学校代码：10246
学　　号：17210240256

硕 士 学 位 论 文

（专业学位）

基于生成对抗网络改进的场景文本识别算法

Improved Scene Text Recognition Algorithm Based on Generative Adversarial Network

院　　系：　　计算机科学技术学院
专业学位类别（领域）：　　计算机应用技术
姓　　名：　　张培尧
指 导 教 师：　　薛向阳　教授
完 成 日 期：　　2019 年 9 月 8 日

指导小组成员名单

薛向阳 教授

张玥杰 教授

金城 教授

李斌 青年研究员

目 录

摘 要

场景文本识别在实际生产生活中有着非常丰富的应用场景，是无人驾驶、传图翻译、车牌识别、广告识别等应用的重要环节，这一方向的研究需要人工智能、计算机视觉、图像处理、深度学习、模式识别等相关技术的综合应用，具有重要的研究价值。本文利用计算机强大的计算能力，设计基于生成对抗网络改进的场景文本识别算法，对场景文本的自动识别研究有着重大的理论意义。

本文回顾了场景文本识别的发展历程和现阶段场景文本识别的技术难点，着重解决复杂背景的干扰和不规则文本的识别问题，对文本识别这一课题做了探索，构想出以生成模型和识别模型相结合的场景文本识别算法。将场景文本识别分为图像预处理和文本识别两部分，实现图像的智能识别。

图像预处理部分使用生成模型对图像进行转换，将复杂背景的文本图像转换为背景干净、易于识别的图像。生成模型采用条件生成对抗网络来进行训练，可以应对图像中由于自然环境的光线、气候等变化或者拍摄角度造成的字符模糊不清、图像分辨率较低等问题。文本识别部分采用基于深度卷积循环神经网络的识别模型，该模型包括特征提取层、序列建模层和转录层三个部分。特征提取层经过特殊设计，在标准的卷积神经网络中加入可以对卷积通道学习权重分布的模块和可以学习输入输出残差的残差学习单元，更好的提取图像特征，并且加入可变形卷积，自适应的改变卷积核采样的方式，从图像中学习到局部密集的形变情况，可以应对图像中各种各样的形变问题，更好的解决不规则文本的识别任务。

在三个文本识别公开数据集上的实验证明，本文提出的文本识别算法在复杂背景和不规则文本的识别任务中性能有很大的提升，实验结果显示取得了更高的识别准确率和更低的平均编辑距离，验证了生成器和可变形卷积在此任务中的有效性。除此之外，在实验过程中，为了丰富数据分布的多样性，对数据集进行平移、旋转、模糊、噪声等数据增强操作，也被证明可以使神经网络具有更好的识别效果。

关键词：场景文本识别，生成对抗网络，可变形卷积，循环神经网络

Abstract

Scene text recognition has a very rich application scene in actual production and life. It is an important part of applications such as driverless, map translation, license plate recognition, and advertisement recognition. The research in this direction requires artificial intelligence, computer vision, image processing, The comprehensive application of related technologies such as deep learning and pattern recognition has important research value. This paper makes use of the powerful computing power of the computer to design a scene text recognition algorithm based on the generation of anti-network improvement. It has great theoretical significance for the automatic recognition of scene text.

This paper reviews the development process of scene text recognition and the technical difficulties of scene text recognition at present. It focuses on solving the problem of complex background interference and irregular text recognition. It explores the topic of text recognition and conceives to generate models and identify them. A scene text recognition algorithm combined with a model. The scene text recognition is divided into two parts: image preprocessing and text recognition to realize intelligent recognition of images.

The image preprocessing section uses the generated model to transform the image, transforming the text image of the complex background into a clean, easily recognizable image. The generated model uses conditional generation to combat the network for training, and can cope with problems such as blurring of characters due to changes in light, climate, etc. of the natural environment or shooting angles, and low image resolution. The text recognition part adopts a recognition model based on deep convolutional cyclic neural network, which includes three parts: feature extraction layer, sequence modeling layer and transcription layer. The feature extraction layer is specially designed to add a module that can learn the weight distribution to the convolution channel and a residual learning unit

that can learn the input and output residuals in the standard convolutional neural network, to better extract image features, and to add deformable Convolution, adaptively change the way of convolution kernel sampling, learn from the image to local dense deformation, can deal with various deformation problems in the image, and better solve the recognition task of irregular text.

The experiments on the three text recognition public datasets prove that the text recognition algorithm proposed in this paper has greatly improved the performance in the recognition task of complex background and irregular text. The experimental results show that the recognition accuracy is higher and more. A low average edit distance verifies the effectiveness of the generator and deformable convolution in this task. In addition, in the process of experimentation, in order to enrich the diversity of data distribution, perform data enhancement operations such as translation, rotation, blur, noise, etc. on the data set, it also proved that neural networks have better recognition effects.

