



环境不可控场景下 拍照文档地址文字识别



DEECAMP



CONTENTS

1

应用场景及挑战

OCR Applications and Challenges

2

技术路线

Technical Routine: Detection, Recognition & Calibration

3

成果展示

Results Show

4

未来改进

Future Work & Some thinking



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

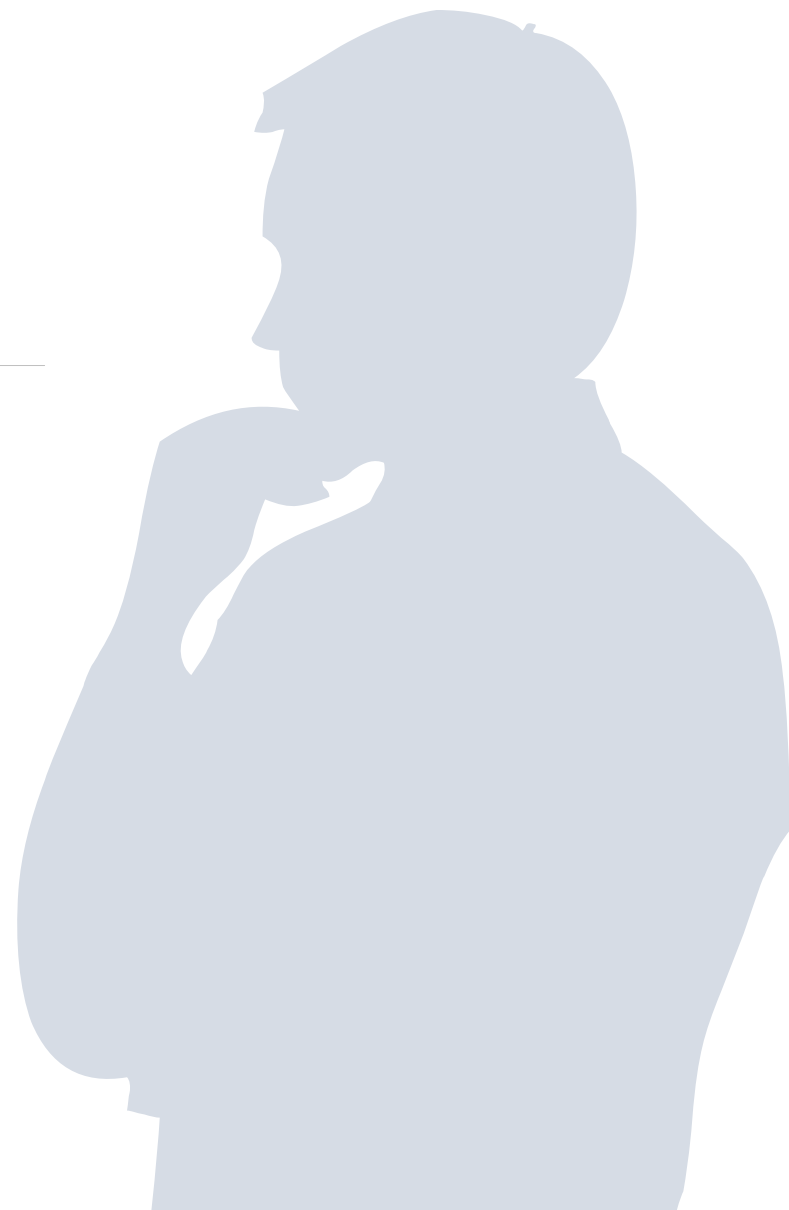
挑战

+ 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





DEECAMP

应用场景及挑战



+ YOLO v2
自然场景





DEECAMP

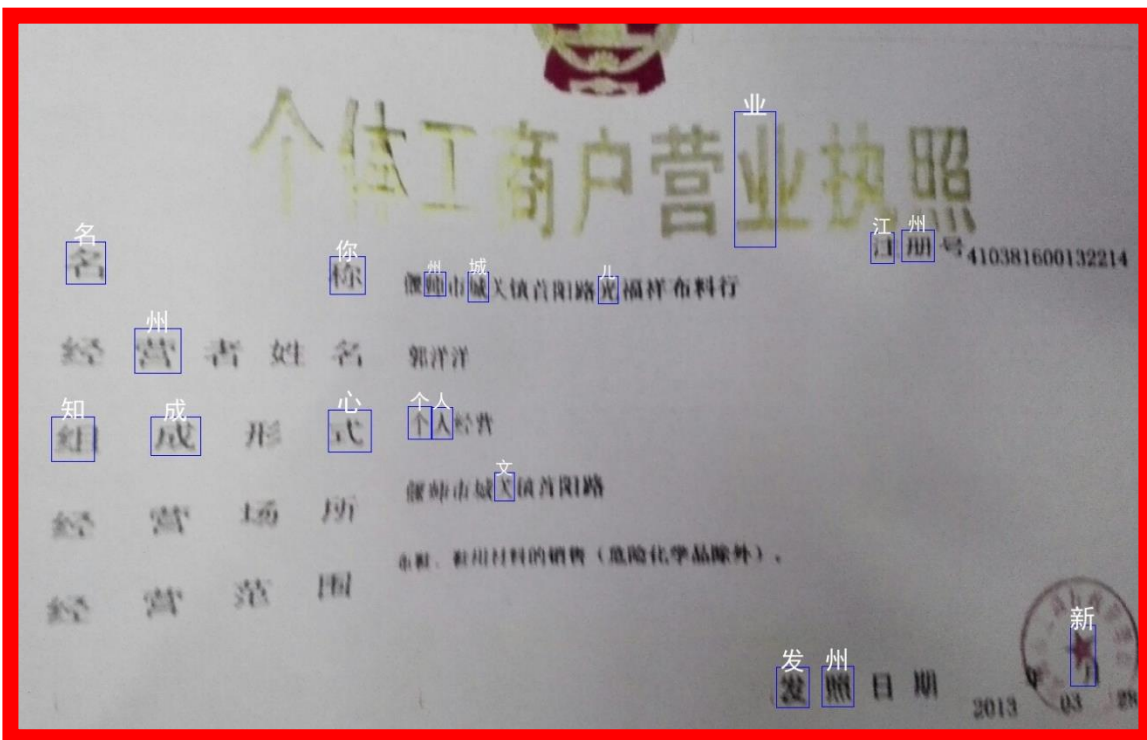
应用场景及挑战



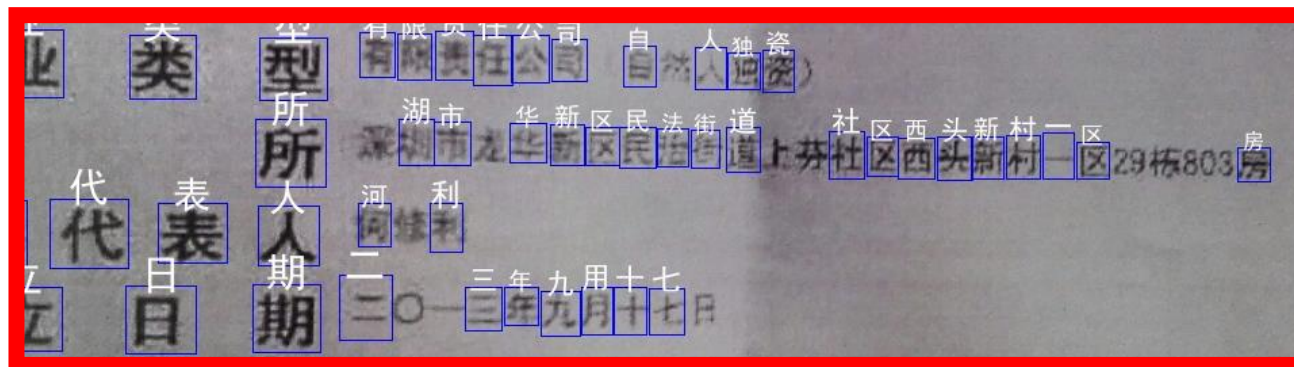
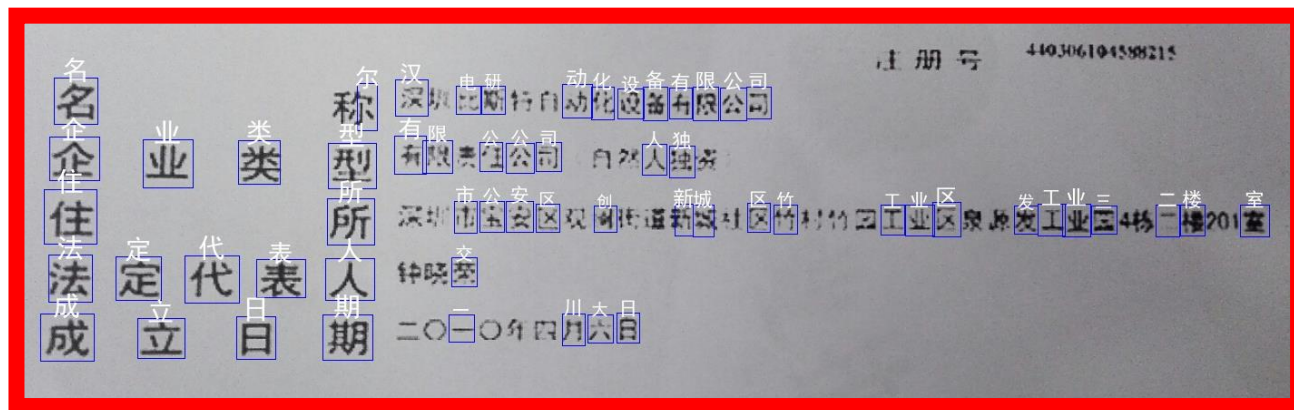
+ YOLO v2
自然场景



