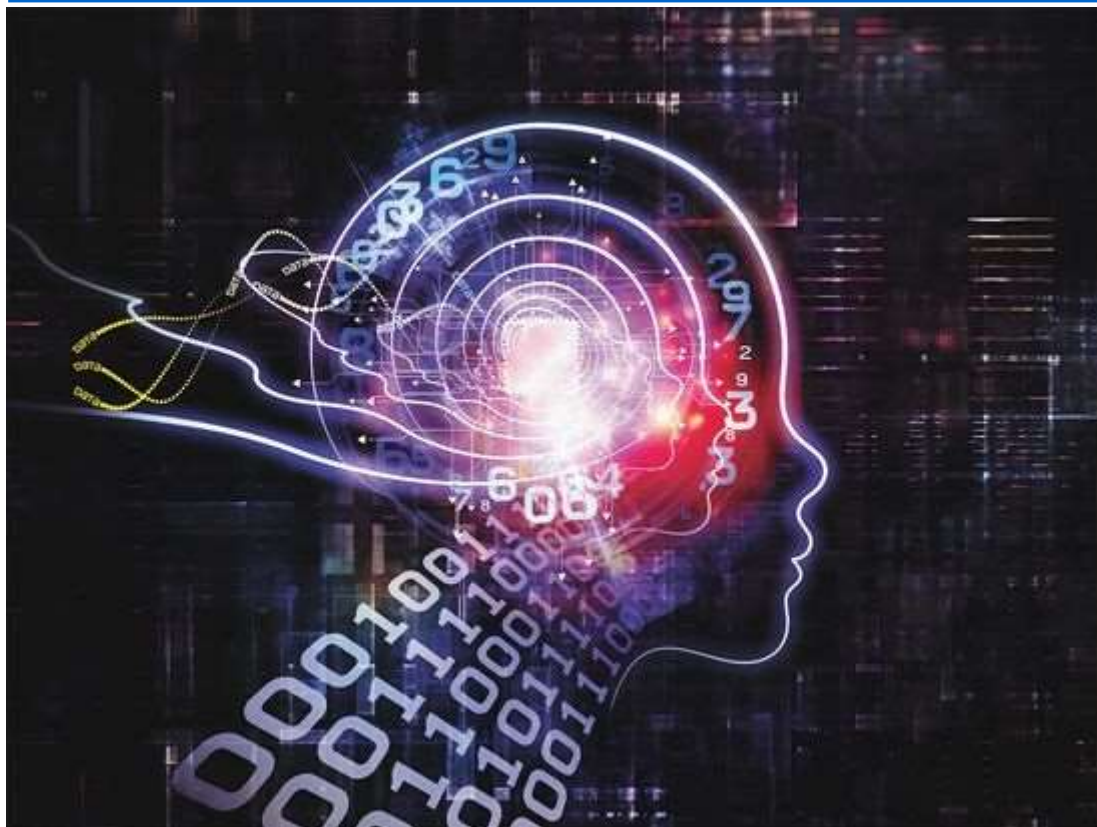


周翊民老师课题组洞察月报

——专注人工智能和多旋翼无人机研究



2018 年 8 月

第二期

专题：OCR 的研究现状调研

编辑：陈鹏

指导：周翊民

校对：吴庆甜

目录

一、	绪论.....	4
1.	研究背景和意义.....	4
2.	需求分析.....	5
二、	典型应用场景.....	6
	工业自动化-物流.....	6
	二维码.....	7
	视觉输入和访问.....	8
	车牌识别.....	9
	火车站闸机身份检测.....	11
	无人驾驶.....	11
三、	OCR 技术相关综述	12
1.	文档文本检测和场景文本检测.....	12
2.	面临的技术挑战.....	13
3.	方法.....	14
4.	原理.....	15
5.	步骤.....	15
6.	基础网络.....	17
	FCN 网络	17
	STN 网络	18
7.	近三年出现的模型介绍.....	19
	文本检测模型.....	19
	文本识别模型.....	28
	端到端模型.....	31
8.	开源数据集介绍.....	32
四、	发展趋势.....	35
五、	参考文献.....	36

OCR 技术调研

OCR（Optical Character Recognition，光学字符识别）是指电子设备（例如扫描仪或数码相机）检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。

我们不得不忽视这样一个事实：我们每天都被文字所包围，像每天办公的文件、上课的板书、商品的介绍等等都是由文字组成的，并且这些文字在某一程度上也是语音交互的基础，而这其中关乎一个关键的技术——OCR，光学字符识别。由于 OCR 是一门与识别率拔河的技术，因此如何除错或利用辅助信息提高识别正确率，是 OCR 最重要的课题。而根据文字资料存在的媒体介质不同，及取得这些资料的方式不同，就衍生出各式各样、各种不同的应用。OCR 可以说是一种不确定的技术研究，正确率就像是一个无穷趋近函数，知道其趋近值，却只能靠近而无法达到，永远在与 100% 作拉锯战！

本文主要叙述了 OCR 起源至现状的基本情形，以综述形式介绍了场景文本识别的原理方法与识别的各大框架和网络模型。下面，我将从以下几个方面对 OCR 技术进行介绍：

- 一、OCR 的兴起背景与研究意义
- 二、OCR 常用的神经网络模型
- 三、现今 OCR 研究的近几年检测与识别模型
- 四、OCR 识别的经典原理与步骤
- 五、场景文字识别训练各大数据集
- 六、OCR 现今应用



一、绪论

1. 研究背景和意义

OCR 的概念诞生于 1929 年，由德国的科学家 Tausheck 首先提出[ref]，并且申请了专利。4 年后，美国科学家 Handel 也提出了对文字进行识别的想法，但这种梦想直到计算机诞生后才成为现实。现在这一技术已经由计算机来实现，OCR 的意思也就演变成利用光学技术对文字和字符进行扫描识别，并将其转化为计算机内码，OCR 技术让大家减少了设备配置，降低了人力成本，提高了工作效率。



图 1.1 生活中的 OCR

OCR 技术发展多年来，从应用场景来看，已经在图像识别，身份证识别，包括银行保险的票据等方面都有应用，从技术层面来看，早先的传统文字识别手法基本都采用基于模板匹配的方式，对特征描述要求非常苛刻，很难满足复杂场景下的识别任务。而自从第三次人工智能浪潮兴起，在算法以及算力都有大幅度突破的情况下，深度学习抛弃了传统人工设计特征的方式，利用海量标定样本数据以及大规模 GPU 集群的优势让机器自动学习特征和模型参数，能一定程度上弥补底层特征与高层语义之间的不足。就在最近这几年，基于深度学习的图像识别达到了前所未有的高度，这也让 OCR 技术有了广阔的场景。

2. 需求分析

OCR 技术的应用前景广阔，就目前从行业需求来看，金融、保险、税务、工商、电子商务等行业对信息识别的需求已经越来越广泛，促进了识别技术的大规模应用。而个人消费者对资料电子化、手写识别技术等各方面需求则拓展了 OCR 识别技术在这一领域的应用之路，另一方面，网络时代的高速发展使个人资料电子化、商务办公自动化等需求的呼声也变得越来越高。



图 1.2 OCR 场景识别

OCR 技术确实也在改变着我们的生活：比如一个手机 APP 就能帮忙扫描名片、身份证，并识别出里面的信息；汽车进入停车场、收费站都不需要人工登记了，都是用车牌识别技术；我们看书时看到不懂的题，拿个手机一扫，APP 就能在网上帮你找到这题的答案。太多太多的应用了，相信，随着行业发展的不断深入。OCR 一定会为人们带来越来越多的惊喜。



图 1.3 OCR 概况

二、 典型应用场景

工业自动化-物流

当快递员从仓库分拣到当日要送的货时，都是用 OCR 扫码移动终端设备进行扫码操作，但也仅仅限于货物外包装上的物流单的条码号而已，系统是没有关联用户的手机号码的，所以经常快递员都要在这个终端上或者自己的手机上拨、按手机号码来通知收件人物流派送信息；现在有一款 APP 程序可以被集成到快递员的移动终端上（或者被安装在快递人员的手机中），当快递员分拣快件的时候就可以同时把单据上的手机号码 OCR 识别提取出来，使得通知收件人派送信息这个流程更加方便、自动快捷。



图 2.1 顺丰 OCR

顺丰应用腾讯优图 OCR 技术，可快速识别手写体的快递单，3 小时可识别 2000 万张。



图 2.2 OCR 扫码

二维码

二维码支付听起来似乎是一项十分新鲜的技术，其实，这个跟手机报差不多，算不上新颖的技术。早在上世纪 90 年代，二维码支付技术就已经形成，其中，韩国与日本是使用二维码支付比较早的国家，日韩二维码支付技术已经普及了 95% 以上，而在国内才刚刚兴起。

二维码支付手段在国内兴起并不是偶然，形成背景主要与我国 IT 技术的快速发展以及电子商务的快速推进有关。IT 技术的日渐成熟，推动了智能手机、平板电脑等移动终端的诞生，这使得人们的移动生活变得更加丰富多彩。与此同时，国内电商也紧紧与“移动”相关，尤其是 O2O 的发展。有了大批的移动设备，也有了大量的移动消费，支付成本就变得尤为关键。因此，二维码支付解决方案便应运而生。



图 2.3 二维码移动支付

条形码（一维码）和二维码主要是作为物品的标识来使用的。条形码主要应用在商品标识、防伪、医药监管、超市收银等场合，二维码可以包含更多的信息，像网址、文字、图片等等，甚至是一首诗，一篇文章等，在网络时代应用逐渐普及很快。



图 2.4 二维码扫描

视觉输入和访问

随着包括数码相机在内的移动设备的发展，成像设备得到了广泛的应用。有了嵌入式模块，移动设备可以自动进入名片、白板和幻灯片演示。不需要键盘输入，用户感觉更舒适，工作效率更高。



图 2.5 文档扫描

在日常生活中，不免有许多工作或者科研学习，要应对大量的纸质文件、书刊杂志、PDF 格式的电子资料，很难对这些格式的资料进行编辑和整理；读书时，看到不错的文章段落想要做书摘笔记，但又懒得去打字或是手抄整理；此时，我们可以看到现在 OCR 领域这方面的技术已经很成熟了，直接使用相应公司的软件就可以了，例如腾讯云文字识别 OCR，它可自动从图片中定位并识别字段，印刷体的平均准确率可达 90% 以上，手写体的识别平均准确率高达 85% 以上，鲁棒性强。