Key words: Scene Text Recognition, Generative Adversarial Network, Deformable Convolution, Recurrent Neural Network

第一章 绪 论

1.1 课题背景与研究意义

人类从点燃了第一堆篝火开始，开始了通过制造和使用工具来代替手工完成任务的不懈追求。从神话传说“潘多拉魔盒”到步入信息时代的人工智能，向我们展示了人类丰富的想象力，人类一直没有放弃对智能化工具的追求。

人类通过眼睛这一视觉感知器感知周围的环境，眼睛接收到来自外界的视觉刺激，视觉皮层对这些视觉信息进行抽象处理，并传递给我们的大脑[45]，帮助大脑做出更加精准科学的判断和决策[28]。受此启发，国内外的研究学者们一直致力于让计算机也拥有如此智能的视觉系统，可以自动感知场景信息，更好的解释和理解视觉世界，并做出相应的决策和反应。计算机视觉使用摄像头模拟人类的眼睛来捕捉现实场景的状态，用图像和视频的形式记录下来[44]，通过对图像或视频中的信息进行分析和处理，抽象出计算机可以理解的特征，完成场景分类[1][2]、检测[3][4]、分割[5][6]等任务。这一研究领域需要多个学科技术的结合，为了更好的收集和研究视觉数据，需要研究成像技术和信号处理，运用数学和机器学习的统计分析方法对数据进行抽象建模。

最近几年，海量数据的涌现、计算资源的大幅度提升、理论知识的不断探索，都促进了深度学习的蓬勃发展。将深度学习理论技术应用在计算机视觉这一方向，用深度神经网络模拟大脑的视觉皮层处理视觉刺激的过程，从数据中建模特征，用数据驱动网络不断学习不断优化，目前这类方法在很多视觉任务上取得了惊人的成果，甚至性能远远超越人类的水平，相关算法在很多实际场景中落地应用，有非常广泛的工业应用前景。

文本在自然场景中起着传播信息的重要作用，扮演传递消息的基本工具标志，这类信息的自主感应与处理也至关重要。不仅如此，图像中的文本识别对后续的高层语义的分析与处理的任务也有很大的帮助[46]。在传图翻译的场景中[37][40]，业务逻辑是用户上传场景图像，计算机系统对图像中的文本进行检测和识别，再将识别出的文本进行翻译，将最后的结果返回给用户。在这个流程里，文本识别是后续翻译的基础。在无人驾驶场景，使用计算机系统自动感知周围的环境，抽取环境中的信息进行分析并处理，做出相应的科学决策。这其中就需要从摄像头捕捉到的场景图像中，分析目前的环境状况，识别沿途的指示路牌[38]。作为计算机视觉领域的重要研究任务，对自然场景下文本识别的研究有着非常重要的理论价值和应用意义。

图 1-1 场景文本示例

1.2 场景文本识别任务及挑战

在场景图像中读取和识别文本序列是计算机视觉领域研究的典型任务，一直以来受到很多学者和研究机构的关注，相关理论技术逐渐成熟，并且在实际场景中得到了应用。

1.2.1 任务描述

场景文本识别技术的主要任务是：自动高效地对图像中的文字序列进行识别。场景文本识别的一般流程是：先从输入的场景图像中检测到文本区域，完成文本区域的定位，并将对应的文本区域从原图中分离出来，这是文本识别的前序任务。紧接着对文本区域图像中的字符序列进行分类与识别，保证正确识别图像中的文字序列。

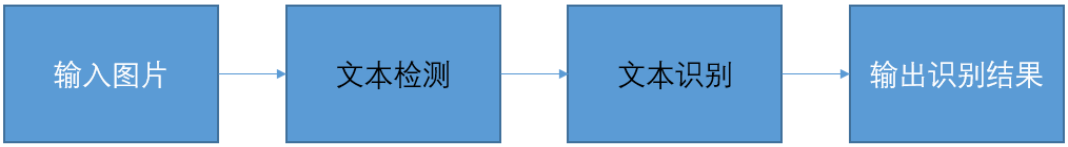

图 1-2 场景文本识别的流程

场景文本识别的方法有很多实际应用的场景，如车牌识别[22][23][24]、验证码识别[30][49][50]、准考证号识别[27]、单据识别[35]、手写字符识别[25][33][42]、票据识别[29]等。不同场景下的场景文本识别方法也不尽相同。本文重点研究文本识别的方法，也就是说完成从检测到的文本区域图像中识别出正确的文本序列的任务。与一般对象识别不同，场景文本基本是以序列数据的形式出现，对图像中出现的字符串序列进行识别需要对输入的图像系统地预测一系列的对象标签。因此，一般情况下

可以将场景文本识别问题当作序列任务去解决。

文本识别从最初的简单的邮票数字识别到现在深入到交通、安防、海关、餐饮、无人驾驶等众多领域，成为丰富多样的场景下不可或缺的智能化工具，应用场景越来越丰富，智能化要求也越来越高。文本识别相关的领域技术在过去短短数十年内的时间里飞速发展并日趋成熟，众多研究成果将深度学习领域的相关技术应用在文本识别任务，取得了不错的性能表现。

1.2.2 任务分类

经过多年的发展，场景文本识别任务领域有很多公开的数据集，既有英文数据集，也有中英文混合的数据集。这些数据集都有各自的特点和适用场景，可以提供给研究学者参考和评测。在公开数据集上对算法模型进行测试，来验证和比较不同算法的性能表现。场景文本识别任务的公开数据集可以根据是否为图像提供语义字典，分为以下三种类型：

（1）强语境约束

在强语境约束文本识别任务中，用于测试的数据集中包含与每张测试图像相匹配的已知长度的固定词典，并且词典中包含的单词个数较少[11]。推理期间将模型预测序列与词典中的单词进行比较和匹配，完成从每张图像对应的词典中找出图像文本所属类别的任务。

（2）弱语境约束

在弱语境约束文本识别任务中，数据集中没有为每一张图像都配置对应的词典，而是为整个数据集包含的图像准备一个字典的合集，里面包括所有可能会出现的字符类别。在推理期间模型会为每张图像从字典中挑选出合适的预测类别并组合在一起。

（3）无语境约束

与上述两种类型相反，在无约束的文本识别任务中，测试数据集里并没有为图像准备对应的字典，推理阶段要在没有字典的情况下，用训练好的模型直接对文本序列进行预测，来识别图像中的每个单词。