+ 拍摄环境不可控



+ 地址字段识别



+ 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

Tencent 腾讯

售地 河南省封丘县潘店镇赛东 村785号

Baidu 百度

是 河 省封压 福店镇赛东 村7B5号



DEECAMP

应用场景及挑战

复地.净月国际4.1期E3-16栋112号门市

Tencent 腾讯:复地 净月国际4.1 湘 3 16栋112号门市



: 复地 净月国际4.1 用 B -16 12号门市



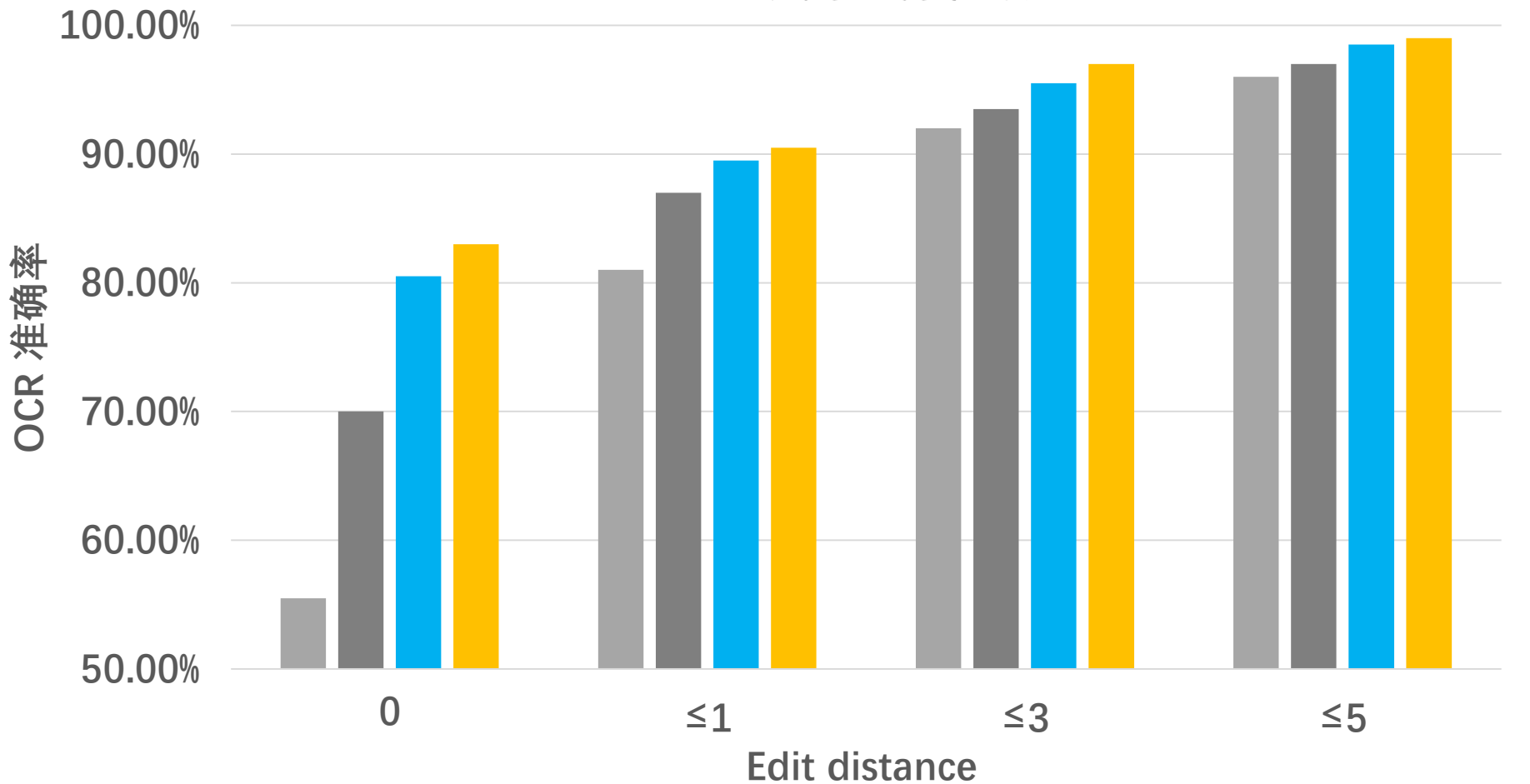
: 复地.净月国际4.1期E3-16栋112号门市



DEECAMP

正确率对比

