DEECAMP

环境不可控场景下
拍照文档地址文字识别

教育部中外人文交流中心 CCIPE

创新工场 人工智能工程院 SINOVATION VENTURES AI INSTITUTE

北京大学 PEKING UNIVERSITY

DEECAMP

CONTENTS

应用场景及挑战

文字识别 OCR

证件识别（身份证、营业执照）

外卖单、快递单扫描

手机扫描线下导航

挑战

01.拍摄环境不可控

- 拍照角度，字段倾斜
- 拍照设备，清晰度不可控

02.地址字段识别

- 字段长、多行
- 地址信息复杂，识别困难

应用场景及挑战

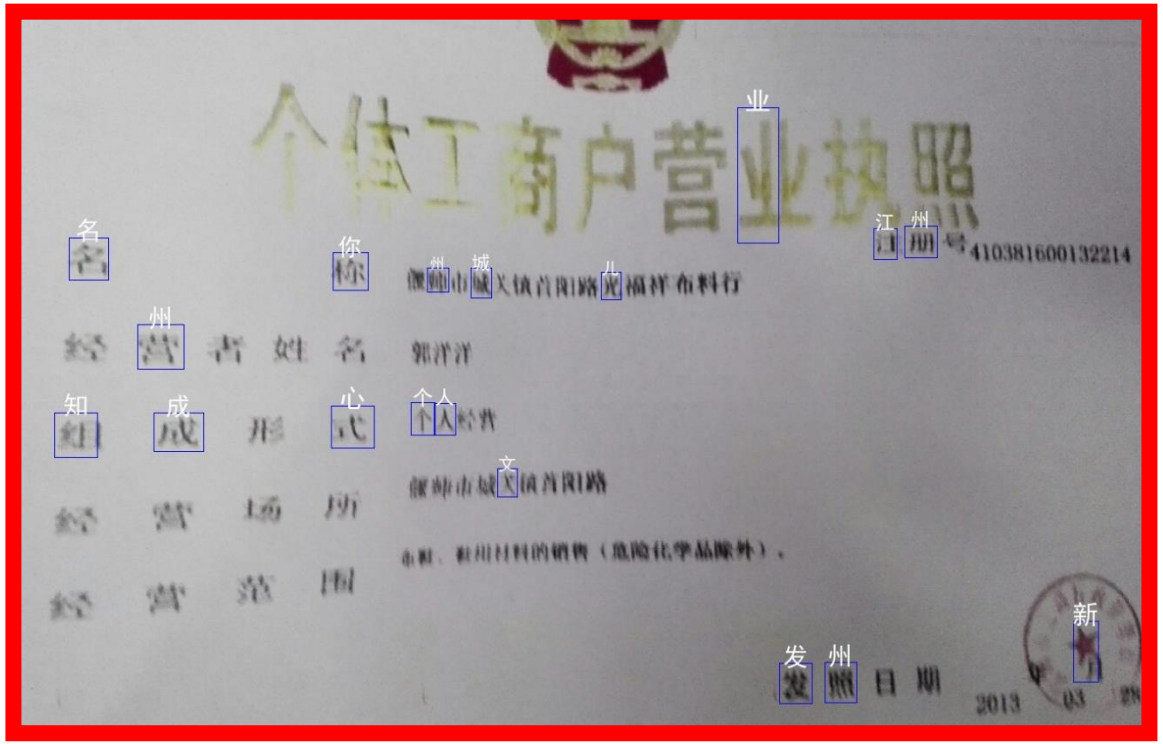
YOLO v2
自然场景

应用场景及挑战

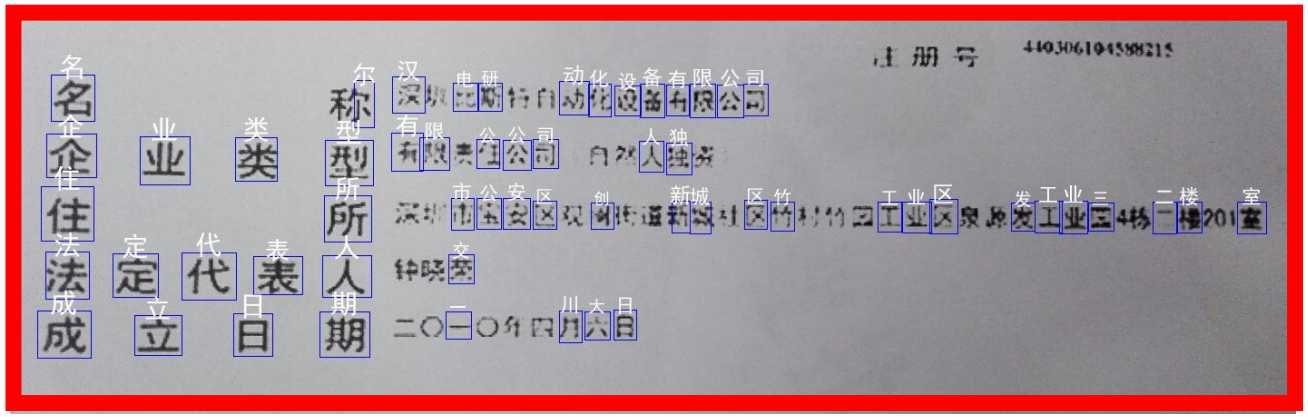
YOLO v2