1.2.3 问题与挑战

不同场景下的文本识别任务有着不同的特点，解决方案也不尽相同。在规则的文本识别任务中，文本识别技术主要关注文档票据类的版面文字，这类场景下通常文本背景较为单一，文本规整，识别难度较低，目前有很多开源软件可以在

这类场景下有很好的性能表现。但是在自然场景识别任务中，这类软件的识别率就大大降低了，识别效果不太理想。这也是目前场景文本识别任务面临的难点和挑战。这些挑战可以分为以下几个方面：

（1）复杂背景

场景文本识别的实际应用场景环境复杂，用于文本识别任务的图像会受到光照、气候、周围遮挡物等自然环境因素的干扰，不同光照条件下图像表面的纹理特征会发生改变，增加识别难度。雨雪天气以及有雾的情况下采集到的图像会在图像前景造成干扰。另外自然场景下不可避免的会发生文本区域附近的树叶、建筑等对字符进行遮挡，造成字符结构缺失的现象。

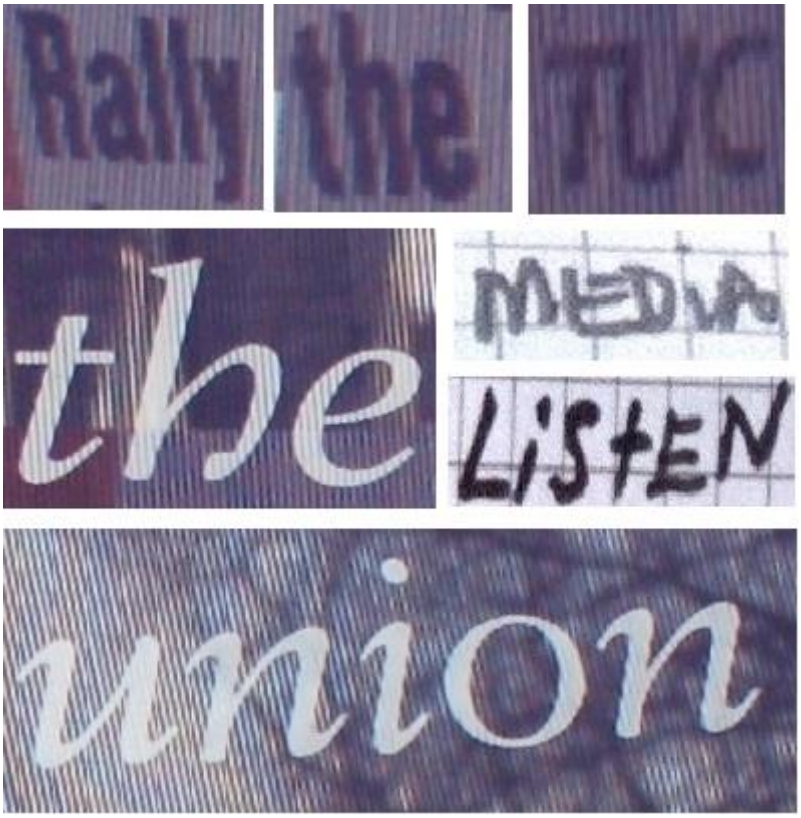

图 1-3 背景干扰示例

（2）文本自身因素

场景下的文本形式多样，变化万千，可能有不同的语言[51]、字体、颜色、大小等。除此之外，场景文本可能存在不同程度的角度旋转、不同比例的缩放、仿射变换等形变情况，字符之间也可能出现粘连的情况[36]。这些不规则文本变化多样，会有更大的概率被误检。

图 1-4 弯曲文本示例

（3）图像质量

图像采集过程中不同设备和不同的采集方式会造成图像大小不一[34]，不同图像的分辨率不同的现象[26]，其中低分辨率的图像损失了很多重要信息，识别难度会更大。采集过程中如果出现采集设备抖动的情况，采集到的图像会出现模糊和重影的情况[32]。不仅如此，采集设备与文本区域的角度不同，会造成采集到的图像中文本区域存在仿射变换的情况，有一定程度的形变，会加大识别难度。

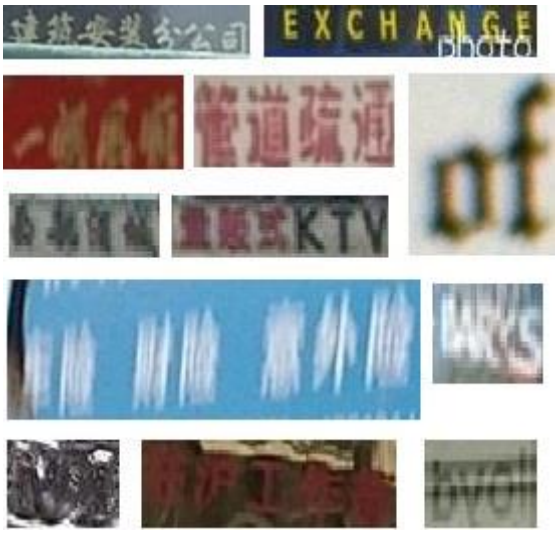

图 1-5 低质量图像示例

图 1-6 仿射变换示例

综上所述，场景文本识别技术主要面临复杂背景、低分辨率、不规则弯曲文本、多语言多角度等问题，难度大大提升。

1.3 本文主要研究内容

目前场景文本识别任务存在很多挑战，本文重点研究目前场景文本识别任务中存在的复杂场景和形变文本的问题，提高场景文本识别的准确率。具体来说，本文提出基于生成对抗网络改进的场景文本识别算法，该算法由两大模块组成-图像生成模块和文本识别模块。

