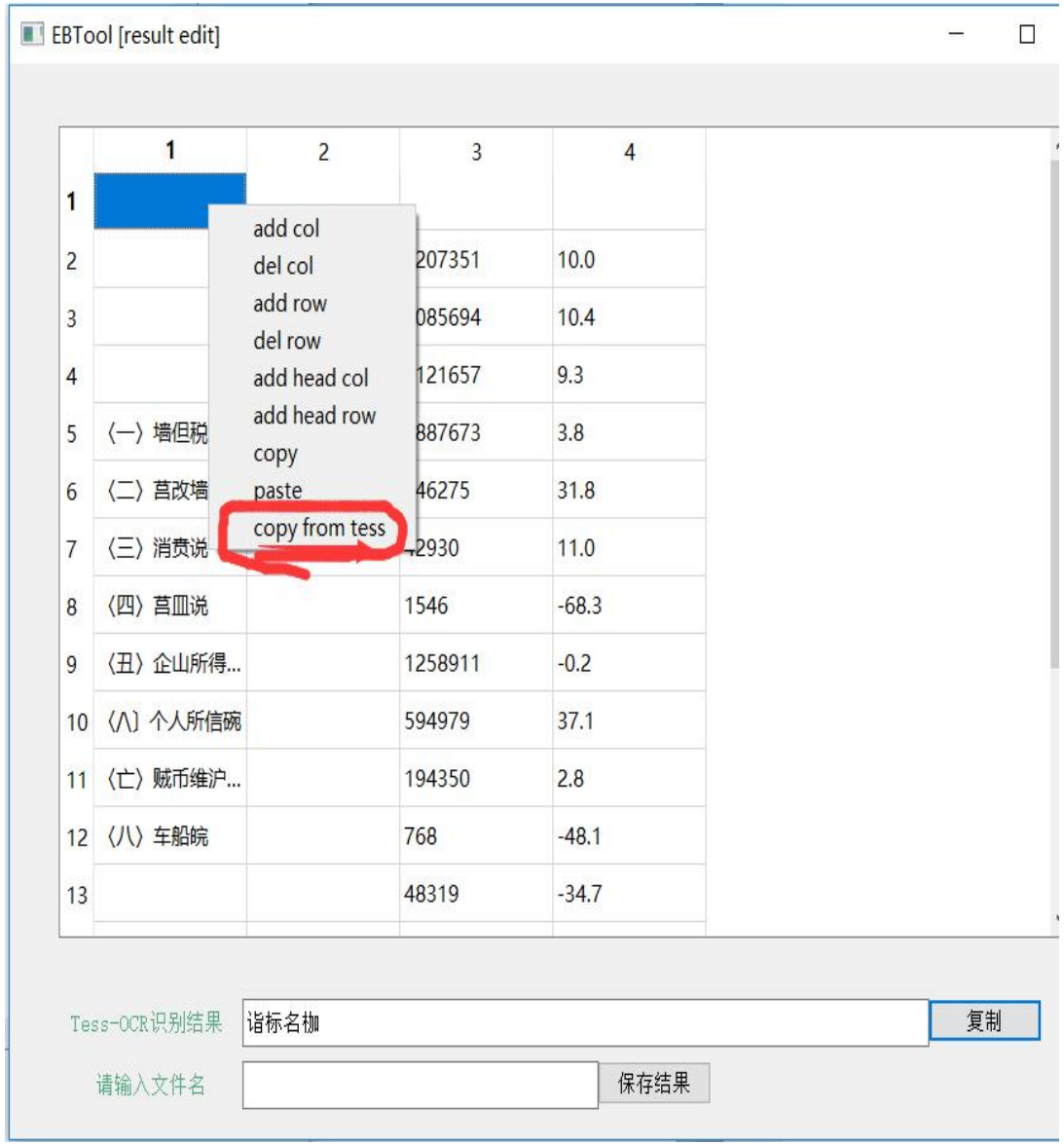


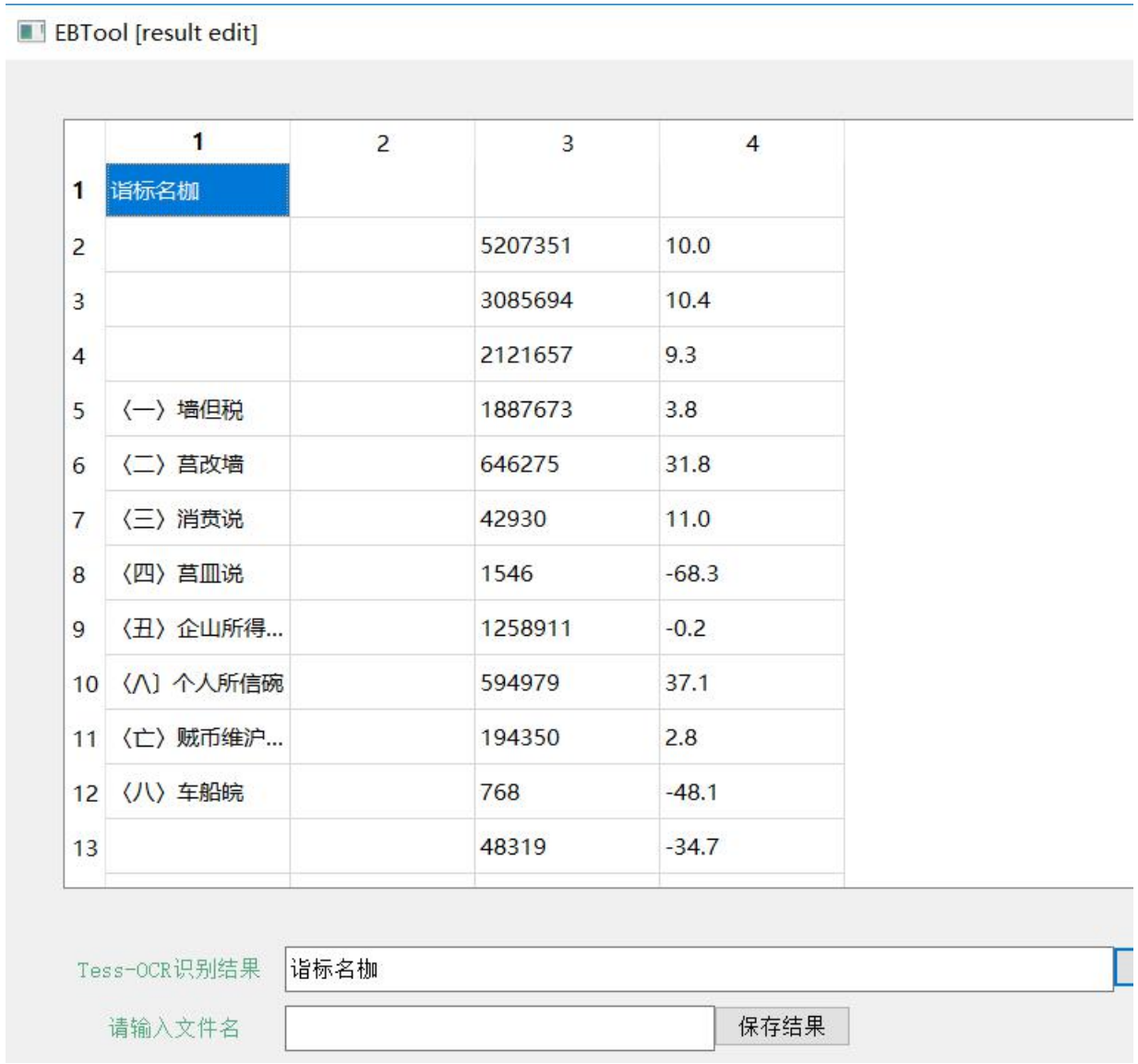