自然场景

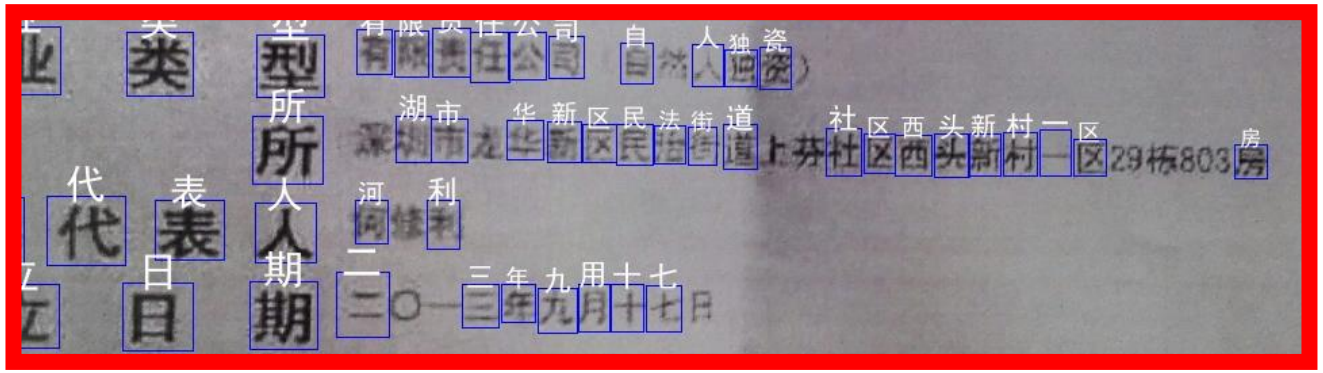
应用场景及挑战

拍摄环境不可控



地址字段识别





应用场景及挑战

身份证识别

识别结果

住址 河南省封丘县潘店镇蔡东村785号

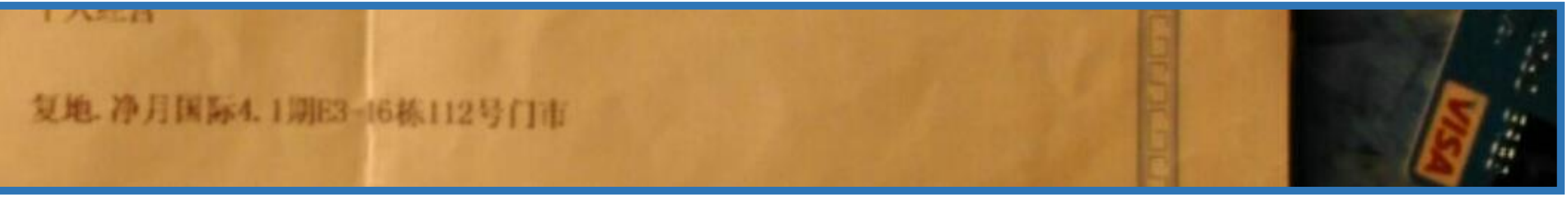
某OCR平台X

售地 河南省封丘县潘店镇赛东 村785号

某OCR平台Y

_是 河__省封压__福店镇赛东 村7B5号

应用场景及挑战

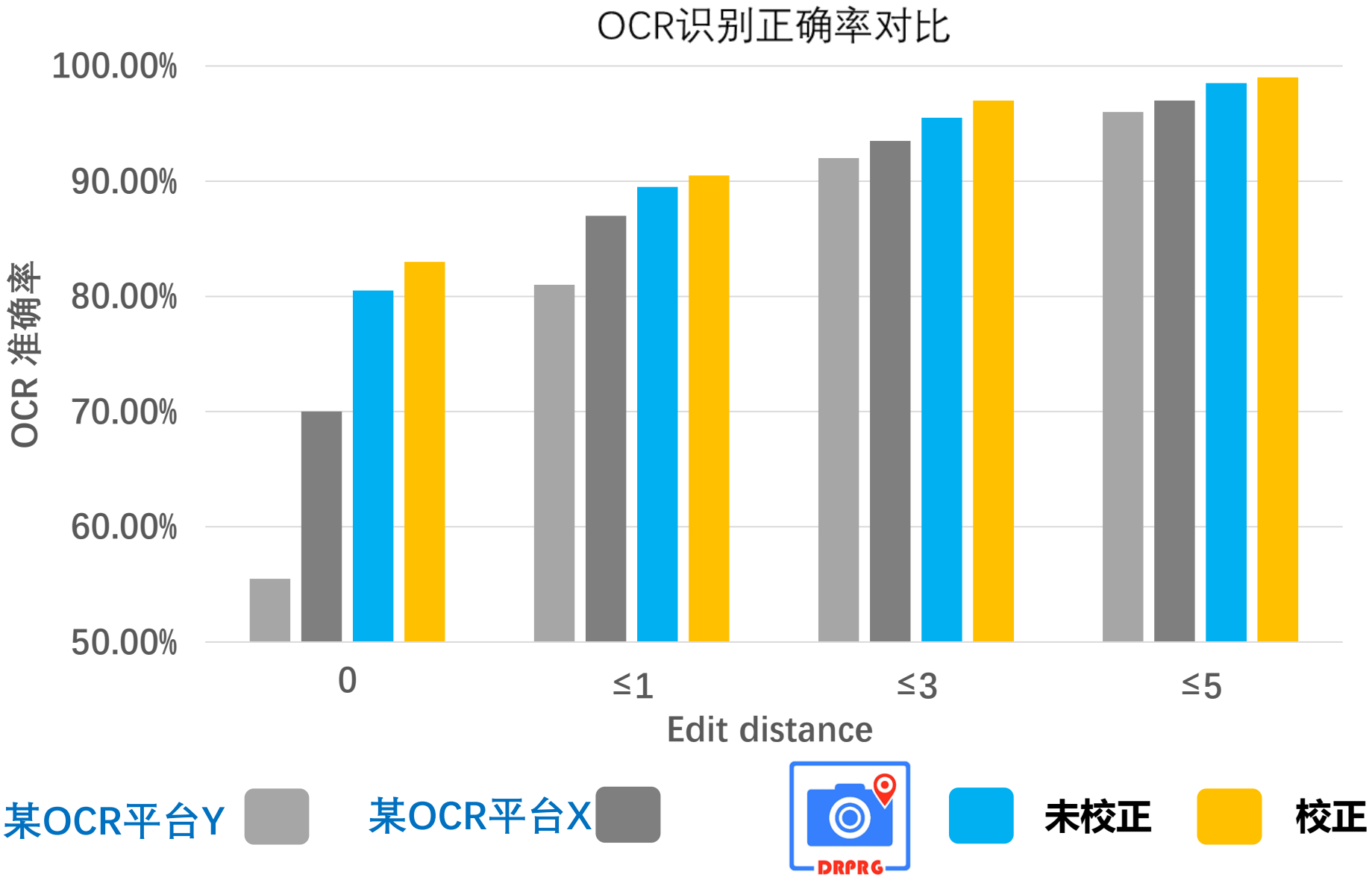


某OCR平台X :复地□净月国际4.1湘□3□16栋112号门市

某OCR平台Y ： 复地□净月国际4.1用B□-16□□12号门市

DRPRG ： 复地.净月国际4.1期E3-16栋112号门市

正确率对比


OCR识别正确率对比