（1）图像生成模块完成对输入的原始图像进行数据预处理的任务。这一模块的加入是为了解决复杂场景下的背景干扰问题，对输入图像进行预处理，将用于文本识别的图像转换为背景干净、易于识别的图像。为此，设计一个条件对抗网络进行训练，生成器使用神经网络学习对输入的图像进行预处理，将图像转换为背景干净的文本图像，判别器使用神经网络学习对图像进行分类判别，生成器和判别器进行对抗学习，在这个过程中，生成器和判别器的性能都逐渐优化，通过训练，生成器可以完成对图像进行预处理的任务。

（2）文本识别模块对图像生成模块的输出图像中的文本序列进行识别。为了解决不规则文本的形变问题，在特征提取器模块引入可变形卷积，可以自适应图像中复杂多样的形变情况。文本识别模块由三部分组成，第一部分使用卷积神经网络完成特征提取的任务，从生成器的输出图像中提取相应的特征，第二部分使用循环神经网络来完成序列建模的任务，输入前一部分提取到的图像特征，将特征转换为特征序列，经过网络的学习得到对应的序列输出。第三部分是基于时

序连接的分类（Connectionist Temporal Classification，简称 CTC）[7]模块对第二部分输出的序列结果进行处理，将循环神经网络输出的序列对应的图像中的位置。CTC 引入空白字段的占位符，对于循环神经网络的输出序列，删除掉其中识别为空白符的部分，并且去除重复的冗余字母，解决神经网络输出序列和输入序列不能一一对应的问题。

1.4 本文组织

本文内容总共分为五章，其中第一章为绪论，介绍场景文本识别任务的研究背景和研究意义，阐述该任务目前面临的主要挑战，并引出本文的主要研究内容，对模型进行创新。

第二章对介绍场景文本识别任务相关的算法研究工作，并对现有的算法进行阐述，分析现有模型的不足之处，并做出总结。介绍与本文设计的算法相关的先进神经网络，为后面算法的设计与实现打下理论基础。

第三章具体介绍本文提出的场景文本识别算法，包括两大模块，训练数据处理模块和文本识别模块。加入生成对抗网络对数据进行转换，作为数据预处理模块。文本识别模块在主流算法的基础上对特征提取器进行改进，可以更好的提取图像特征，并且自适应图像中各式各样的形变问题。

第四章设计消融实验，验证算法创新点的有效性，介绍实验所用的数据集以及评测指标、实验设置以及相应的实验结果，并对一系列对比实验进行分析与总结。

第五章会对全文做一次总结，总结本文对场景文本识别算法做出的改进，并展望未来。

第二章 相关研究工作

场景文本识别任务的快速发展离不开深度学习在计算机视觉领域的大规模应用和背后的理论支撑，本章将介绍近几年场景文本识别任务的相关研究工作，对已有的文本识别算法模型进行分类，分析这些方法的解决思路和性能表现。本文设计的基于生成对抗网络改进的场景文本识别算法在识别模型前加入生成器对输入图像进行转换，生成器采用条件生成对抗网络进行训练。识别模型包括特征提取层、序列建模层和转录层，其中特征提取层和序列建模层分别选择了先进的卷积神经网络和循环神经网络，本章也会对识别模型中用到的现今网络结构进行介绍，为接下来对本文提出的场景文本识别算法的设计与改进打下坚实的理论基础。

2.1 场景文本识别相关算法

在实际生产生活中，文本在信息交流与传输等众多场景下扮演着重要角色。图像中的文本识别有很多应用场景，也一直受到国内外研究学者们的关注。近年来，深度神经网络在许多计算机视觉任务中取得了巨大的成功。深度神经网络的蓬勃发展引起了研究人员的关注和兴趣，并利用它们来解决场景文本识别问题。

相较于传统的文本识别算法[41][43]，使用深度学习来完成场景文本识别任务的新兴算法不需要人工设计和提取特征，通过神经网络层不断从输入图像中抽象出视觉特征，浅层网络学习的图像分辨率高，可以学习到图像的纹理、局部属性等细节特征，深层网络通过不断对图像特征进行抽取，可以学习到图像中目标物体的轮廓等高层语义信息，不同的网络学习到不同层次的视觉特征，最后由输出层综合特征进行决策判断或者优化，完成对应的视觉任务[39]。这些神经网络从大量数据中进行学习，目前大多数研究使用的是有标注信息的数据集，数据集的每张图像都有对应的图像文本序列的标注，通过反向传播算法[47]，设计适合该任务和网络的目标函数，来约束网络的学习过程和模型的训练，动态调整神经元的连接和权重，学习输入图像到输出类别的映射关系。

2.1.1 基于字符切割的识别方法

谷歌公司于 2013 年提出的 PhotoOCR 方法[8]可先对图像中的文本序列进行切割，将字符串切割成一个一个的字符，再对切割好的字符进行分类。该方法首先

使用滑动窗口从图像中找到包含有文字的区域，完成文本检测任务，接着将文本序列进行分割，得到多个包括单个字符的图像，然后训练深度卷积神经网络作为分类器对图像特征进行建模，对分割得到的每个字符进行分类，将多个字符的识别结果通过定向搜索的方式进行合并，得到图像文本识别结果。

这种方法将文本识别任务拆分成了多个子任务来解决，算法模型由多个组件拼接在一起，不能够将不同组件在同一个目标函数的约束下对网络参数进行训练和优化，而且该方法要先对图像中的字符进行分割，得到一个一个的字符区域再进行识别，识别效果一定程度上依赖于文本区域的定位和分割的效果，会损失一部分精度，识别速度也很慢，不能满足实际应用场景的需求。