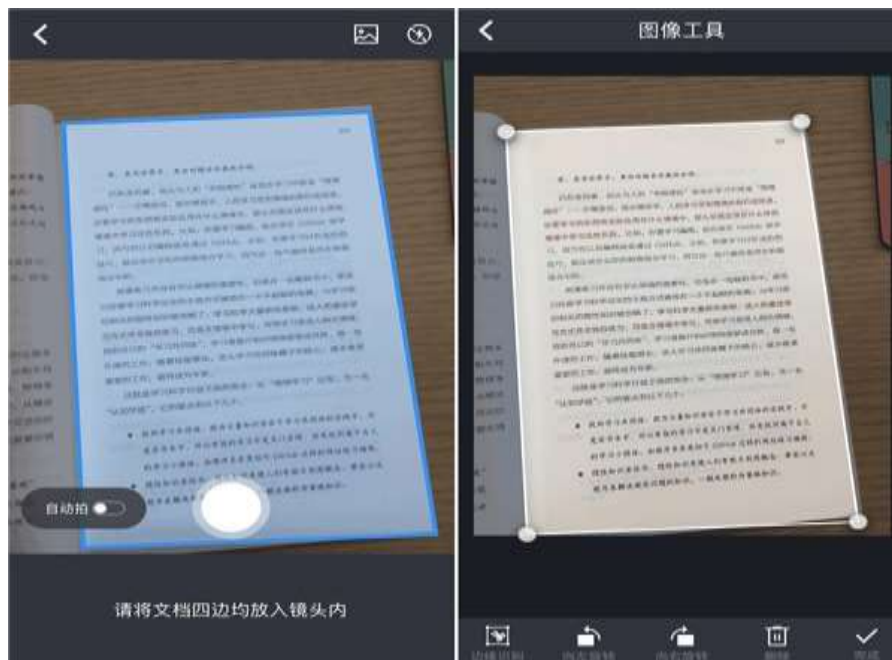


图 2.6 OCR 扫描软件

车牌识别

随着城市建设的发展，我们现在可以注意到小区，商场，办公大楼等地方的出入口

中大多都有一套智能停车场系统。基于车牌识别的智能停车场系统作为交通自动化管理的重要手段，以及车辆检测系统的一个重要环节。智能停车场管理系统车牌识别技术能够为城市道路规划设计提供精确、详尽的分类车流统计数据，实现道路规划管理的最优化设计，减少交通阻塞黑洞。



图 2.7 车牌识别

车牌识别在复杂场景下的识别是一个很大的挑战，例如，如果车牌识别相机安装角度掌握不好，在大灯的作用下，车牌的图像可能会变得一团黑或一团白，为解决这个问题的通常会用到宽动态算法，增加补光灯，车牌识别相机自带补光灯，由算法控制依据车牌局部的识别需要而控制的补光技术，抓拍到清晰的图片，进而保证识别率。车牌识别相机在恶劣天气下的识别能力将接受考验，在或者：雨雾天气的能见度较低，所以获取的车牌图片质量会有严重的衰减。那么在雨雾天车牌识别处理中图像复原功能实现对雨雾天退化图像场景的再现就发挥着很大的作用，其学名是图像复原算法，只有车牌识别相机的核心算法得到充分优化，才能保证成像和识别率。



图 2.8 不同复杂场景下的车牌识别

火车站闸机身份检测

火车站建立“人像是明身份验证系统份证”，利用生物识别技术，实现实名制车票与进站乘客的自动身份核查验证功能，可有效地减少火车站的人力投入，提高进站乘客的身份验证效率和进站速度，保障“购票实名制”制度的顺利实施，提高火车站服务满意度。



图 2.9 火车站身份检测

无人驾驶

无人驾驶，是每个人心中的理想驾驶状态，当我们聚会饮酒或者极度疲劳的时候，我们依旧可以通过无人驾驶汽车去到我们想去的地方。当然，在这个行驶时间内，我们还可以干点其他的事，譬如，打开笔记本开始办公，或许还可以和朋友继续畅饮。对于喜欢饮酒的司机来讲，他们一定很期待无人驾驶汽车售卖日子的及早到来。



图 2.10 无人驾驶

可以想象一下，没有驾驶员，那就意味着我们的汽车中央大脑要时刻关注路面信息、交通情况、前后车距离还有交通标识以及信号灯等情况，而这些数据将要通过激光雷达、红外相机、摄像头、GPS 和传感器等设备不断搜集反馈。当然，我们还需要通过人为的记录所有道路的物理特点数据，在汽车上路时，通过传感器和摄像头收集数据，和系统已有的数据进行对比和分析，以便快速的定位自己的方位和位置。

无人驾驶现在还没有进行推广，仍在提升各方面安全性的研究之中。有权威报告预测，到 2040 年全球上路的汽车总量中，75%将会是无人驾驶汽车。不过我们更需要关注的是安全问题，未来更多人工智能被应用，而我们将享受更多机器智所带来的生活便利。

三、OCR 技术相关综述

1.文档文本检测和场景文本检测

OCR 传统上指对输入扫描文档图像进行分析处理，识别出图像中文字信息。场景文字识别（Scene Text Recognition, STR）指识别自然场景图片中的文字信息。自然场景图像中的文字识别，其难度远大于扫描文档图像中的文字识别，因为它的文字展现形式极其丰富：

- 允许多种语言文本混合，字符可以有不同的大小、字体、颜色、亮度、对比度等。
- 文本行可能有横向、竖向、弯曲、旋转、扭曲等式样。
- 图像中的文字区域还可能会产生变形(透视、仿射变换)、残缺、模糊等现象。
- 自然场景图像的背景极其多样。如文字可以出现在平面、曲面或折皱面上；文字区域附近有复杂的干扰纹理、或者非文字区域有近似文字的纹理，比如沙地、草丛、栅栏、砖墙等。



图 3.1 文档文本与场景文本

也有人用 OCR 技术泛指所有图像文字检测和识别技术，包括传统 OCR 技术与场景文字识别技术。这是因为，场景文字识别技术可以被看成是传统 OCR 技术的自然演进与升级换代。

2. 面临的技术挑战

在 OCR 场景文字识别领域，目前的研究遇到相当大的阻力。受制于自然场景中的多种图像退化以及多变的字体和风格等因素，STR 的识别率一直较低。现存的挑战如下：



图 3.2 自然场景图片中的文字多样性示例

复杂性：在自然环境中，出现了许多人造物体，例如建筑物，符号和绘画，其具有与文本相似的结构和外观。文本本身通常被布置以便于阅读。场景复杂性的挑战在于周围的场景使得难以将文本与非文本区分开。

照明不均匀：当在野外捕捉图像时，由于照明和感觉装置的不均匀响应，不均匀的照明是常见的。不均匀的照明引入了颜色失真和视觉特征的恶化，因此引入了错误的检测，分割和识别结果。

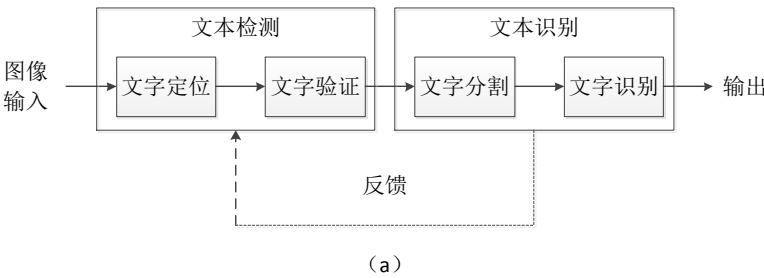
模糊和退化：灵活的工作条件和无焦点的相机会出现文本图像的散焦和模糊。图像/视频压缩和解压缩过程也会降低文本质量，特别是图形视频文本。散焦，模糊和退化的典型影响是它们减少了字符的清晰度并引入了触摸字符，这使得诸如分割等基本任务变得困难。

纵横比：交通标志等文字可能很简短，而其他文字（如视频字幕）可能会更长。换句话说，文本具有不同的宽高比。在检测文本时，需要考虑关于位置，比例和长度的搜索过程，这引入了高计算复杂度。

多语言环境：虽然大多数拉丁语言都有数十个字符，但是中文，日文和韩文（CJK）等语言有数千个字符类。在多语言环境中，OCR 在扫描中文档仍然是一个研究问题，而复杂图像中的文本识别则更加困难。

3. 方法

我们分析了完整文本检测和识别系统中两种常用的方法：逐步和集成[4]。如图 2.3a 所示，逐步方法具有分离的检测和识别模块，并使用前馈管道来检测，分割和识别文本区域。相比之下，综合方法的目标是识别检测和识别程序与字符分类共享信息和/或使用联合优化策略的单词，如图 2.3b 所示。



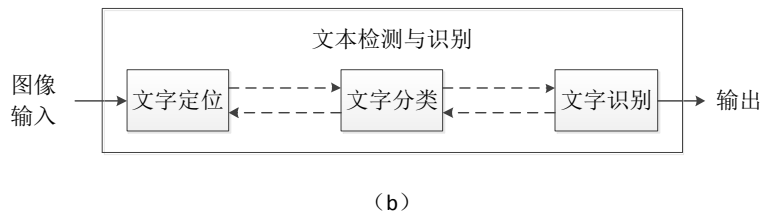


图 3.3 文本检测与识别方法：逐步和集成

4. 原理

场景文本识别一般被分割成两个独立的子问题：检测和识别。前者的目标是从图片中尽可能准确的找出文字所在区域，后者的目标则是在前者的基础上，将区域中的单个字符识别出来。当然也有一些研究者把两个问题用一套统一的框架解决，通常我们称这种做法叫 End-to-end 系统。上面我们介绍的模型框架结合使用基本就是 End-to-end 系统。

而说到底，不论是不是 End-to-end，原理都是检测和识别，只是手段不同，从前对于文档文本的检测用的是比较经典的方法，而现在神经网络的应用用的都是端对端的集成的方法。