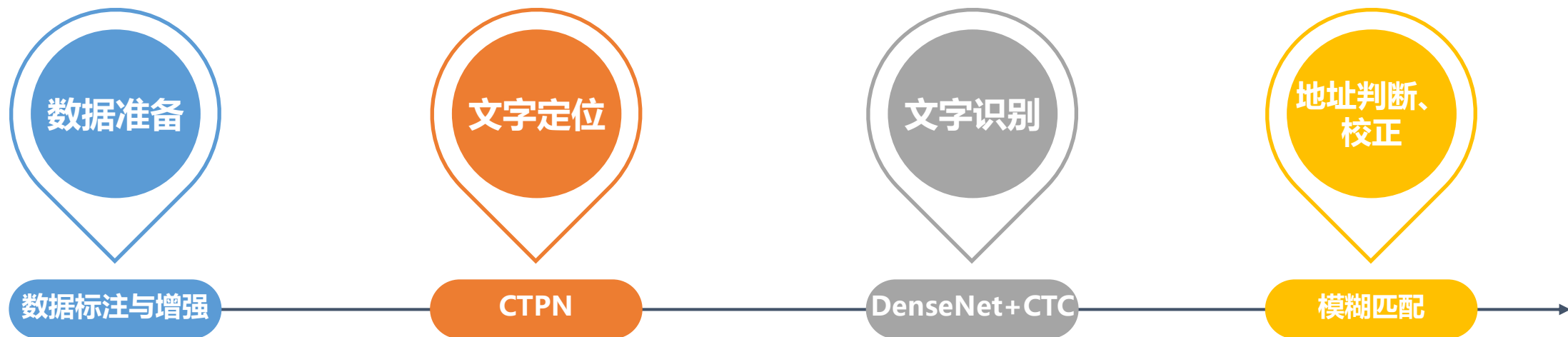
OCR识别正确率对比



未校正



校正



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

➤ 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

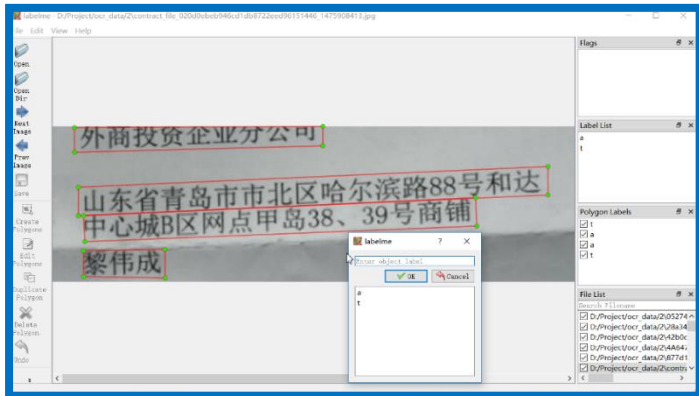
➤ ~~个体土商厂~~
➤ 深圳市光明新区公明街道将石社区南庄村中心街十排三号101



01

人工标注2000张图片

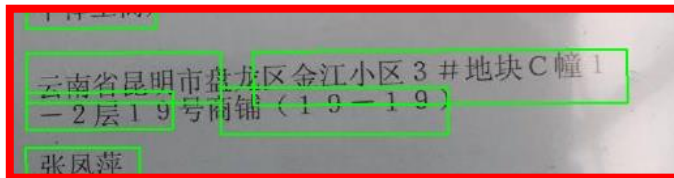
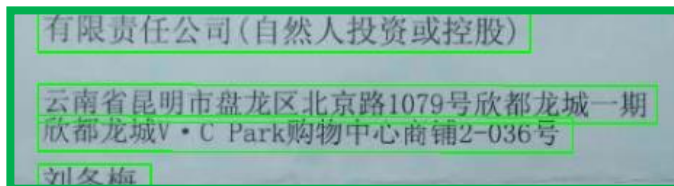
文本框坐标、信息分类