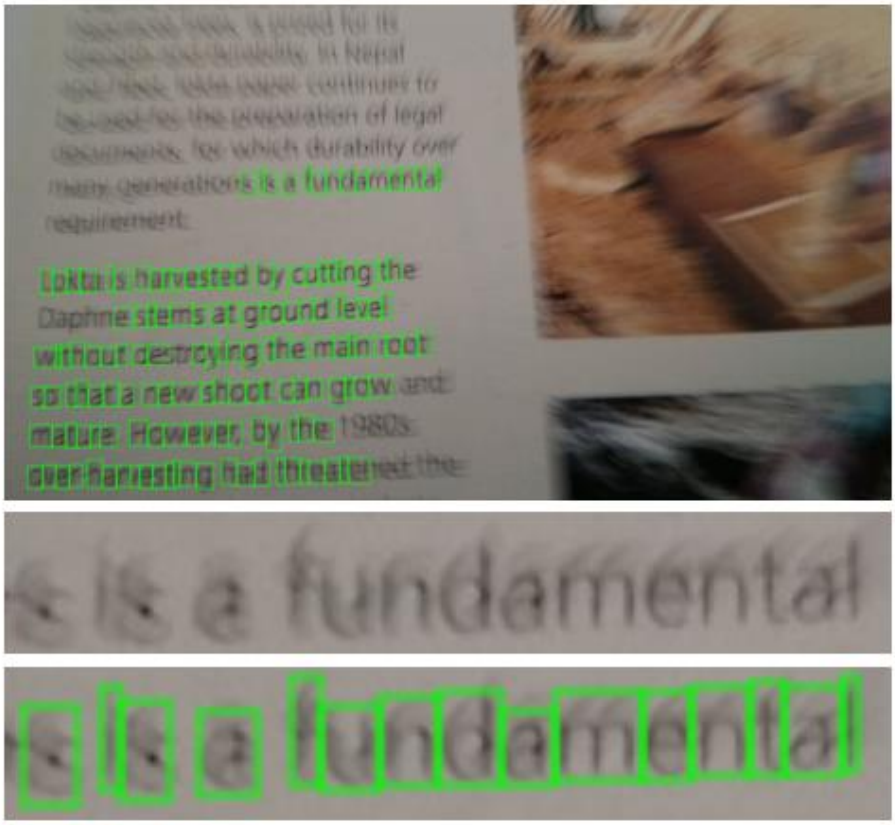

图 2-1 PhotoOCR[8]

2.1.2 单词级别的分类方法

牛津大学的研究学者于 2014 年提出一种利用深度学习和神经网络来进行场景文本识别的方法[9]，将文本识别任务当作图像分类任务来完成，不需要对图像中的字符进行拆分，直接对图像中的单词进行预测。算法模型采用卷积神经网络自动高效地从图像中不断抽取和学习视觉特征，再用全连接层对图像中的文本序

列直接进行分类，对图像中的单词进行预测。用于训练的每一张输入图像都对图像中的单词进行了标注，使用这种标注数据来对网络参数进行训练和更新。这种方法降低了提取特征的难度，效率更高。

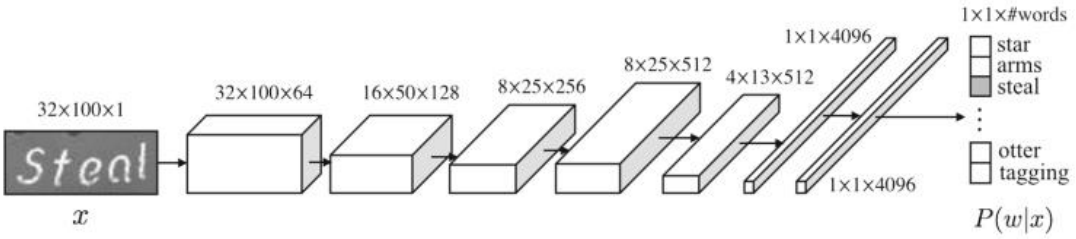

图 2-2 用于文本识别的卷积神经网络[9]

这种方法有两个缺点，首先因为全连接层参数较多，这类方法不适合类别数较多的任务场景，所以这种方法不适合解决中英文混合等字符类别数目众多的任务，因为这类序列的基本组合的数量可以超过百万。其次，文本序列的独特属性是它们的长度可以发生很大变化。例如，英语单词可以由两个字符组成，例如“OK”或十五个字符，例如“Congratulations”。卷积神经网络在图像分类任务上表现卓越，但是这种网络要求有固定长度的预测序列，输出分类结果的全连接层的数目在训练网络过程中要保持固定不变，事实证明，用特定数据集训练出的模型很难推广到其他类型的序列式对象，如中文文本，乐谱等，因此不能应对任意长度的文本序列的识别任务

2.1.3 序列级别的文本识别方法

上述两种方法都是将文本识别任务转化成了图像分类任务，第一种方法将图像中的字符串进行切割，对单个字符进行分类，第二种方法直接对图像中的单词进行分类，这种思路只利用了图像中的背景信息和字符的结构信息，忽略了文本序列特有的语义相关性，不能够灵活应对场景中的文本序列识别问题。

