摘 要

场景文本的检测与识别是计算机视觉任务中聚焦的热点，目前已经广泛应用于自动驾驶、实时翻译、智能识别等领域。与传统的印刷文档相比，场景文本图像主要难点在于其图像背景复杂、文本形式不一、文本模糊失真等，这给自然场景下的文本检测与识别工作带来一定的挑战。相较于传统算法，基于深度学习的算法在图像特征提取、网络训练稳定性方面更有优势。但是，基于深度学习的方法目前存在着一些问题：当场景文本图像的文本尺度差距较大、文本的形状各异时，文本检测模型性能下降，检测出的文本区域的图像分辨率也比较低，对后续的文本识别工作造成影响。为了进一步提升自然场景下的文本检测与识别的性能，本文的主要工作如下：

（1）提出一种基于特征引导和自适应融合机制的不规则场景文本检测算法，该算法解决了文本检测模型定位不好的问题。本文针对网络特征提取的不足，设计了特征聚合引导模块，该模块在骨干特征提取网络之后，实现高维通道转换低维通道的同时增强了网络的特征表达能力。在多尺度特征融合阶段，设计了自适应特征融合模块，该模块可以较好地融合不同层次的特征，达到关注不同尺寸的文本的目的，通过引入本文提出的模块，模型的检测性能得到了进一步提升。

（2）提出了一种基于全局与局部特征信息结合的场景文本识别算法，该算法针对图像文本质量低下、模糊失真、扭曲等问题，提出了尺度感知与语义增强模块。在尺度感知模块中，首先通过全局平均池化聚合多尺度特征，其次利用 Transformer 强大的自注意力机制，捕获图像特征的全局上下文信息，然后将全局特征信息与局部特征信息输送到语义增强模块，通过该模块缓解不同特征之间的差异，最后采用结合注意力机制的 GRU 网络对文本序列进行识别。

本文提出的文本检测模型在 ICDAR2015、CTW1500、MSRA-TD500 和 TotalText 四个基准数据集上的 F 值，与 PANNet 文本检测算法相比，分别提升了 2.4%、1.3%、1.8%和 1.4%，性能提升较为显著。本文提出的文本识别模型在数据集 IIIT5K、SVT、IC13、IC15、SVTP、CUTE 的识别准确率分别是 95.0%、91.2%、94.9%、81.7%、83.9%、88.5%，与 Aster 文本识别算法相比，识别的准确率更高。

关键词：文本检测，文本识别，特征引导，自适应融合，全局与局部特征

目 录

第 1 章 绪论

文字在人类文明中扮演着重要角色，对人类的传承有着深远影响，例如文字使得人们能够更好地表达自己的想法，使得信息可以传播得更加迅速、准确、有效，也使得人们可以方便地记录历史的发展，文字可以说是人类历史上最伟大、最具影响力的发明之一。随着信息时代的到来，利用深度学习技术自动提取图像中的文字信息，可以更好地提高人类的工作效率，因此，对场景下的文字进行信息提取是计算机视觉领域中的一个重要研究方向。

