北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

深度学习图像算法在内容安全领域的应用

网易易盾 李雨珂
2019.8.24

出品：LiveVideoStack 音视频技术社区 CSDN

深圳 2019

遨游“视”界 做你所想
Explore World, Do What You Want

LiveVideoStackCon 2019 深圳

2019.12.13-14

出品：LiveVideoStack 音视频技术社区

成为讲师： speaker@livevideostack.com

成为志愿者： volunteer@livevideostack.com

赞助、商务合作： kathy@livevideostack.com

目标

展示人工智能商业**落地案例**

分享深度学习算法**优化经验**

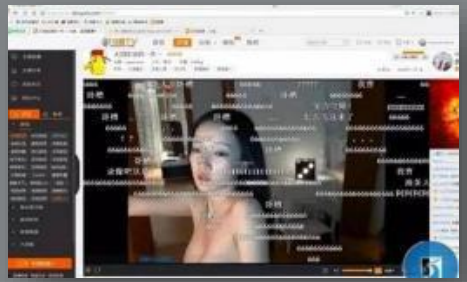
1. 背景介绍

2. 初期探索

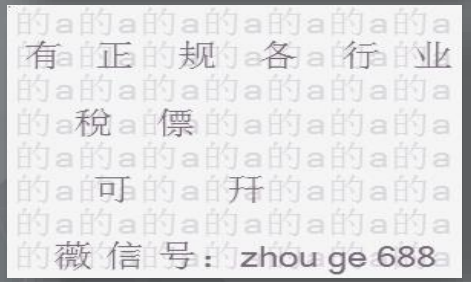
3. 优化过程

4. 延伸与总结

互联网内容安全




手机轰炸机
微信wnqmqa


有 正 规 各 行 业
税 僄
可 扦
薇 信 号： zhou ge 688


暴恐
色情
低俗
违禁
有害信息
垃圾
广告
涉政

互联网内容安全

国家网信办针对网络乱象启动专项整治行动

大量社交类、音视频类APP下架

互联网内容安全

《中华人民共和国网络安全法》
《出版管理条例》
《信息网络传播权保护条例》
《互联网新闻信息服务管理规定》
《互联网信息服务管理办法》
《网络出版服务管理规定》

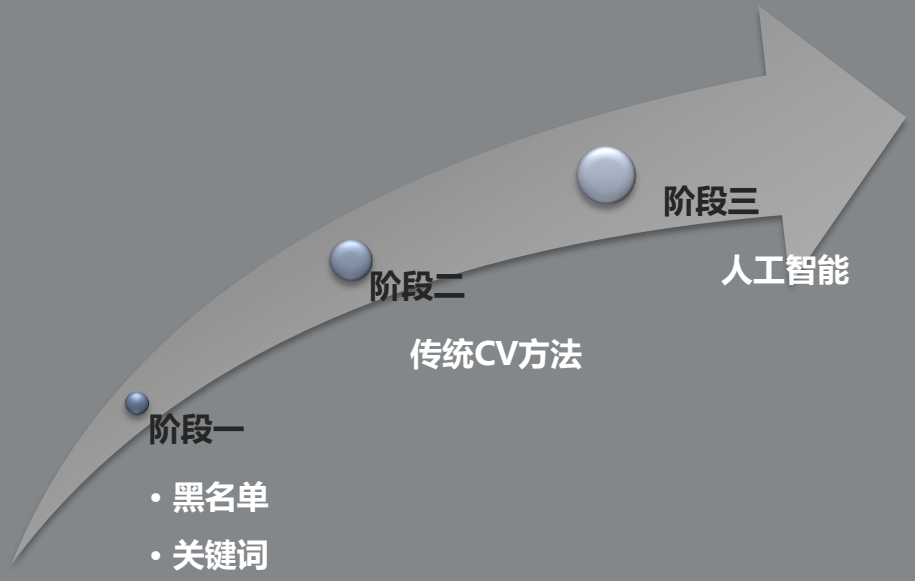
“没有网络安全就没有国家安全”
——习近平总书记

随着“互联网+”战略的提出，各行各业与互联网的结合越来越紧密。融合创新的新业态使得各关键行业和重要系统对网络安全保障的需求不断增加，安全已成为网络强国建设的基础保障，习近平总书记倡导建设“和平、安全、开放、合作”的网络空间，更加把网络安全和数据安全上升到国家战略的高度。