02

冷启动检测模型

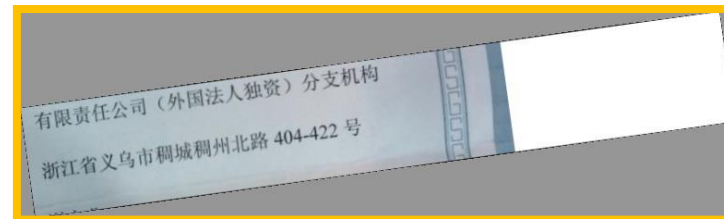
处理7000张图片，筛选出效果较差样本，重新标注



03

数据增强

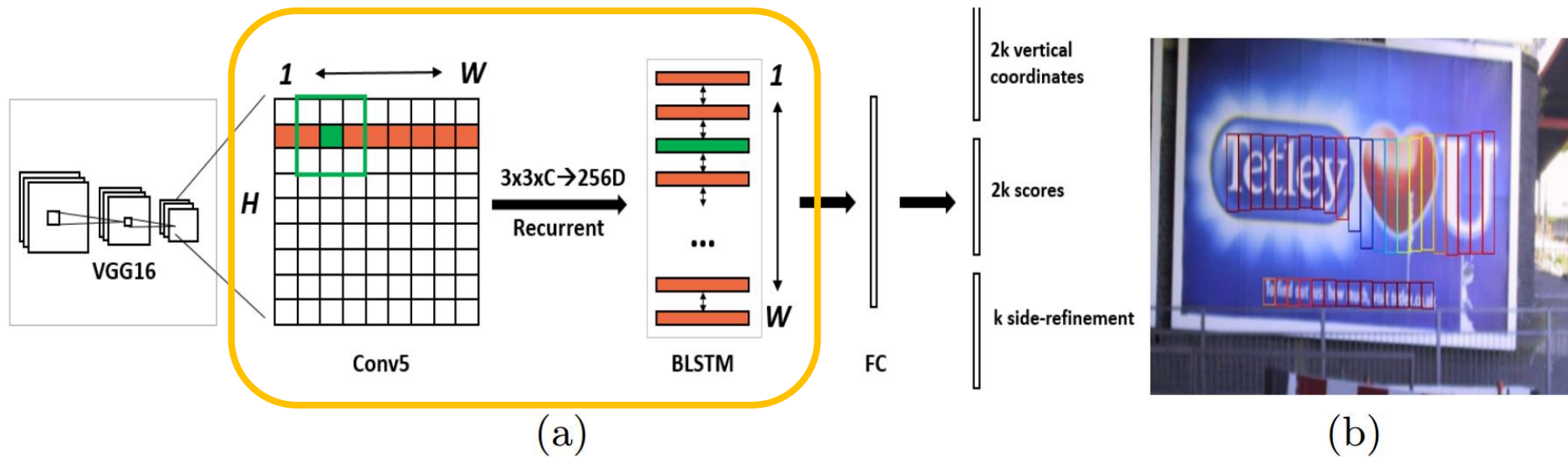
旋转、翻转差样本，加入训练数据再次训练





DEECAMP

技术路线——文字定位



+ Bounding Box 改进前

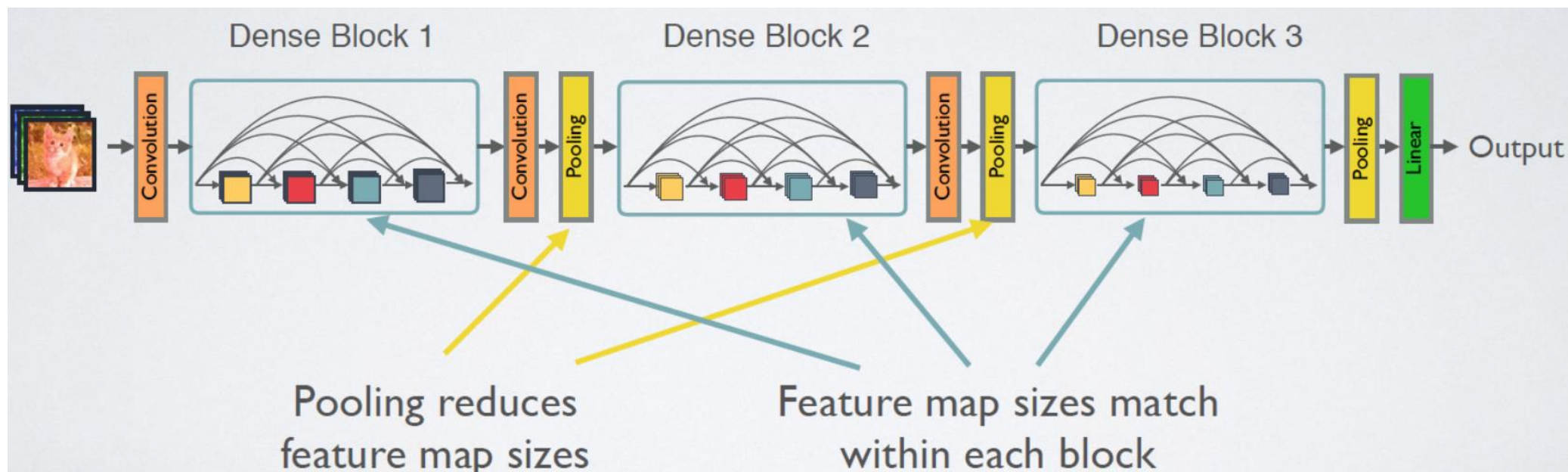


+ Bounding Box 改进后



DEECAMP | 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



DEECAMP

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

➤ 拼接：

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

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**

所、江苏昊洲市绅楼区综合市场

剔除干扰

(re.*?)