- **深层文本循环网络**（Deep-Text Recurrent Network[10]）

2015 年国内的研究学者提出一种深层文本循环网络（Deep-Text Recurrent Network）的方法[10]，将场景文本识别任务当作是序列识别任务去解决，加入循环神经网络学习文本序列的语义信息和依赖关系。该方法使用滑动窗口从图像中按照从左到右的顺序进行截取，将输入图像编码成为有序的图像序列，使用卷积神经网络对图像进行特征提取，抽象出与文本识别任务相关的视觉特征，然后将图像序列对应的特征序列输入到循环神经网络，学习图像序列中包含的语义信息和序列之间的相互依赖关系，将特征序列解码成为图像对应的字符串，完成文本

识别的任务。

这种方法不仅使用了卷积神经网络从图像中提取有用的背景信息和字符的结构信息，还加入了循环神经网络提取文本序列的语义信息，可以对任意长度的序列进行识别。整个网络可以端到端进行训练和优化，用于网络训练的数据只需要对图像中的字符串进行标注，不需要单个字符的标注。但是该方法也存在一些不足之处，输入图像要先使用滑动窗口的方法得到图像序列，再对图像序列的每张图像进行特征建模，过程比较繁琐，网络训练效率低。而且循环神经网络输出的字符序列进行组合的过程中会存在冗余的现象。

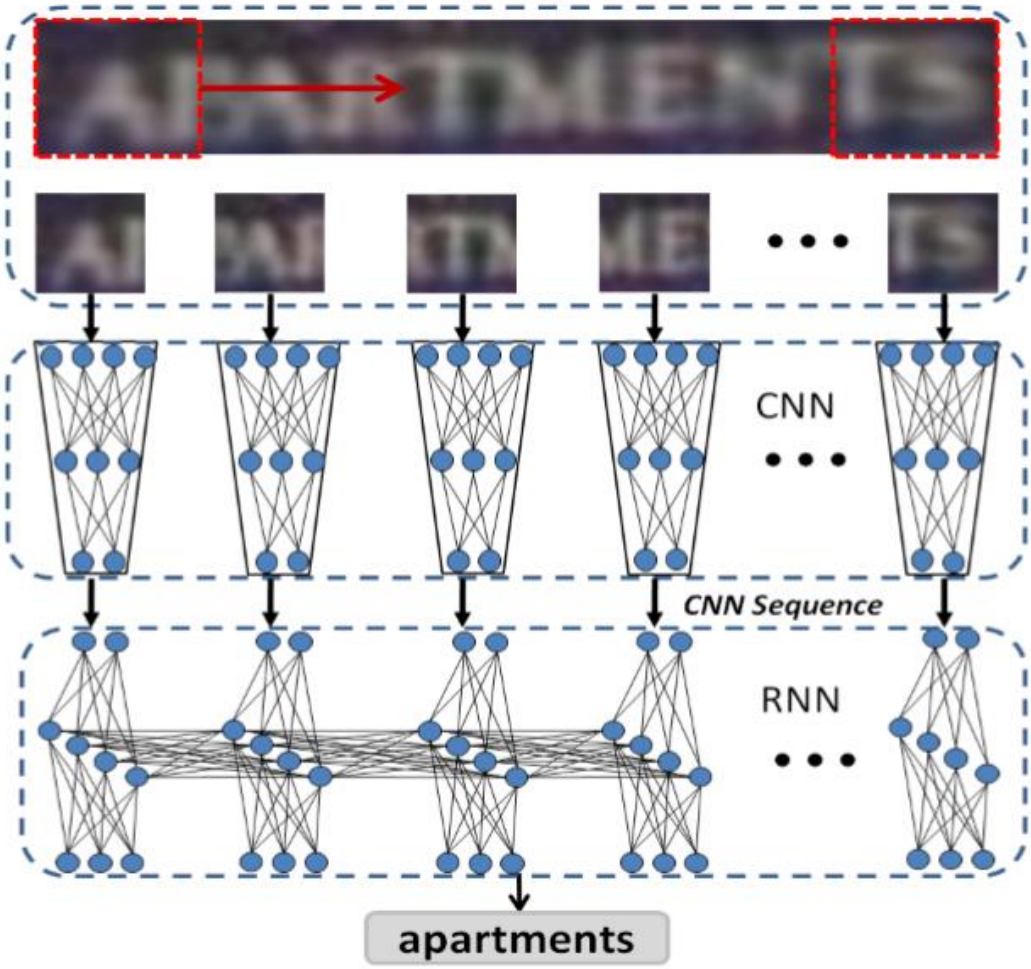

图 2-3 Deep-Text Recurrent Network[10]

- **卷积循环神经网络**

由华中科技大学的学者提出的卷积循环神经网络（Convolutional Recurrent Neural Network，简称为 CRNN[11]）方法将文本识别任务转换为序列识别任务去解决，与上述方法的思路类似，在对图像进行特征建模之后通过循环神经网络学习特征序列中包含的语义信息。这个模型将特征提取、序列建模和转

录层集成到一个统一的框架中，完成场景文本识别任务。

如下图 2-4 所示该模型的输入是要对文本内容进行识别的图像，首先特征提取层选取普通的卷积神经网络对输入图像进行特征提取，分层卷积提取语义特征和信息，最后输出网络从输入图像中提取到的卷积特征图，之后将卷积特征图切分成为特征序列送入序列建模层，这一层选取的是循环神经网络，具体来说，这里选取的是深层双向长短期记忆模型，学习特征序列的上下文依赖关系以及文本序列的语义信息，输入是特征提取层输出的特征序列，输出是对特征序列每一个时间步的分类预测结果，这个结果因为是每一个时间步对应一个预测值，存在冗余以及字符序列不能对齐的问题，所以转录层对序列建模层输出的预测结果进行纠正，去除识别为空白的字符，将重复识别的字符只保留一位，达到去冗余的作用，这样转录层的输出就是最后的文本识别结果。