选择单元格不要双击进入单元格编辑模式，不然相
copy from tess 操作

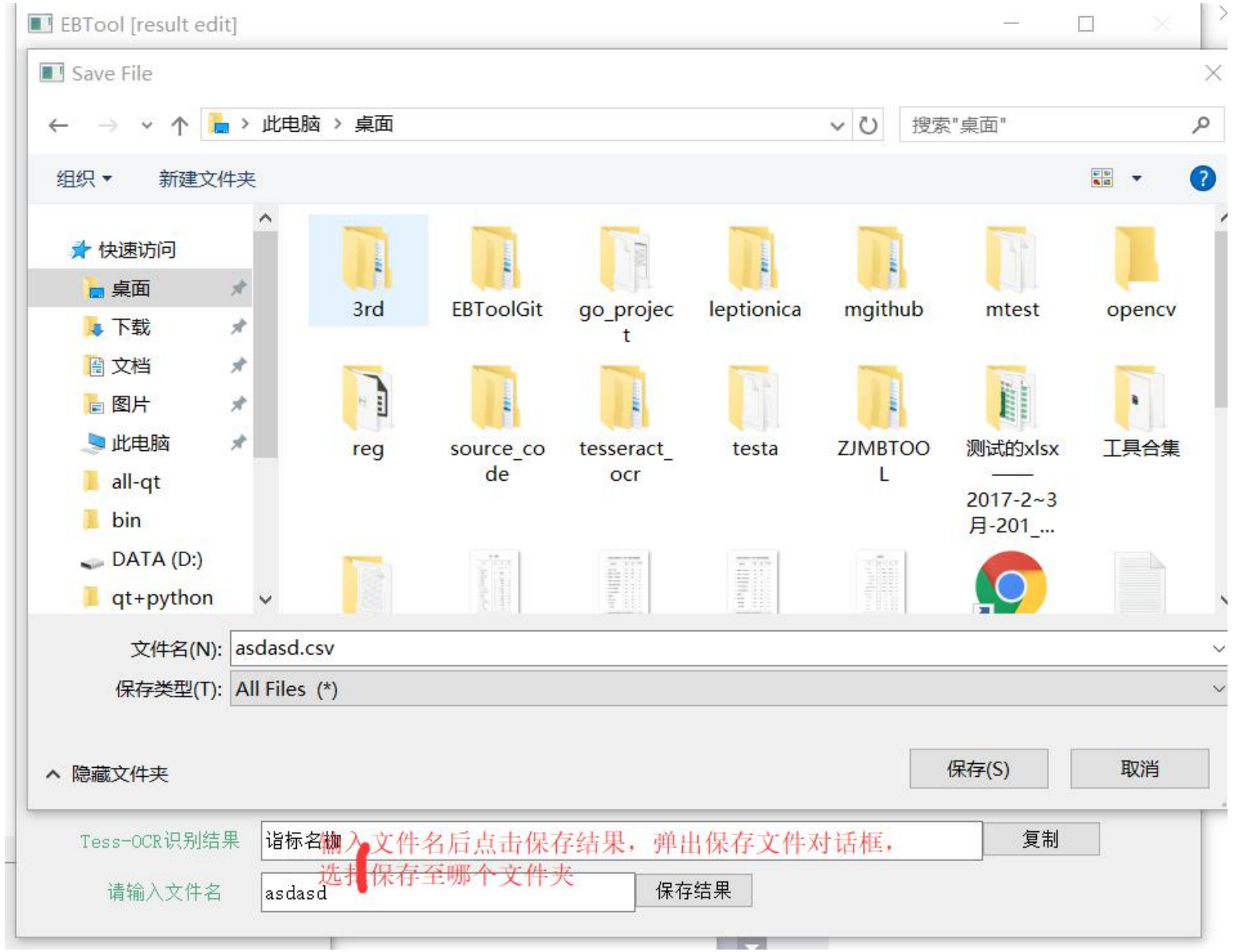
结果，需要人工校正。
目前只支持竖直批量处理tesseract结果。横向请逐个操作.