江苏昊洲市绅楼区

模糊匹配

地址库

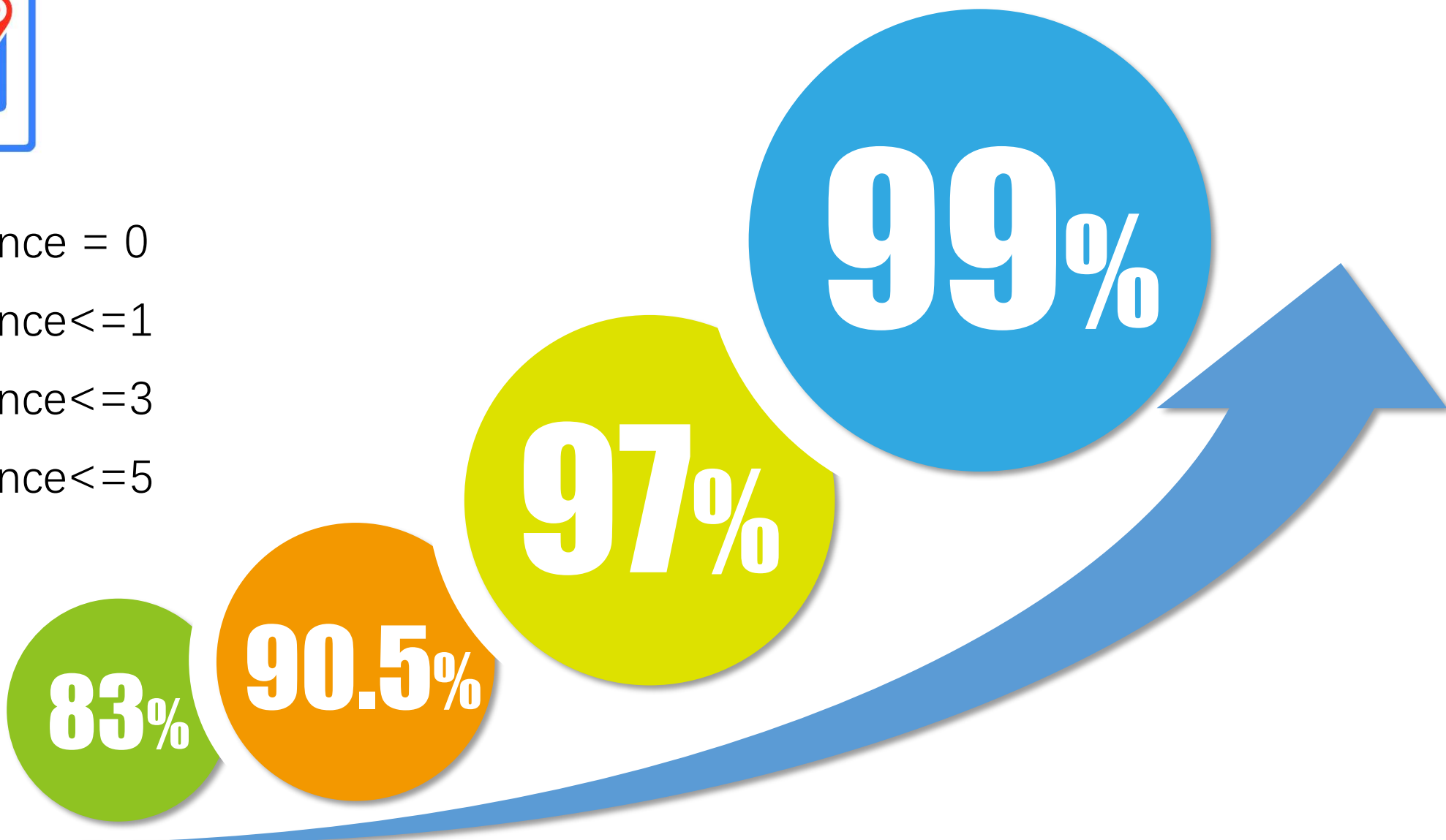
。 。 。
江苏徐州市鼓楼区
江苏常州市钟楼区
。 。 。

校正

江苏常州市钟楼区综合市场



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





DEECAMP

成果展示

个体户

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

阿拉腾其和

Tencent 腾讯：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



：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

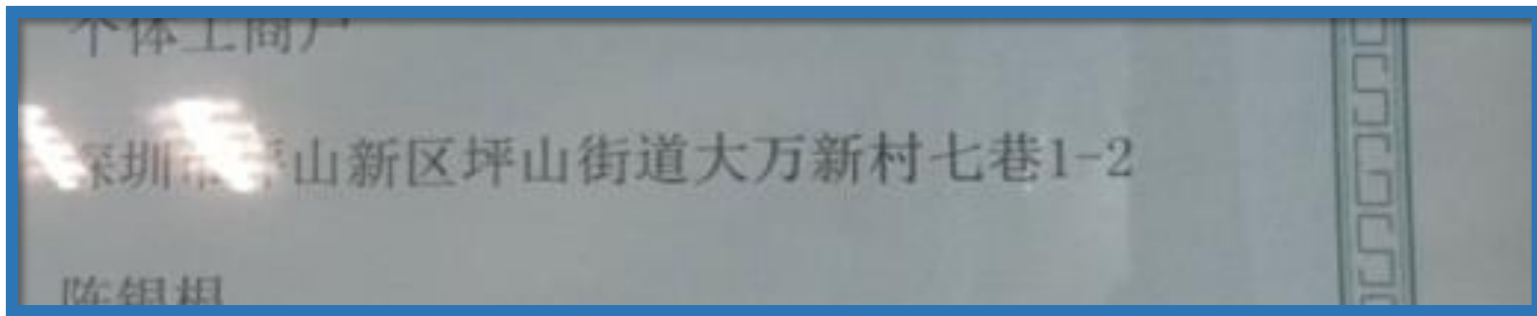


：内蒙古自治区呼和浩特市玉泉区文博路香
榭澜湾7号商业楼三层8号



DEECAMP

成果展示