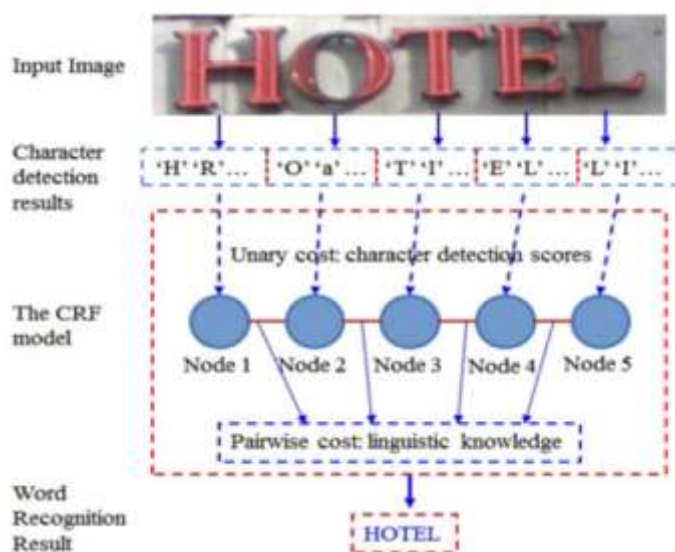


图 3.4 OCR 经典原理

5. 步骤

OCR 经典原理可以分为以下几个步骤：预处理、特征提取和降维、分类器设计、训练和实际识别和后处理。

预处理：对包含文字的图像进行处理以便后续进行特征提取、学习。这个过程的主

要目的是减少图像中的无用信息，以便方便后面的处理。在这个步骤通常有：灰度化（如果是彩色图像）、降噪、二值化、字符切分以及归一化这些子步骤。经过二值化后，图像只剩下两种颜色，即黑和白，其中一个为图像背景，另一个颜色就是要识别的文字了。降噪在这个阶段非常重要，降噪算法的好坏对特征提取的影响很大。字符切分则是将图像中的文字分割成单个文字——识别的时候是一个字一个字识别的。如果文字行有倾斜的话往往还要进行倾斜校正。归一化则是将单个的文字图像规整到同样的尺寸，在同一个规格下，才能应用统一的算法。

特征提取和降维：特征是用来识别文字的关键信息，每个不同的文字都能通过特征来和其他文字进行区分。

文字块：完全由字符构成，它包含中文、字母、数字以及各种标点符号等。

线条块：有些名片图像中含有线条，用来醒目单位名称以及有关信息等，通常位于姓名与单位名称间或单位名称与地址之间。

图形图片块：由各种线条构成的图案，如单位的标识等。有时，图片块中包含字符或线条，本文将其与图形块和线条分开处理，因为图片块中的信息是偶尔也是有用的。对于数字和英文字母来说，这个特征提取是比较容易的，因为数字只有 10 个，英文字母只有 52 个，都是小字符集。对于汉字来说，特征提取比较困难，因为首先汉字是大字符集，国标中光是最常用的第一级汉字就有 3755 个；第二个汉字结构复杂，形近字多。在确定了使用何种特征后，视情况而定，还有可能要进行特征降维，这种情况就是如果特征的维数太高（特征一般用一个向量表示，维数即该向量的分量数），分类器的效率会受到很大的影响，为了提高识别速率，往往就要进行降维，这个过程也很重要，既要降低维数吧，又得使得减少维数后的特征向量还保留了足够的信息量（以区分不同的文字）。

分类器设计、训练和实际识别：分类器是用来进行识别的，就是对于第二步，对一个文字图像，提取出特征给，给分类器，分类器就对其进行分类，告诉你这个特征该识别成哪个文字。在进行实际识别前，往往还要对分类器进行训练，这是一个监督学习的案例。成熟的分类器也很多，SVM，KNN，神经网络。每棵决策树都是一个分类器（假设现在针对的是分类问题），那么对于一个输入样本，N 棵树会有 N 个分类结果。而随机森林集成了所有的分类投票结果，将投票次数最多的类别指定为最终的输出。

后处理：后处理是用来对分类结果进行优化的，第一个，分类器的分类有时候不一定是完全正确的，比如对汉字的识别，由于汉字中形近字的存在，很容易将一个字识别

成其形近字。后处理中可以去解决这个问题，比如通过语言模型来进行校正——如果分类器将“在哪里”识别成“存哪里”，通过语言模型会发现“存哪里”是错误的，然后进行校正。第二个，OCR 的识别图像往往是有大量文字的，而且这些文字存在排版、字体大小等复杂情况，后处理中可以尝试去对识别结果进行格式化，比如按照图像中的排版排列什么的，举个例子，一张图像，其左半部分的文字和右半部分的文字毫无关系，而在字符切分过程中，往往是按行切分的，那么识别结果中左半部分的第一行后面会跟着右半部分的第一行诸如此类。

6. 基础网络

图文识别任务中充当特征提取模块的基础网络，可以来源于通用场景的图像分类模型。同样也可以来源于特定场景的专用网络模型。例如，擅长提取图像细节特征的 FCN 网络，擅长做图形矫正的 STN 网络。

FCN 网络

Evan Shelhamer 在 2015 年的 CVPR 上提出了全卷积网络 (FCN, fully convolutional network) [12]，其是去除了全连接(fc)层的基础网络，最初是用于实现语义分割任务。FC 的优势在于利用反卷积(deconvolution)、上池化(unpooling)等上采样(upsampling)操作，将特征矩阵恢复到接近原图尺寸，然后对每一个位置上的像素做类别预测，从而能识别出更清晰的物体边界。基于 FCN 的检测网络，不再经过候选区域回归出物体边框，而是根据高分辨率的特征图直接预测物体边框。因为不需要像 Faster-RCNN 那样在训练前定义好候选框长宽比例，FCN 在预测不规则物体边界时更加鲁棒。由于 FCN 网络最后一层特征图的像素分辨率较高，而图文识别任务中需要依赖清晰的文字笔画来区分不同字符（特别是汉字），所以 FCN 网络很适合用来提取文本特征。当 FCN 被用于图文识别任务时，最后一层特征图中每个像素将被分成文字行（前景）和非文字行（背景）两个类别。感兴趣的读者可参考文献[12]。

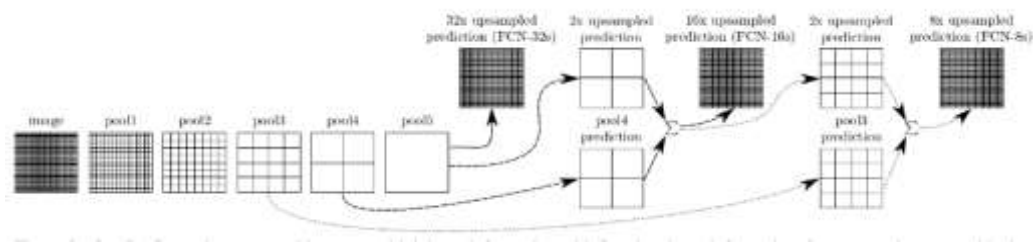


图 3.5 全卷积网络

表 4 给出了我们的 FCN-8 在 PASCAL VOC 2011 和 2012 测试装置上的性能，并将其与之前的最佳 SDS [14]和众所周知的进行了比较 R-CNN [15]。我们在平均 IU 上取得了最好的结果相对于 30%。

表 3.5 PASCAL VOC 比较结果

	mean IU VOC2011 test	mean IU VOC2012 test	inference time
R-CNN [5]	47.9	-	-
SDS [14]	52.6	51.6	~ 50 s
FCN-8s	67.5	67.2	~ 100 ms

STN 网络

Max Jaderberg 等人在 2016 年提出了空间变换网络（STN，Spatial Transformer Networks）[16]的作用是对输入特征图进行空间位置矫正得到输出特征图，这个矫正过程是可以进行梯度传导的，从而能够支持端到端的模型训练。

如下图所示，STN 网络由定位网络（Localization Network），网格生成器（Grid generator），采样器（Sampler）共 3 个部分组成。定位网络根据原始特征图 U 计算出一套控制参数，网格生成器这套控制参数产生采样网格（sampling grid），采样器根据采样网格核函数将原始图 U 中像素对应采样到目标图 V 中。感兴趣的读者可参考文献[13]

空间变换的控制参数是根据原始特征图 U 动态生成的，生成空间变换控制参数的元参数则是在模型训练阶段学习到的、并且存放于定位网络的权重（weights）矩阵中。

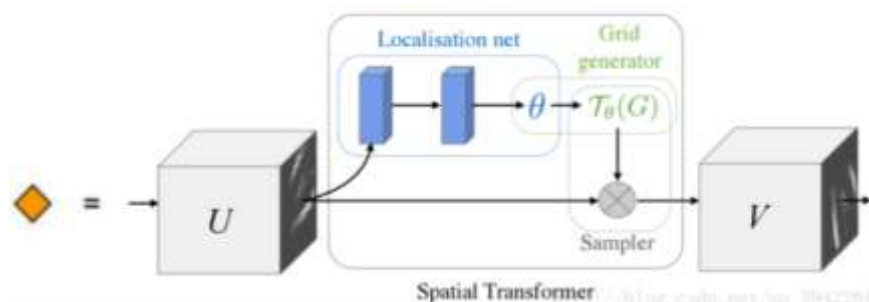



图 3.6 空间变换网络

如下图表，左边是 CUB-200-2011 鸟类分类数据集的准确性，右边是由输入图像上的 2*ST-CNN（顶行）和 4*ST-CNN（底行）的空间变换器预测的变换。ST-CNN 的准确率达到 84.1%，优于基线 1.8%。ImageNet 训练集和 CUB-200-2011 测试集 1 之间存在小的(22/5794)重叠 - 从测试集中移除这些图像导致相同 ST-CNN 的准确度为 84.0%。在 2STCNN 预测的变换的可视化中（表 3（右）），人们可以看到有趣的行为：一个空间变换器（红色）已经学会成为头部探测器，而另一个（绿色）则固定在探测器的中央部分一只鸟的身体。来自分类网络的空间变换器的最终输出是一种有点姿势标准化的鸟类表现形式。

表 3.6 在 CUB-200-2011 鸟类分类数据集的准确性及效果图

Model	
Cimpoi '15 [5]	66.7
Zhang '14 [40]	74.9
Branson '14 [3]	75.7
Lin '15 [23]	80.9
Simon '15 [30]	81.0
CNN (ours) 224px	82.3
2×ST-CNN 224px	83.1
2×ST-CNN 448px	83.9
4×ST-CNN 448px	84.1