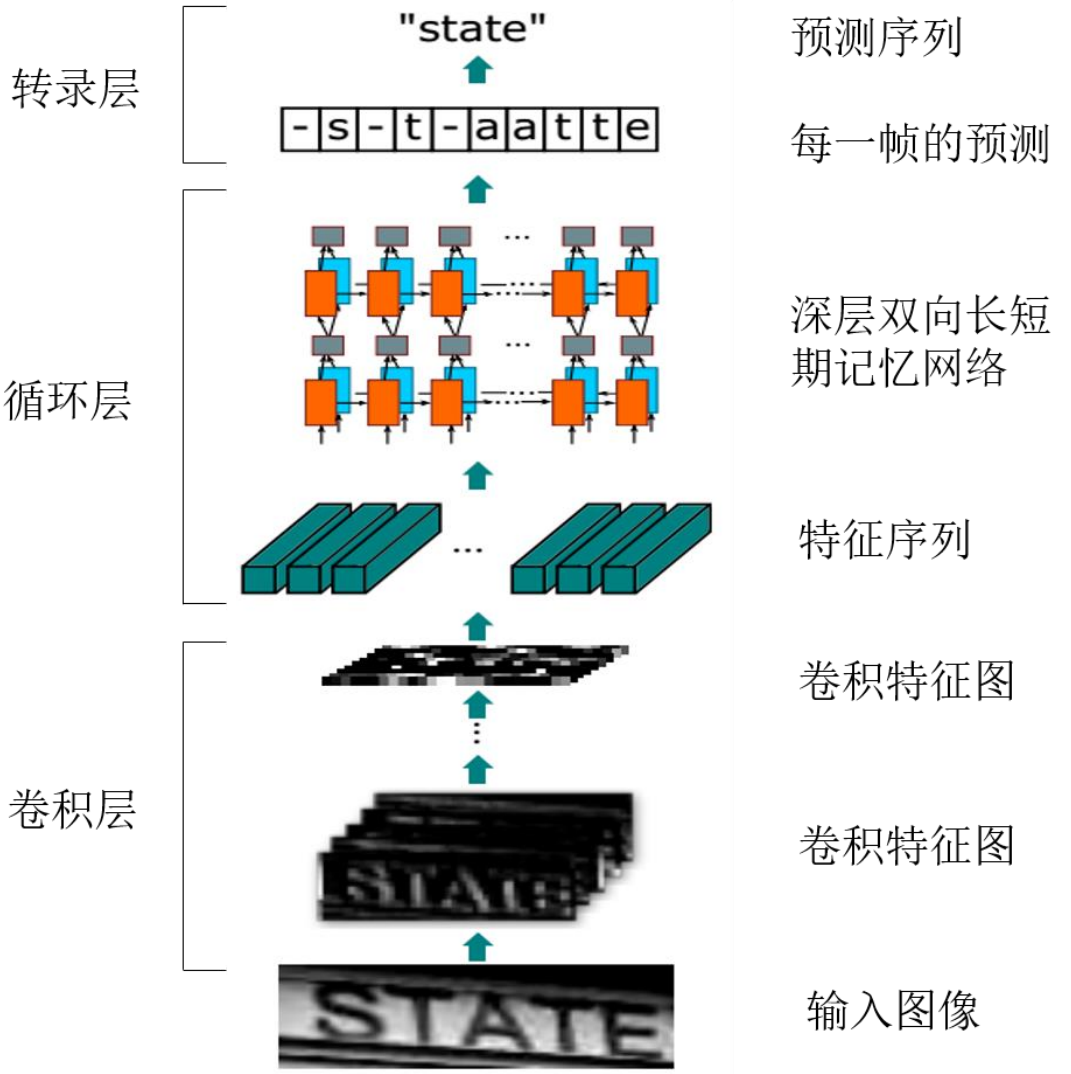

图 2-4 CRNN 架构[11]

与先前用于场景文本识别的系统相比，该算法所提出的架构具有四个独特的

属性：首先，该模型的不同模块可以通过同一个目标函数进行约束训练，不同网络模块在目标函数的约束下同时进行训练和优化，而这之前的大多数算法，在训练过程中采用的是将网络的不同组件进行单独训练。其次，这个模型可以对任意长度的文本序列进行识别，不受文本长度的限制，不需要对字符进行分割或者采取水平缩放、归一化等前处理过程，因此不只局限于固定长度的文本序列的识别任务，适用范围更加广泛。然后，该模型并不局限于有约束的文本识别任务，也就是说对于要识别的图像，不要求提供图像对应的预定义的词典。该模型在无约束和有约束的场景文本识别任务中都取得了优异的表现。最后，该模型在场景识别任务公开数据集上进行验证，实验证明该模型在这些数据集上的识别准确率超越了前人的研究工作，而且该模型不仅性能优异并且更加小巧精炼，更适用于实际应用场景，通用性也更好。

该算法性能超越了之前其他研究学者们提出的场景文本识别算法，并且可以识别长度不一的文本序列。但是随着计算机性能的不断改进提升以及硬件计算能力的快速提升，场景文本识别任务的应用场景日渐丰富，该算法的性能表现受到了挑战，在复杂场景下的文本识别任务中表现一般。其中有两个主要原因：

（1）复杂场景下的图像受不同光照、不同气候等的影响，不同自然环境下呈现出不一样的数据特点。图像中经常有过多的背景干扰和前景遮挡，有较多的噪点，给文本识别带来很大的挑战。图像质量也层次不齐，存在一部分模糊和低分辨率的情况。总之自然场景下的图像分布复杂多样，特征丰富多变，而该方法的特征提取器过于简单，使用简单的卷积网络来对图像进行特征提取，不能有效的提取到图像的重要信息，也不能有效过滤图像中的噪音以及干扰信息。