Tencent 腾讯 : 深圳市 **手** 山新区坪山街道大万新村七巷1-2



Baidu 百度 : **□□□□** 山新区坪山街道大万新村七巷1-2

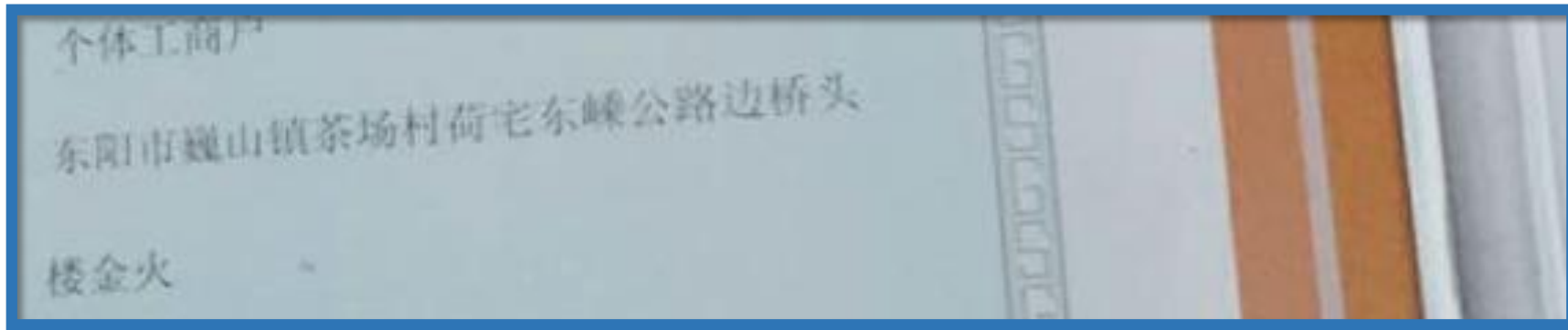


DRPG : 深圳 **□** 坪山新区坪山街道大万新村七巷1-2



DEECAMP

成果展示



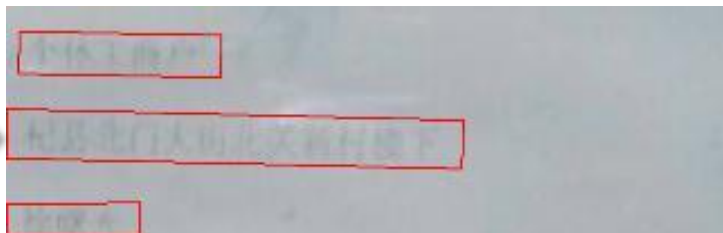
Tencent 腾讯：东阳市桃山镇茶场村街宅东限公路边桥头

Baidu 百度：东阳市□山镇茶场村荷宅东□公路边桥头



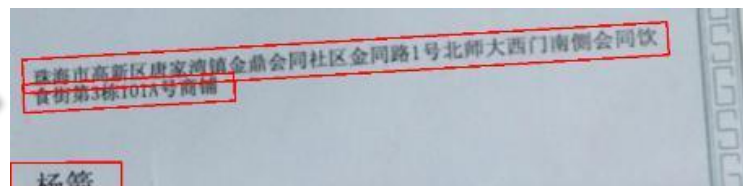
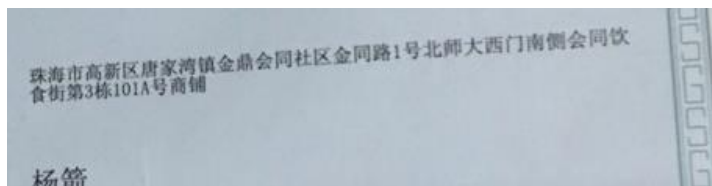
：东阳市巍山镇茶场村荷宅东岷公路边桥头

+ 非常模糊



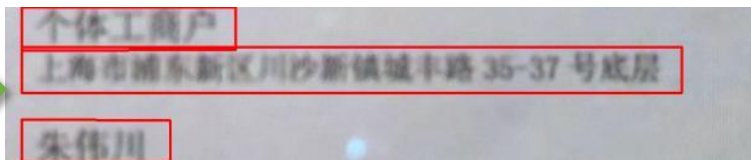
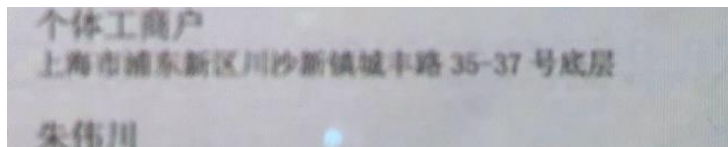
真实标签：
杞县北门大街北关新村楼下
识别：
杞县北门大街组关新村楼6