7. 近三年出现的模型介绍

文本检测模型

文本检测模型的目标是从图片中尽可能准确地找出文字所在区域。

但是，视觉领域常规物体检测方法(SSD, YOLO, Faster-RCNN 等)直接套用于文字检测任务效果并不理想，主要原因如下：

- 相比于常规物体，文字行长度、长宽比例变化范围很大。
- 文本行是有方向性的。常规物体边框 BBox 的四元组描述方式信息量不充足。
- 自然场景中某些物体局部图像与字母形状相似，如果不参考图像全局信息将有误报。
- 有些艺术字体使用了弯曲的文本行，而手写字体变化模式也很多。
- 由于丰富的背景图像干扰，手工设计特征在自然场景文本识别任务中不够鲁棒。

针对上述问题根因，近年来出现了各种基于深度学习的技术解决方案。它们从特征提取、区域建议网络(RPN)、多目标协同训练、Loss 改进、非极大值抑制（NMS）、半监督学习等角度对常规物体检测方法进行改造，极大提升了自然场景图像中文本检测

的准确率。例如：

CTPN 方案中，用 BLSTM 模块提取字符所在图像上下文特征，以提高文本块识别精度。

RRPN 等方案中，文本框标注采用 BBOX + 方向角度值的形式，模型中产生出可旋转的文字区域候选框，并在边框回归计算过程中找到待测文本行的倾斜角度。

DMPNet 等方案中，使用四边形（非矩形）标注文本框，来更紧凑的包围文本区域。

SegLink 将单词切割为更易检测的小文字块，再预测邻近连接将小文字块连成词。

FTSN 方案中，作者使用 Mask-NMS 代替传统 BBOX 的 NMS 算法来过滤候框。

WordSup 方案中，采用半监督学习策略，用单词级标注数据来训练字符级文本检测模型。

下面用近年来出现的多个模型案例，介绍如何应用上述各方法提升图像文本检测的效果。

CTPN 模型

Tian, huang 等人在 2016 年提出 CTPN[3]，这是目前流传最广、影响最大的开源文本检测模型，可以检测水平或微斜的文本行。文本行可以被看成一个字符 sequence，而不是一般物体检测中单个独立的目标。同一文本行上各个字符图像间可以互为上下文，在训练阶段让检测模型学习图像中蕴含的这种上下文统计规律，可以使得预测阶段有效提升文本块预测准确率。CTPN 模型的图像预测流程中，前端使用当时流行的 VGG16 做基础网络来提取各字符的局部图像特征，中间使用 BLSTM 层提取字符序列上下文特征，然后通过 FC 全连接层，末端经过预测分支输出各个文字块的坐标值和分类结果概率值。在数据后处理阶段，将合并相邻的小文字块为文本行。

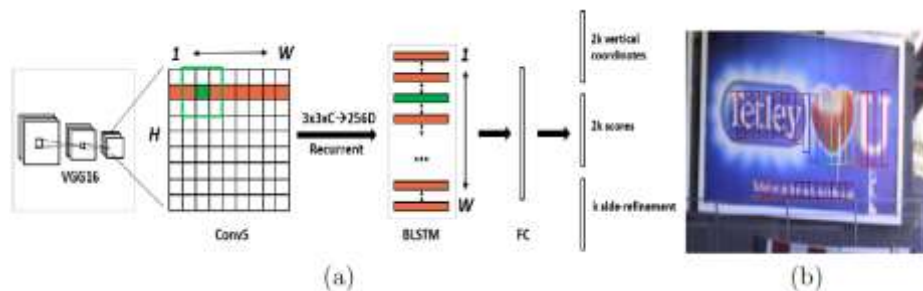


图 3.7 CTPN 检测模型

在 ICDAR 2013 和 2015 becnmarks 上分别取得 0.88 和 0.61 的 F-measure 值，比最近

的较好结果有很大的提升。CTPN 的计算效率是 0.14s/image，它使用了 very deep VGG16 model。

表 3.7 ICDAR 2011，2013 和 2015 上的最新结果

ICDAR 2011				ICDAR 2013					ICDAR 2015			
Method	P	R	F	Method	P	R	F	T(s)	Method	P	R	F
Huang [13]	0.82	0.75	0.73	Yin [33]	0.88	0.66	0.76	0.43	CNN Pro.	0.35	0.34	0.35
Yao [31]	0.82	0.66	0.73	Neumann [22]	0.82	0.72	0.77	0.40	Deep2Text	0.50	0.32	0.39
Huang [14]	0.88	0.71	0.78	Neumann [23]	0.82	0.71	0.76	0.40	HUST	0.44	0.38	0.41
Yin [33]	0.86	0.68	0.76	FASText [1]	0.84	0.69	0.77	0.15	AJOU	0.47	0.47	0.47
Zhang [34]	0.84	0.76	0.80	Zhang [34]	0.88	0.74	0.80	60.0	NJU-Text	0.79	0.36	0.47
TextFlow [28]	0.86	0.76	0.81	TextFlow [28]	0.85	0.76	0.80	0.94	StradVision1	0.53	0.46	0.50
Text-CNN [11]	0.91	0.74	0.82	Text-CNN [11]	0.93	0.73	0.82	4.6	StradVision2	0.77	0.37	0.50
Gupta [8]	0.92	0.75	0.82	Gupta [8]	0.92	0.76	0.83	0.07	Zhang [35]	0.71	0.43	0.54
CTPN	0.89	0.79	0.84	CTPN	0.93	0.83	0.88	0.14 *	CTPN	0.74	0.52	0.61

FTSN 模型

Dai , Huang 等人在 2018 年提出了 FTSN（Fused Text Segmentation Networks）模型 [7]，使用分割网络支持倾斜文本检测。它使用 Resnet-101 做基础网络，使用了多尺度融合的特征图。标注数据包括文本实例的像素掩码和边框，使用像素预测与边框检测多目标联合训练。

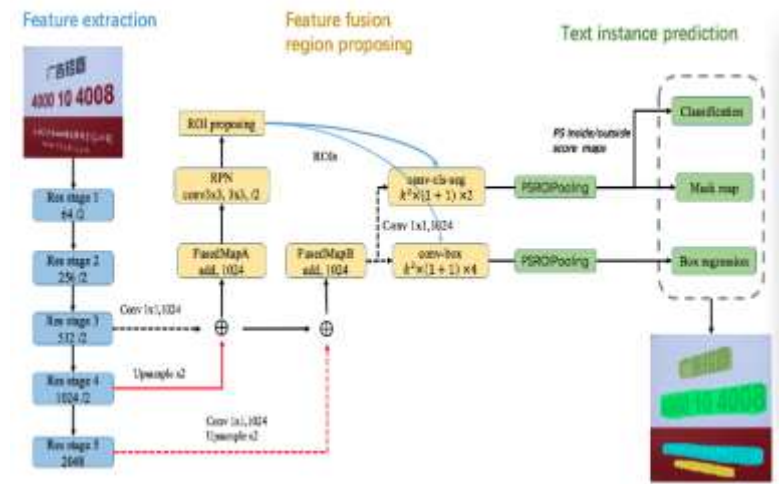


图 3.8 FTSN 检测模型

ICDAR 2015 强力阅读竞赛的挑战 4 [17]。IC15 包含 1000 次训练和 500 次测试 Google 眼镜拍摄的附带图像，而不关注视点和图像质量。因此，文本比例，方向和分辨率的大的变化导致文本检测的困难。下表给出 FTSN 在 ICDAR2015 数据集上的性能指标。

表 3.8 FTSN 在 ICDAR2015 数据集上的性能

Method	Precision (%)	Recall (%)	Hmean (%)
HUST[16]	44.0	37.8	40.7
Zhang <i>et al.</i> [25]	71.0	43.0	54.0
DMPNet[21]	73.2	68.2	70.6
Qin <i>et al.</i> [17]	79.0	65.0	71.0
SegLink[19]	73.1	76.8	75.0
RRPN[22]	73.2	82.2	77.4
EAST[20]	83.3	78.3	80.7
He <i>et al.</i> [18]	82.0	80.0	81.0
Proposed FTSN+SNMS	87.1	80.0	83.4
Proposed FTSN+MNMS	88.6	80.0	84.1

RRPN 模型

Ma, Shao 等人在 2018 年提出了基于旋转区域候选网络 (RRPN, Rotation Region Proposal Networks) [19]的方案, 将旋转因素并入经典区域候选网络 (如 Faster RCNN)。这种方案中, 一个文本区域的 ground truth 被表示为具有 5 元组(x,y,h,w, θ)的旋转边框, 坐标(x,y)表示边框的几何中心, 高度 h 设定为边框的短边, 宽度 w 为长边, 方向是长边的方向。训练时, 首先生成含有文本方向角的倾斜候选框, 然后在边框回归过程中学习文本方向角。

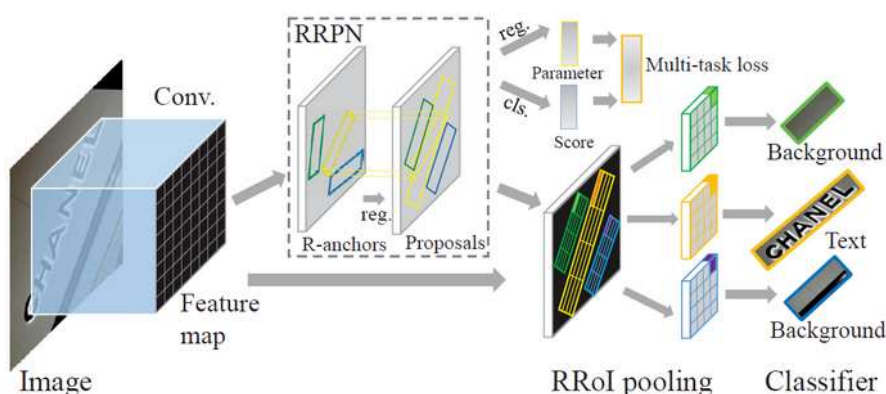


图 3.9 RRPN 模型

RRPN 中方案中提出了旋转感兴趣区域(RRoI, Rotation Region-of-Interest)池化层, 将任意方向的区域建议先划分成子区域, 然后对这些子区域分别做 max pooling、并将结果投影到具有固定空间尺寸小特征图上。