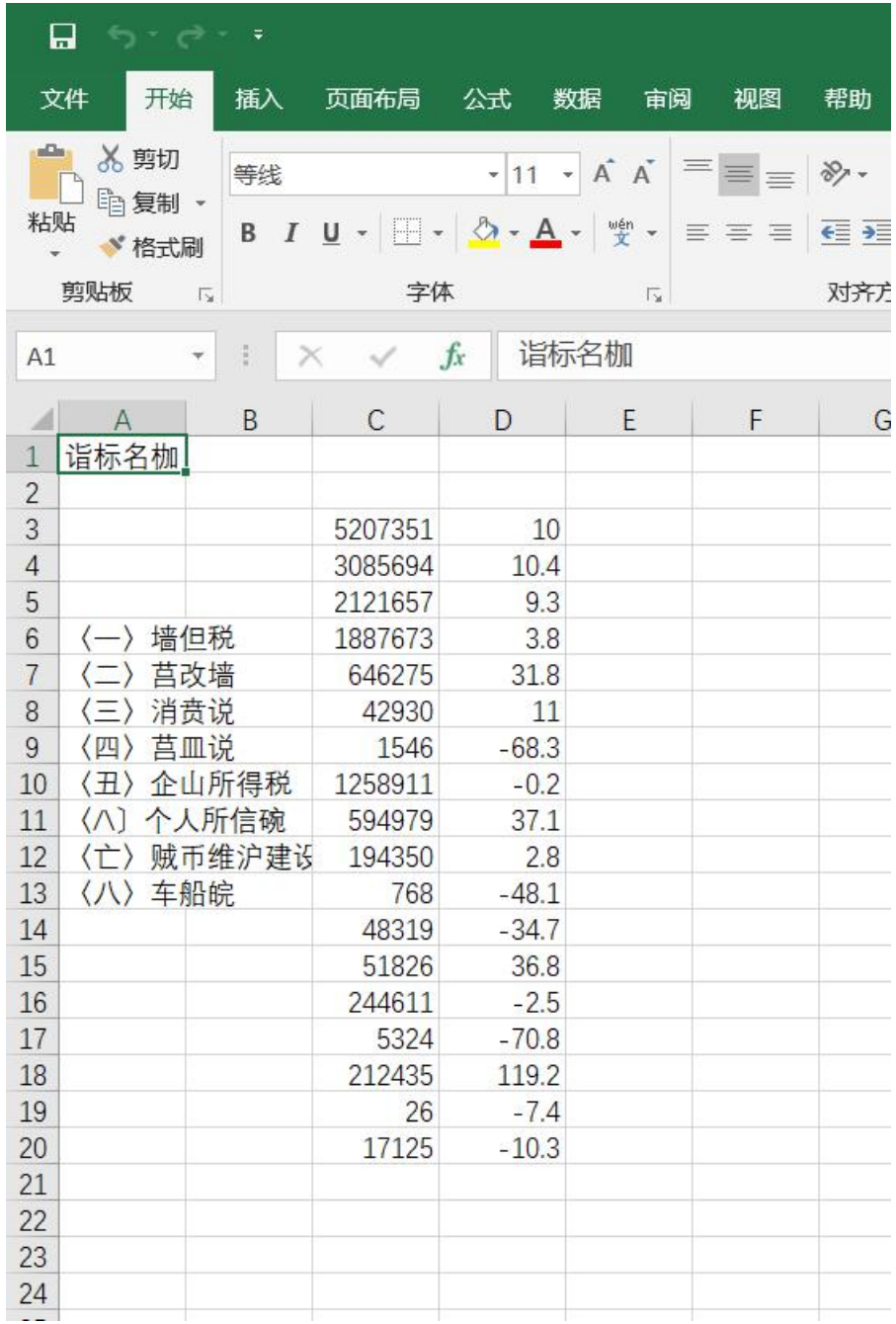
该单元格复制粘贴类似excel内形式。可与通过复制粘贴excel交互。





直接从识别结果复制到单元格内(忽略换行，换列符号)


EBTool [result edit]
Save File
此电脑 › 桌面
搜索"桌面"
组织
新建文件夹
快速访问
桌面
下载
文档
图片
此电脑
all-qt
bin
DATA (D:)
qt+python
3rd
EBToolGit
go_project
leptionica
mgithub
mtest
opencv
reg
source_code
tesseract_ocr
testa
ZJMBTOOL
测试的xlsx——2017-2~3月-201_...
工具合集
文件名(N): asdasd.csv
保存类型(T): All Files (*)
隐藏文件夹
保存(S)
取消
Tess-OCR识别结果
复制
请输入文件名
asdasd
保存结果
输入文件名后点击保存结果，弹出保存文件对话框，
选择保存至哪个文件夹



	A	B	C	D	E	F	G
1	诣标名枷						
2							
3			5207351	10			
4			3085694	10.4			
5			2121657	9.3			
6	〈一〉墙但税		1887673	3.8			
7	〈二〉莒改墙		646275	31.8			
8	〈三〉消贲说		42930	11			
9	〈四〉莒皿说		1546	-68.3			
10	〈丑〉企山所得税		1258911	-0.2			
11	〈八〕个人所信碗		594979	37.1			
12	〈亡〉贼币维沪建设		194350	2.8			
13	〈八〉车船皖		768	-48.1			
14			48319	-34.7			
15			51826	36.8			
16			244611	-2.5			
17			5324	-70.8			
18			212435	119.2			
19			26	-7.4			
20			17125	-10.3			
21							
22							
23							
24							