1.1 研究背景及意义

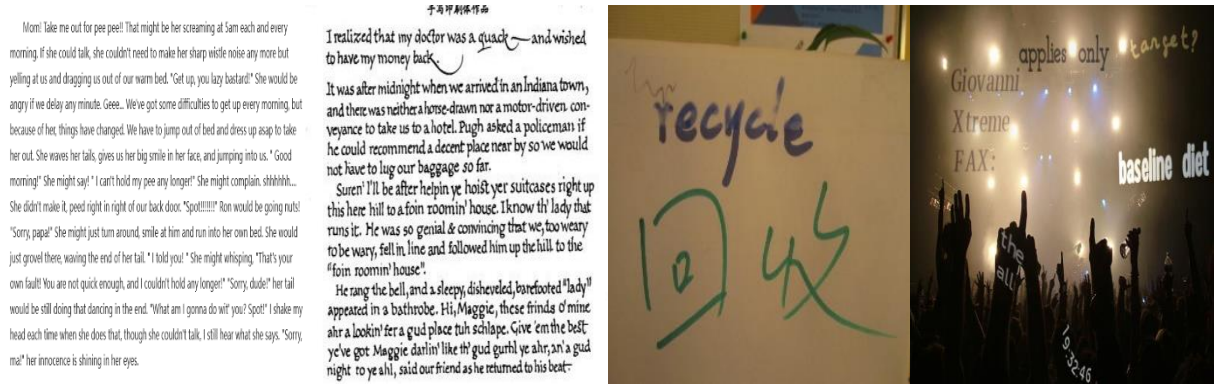
1.1.1 研究背景

文字是我们生活中不可或缺的组成部分，以各种形式存在于我们的日常生活中，例如期刊杂志，高速标志牌，报纸信件以及快递包装盒子等。近些年来，随着移动互联网的普及和多媒体技术水平的发展，人们可以利用多种电子设备将自然场景中的图像保存和传播，对这些图像中包含的文字进行检测和识别，可以帮助计算机更好地理解文本内容，从而提高智能应用的效率。在早期的文字研究中，主要的研究对象是标准格式的纸质载体，如报纸、图书、期刊等。由于该载体的印刷文字具有排版统一、字体稳定、颜色单一等特点，因此光学字符识别（Optical Character Recognition，OCR）[1]技术在上述对象已经取得了不错的成绩，然而如今的研究趋近于工业方向发展，自然场景下的文本检测与识别技术成为主要研究方向。相对来讲，自然场景下文本检测与识别比较困难：一方面容易受到自身背景复杂多变、字体形式不一、排版错综复杂等的影响，另一方面也容易受自然条件的干扰，如强光直射、树叶遮挡、物体与文字相似等因素。一般来讲，自然场景下的文本检测与识别存在着以下几个方面的困难：

（1）文本格式多样化：自然场景中的文本和标准的文档在格式方面有着很大的不同，在标准文档中，印刷体文字的排列是干净整齐的，如字符之间的距离、字体的颜色、大小、乃至文档的背景等，都是遵循标准格式的，如图 1-1（a），（b）所示，然而处于自然场景下的文本格式却是未知的，同一个场景下的文本会也会因角度、艺术字体等因素的影响，形成不同的文本格式，如图 1-1（c），（d）所示。上述问题无疑对文本检测和识别带来了巨大的挑战。

（2）文本背景复杂性：自然场景下的图像背景多种多样，难免会出现有一些图像的背景颜色和字体颜色相近的情况，如图 1-2（a）所示，或者图像背景中存在许多与文本相似的对象（栅栏，窗帘等），如图 1-2（b）所示，这样就很容易造成误检，如此多的噪声信息给检测与识别工作造成了一定的困难，对模型算法的鲁棒性提出了不小的挑战。

（3）图像文本质量低下：自然场景中的文本图像一般来自外接设备，如手机摄像头、数码相机拍摄、监控设备截取等，获取的图像难免存在着透视、模糊、畸变等问题，如图 1-2（c）所示，或者受场景图像的分辨率影响，导致电子设备获取的图像质量较低，如图 1-2（d）所示，上述图像即便是人类用肉眼，也很难精准地对文本进行定位和识别，对模型来说更是一个巨大的挑战。



（a） （b） （c） （d）

图 1-1 标准文本和自然场景文本的格式

（a） （b） （c） （d）

图 1-2 背景复杂和图像质量低下的自然场景文本

综上所述，文本格式的多样化、图像背景的复杂性、以及图像文本质量低下等问题，对自然场景下的文本检测与识别技术提出了更高的要求，传统的 OCR 技术已经难以满足自然场景下的文本检测与识别任务，于是，研究人员基于深度学习技术提出了很多解决方法和思路。

1.1.2 研究意义

对于图像中包含的诸多信息来说，文本信息的内容表达更为丰富，能够更好地帮助人们或者计算机理解图像的表达含义，因此准确定位且识别出自然场景图像中的文本信息，有利于进一步理解图像。目前，自然场景文本检测与识别技术，已经广泛应用于自动驾驶、实时翻译、智能识别等领域，人们的生活和工作也因此变得更加方便、快捷。因此，场景文本的检测与识别具有重要的发展意义，潜藏着巨大的实用价值，并且推动了其他领域的发展，下面是一些领域的简单应用：

（1）自动驾驶：随着汽车走进人们的生活，车辆的自动辅助驾驶逐渐成为社会关注的话题，目前越来越多的汽车制造商都希望在车辆中安装辅助系统，该辅助系统可以通过车载摄像头和传感器，获取大量的自然场景信息，例如在城乡道路、公园道路、人行斑马线等道路的侧方，通常设置了大量的指路标志、限速标志、注意标志等道路标志牌，如果采用场景文本检测与识别技术提取标志牌上的内容，就可以辅助自动驾驶导航系统的建立。

（2）即时翻译：随着全球化时代的到来，人们有了更多接触不同国家语言的机会。如果在日常生活中碰到了不同语言切换的情形，那么通过手机拍摄或者截取需要翻译的照片，利用文本检测与识别技术就可以将图像中的文本检测并识别出来，方便快速地实现不同国家语言之间的翻译，在一定程度上解决了人们在生活中语言不通的难题。