人工审核与机器审核

人工主导

人工审核与机器审核



机器配合

机器审核面临的挑战

前期海量数据资源的要求

- 垃圾类型数据收集难度大，覆盖类型有限

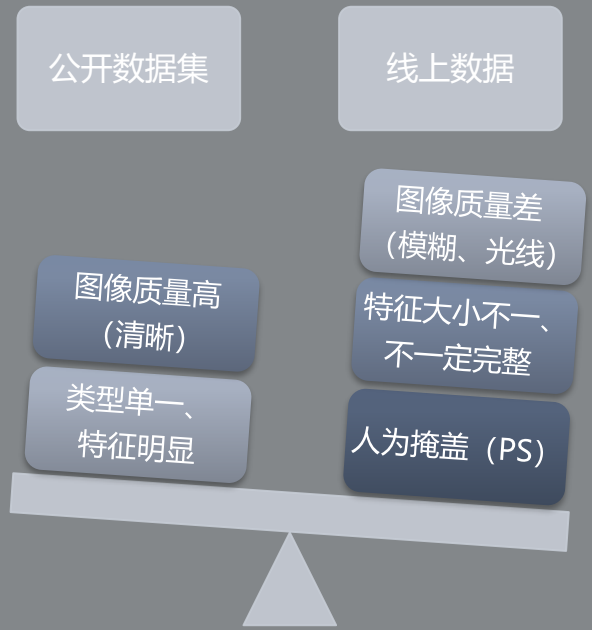
后期投入和运营维护

- 无底洞：随着业务和形态的发展，以及黑灰产攻防的升级，需要不断投入大量人力、物力

前期数据壁垒

数据成本高

数据分布不均



后期样本攻防

以广告数据为例

- 垃圾内容变种快
- 特征趋近模糊、小、不完整

常规广告

文字无涂鸦

印刷体广告

文字清晰

横排文字、文字区域大

变种广告

涂鸦、文字乱序

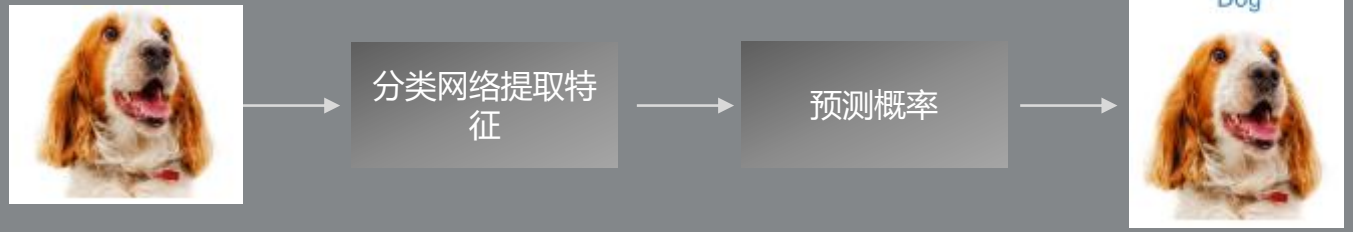
手写体广告

文字透明处理

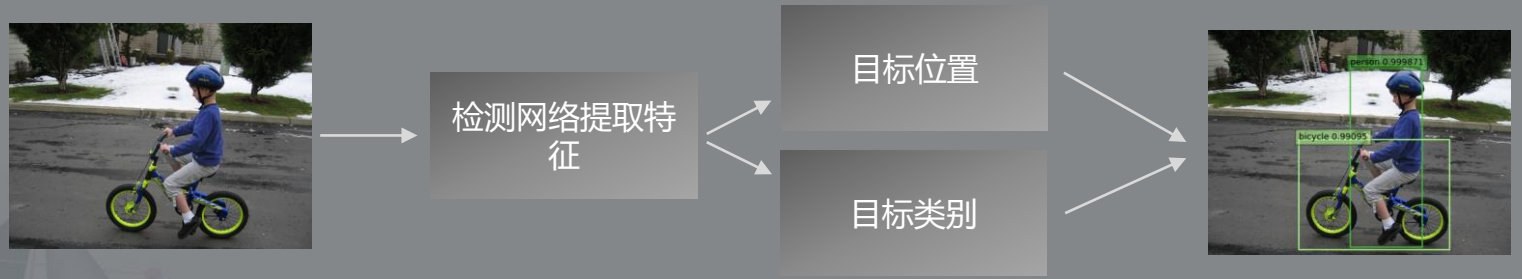
文字扭曲、区域小

采用的技术

图像分类算法：ResNet、DenseNet、SENet等



目标检测算法：YOLO-v3, SSD, RefineDet等



初期探索

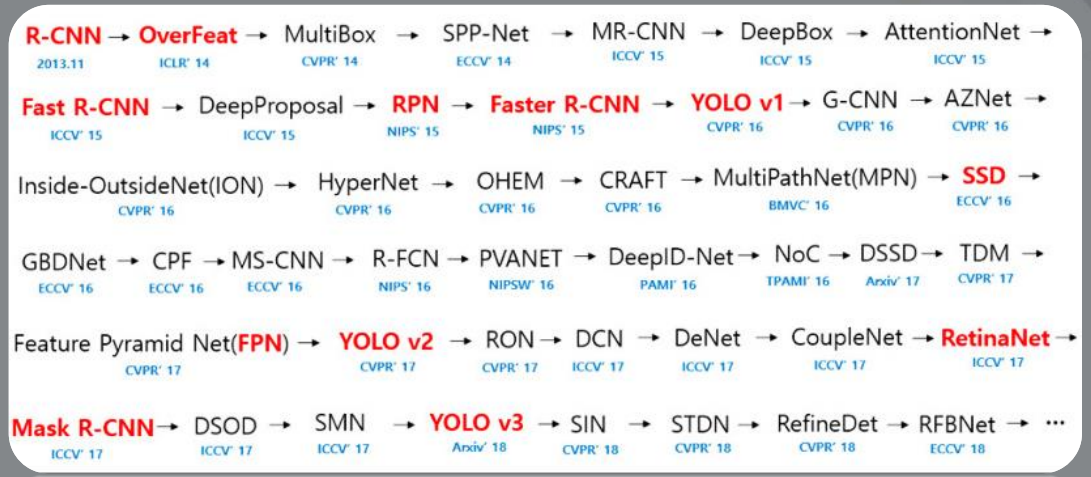