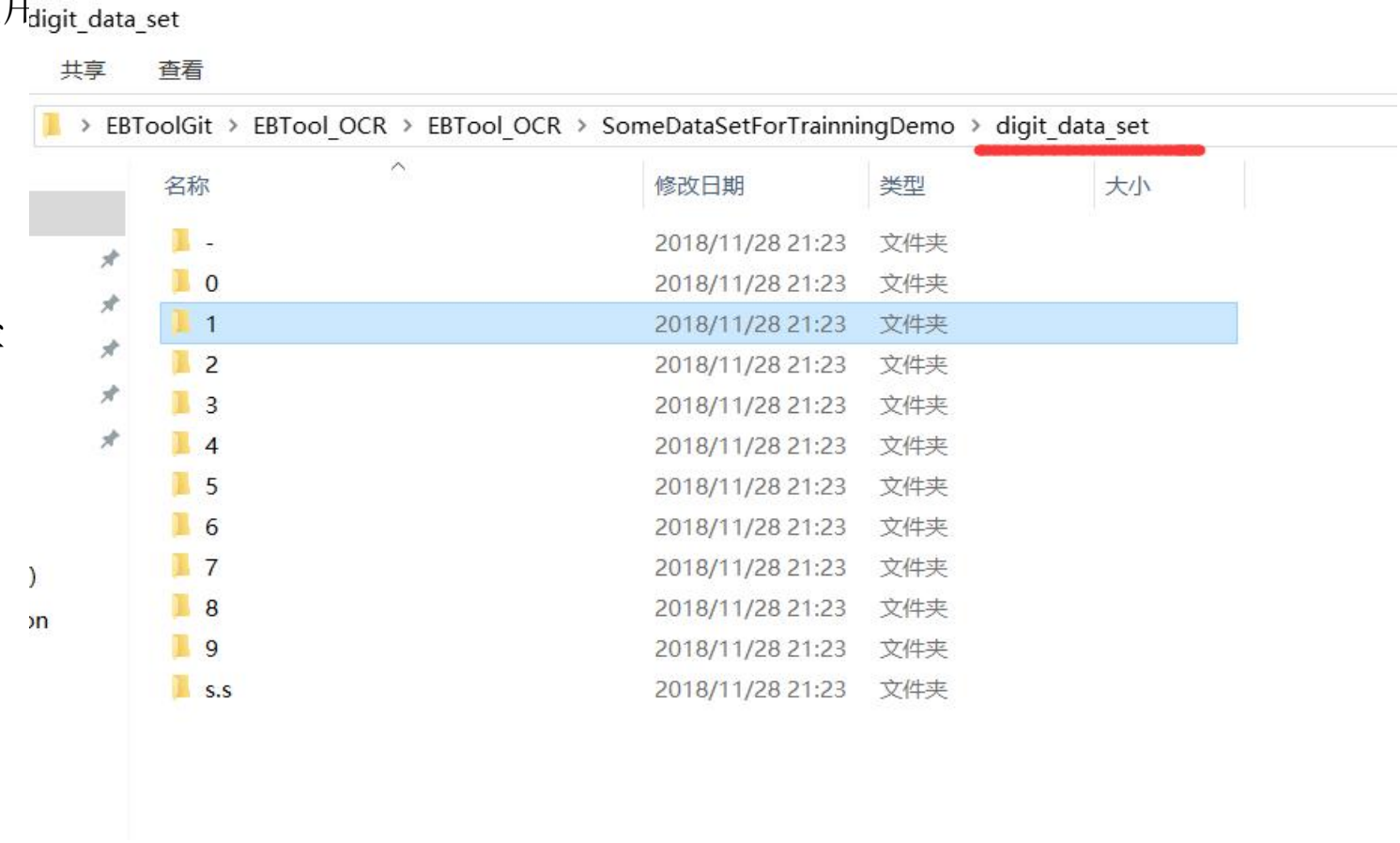
结果

TrainningTool使用：【很多参数写死，使用意义不大】

打开cmd，
cd 至训练结果需要保持的文件夹

在cmd输入指令
路径/TrainningTool.exe　路径/数据集路径

项目中有部分数据集样本。位置如右图



程序配置

config.ini - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
TCP_SERVER_HOST=127.0.0.1
TCP_SERVER_PORT=8034

#MLAPI   DEFAULT: 使用默认内置ANN，  TCP:采用tcp方式与ML服务端交互进行识别
MLAPI=DEFAULT
# 如果采用DEFAULT模型，则给出输出标签种类个数
MLAPI_LABELS=12

#IMAGE_SAVE_DIR  DEFAULT:不保存扣取的图片，  指定路径: 保存图片（采集数据集）到指定路径，用于机器学习训练
#IMAGE_SAVE_DIR=xx/im
# xx/im 图片保存的位置，当点击识别时，识别之前会把需要识别的图片保存到该指定文件夹
IMAGE_SAVE_DIR=DEFAULT
```

TCP交互方式，在EBTool_OCR\mlServiceDemon，有个python版的demo

数字识别模型的通用性

该工具在使用时，涉及图片放大缩小，以及阈值改变。
若图片不黏连，能切分出来，实际对数字识别模型影响有限，因为这些字体都非手写，特征明显。
非抱着0识别率的态度，怎样放大缩小以及阈值设置之后每个数字的特征还是会相当明显。故模型在年鉴识别等通用性较强，无需频繁更新模型。