（3）智能识别：随着信息化时代的到来，深度学习技术的不断应用，促进了智能办公系统的建立。在现代化的办公场景中，对于名片、发票、税务票等纸质票据的管理，财务管理人员需要将其手工录入系统，录入过程不仅低效枯燥，而且在操作过程中很难避免出错，但是，通过扫描或者拍照待处理的纸质票据，然后把图片上传到文本检测与识别软件，就可以快速地获取到文字内容，不仅节省了财务人员的时间成本，而且提升了公司的工作效率。

自然场景图像的文本检测和识别的应用场景远不止如此，目前，随着电脑的计算能力越来越强大，人们对文本检测与识别算法提出了更高的要求，因此，发展更高性能的场景文本检测与识别的算法越来越重要。

1.2 国内外研究现状

在上个世纪 30 年代，德国学者首先提出了 OCR 的概念，紧接着从上个世纪的 50 年代起，欧美国家的学者开始了对英文字符识别技术的研究，英文 OCR 技术的发展随

着计算机技术的进步而逐渐完善，并且投入了商业应用。而国内从 70 年代末期和 80 年代初期，开始进行中文 OCR 相关的研究，经过数十年的发展，也取得了不错的成绩。这个时期的 OCR 解决的是简单场景下的文字识别，但是，随着各种行业的应用场景的逐渐发展，基于传统的 OCR 算法很难迁移到各个场景中，而深度学习技术的在各个领域都获取了不错的效果，并且可以方便的进行迁移学习，于是利用深度技术对自然场景下的文本进行检测与识别，便成为了研究人员关注的热点。

1.2.1 自然场景文字检测算法

自然场景文本检测是对场景图像的文本区域进行准确定位，且定位到的文本区域尽量不包含背景，是一个比较具有挑战性的任务。当前的场景文本检测分为传统算法和深度学习算法，传统算法利用手工设计和理论经验提取图像特征，给时间与人力成本等带来了负担，而深度学习算法可以自动提取图像的特征，省去了人工设计的繁琐操作，节约了时间和人力成本。目前基于深度学习技术的场景文本检测算法已经成为了主流，传统的场景文本检测算法慢慢退出了历史的舞台。基于深度学习的文本检测一般可以分为两类：基于边框回归与基于语义分割，两者的不同在于基于边框回归的场景文本检测方法把文本作为物体进行定位得到边界坐标，根据坐标计算图像中文本的位置，把该位置表示文本的检测结果，基于语义分割的文本检测方法从全卷积网络 FCN（Fully Convolutional Network，FCN）[2]生成的分割图中提取文本区域，然后把文本区域拟合成完整文本实例，得到最终的检测结果。以下是两种不同算法的具体介绍。

（1）基于边框回归的文本检测方法

这类算法经常在通用的目标检测框架 Fast R-CNN[3,4]系列和 YOLO[5,6]系列的基础上进行改进，与目标检测算法一样，文本检测算法同样采用区域提议网络（Region Proposal Network，RPN），也是需要预先生成大量的锚框，锚框的数量与尺寸利用人工手动设置，经过网络训练之后判断锚框中是否存在文本实例，假设锚框中存在文本实例，把锚框的坐标作为输出结果。根据文本在自然场景图像中存在的特点，许多基于边框回归的文本检测算法被提出，如 Tian 等人[7]在 Fast R-CNN 基础上提出了 CTPN（Connectionist Text Proposal Network，CTPN），该算法采用分治法的策略，预先设置固定宽度的锚框，利用网络模型对文本行分段预测锚框，然后对符合筛选条件的锚框串联成文本行，同时为了高效地利用图像特征的上下文信息，在卷积层与全连接层之间引入了双向循环神经网络，一定程度上提升了模型检测文本方面的能力，该算法虽

然在任意长度上的水平文本中表现良好，但是对倾斜文本的效果表现稍弱。Ma 等人[8]同样是在 Fast R-CNN 的基础上提出了 RRPN（Rotation Region Proposal Networks，RRPN）算法，该模型使用旋转区域提议网络，生成含有文本角度信息的候选框，同时提出了旋转感兴趣区域 RRoI（Rotation Region-of-Interest，RRoI）池化层将任意方向的候选框和特征图进行映射，经过实验表明，该算法提升了任意方向文本检测的效果。Liao 等人[9]在 SSD[10]算法的基础上提出了 TextBoxes 算法，该算法认为自然场景下的文本纵横比变化较大，不仅修改了 SSD 算法网络中的卷积核的大小，而且也更改了预设锚框的长宽比例，经过改进后的算法更好地适应了文本检测的任务。为了提升多角度文本的检测效果，Liao 等人[11]继而提出 TextBoxes 的改进算法 TextBoxes++，不仅重新调整了候选框的比例，而且在回归阶段增添了边框的角度信息，使得边框可以根据角度信息进行旋转，改进后的算法可以更好适应倾斜文本。