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
OCR 准确率
0
≤1
≤3
≤5
Edit distance
某OCR平台Y
某OCR平台X
未校正
校正

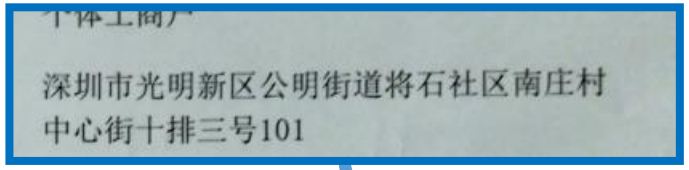
技术路线

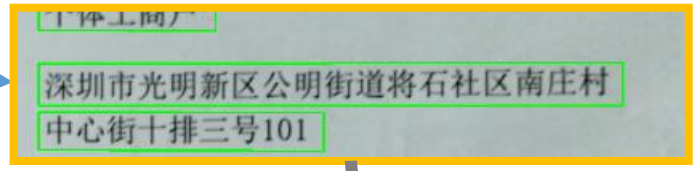
数据准备 → 数据标注与增强

文字定位 → CTPN

文字识别 → DenseNet+CTC

地址判断、校正 → 模糊匹配





- 个体土商厂
- 深圳市光明新区公明街道将石社区南庄村
- 中心街十排三号101

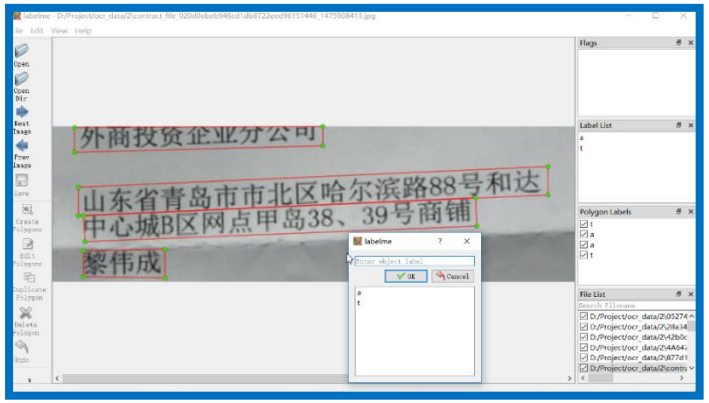
- ~~个体土商厂~~
- 深圳市光明新区公明街道将石社区南庄村中心街十排三号101