+ 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

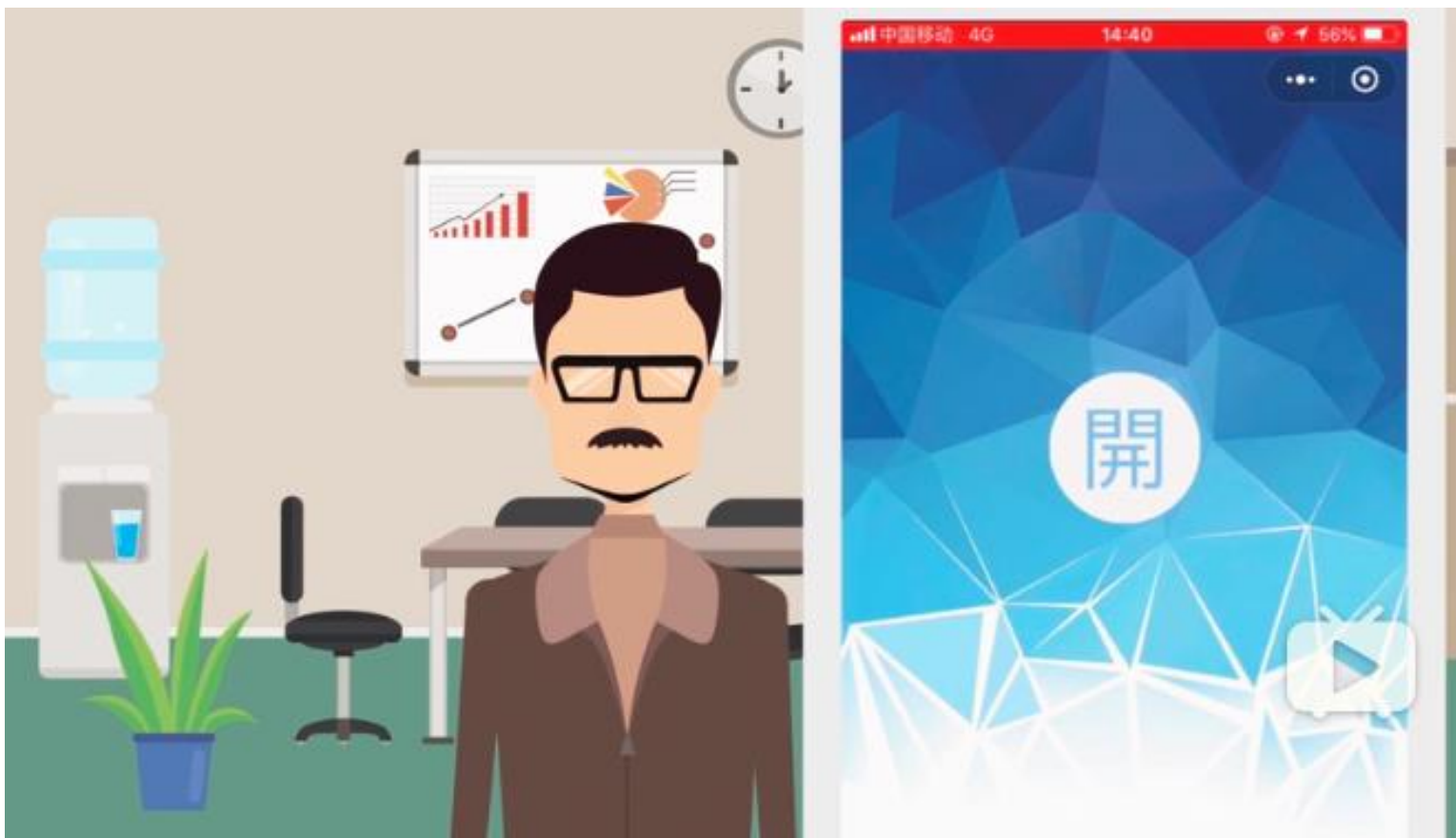
+ 场景干扰



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

Demo 视频

<https://www.bilibili.com/video/av30081208>

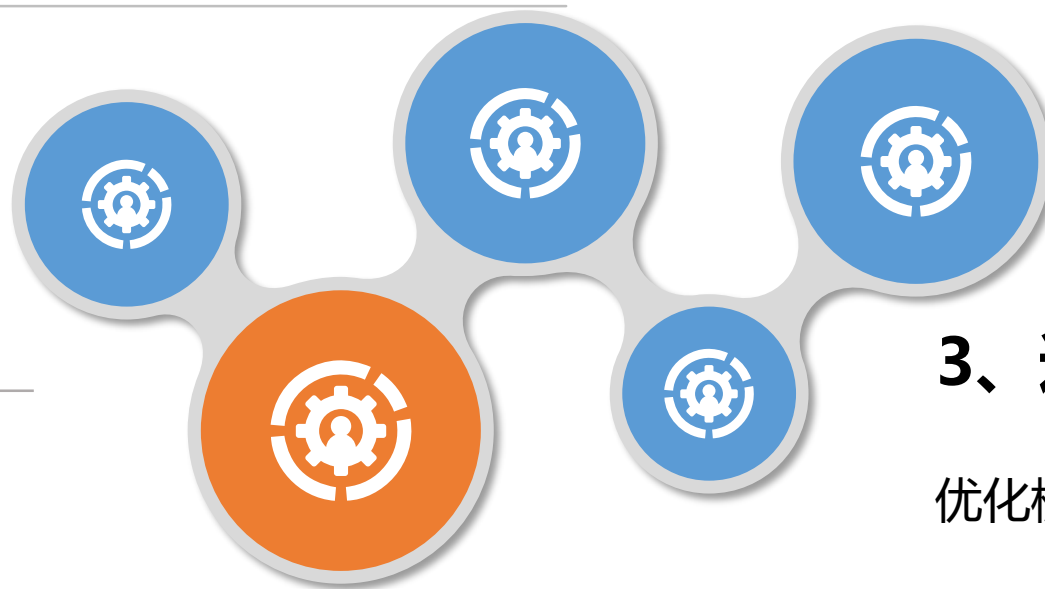


2、更多数据

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