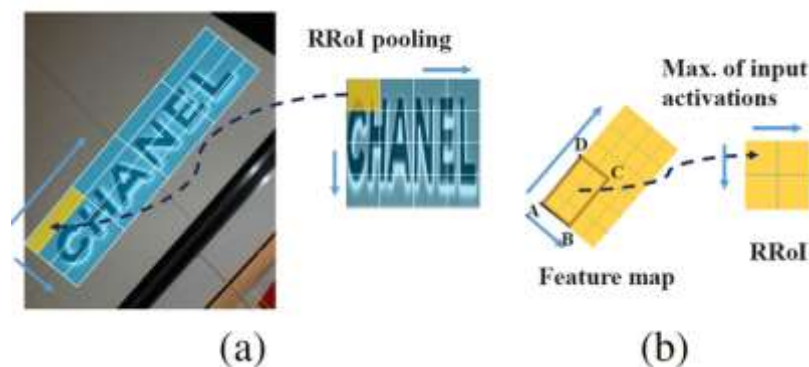


图 3.10 RRoI 区域

如下表 2.11 是 RRPN 与三个基准上最先进的方法进行比较。基于 ICDAR2013 的 FASTER-RCNN 结果报告。ICDAR2015 和 5 元组旋转建议适用于水平对齐的矩形。根据 ICDAR 2013 评估协议，结果是精确度为 90.22%，召回率为 71.89%，F 值为 80.02%。与 Faster-RCNN 相比，有 7% 的改进，这证实了我们的检测框架与旋转因子的稳健性。

表 3.10 RRPN 在 MSRA-TD500、ICDAR2013、2015 数据集上的性能

MSRA-TD500					ICDAR2015				ICDAR2013			
Approach	P	R	F	Time	Approach	P	R	F	Approach	P	R	F
Yin et al. [53]	71	61	65	0.8 s	CTPN [42]	74	52	61	Faster-RCNN [20]	75	71	73
Kang et al. [54]	71	62	66	-	Yao et al. [18]	72	59	65	Gupta et al. [55]	92	76	83
Yin et al. [56]	81	63	71	1.4 s	SCUT_DMPNet [57]	68	73	71	Yao et al. [18]	89	80	84
Zhang et al. [117]	83	67	74	2.1 s	UCSC_TextSpotter [58]	65	79	71	DeepText [43]	85	81	85
Yao et al. [18]	77	75	76	0.6 s	hust_orientedText [59]	77	75	76	CTPN [42]	93	83	88
RRPN	82	68	74	0.3 s	RRPN	82	73	77	RRPN	90	72	80
RRPN*	82	69	75	0.3 s	RRPN*	84	77	80	RRPN*	95	88	91

DMPNet 模型

Liu, Jin 等人在 2017 年提出了 DMPNet (Deep Matching Prior Network) [8]模型，使用四边形（非矩形）来更紧凑地标注文本区域边界，其训练出的模型对倾斜文本块检测效果更好。

如下图所示，它使用滑动窗口在特征图上获取文本区域候选框，候选框既有正方形的、也有倾斜四边形的。接着，使用基于像素点采样的 Monte-Carlo 方法，来快速计算四边形候选框与标注框间的面积重合度。然后，计算四个顶点坐标到四边形中心点的距离，将它们与标注值相比计算出目标 loss。文章中推荐用 Ln loss 来取代 L1、L2 loss，从而对大小文本框都有较快的训练回归 (regress) 速度。

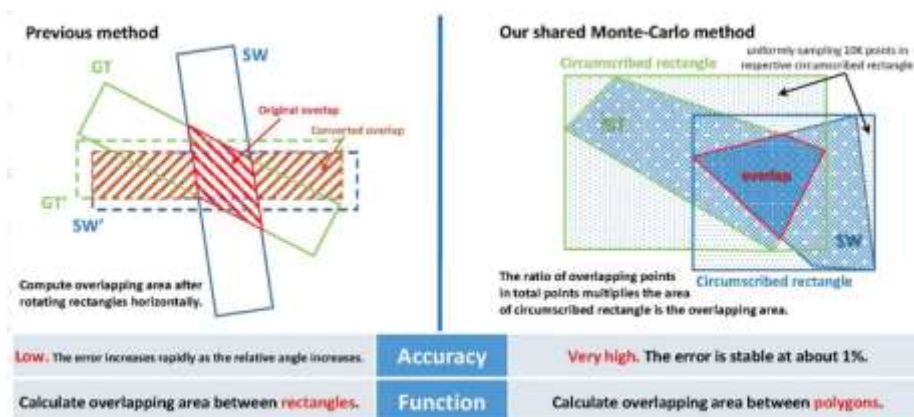


图 3.11 DMPNet 检测模型

如下表 2.11, ICDAR 2015 Challenge 4 是第一个使用四边形的数据集标签, DMPNet 检测法在 ICDAR 2015 的性能证明了利用四边形标签的有效性。

表 3. 11 RRPN 在 ICDAR2015 数据集上的性能

Algorithm	Recall (%)	Precision (%)	Hmean (%)
Baseline (SSD-VGGNet)	25.48	63.25	36.326
Proposed DMPNet	68.22	73.23	70.64
Megvii-Image++ [33]	56.96	72.40	63.76
CTPN [29]	51.56	74.22	60.85
MCLAB_FCN [14]	43.09	70.81	53.58
StardVision-2 [14]	36.74	77.46	49.84
StardVision-1 [14]	46.27	53.39	49.57
CASIA_USTB-Cascaded [14]	39.53	61.68	48.18
NJU_Text [14]	35.82	72.73	48.00
AJOU [16]	46.94	47.26	47.10
HUST_MCLAB [14]	37.79	44.00	40.66
Deep2Text-MO [36]	32.11	49.59	38.98
CNN Proposal [14]	34.42	34.71	34.57
TextCatcher-2 [14]	34.81	24.91	29.04

EAST 模型

Zhou, Yao 等人在 2017 年提出了 EAST (Efficient and Accuracy Scene Text detection pipeline) 模型[1]中, 首先使用全卷积网络 (FCN) 生成多尺度融合的特征图, 然后在此基础上直接进行像素级的文本块预测。该模型中, 支持旋转矩形框、任意四边形两种文本区域标注形式。对应于四边形标注, 模型执行时会特征图中每个像素预测其到四个顶点的坐标差值。对应于旋转矩形框标注, 模型执行时会特征图中每个像素预测其到矩形框四边的距离、以及矩形框的方向角。

根据开源工程中预训练模型的测试, 该模型检测英文单词效果较好、检测中文长文本行效果欠佳。或许, 根据中文数据特点进行针对性训练后, 检测效果还有提升空间。

上述过程中, 省略了其他模型中常见的区域建议、单词分割、子块合并等步骤, 因

此该模型的执行速度很快。

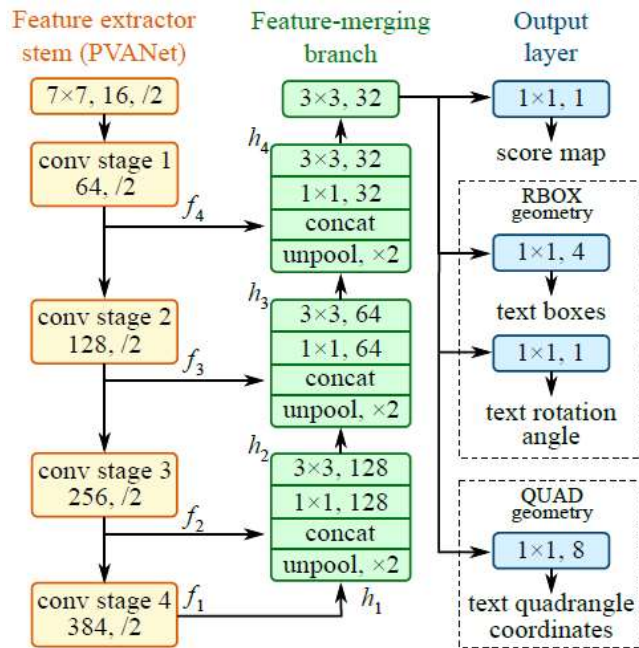


图 3.12 EAST 模型

如表格 2.81，FTSN 的方法在 ICDAR 2015 上大大优于以前最先进的方法。在 ICDAR 2015 Challenge 4 中，当图像以其原始比例进给时，FTSN 实现了 0.7820 的 Fscore。当使用相同网络在多个等级 3 进行测试时，F 分数中达到 0.8072，在绝对值方面比最佳方法高近 0.16（0.8072 对比 0.6477）。

表 3.81 FTSN 在 ICDAR2015 数据集上的性能

Algorithm	Recall	Precision	F-score
Ours + PVANET2x RBOX MS*	0.7833	0.8327	0.8072
Ours + PVANET2x RBOX	0.7347	0.8357	0.7820
Ours + PVANET2x QUAD	0.7419	0.8018	0.7707
Ours + VGG16 RBOX	0.7275	0.8046	0.7641
Ours + PVANET RBOX	0.7135	0.8063	0.7571
Ours + PVANET QUAD	0.6856	0.8119	0.7434
Ours + VGG16 QUAD	0.6895	0.7987	0.7401
Yao <i>et al.</i> [41]	0.5869	0.7226	0.6477
Tian <i>et al.</i> [34]	0.5156	0.7422	0.6085
Zhang <i>et al.</i> [48]	0.4309	0.7081	0.5358
StradVision2 [15]	0.3674	0.7746	0.4984
StradVision1 [15]	0.4627	0.5339	0.4957
NJU [15]	0.3625	0.7044	0.4787
AJOU [20]	0.4694	0.4726	0.4710
Deep2Text-MO [45, 44]	0.3211	0.4959	0.3898
CNN MSER [15]	0.3442	0.3471	0.3457

SegLink 模型

Shi, Bai 等人在 2017 年提出了 SegLink 模型[20]，在 SegLink 模型的标注数据中，

先将每个单词切割为更易检测的有方向的小文字块(segment)，然后用邻近连接(link)将各个小文字块连接成单词。这种方案方便于识别长度变化范围很大的、带方向的单词和文本行，它不会象 Faster-RCNN 等方案因为候选框长宽比例原因检测不出长文本行。相比于 CTPN 等文本检测模型，SegLink 的图片处理速度快很多。