技术路线——数据准备

01

人工标注2000张图片

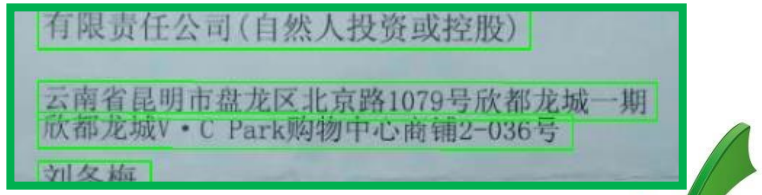
文本框坐标、信息分类

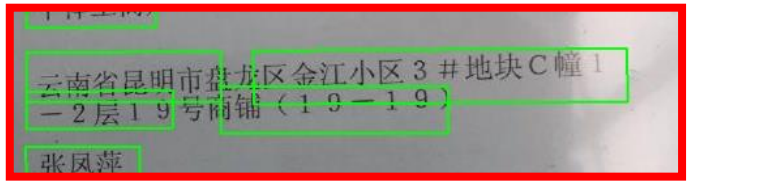


02

冷启动检测模型

处理7000张图片，筛选出效果较差样本，重新标注

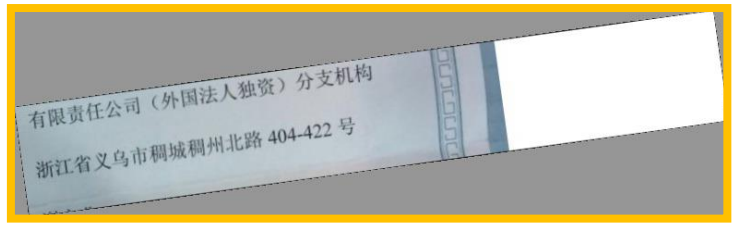




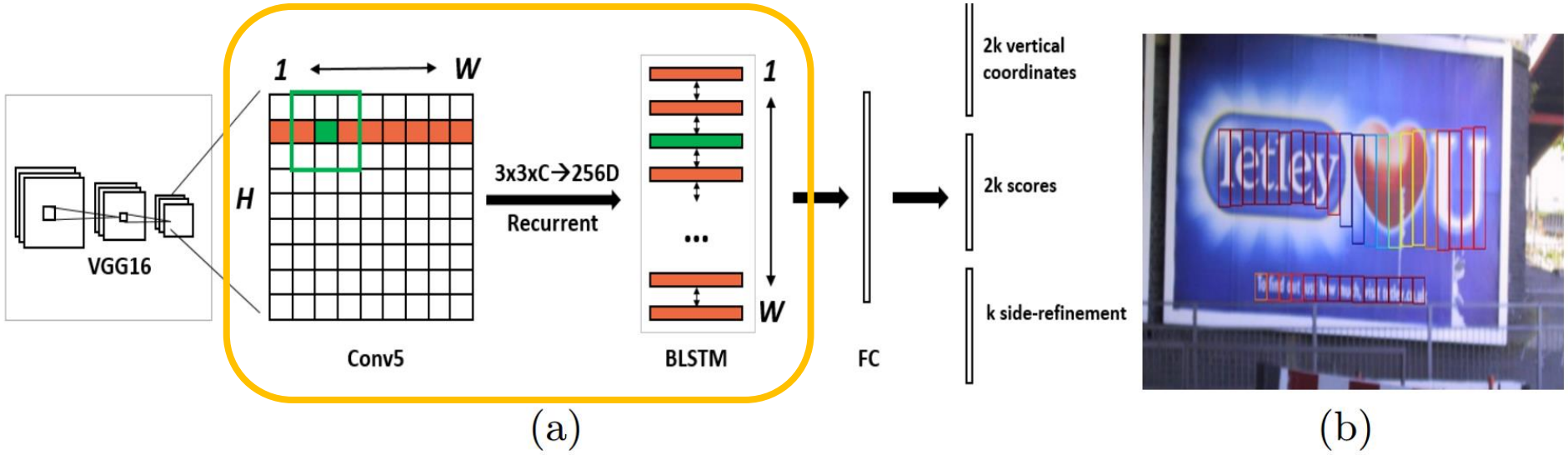
03

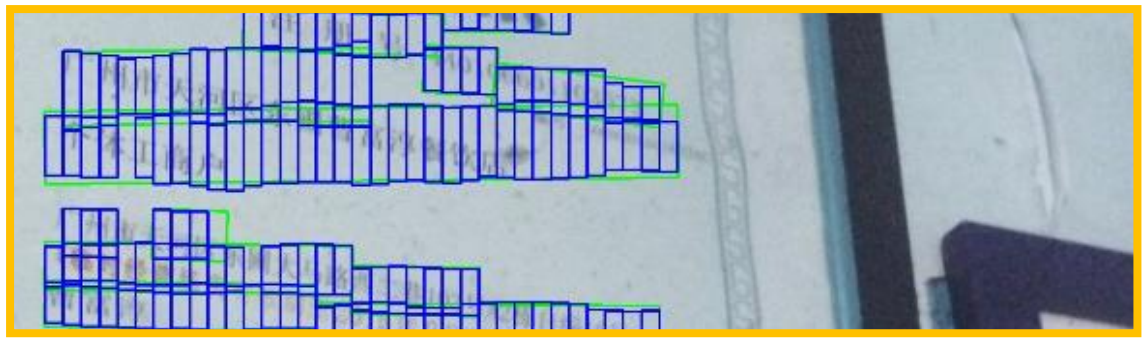
数据增强

旋转、翻转差样本，加入训练数据再次训练

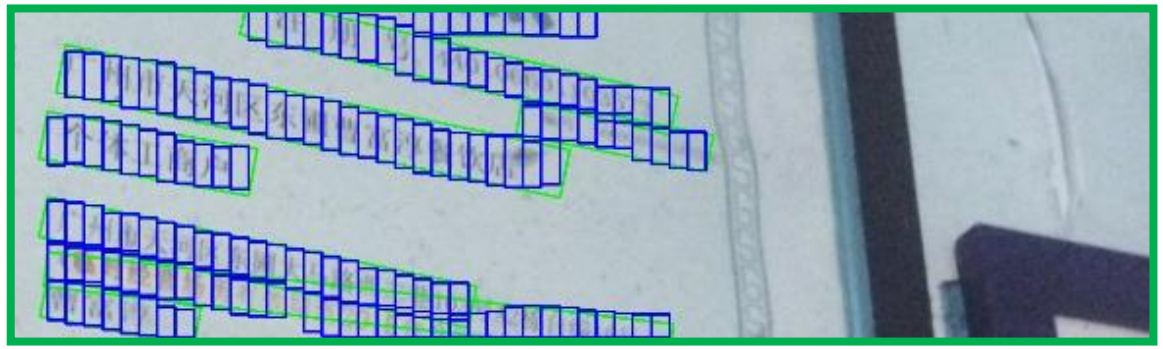


技术路线——文字定位