沉浸于

- 模型搭建，跟进前沿方法
- 炼丹（模型调参）

实际上

- 方法带来的优势往往不大
- 公开数据集上有优势的方法不一定适用数据集

（*在前期数据积累有限的情况下*）



初期探索

算法指标 *vs* 用户诉求（重方法，轻诉求）

主要出现问题

- ***误判**问题*：大量不可解释的误判
- ***漏判**问题*：特征不明显、较模糊的样例不能召回

优化过程

定义业务标准

- 每一个细分类都需要有明确的描述，标注作出判断时有明确依据。（人无法分清，机器更加糊涂）

明确重要程度

- 全局角度，放弃一些零碎的偶发样例，集中解决某一类型的问题。

广告为例

美女广告

垃圾广告

优化过程

测试标准，挑选更有代表性的测试

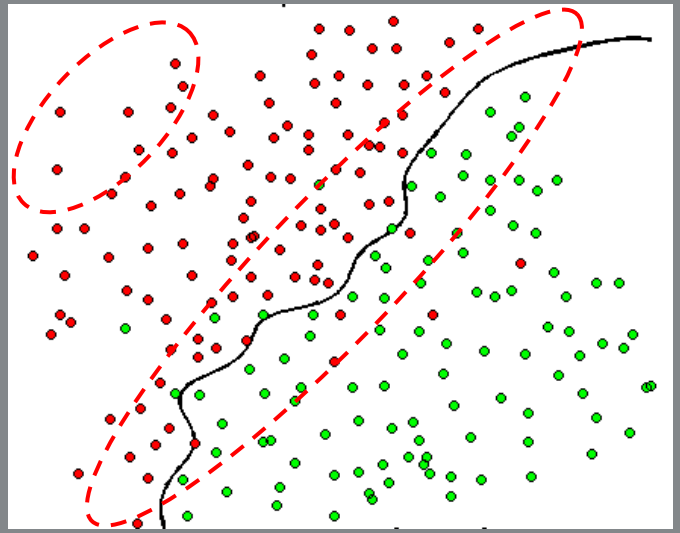
基础测试集	**线上**数据	**特定类型**集	**历史反馈**集	预发测试
• 至少十万级别，基准效果	• 千万级别、本地测试，误判评估	• 针对高频出现类型，漏判评估	• 收集历史反馈，效果提升评估	• 模拟上线情况，减少突发问题