图 3.13 SegLink 效果图

如下图所示，该模型能够同时从 6 种尺度的特征图中检测小文字块。同一层特征图、或者相邻层特征图上的小文字块都有可能被连接入同一个单词中。换句话说，位置邻近、并且尺寸接近的文字块都有可能被预测到同一单词中。

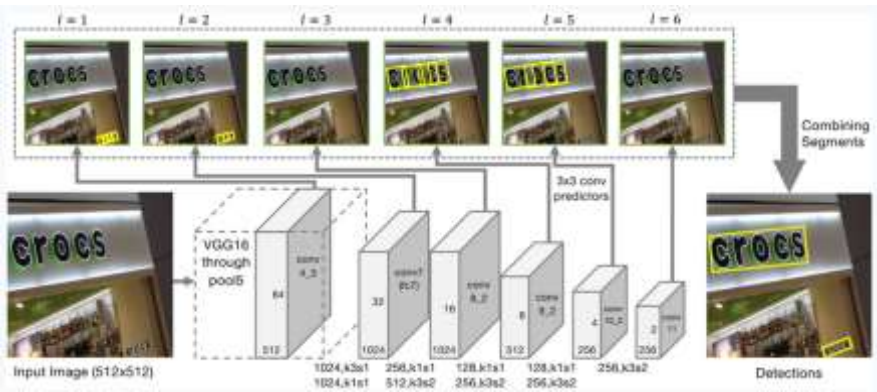


图 3.14 SegLink 检测模型

表 2.13 列出并比较了 SegLink 和其他最先进方法的结果。一些结果是从在线排行榜获得的。SegLink 的表现远远超过其他人。就 f-measure 而言，它的表现优于第二好的 10.2%。考虑到某些方法比 SegLink 具有接近甚至更高的精度，改进主要来自召回。

表 3.13 FTSN 在 ICDAR2015 数据集上的性能

Method	Precision	Recall	F-measure
HUST_MCLAB	47.5	34.8	40.2
NJU_Text	72.7	35.8	48.0
StradVision-2	77.5	36.7	49.8
MCLAB_FCN [30]	70.8	43.0	53.6
CTPN [22]	51.6	74.2	60.9
Megvii-Image++	72.4	57.0	63.8
Yao et al. [26]	72.3	58.7	64.8
SegLink	73.1	76.8	75.0

WordSup 模型

Hu,Zhang 等人在 2017 年提出了 WordSup 框架[9], WordSup 是一种弱监督的训练框架, 可以文本行、单词级标注数据集上训练出字符级检测模型。如下图 2.15 所示, 在数学公式图文识别、不规则形变文本行识别等应用中, 字符级检测模型是一个关键基础模块。由于字符级自然场景图文标注成本很高、相关公开数据集稀少, 导致现在多数图文检测模型只能在文本行、单词级标注数据上做训练。

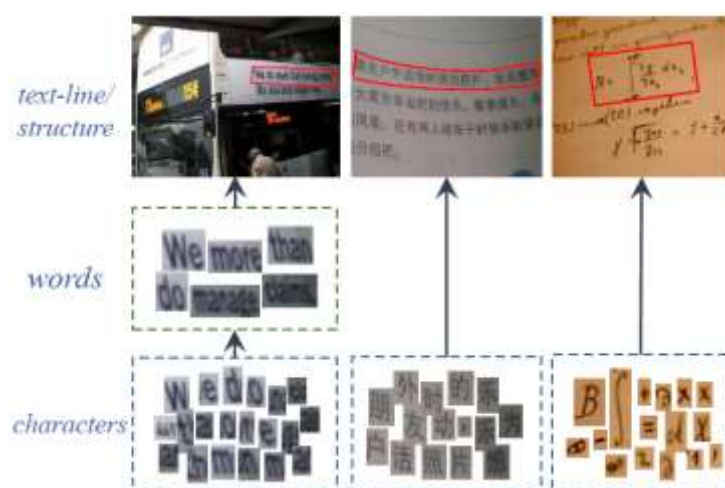


图 3.15 WordSup 模型

如下图所示, WordSup 弱监督训练框架中, 两个训练步骤被交替执行: 给定当前字符检测模型, 并结合单词级标注数据, 计算出字符中心点掩码图; 给定字符中心点掩码图, 有监督地训练字符级检测模型。

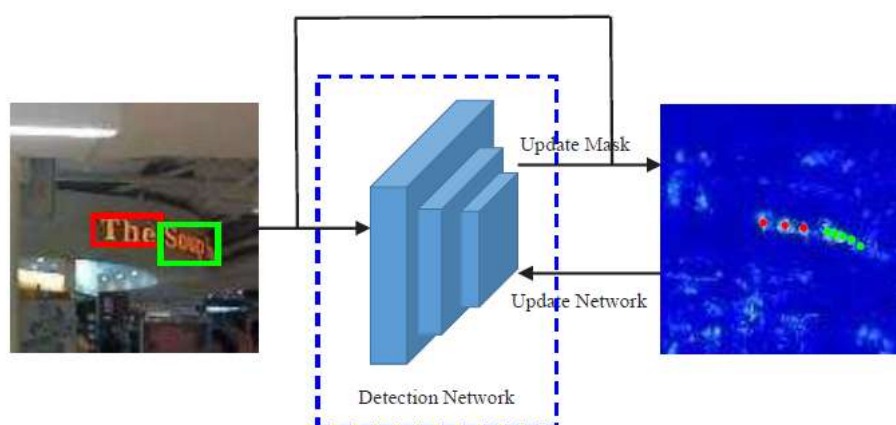


图 3.16 训练步骤

如下图, 训练好字符检测器后, 可以在数据流水线中加入合适的文本结构分析模块, 以输出符合应用场景格式要求的文本内容。该文作者例举了多种文本结构分析模块的实现方式。

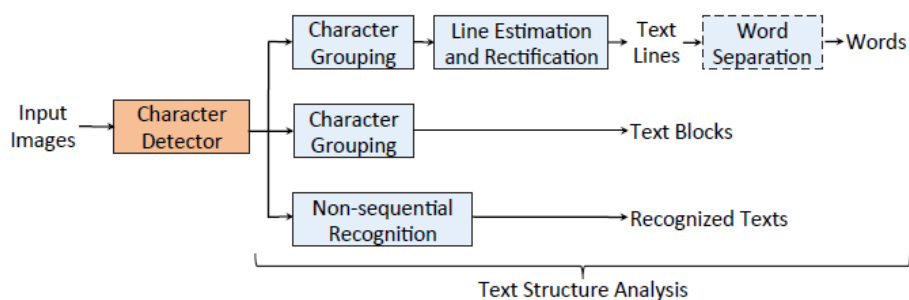


图 3.17 管道

在更具挑战性的 ICDAR15 数据集中，图像更容易受到模糊，透视失真，极端照明等的影响。如下表 2.17 所示 WordSup 模型实现了 78.16% 的测量。比较我们使用不同角色模型的方法，VGG-synth-icdar 比 VGG-synth 模型表现更好（78.16% 对 69.18%）。与 VGG-synth 模型（50k 训练图像）相比，VGG-synth-icdar 仅增加 1k 训练图像。这表明增益来自更多真实数据，而不是更多数据。

表 3.17 WordSup 在 ICDAR2015 数据集上的性能

Method	Recall	Precision	F-measure
MCLAB-FCN [49]	43.09	70.81	53.58
CTPN [38]	51.56	74.22	60.85
Yao et al. [44]	58.69	72.40	64.77
SCUT-DMPNet [25]	68.22	73.23	70.64
RRPN-2 [27]	72.65	68.53	70.53
our (VGG16-synth)	64.37	74.79	69.18
our (VGG16-synth-icdar)	77.03	79.33	78.16

文本识别模型

文本识别模型的目标是从已分割出的文字区域中识别出文本内容。

CRNN 模型

Shi, Bai 和 Yao 等人在 2015 年提出了 CRNN(Convolutional Recurrent Neural Network) [6]是目前较为流行的图文识别模型，可识别较长的文本序列。它包含 CNN 特征提取层和 BLSTM 序列特征提取层，能够进行端到端的联合训练。它利用 BLSTM 和 CTC 部件学习字符图像中的上下文关系，从而有效提升文本识别准确率，使得模型更加鲁棒。预测过程中，前端使用标准的 CNN 网络提取文本图像的特征，利用 BLSTM 将特征向量进行融合以提取字符序列的上下文特征，然后得到每列特征的概率分布，最后通过转录层(CTC rule)进行预测得到文本序列。