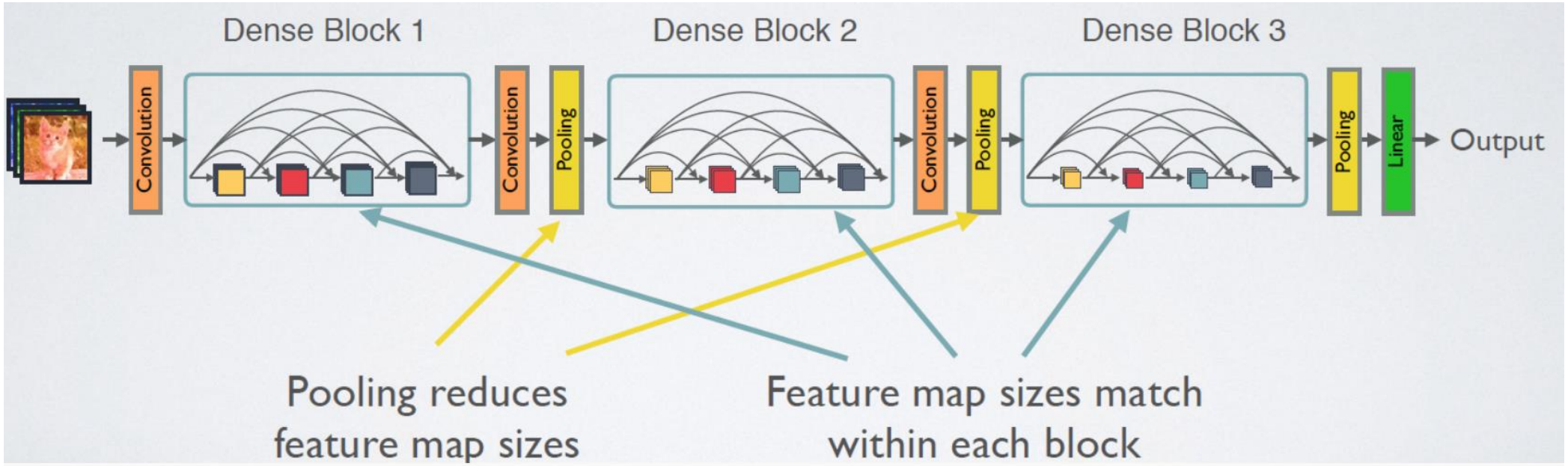
Bounding Box 改进前



Bounding Box 改进后

技术路线——文字识别

- **DenseNet** 图像特征提取



- **CTC** loss预测
- 使用DenseNet替换传统CRNN特征提取过程
- 不使用LSTM层，提升速度

技术路线——地址判断校正

地址判断： 基于LibLinear和结巴分词的短文本分类:**TextGrocery**

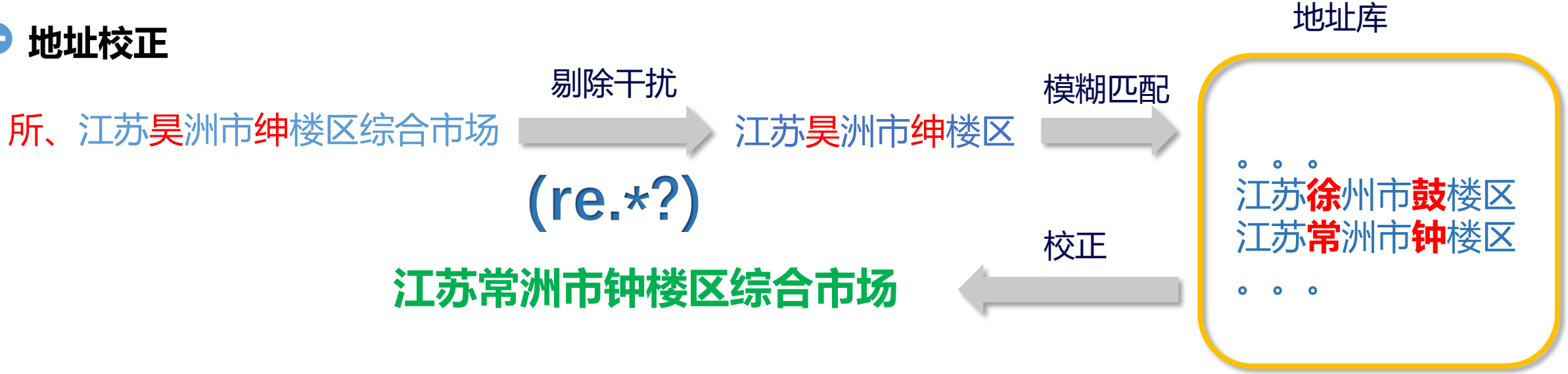
- 类别预测

Others ： 个体工商户
Address： 山东省威海市文登区龙山办富尔佳商场
Address： 一楼3号门市
Others ： 赵洪明

- 拼接：

山东省威海市文登区龙山办富尔佳商场一楼3号门市