优化过程

数据层面

- 闭环迭代，数据回流，同时确定性 和 解释性

明确数据回流
（增加确定性）



边界数据回流
（增加解释性）

TIPS

- ✓ 先测试后训练
- ✓ 利用模型工具捞取误漏判数据
- ✓ 重视数据分布、多样性

优化过程

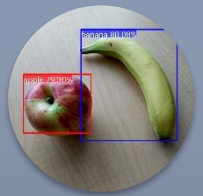
漏判优化

数据回流扩充正样本

（内部）

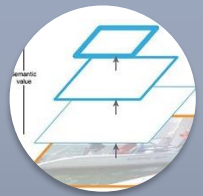
定向收集数据

（外部）



目标检测辅助

（特写区域）



FPN + ATTENTION

（多尺度、小目标）

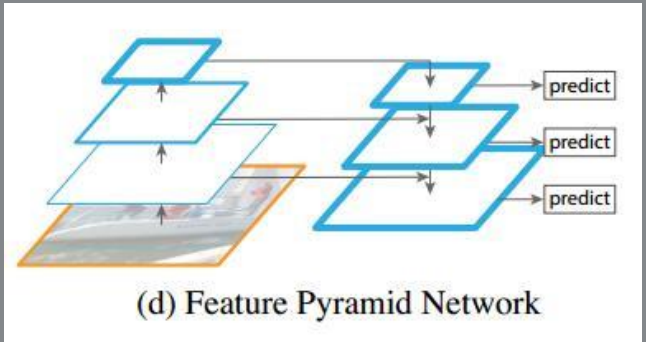
OCR

其他技术辅助

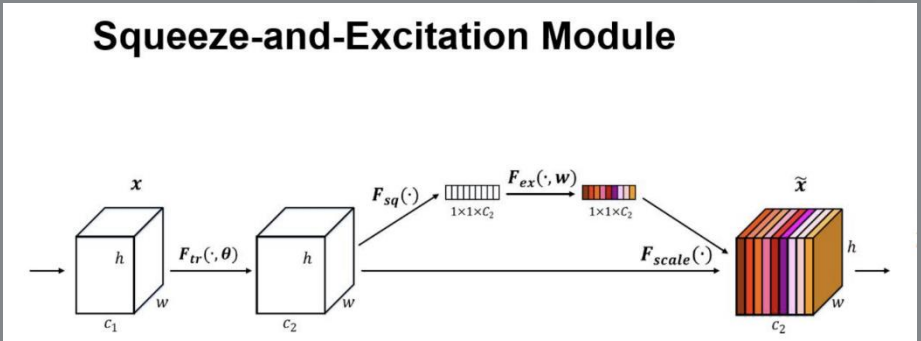
（OCR，二维码，图片库，图像聚类，用户维度）

漏判优化

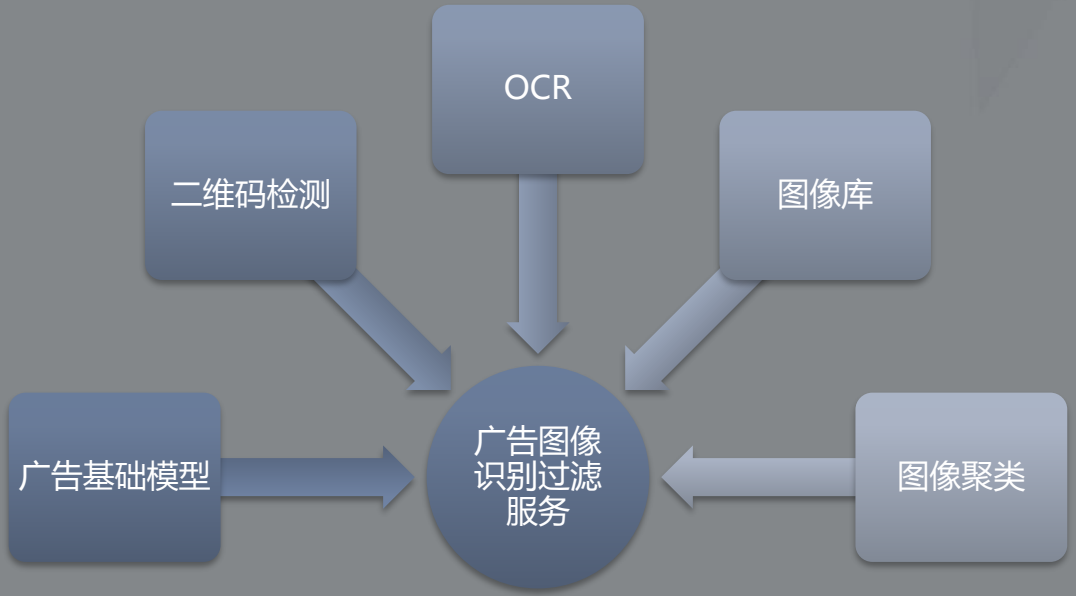
Detection: SSD + FPN



Classification: Attention



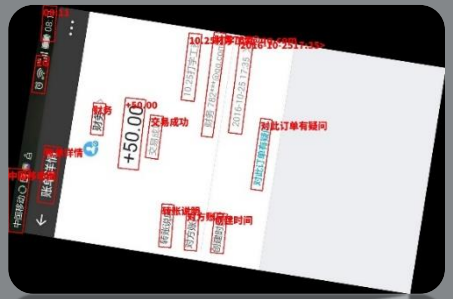
多技术手段辅助

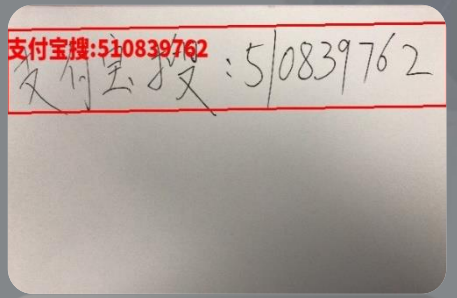


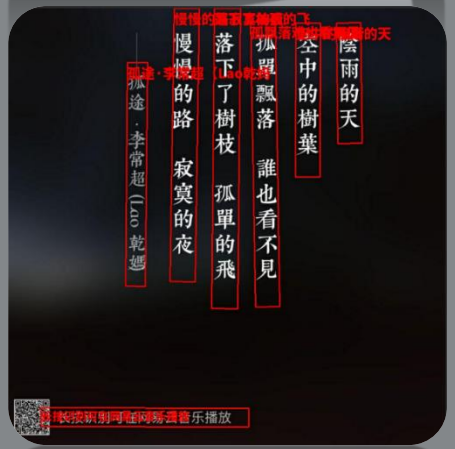
OCR辅助

主要解决

- 倾斜，倒立，仿射变换
- 竖排
- 特殊字体和排版
- 手写体



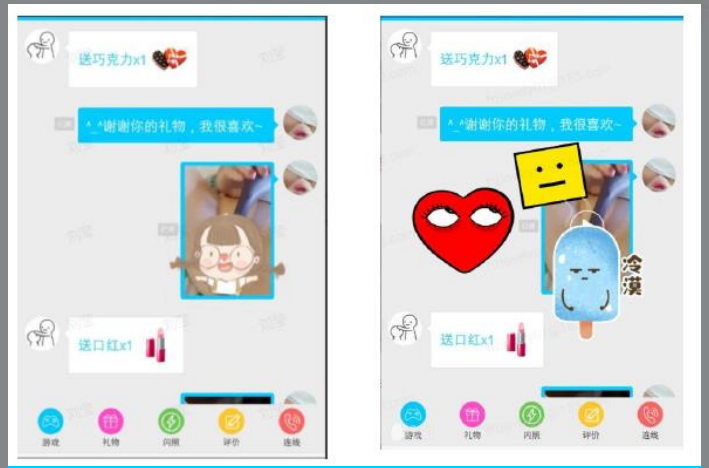




图片库辅助