（2）基于语义分割的文本检测方法

这类算法一般是以全卷积网络为基础，结合特征金字塔网络（Feature Pyramid Network，FPN）[12]组成对称的编解码结构。其主要流程是：首先利用卷积神经网络对输入图像进行特征提取，以得到不同尺度的特征图，然后使用 FPN 融合不同尺度的特征，得到具有丰富信息的特征图，最后对当前特征图的每个像素分类，若同属于当前文本的像素，则把当前像素区域拟合成文本框，否则就把该像素区域归为背景。对于不规则的文本检测来说，基于语义分割的检测效果明显优于基于边框回归算法，例如 Zhang 等人[13]首次将全卷积网络应用于文本检测任务，该算法首先通过 VGG[14]提取输入图像的多个尺度特征图，并且融合这些尺度的特征图，根据融合后的特征对像素分类，得到文本区域的概率分割图，其次利用 MSER 算法提取候选字符区域，最后通过后处理手段生成文本结果。Deng 等人[15]采用实例分割的方式提出了 PixelLink 算法，该算法采用预测像素的类别与判断相邻像素之间的连通关系，对同一个类的文本区域的像素进行连通，最后将文本区域分割成不同文本实例。Wang 等人[16]提出了渐进尺度扩张的后处理手段算法 PSENet，该算法根据各种文本实例预先生成不同比例的内核，在后处理阶段，采用内核合并外核的方式进行逐步扩张，最后形成一个完整的文本实例，由于不同尺寸内核之间具有一定的约束，因此采用渐进扩张的方式可以有效地检测任何形状的文本实例。对于基于语义分割的场景文本检测算法而言，在区分像素的类别是文本区域还是背景区域的过程中，普遍采用手动设置阈值干预像素类别，一定程度上影响了模型的性能，为了

能够解决这个问题。Liao 等人[17]提出的 DBNet（Differentiable Binarization Network，DBNet）算法，该算法可以结合训练网络动态地学习每个像素的阈值，利用近似可微分的二值化阈值替换固定阈值，自适应的学习每个像素的阈值，然后预测特征的概率图和阈值图，进而推断出近似二值图，用于区分文本区域与背景区域，最后生成完整的文本实例。考虑到不同特征信息对网络分割的重要性，Liao 等人[18]提出的 DBNet 改进算法的 DBNet++，该算法提出了一种自适应尺度融合 ASF 模块，在多尺度特征融合阶段采用自适应的手段实现特征的融合，在 ASF 模块中引入了空间注意力机制[19]，用于得到不同尺度特征的权重因子，然后将权重因子乘以对应的尺度特征，这不仅提高了模型的尺度鲁棒性，而且提升了模型的检测效果。

Radford 等人结合视觉与语言知识的预训练提出了 CLIP[20]（Contrastive Language-Image Pre-Training，CLIP）模型，目前已经被越来越多的研究人员所证实，该模型在下游任务中表现出了巨大潜力， 由于文本检测任务也可以表示为视觉与语言组成的模型，于是 Yu 等人[21]提出了基于 CLIP 的文本检测 TCM 算法，该算法首次将 CLIP 与文本检测任务相结合，直接把 CLIP 迁移于文本检测模型中，经过大量的实验表明，在没有使用特定的预训练过程的情况下， 引入 CLIP 到场景文本检测模型中，检测效果优于使用预训练的场景文本检测算法。

1.2.2 自然场景文字识别算法

与传统的算法不同，基于深度学习的文本识别算法无需对字符分割，普遍采用序列预测的方式，对整张文本图像进行识别，目前基于深度学习的识别框架大多是编码器-解码器的结构，在该结构中经常采用连接时间分类技术（Connectionist Temporal Classification，CTC）[22]和注意力机制[23]进行识别模型的搭建。CTC 最初应用于语音识别工作，用来解决神经网络的输入音频符号和输出识别标签长度不同的问题。注意力机制最初应用在自然语言领域中的机器翻译任务中，用来获取不同时刻编码状态的权重，进而解决解码阶段语义编码向量固定的问题。基于深度学习技术的识别算法一般都采用上述方式，例如 Shi 等人[24]首次将 CTC 引入到场景文本识别中，该算法利用 CNN 作为特征提取的编码器，然后将得到的图像特征转换为时间序列，最后联合 BLSTM 与 CTC 作为网络识别端的解码器，预测输入序列的最终结果。针对 CTC 在神经网络训练过程中出现过拟合的问题，Zhang 等人[25]提出基于最大熵的正则化方法，通过平滑整个数值的分布，抑制最大概率路径，增强 CTC 的泛化和探索能力。但是，采用 CTC 的场景文