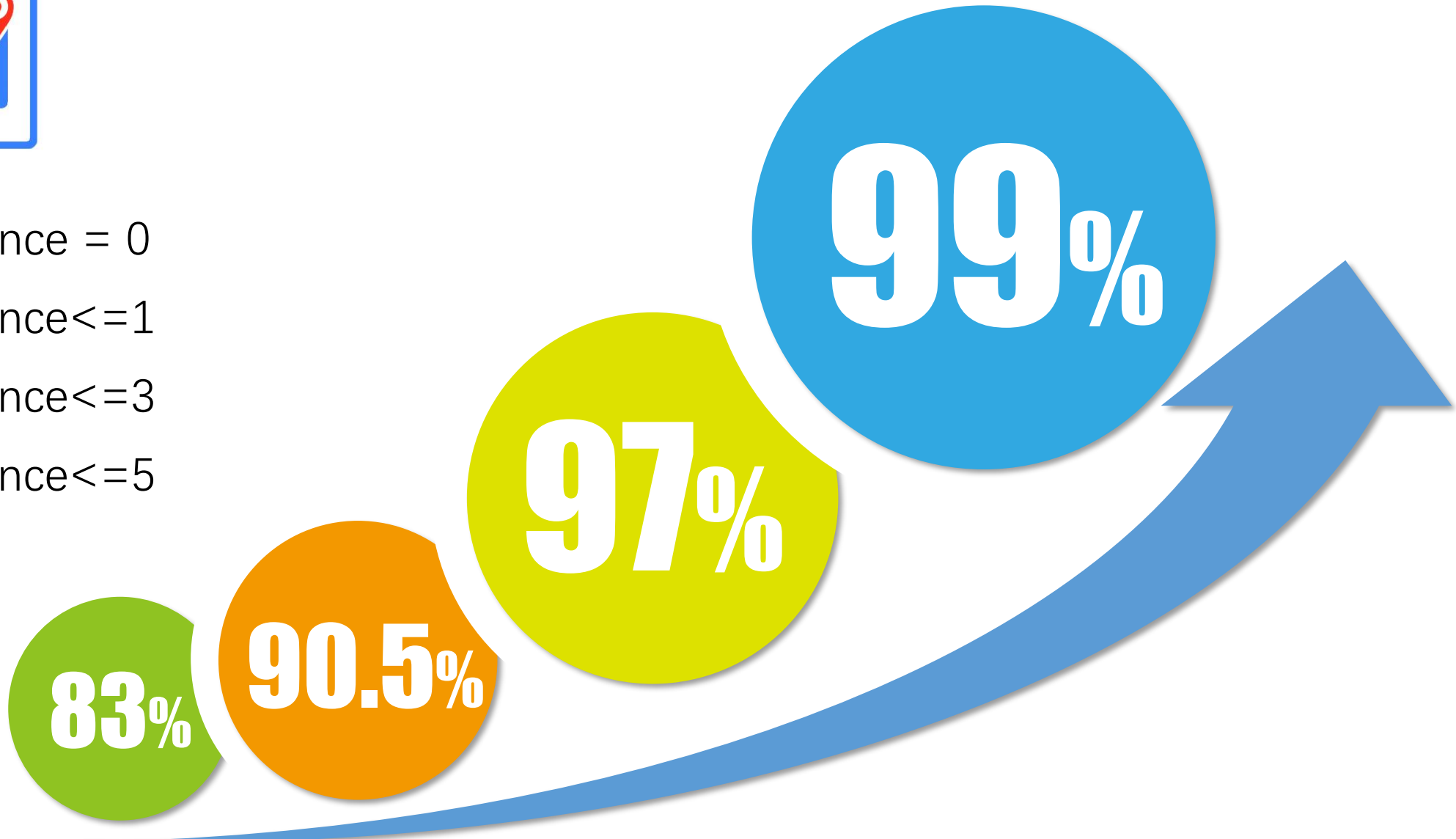
地址校正



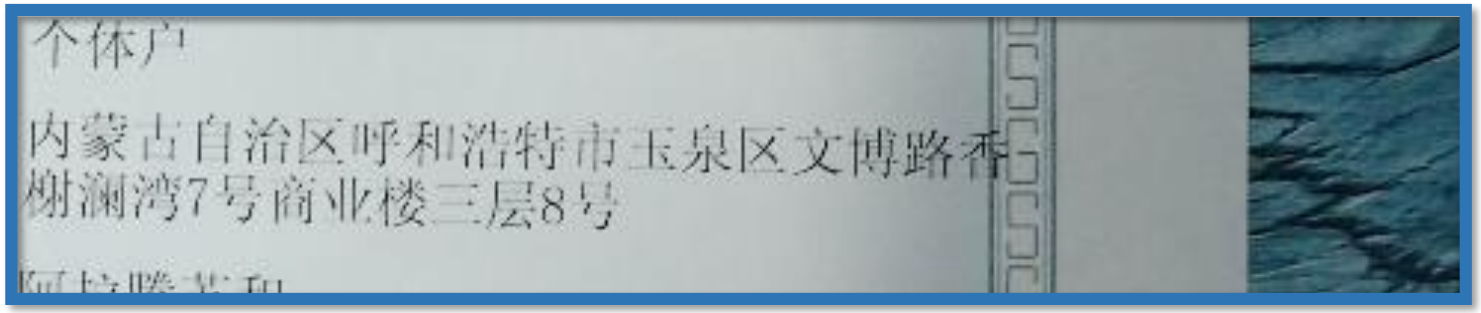
DEECAMP | 成果展示

DRPRG

- Edit Distance = 0
- Edit Distance<=1
- Edit Distance<=3
- Edit Distance<=5



成果展示

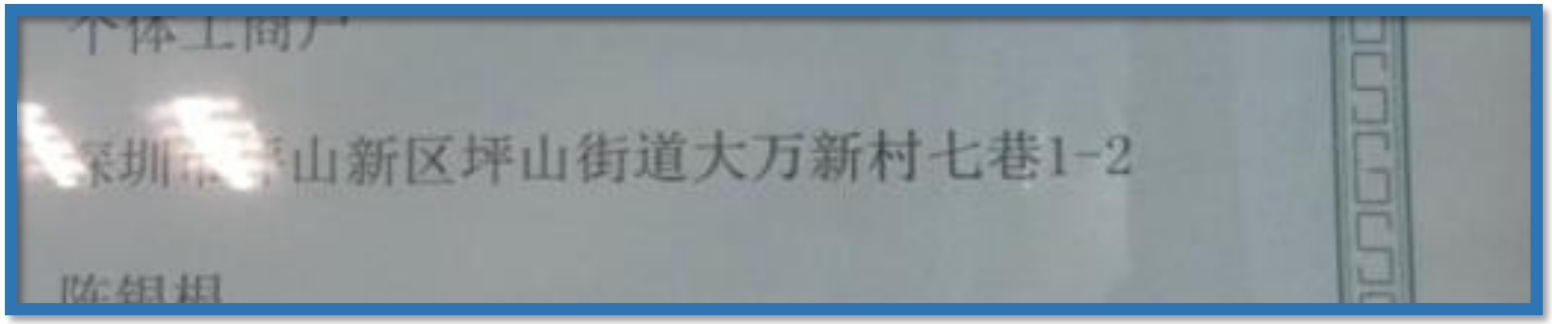


某OCR平台X：内蒙古自治区呼和浩特市玉泉区文博路香5
榭澜湾7号商业楼 层8号

某OCR平台Y：内蒙占自治区呼和沿特市玉泉区文博路
谢澜湾7号商业楼三层8号

DRPRG：内蒙古自治区呼和浩特市玉泉区文博路香
榭澜湾7号商业楼三层8号

成果展示

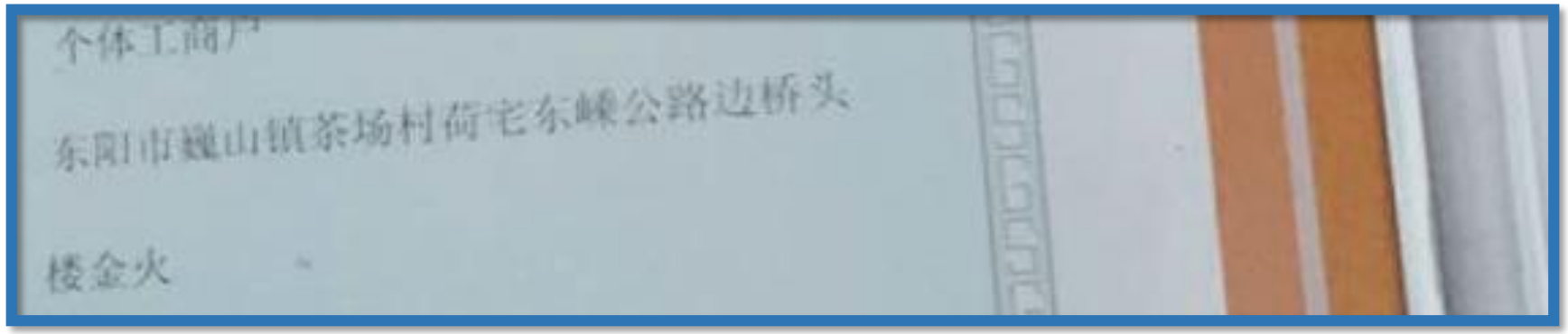