同源图像检索

剪裁



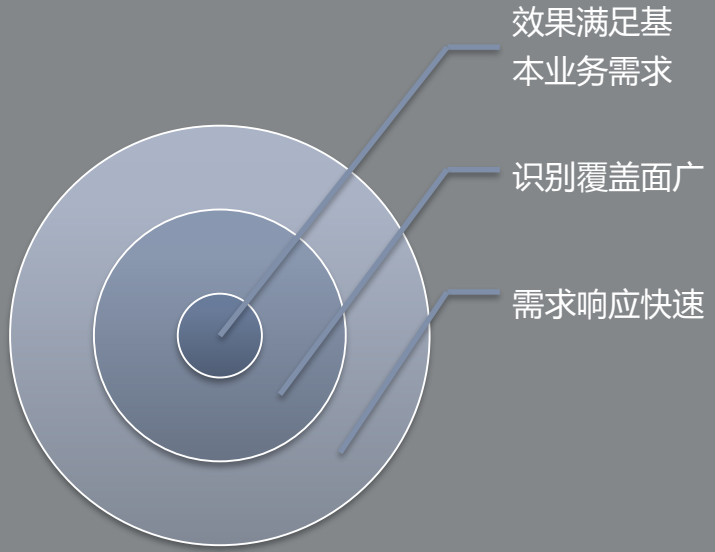
涂改

优化过程小结

优化收益：

- 问题定义与方案
- 数据捞取
- 模型选择与调参

业务效果



核心模块

漏判（整体）：万分之三以内

精度（确定）：97%以上

进一步工作

1 业务输出精细化

2 模型层面精细化

3 模型性能优化

进一步工作

图像业务横向拓展

- Logo识别
- 旗帜识别

平台化支撑 vs 独立精耕细作

- 快速支撑新类型
- 垂直优化已有业务

架构

客户反馈收集 | 线上主动跟踪 | 抽样审核 | 全流程测试

算法运营 & 内容审核

Docker | Kubernetes | 服务监控

算法服务部署

图像预处理 | 图像检索 | 特征库

工程包装

业务反馈集\对抗 | 测试环境QA集 | 本地大规模测试集

算法质量控制

资源占用评估 | 性能瓶颈评估 | 算法压力测试

算法性能控制

特定样例挖掘 | 注意力机制 | 结构设计 | 多尺度训练 | 模型融合

难样例挖掘 | 特征金字塔 | 损失设计 | 多任务训练 | 策略优化

算法效果优化

模型剪枝\量化 | 推理加速

模型压缩与加速

需求

标准

样例

业务输入

图像分类 | 图像检测 | 人脸标注 | 语音标注 | 图中文字

数据标注平台

分类网络 | 检测网络 | 人脸 | OCR | 语音 | 视频 | 图像检索 | 无监督

基础算法储备（2B反垃圾场景）

数据自动回流 | **样本挖掘** | 定向爬虫

数据收集

最佳实践模板 | 自动训练框架 | **基础剪枝模型（主干）**

底层算法框架

延伸

音频技术