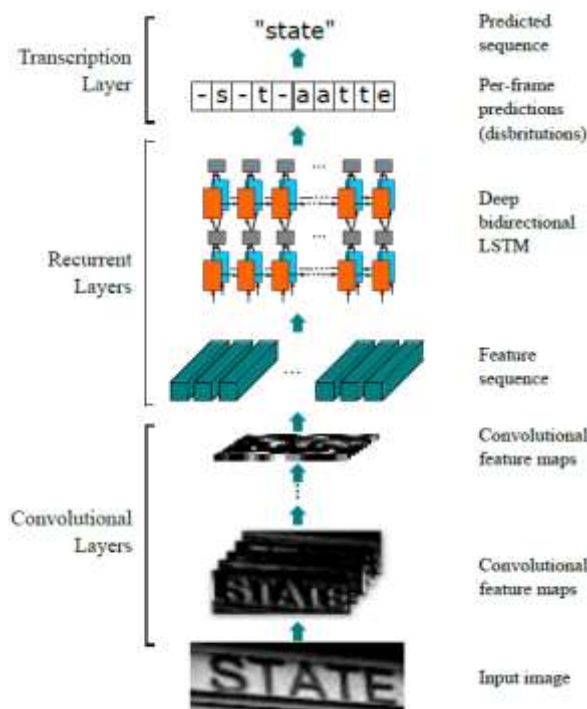


图 3.18 CRNN 模型

以上四种公共数据集的所有识别精度，由所提出的 CRNN 模型和最新的尖端技术 (包括基于深度模型的方法) 获得，如表 2 所示。

表 3.18 CRNN 在四个数据集中的性能

	IIIT5k			SVT		IC03				IC13
	50	1k	None	50	None	50	Full	50k	None	None
ABBY [34]	24.3	-	-	35.0	-	56.0	55.0	-	-	-
Wang <i>et al.</i> [34]	-	-	-	57.0	-	76.0	62.0	-	-	-
Mishra <i>et al.</i> [38]	64.1	57.5	-	73.2	-	81.8	67.8	-	-	-
Wang <i>et al.</i> [35]	-	-	-	70.0	-	90.0	84.0	-	-	-
Goel <i>et al.</i> [13]	-	-	-	77.3	-	89.7	-	-	-	-
Bissacco <i>et al.</i> [4]	-	-	-	90.4	78.0	-	-	-	-	87.6
Alsharif and Piatou [6]	-	-	-	74.3	-	93.1	88.6	85.1	-	-
Almazán <i>et al.</i> [5]	91.2	82.1	-	89.2	-	-	-	-	-	-
Yao <i>et al.</i> [36]	80.2	69.3	-	75.9	-	88.5	80.3	-	-	-
Rodríguez-Serrano <i>et al.</i> [30]	76.1	57.4	-	70.0	-	-	-	-	-	-
Jaderberg <i>et al.</i> [23]	-	-	-	86.1	-	96.2	91.5	-	-	-
Su and Lu [33]	-	-	-	83.0	-	92.0	82.0	-	-	-
Goode [14]	93.3	86.6	-	91.8	-	-	-	-	-	-
Jaderberg <i>et al.</i> [22]	97.1	92.7	-	95.4	80.7*	98.7	98.6	93.3	93.1*	90.8*
Jaderberg <i>et al.</i> [21]	95.5	89.6	-	93.2	71.7	97.8	97.0	93.4	89.6	81.8
CRNN	97.6	94.4	78.2	96.4	80.8	98.7	97.6	95.5	89.4	86.7

RARE 模型

Shi, Bao 等人在 2016 年提出了 RARE (Robust text recognizer with Automatic Rectification) 模型[10]。此模型面向识别变形的图像文本时效果很好。如下图 2.19 所示，模型预测过程中，输入图像首先要被送到一个空间变换网络中做处理，矫正过的图像然后被送入序列识别网络中得到文本预测结果。

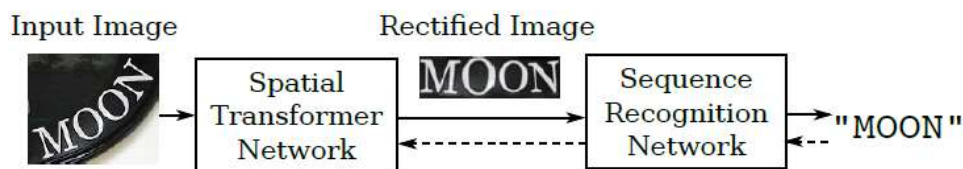


图 3.19 RARE 模型

如下图 2.20 所示，空间变换网络内部包含定位网络、网格生成器、采样器三个部件。经过训练后，它可以根据输入图像的特征图动态地产生空间变换网格，然后采样器根据变换网格核函数从原始图像中采样获得一个矩形的文本图像。RARE 中支持一种称为 TPS（thin-plate splines）的空间变换，从而能够比较准确地识别透视变换过的文本、以及弯曲的文本。

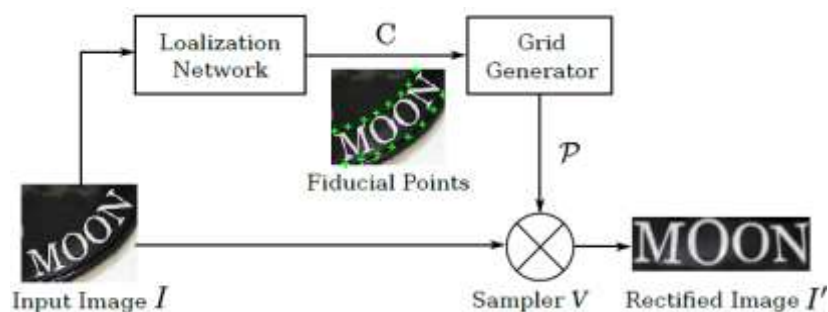


图 3.20 STN 的结构

如下表 2.19 所示我们报告结果，并与其他方法进行比较。在无约束的识别任务（没有词典识别）的情况下，我们的模型在比较中优于所有其他方法。在 IIIT5K 上，RARE 表现优异现有技术的 CRNN [6]增加了近 4 个百分点，表明性能有明显改善。我们观察到 IIIT5K 包含大量不规则文本，尤其是弯曲的文本，而 RARE 在处理不规则文本方面具有优势。请注意，尽管我们的模型在某些数据集上落后于[21]，但我们的模型与[21]的不同之处在于它能够识别随机字符串，如电话号码，而[21]只能识别其 90k 字典中的单词。在受限制的识别任务（用词汇识别）中，RARE 实现了最先进或极具竞争力，并且精度最高。

表 3.19 RARE 在四个数据集中的性能

Method	IIIT5K			SVT		IC03				IC13	
	50	1k	None	50	None	50	Full	50k	None	None	None
ABHY [35]	24.3	-	-	35.0	-	56.0	55.0	-	-	-	-
Wang et al. [35]	-	-	-	57.0	-	76.0	62.0	-	-	-	-
Mishra et al. [25]	64.1	57.5	-	73.2	-	81.8	67.8	-	-	-	-
Wang et al. [37]	-	-	-	70.0	-	90.0	84.0	-	-	-	-
Goel et al. [11]	-	-	-	77.3	-	89.7	-	-	-	-	-
Bisacco et al. [5]	-	-	-	90.4	78.0	-	-	-	-	-	87.6
Alsharif and Pateu [3]	-	-	-	74.3	-	93.1	88.6	85.1	-	-	-
Almazan et al. [2]	91.2	82.1	-	89.2	-	-	-	-	-	-	-
Yao et al. [39]	80.2	69.3	-	75.9	-	88.5	80.3	-	-	-	-
Rodriguez-Serrano et al. [21]	76.1	57.4	-	70.0	-	-	-	-	-	-	-
Jaderberg et al. [19]	-	-	-	86.1	-	96.2	91.5	-	-	-	-
Sa and Lu [34]	-	-	-	83.0	-	92.0	82.0	-	-	-	-
Gordo [32]	93.3	86.6	-	91.8	-	-	-	-	-	-	-
Jaderberg et al. [17]	97.1	92.7	-	95.4	80.7	98.7	98.6	93.3	93.1	90.8	-
Jaderberg et al. [16]	95.5	89.6	-	93.2	71.7	97.8	97.0	93.4	89.6	81.8	-
Shi et al. [32]	97.6	94.4	78.2	96.4	80.8	98.7	97.6	95.5	89.4	86.7	-
RARE	96.2	93.8	81.9	95.5	81.9	98.3	96.2	94.8	90.1	88.6	-
RARE (SRN only)	96.5	92.8	79.7	96.1	81.5	97.8	96.4	93.7	88.7	87.5	-

端到端模型

端到端模型的目标是一站式直接从图片中定位和识别出所有文本内容来。

FOTS Rotation-Sensitive Regression

Liu, Liang 等人在 2018 年提出了 FOTS (Fast Oriented Text Spotting) 模型[5]。FOTS 是图像文本检测与识别同步训练、端到端可学习的网络模型。检测和识别任务共享卷积特征层，既节省了计算时间，也比两阶段训练方式学习到更多图像特征。引入了旋转感兴趣区域 (RoIRotate)，可以从卷积特征图中产生出定向的文本区域，从而支持倾斜文本的识别。

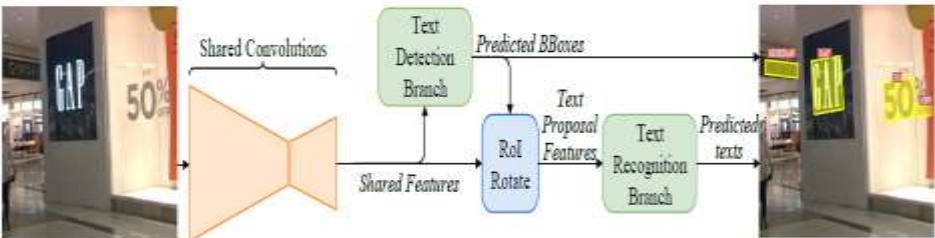


图 3.21 FOTS 模型

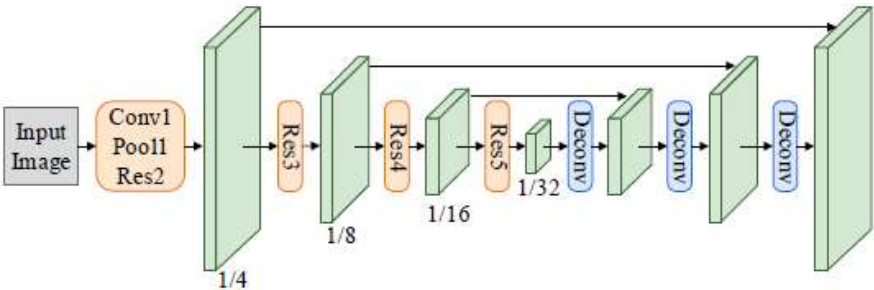


图 3.21 共享卷积模型

如下表 2.21 所示,FOTS 与 ICDAR 2015 上的其他结果进行比较,得分为百分比。“FOTS MS”代表多尺度测试,“FOTS RT”代表 FOTS 现在的实时版本。“端到端”和“Word Spotting”是用于文本定位的两种类型的评估协议。“P”,“R”,“F”分别表示“精度”,“召回”,“F-量度”。