某OCR平台X ：深圳市手山新区坪山街道大万新村七巷1-2

某OCR平台Y ：□□□□山新区坪山街道大万新村七巷1-2

：深圳□坪山新区坪山街道大万新村七巷1-2

成果展示



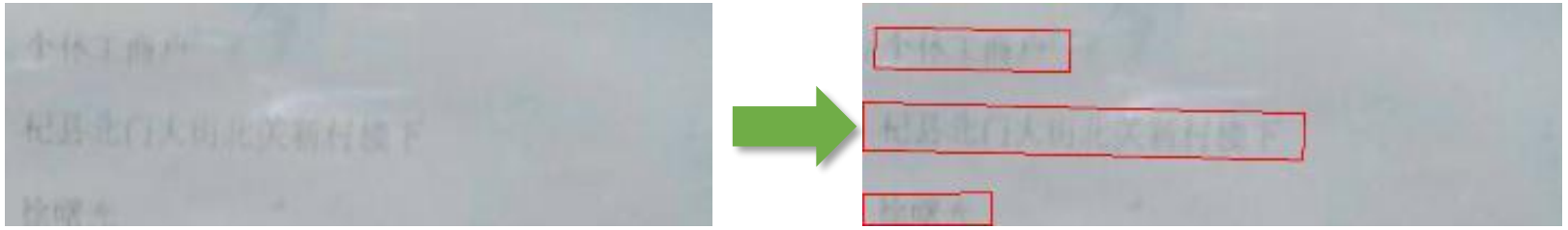
某OCR平台X ：东阳市桃山镇茶场村街宅东限公路边桥头

某OCR平台Y ：东阳市□山镇茶场村荷宅东□公路边桥头

DRPRG ：东阳市巍山镇茶场村荷宅东嵊公路边桥头

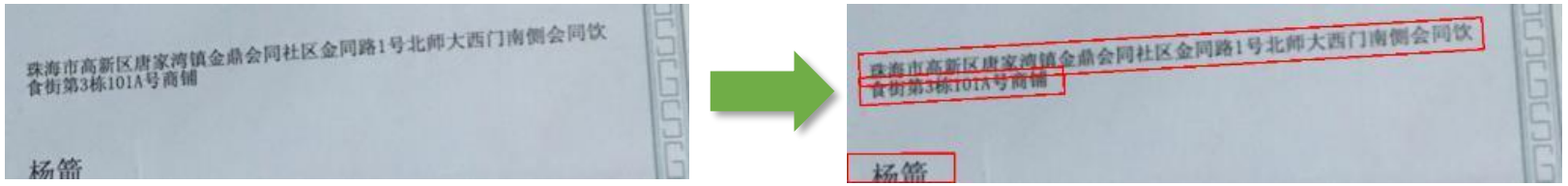
DEECAMP | 成果展示

非常模糊



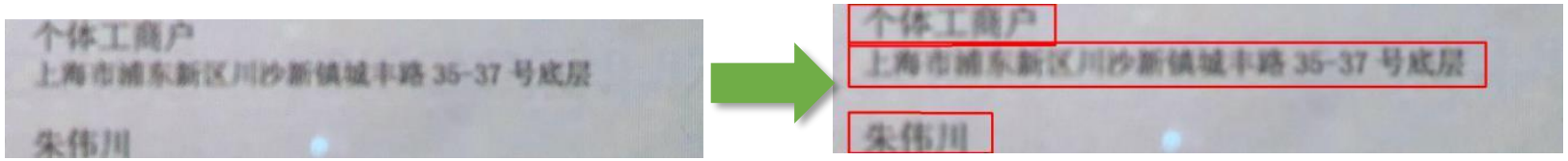
真实标签：
杞县北门大街北关新村楼下
识别：
杞县北门大街组关新村楼6

多行长文本



真实标签：