声音检测

- 娇喘、呻吟
- ASMR
- 枪击爆炸

语言检测

- 语种识别

荣誉

先进单位证书

- 获得公安部授予的【网络安保先进单位】
- 获得杭州市公安局授予的【互联网安全管理工作先进单位】
- 获得杭州市公安局授予的【G20网络安全管理荣誉单位】
- 获得新闻出版局颁发的【网络视听节目审核】准许证书
- 获得2018年雷锋网评选的短视频联盟常务理事奖

2018年，网易检测量**3000亿+**的信息，年度检测数据总量行业内第一

荣誉

中国人工智能竞赛


中国人工智能·多媒体信息识别技术竞赛
北京百度网讯科技有限公司
四川远鉴科技有限公司
上海依图网络科技有限公司
深圳市网心科技有限公司
网易（杭州）网络有限公司
北京奇虎科技有限公司


A级
CERTIFICATE OF
FLAG RECOGNITION
旗帜识别
To 网易（杭州）网络有限公司
中国人工智能·多媒体信息识别技术竞赛组委
China Artificial Intelligence Multimedia Information
Recognition Technology Competition
Organizing Committee
2019 8 9

总结

- **目标**：重视问题定义，保持全局角度
- **数据**：关注数据有效收集，大批量标注一时爽，花费时间精力
- **成本**：重视成本与效率，数据收集成本，问题解决投入成本与机器成本
- **定制**：场景决定精细化程度，通用方案难度较大

北京
2019

遨游“视”界 做你所想
Explore World, Do What You Want

Thank you

出品：LiveVideoStack 音视频技术社区 CSDN