表 3.21 在 ICDAR 2015 上与其他结果的百分比得分比较

Method	Detection			Method	End-to-End			Word Spotting		
	P	R	F		S	W	G	S	W	G
SegLink [43]	74.74	76.50	75.61	Baseline OpenCV3.0+Tesseract [26]	13.84	12.01	8.01	14.65	12.63	8.43
SSTD [13]	80.23	73.86	76.91	Deep2Text-MO [51, 50, 29]	16.77	16.77	16.77	17.58	17.58	17.58
WordSpot [17]	79.33	77.03	78.16	Beam search CUNI+S [26]	22.14	19.80	17.46	23.37	21.07	18.38
RRPN [39]	83.52	77.13	80.20	NJU Text (Version3) [26]	32.63	-	-	34.10	-	-
EAST [54]	83.27	78.33	80.72	StradVision.v1 [26]	33.21	-	-	34.65	-	-
NLPR-CASIA [15]	82	80	81	StradVision-2 [26]	43.70	-	-	45.87	-	-
R ² CNN [25]	85.62	79.68	82.54	TextProposals+DictNet [7, 19]	53.30	49.61	47.18	56.00	52.26	49.73
CCHLAB_FTSN [4]	88.65	80.07	84.14	HUST_MCLAB [43, 44]	67.86	-	-	70.57	-	-
Our Detection	88.84	82.04	85.31	Our Two-Stage	77.11	74.54	58.36	80.38	77.66	58.19
FOTS	91.0	85.17	87.99	FOTS	81.09	75.90	60.80	84.68	79.32	63.29
FOTS RT	85.95	79.83	82.78	FOTS RT	73.45	66.31	51.40	76.74	69.23	53.50
FOTS MS	91.85	87.92	89.84	FOTS MS	83.55	79.11	65.33	87.01	82.39	67.97

STN-OCR 模型

Bartz, Yang 等人在 2017 年提出了 STN-OCR 模型[11]，STN-OCR 是集成了了图文检测和识别功能的端到端可学习模型。在它的检测部分嵌入了一个空间变换网络（STN）来对原始输入图像进行仿射（affine）变换。利用这个空间变换网络，可以对检测到的多个文本块分别执行旋转、缩放和倾斜等图形矫正动作，从而在后续文本识别阶段得到更好的识别精度。在训练上 STN-OCR 属于半监督学习方法，只需要提供文本内容标注，而不要求文本定位信息。作者也提到，如果从头开始训练则网络收敛速度较慢，因此建议渐进地增加训练难度。STN-OCR 已经开放了工程源代码和预训练模型。

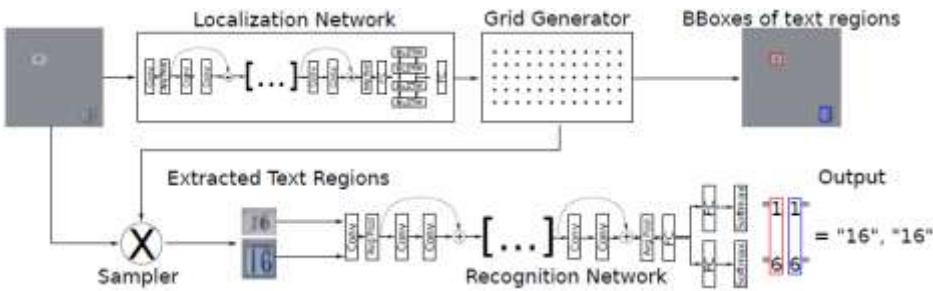


图 3.22 STN-OCR 模型

在表 2.22 中,展示了 STN-OCR 模型在 ICDAR 2013 强健读数,街景文本(SVT)和 IIIT5K 基准数据集上的识别结果。 为了评估 ICDAR 2013 和 SVT 数据集，其中过滤了包含非字母数字字符的所有图像，并通过后处理获得了最终结果。

表 3.22 STN-OCR 在三个数据集上上与其他结果比较

Method	ICDAR 2013	SVT	IIIT5K
Photo-OCR [1]	87.6	78.0	-
CharNet [18]	81.8	71.7	-
DictNet* [15]	90.8	80.7	-
CRNN [27]	86.7	80.8	78.2
RARE [28]	87.5	81.9	81.9
Ours	90.3	79.8	86

8. 开源数据集介绍

本章将列举可用于文本检测和识别领域模型训练的一些大型公开数据集， 不涉及仅用于模型 fine-tune 任务的小型数据集。

Google FSNS

该数据集是从谷歌法国街景图片上获得的一百多万张街道名字标志，每一张包含同

一街道标志牌的不同视角，图像大小为 600*150，训练集 1044868 张，验证集 16150 张，测试集 20404 张。



图 3.23 Google FSNS 数据集样本

ICDAR

ICDAR 数据集每届呈现的特点不同，难度和复杂性也逐渐提高。从 ICDAR 2013 起，主要针对在自然场景下的水平文字检测，随后在 ICDAR2015 和 ICDAR2017 中，数据集的难度不断增加，文字检测不仅要在自然场景下完成，字体变化如模糊、倾斜、背景干扰等也增加复杂性和识别难度，带来更大挑战。ICDAR2017 又增加了多国语言的文字检测，难度进一步加深。



图 3.24 ICDAR 数据集样本

所以，由于 ICDAR 的难度复杂性和专业度，无可厚非成为了 OCR 领域最受关注的竞赛，每届 ICDAR 公布的数据集，吸引世界各国的深度学习、OCR 领域技术强队前来刷榜、挑战。由 IAPR 主办的 ICDAR 文档分析与识别国际会议，自 1991 年创办以来，每两年举行一次，每届的东道国也在变化。

COCO-TEXT

该数据集，包括 63686 幅图像，173589 个文本实例，包括手写版和打印版，清晰版和非清晰版。文件大小 12.58GB，训练集：43686 张，测试集：10000 张，验证集：10000 张。



图 3.25 COCO-TEXT 数据集样本

Synthetic Data for Text Localisation

在复杂背景下人工合成的自然场景文本数据。包含 858750 张图像，共 7266866 个单词实例，28971487 个字符，文件大小为 41GB。该合成算法，不需要人工标注就可知道文字的 label 信息和位置信息，可得到大量自然场景文本标注数据。



图 3.26 Synthetic Data for Text Localisation 数据集样本

Chinese Text in the Wild (CTW)

该数据集包含 32285 张图像，1018402 个中文字符(来自于腾讯街景)，包含平面文本，

凸起文本，城市文本，农村文本，低亮度文本，远处文本，部分遮挡文本。图像大小 2048*2048，数据集大小为 31GB。以(8:1:1)的比例将数据集分为训练集(25887 张图像，812872 个汉字)，测试集(3269 张图像，103519 个汉字)，验证集(3129 张图像，103519 个汉字)。



图 3.27 Chinese Text in the Wild(CTW)数据集样本

四、 发展趋势

从 OCR 涉及层面来讲：图像文字检测和识别技术有着广泛的应用场景。已经被互联网公司落地的相关应用涉及了识别名片、识别菜单、识别快递单、识别身份证、识别营业证、识别银行卡、识别车牌、识别路牌、识别商品包装袋、识别会议白板、识别广告主干词、识别试卷、识别单据等等。

从各大巨头企业涉及 OCR 来讲：已经有不少服务商在提供图像文字检测和识别服务，这些服务商既包括了腾讯、百度、阿里、微软、亚马逊、谷歌等大型云服务企业，也包括了一些活跃在物流、教育、安防、视频直播、电子政务、电子商务、旅游导航等垂直细分行业的服务企业。这些企业既可以使用提前训练好的模型直接提供场景图文识别、卡证识别、扫描文档识别等云服务，也可以使用客户提供的数据集训练定制化模型（如票据识别模型），以及提供定制化 AI 服务系统集成等

从 OCR 未来发展来讲：作为近五年来的第一次全面调查，这里分析了最近的研究方法，根据各种标准对它们进行分类，并说明了最具代表性的方法的性能。在过去的十

年中，随着改进方法的出现，这一领域的研究取得了进展。然而，端到端识别性能较低，表明未来的研究还有很大的空间。

五、 参考文献

- [1] Zhou X, Yao C, Wen H, et al. EAST: An Efficient and Accurate Scene Text Detector[J]. 2017:2642-2651.
- [2] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Trans Pattern Anal Mach Intell, 2015, 39(6):1137-1149.
- [3] Tian Z, Huang W, He T, et al. Detecting Text in Natural Image with Connectionist Text Proposal Network[C]// European Conference on Computer Vision. Springer, Cham, 2016:56-72.
- [4] Ye Q, Doermann D. Text Detection and Recognition in Imagery: A Survey[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(7):1480-1500.
- [5] Liu X, Liang D, Yan S, et al. FOTS: Fast Oriented Text Spotting with a Unified Network[J]. 2018.
- [6] Shi B, Bai X, Yao C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(11):2298-2304.
- [7] Dai Y, Huang Z, Gao Y, et al. Fused Text Segmentation Networks for Multi-oriented Scene Text Detection[J]. 2018.
- [8] Liu Y, Jin L. Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:3454-3461.
- [9] Hu H, Zhang C, Luo Y, et al. WordSup: Exploiting Word Annotations for Character Based Text Detection[C]// IEEE International Conference on Computer Vision. IEEE Computer Society, 2017:4950-4959.
- [10] Shi B. Robust Scene Text Recognition with Automatic Rectification[J]. Computer Vision and Pattern Recognition (cs.CV), 2016
- [11] Bartz C, Yang H, Meinel C. STN-OCR: A single Neural Network for Text Detection and Text Recognition[J]. 2017.
- [12] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4):640-651.
- [13] Max Jaderberg. Spatial Transformer Networks Int. J. Computer Vision and Pattern Recognition (cs.CV), PP, 3, 2017 VOL. 39, NO. 4, APRIL 2017
- [14] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in Proc. Eur. Conf.Comput. Vis., 2014, pp. 297–312.
- [15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38,no. 1, pp. 142–158, Jan. 2015.
- [16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Spatial Transformer Networks,"

Computer Vision and Pattern Recognition..., pp. 2–3, 2016.

[17] Dimosthenis Karatzas, Lluís Gomez-Bigorda, Angelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, and Faisal Shafait. Icdar 2015 competition on robust reading. In International Conference on Document Analysis and Recognition, pages 1156–1160, 2015.

[18] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al. Icdar 2015 competition on robust reading. In Document Analysis and Recognition (ICDAR), 2015 13th International Conference on, pages 1156–1160. IEEE, 2015. 2, 6

[19] Jianqi M, Weiyuan S "Arbitrary-Oriented Scene Text Detection via Rotation Proposals," Computer Vision and Pattern Recognition (cs.CV) , 10.1109/TMM.2018.2818020

[20] Shi B, Bai X, Belongie S. Detecting Oriented Text in Natural Images by Linking Segments[J]. 2017:3482-3490.PP.1-8

[21] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Reading text in the wild with convolutional neural networks. Int. J. Comput. Vision, 2015.