珠海市高新区唐家湾镇金鼎会同社区金同路1号北师大西门南侧会同饮食街第3栋101A号商铺
识别：
完全正确

场景干扰

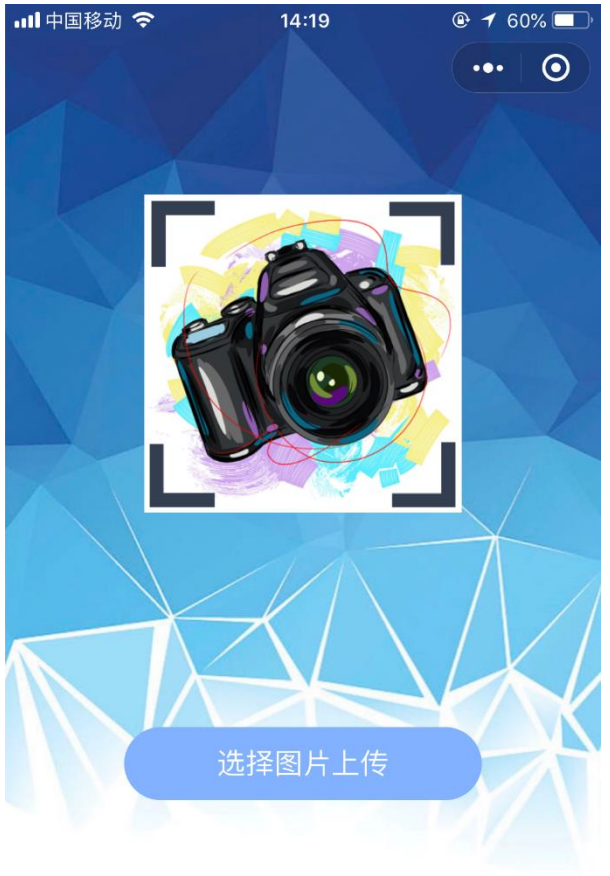


真实标签：
上海市浦东新区川沙新镇城丰路35-37号底层
识别：
完全正确

成果展示

小程序展示

OCRdeecamp



DEECAMP | 未来改进

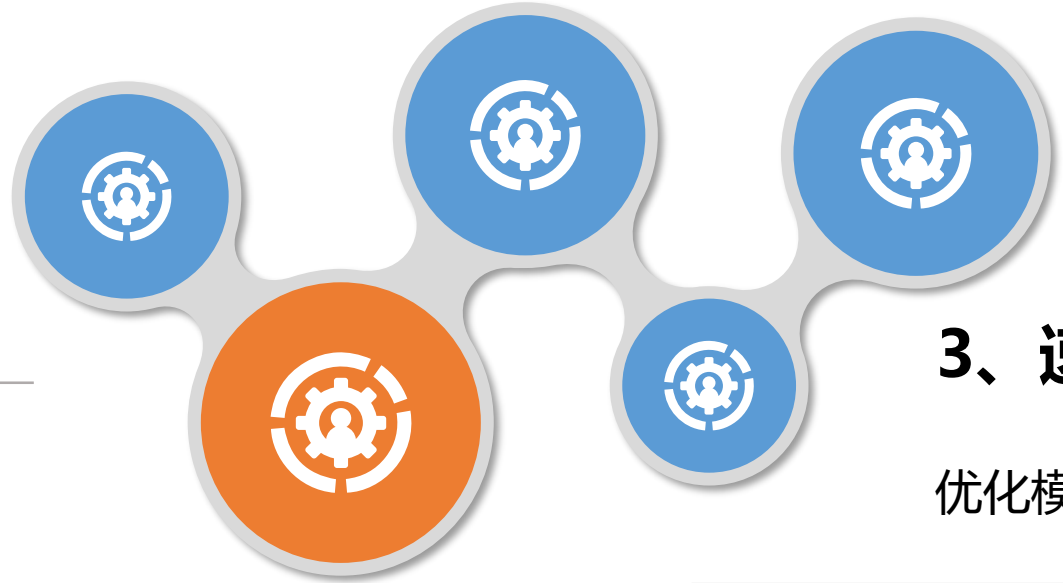
2、更多数据

有多少人工才有多少智能!

1、地址矫正

使用信息更完备的地址库，

模糊匹配



3、速度提升

优化模型、压缩参数

4、OCR未来方向

通用模型 VS 专用模型

谢谢！

Deecamp 25

阿不都维力　陈云天
段　晨　顾笑风
李沐辰　李新慧
孟　晋　裴胜兵
田寅兵　张继元
郑雅文　庄新瑞