（2）复杂场景下的文本存在各式各样的形变问题，不同于电子文档类型的规则文本，场景图像中的文本可能会有不同的大小，部分图像中的文本是弯曲排布，或者文本存在倾斜问题，有一定角度的旋转。另外由于拍摄角度的问题，部门图像中的文本会出现仿射变换的情况。这些不规则文本都给识别任务带来挑战，识别难度很大。

• 基于注意力的场景文本识别器

前面讲到的 CRNN 方法[11]可以对任意长度的文本序列进行识别，在卷积神经网络对图像提取特征之后，将特征图转换为特征序列并使用循环神经网络学习序列的语义信息和依赖关系，有很好的性能表现。但是在实际业务场景中，存在很多不规则的文本给识别任务带来很大的挑战。先前已经发表的一篇论文提出了一种基于注意力的场景文本识别器（An Attentional Scene Text Recognizer，简称为 ASTER[12]）的方法，该方法针对不规则文本的识别问题，在识别网络前加入文本矫正网络，额外训练一个矫正网络来对图像中的不规则文本进行转换，通过

对图像进行转换从而将不规则文本矫正成为较为规整、更加易于识别的文本序列，网络架构如下图 2-5 所示。

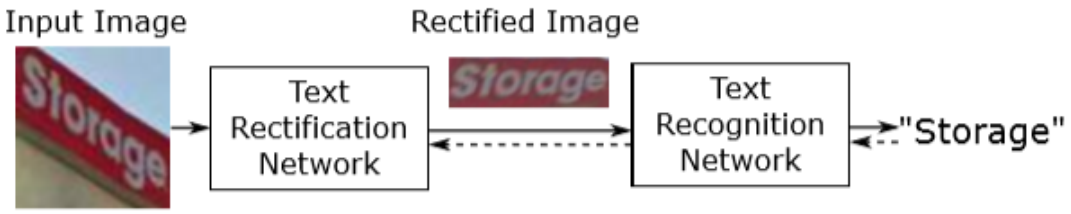

图 2-5 ASTER[12]

具体来说，网络主要包括两个模块，第一个模块使用空间变换网络(Spatial Transformer Network，简称为 STN)[13]作为文本矫正模块，通过训练可以学习到不规则文本的空间变换信息，对这类文本进行矫正，矫正成为较为规则整齐的易于识别的文本。第二个模块对已经矫正好的文本图像进行识别，对图像中的文本序列进行预测。识别网络由编码器和解码器组成，编码器采用卷积神经网络对图像进行特征提取，并用循环神经网络学习特征序列包含的语义信息，得到图像的预测，解码器采用循环神经网络对预测结果进行路径搜索，找到最佳的字符组合路径，得到最后的识别结果。

该方法通过空间变换网络学习不规则文本的变换信息，使用双线性插值的方法对图像的特征矩阵进行转换，对图像中的文本进行矫正，并加入注意力机制引导网络重点关注图像中具有分辨力的区域，在不规则文本的数据集上表现良好，可以提升识别性能，是对不规则文本识别问题的一个探索。但是空间变换网络的训练过程严重依赖人工设定的网络参数初始值，并且网络学习到的是图像的全局变换，而自然场景中的文本变换万千，同一张图像中的文本可能存在多种不同的变换。所以本文设计的场景文本识别算法会针对不规则文本的识别问题，对网络的特征提取器进行改进，通过可变形卷积学习到不规则文本的局部密集形变特征，更好的提升网络性能，之后的章节中会重点介绍。

回顾场景文本识别算法的发展历程，问题解决思路从最初的图像分类变成了序列识别，网络从用卷积神经网络抽取有用特征然后全连接层进行单个字符或者整个单词的分类，到之后的加入循环神经网络学习特征序列的语义信息，网络学习到的特征越来越丰富，网络结构更加多样化。目前的主流场景文本识别算法通常将该问题当作序列建模的任务去完成，使用卷积神经网络和循环神经网络等结构学习任务特征，并且加入注意力机制等先进思想，引导网络更好的对图像进行特征建模。但是自然场景下的文本识别依然存在很多挑战，主要表现在复杂场景带来的干扰和不规则文本的形变问题，本文会在之后的章节中针对这两个问题对算法进行改进，提升模型性能。

2.2 生成对抗网络

目前扫描文档的识别技术已经成熟，但场景文本识别任务的性能表现还不理想.其中的一个原因是自然场景图像和网络图像的背景丰富，会干扰文本内容的识别，给文本识别任务带来很大的挑战。本文提出的算法采用生成对抗网络这一深度学习模型对输入图像进行转换，生成对应的背景干净、更加易于识别的图像，减小图像的识别难度。

2.2.1 标准生成对抗网络

生成对抗网络（Generative Adversarial Network，简称为GAN）[14]的同一个架构下的两个网络组件相互制约，彼此促进，达到优化学习的目的，最终可以模拟数据分布来生成数据。GAN本质上对两个对抗组件没有结构上的限制，不过目前主流的组件都选取了深度学习领域性能优异的网络结构。

标准的生成对抗网络由生成器（Generator）和判别器（Discriminator）组成，其中，生成器G的输入是随机噪声z，通过神经网络的训练生成器G自动生成图像G(z)，并作为判别器D的输入，判别器D由神经网络组成，通过训练对输入图像进行分类判断，判断输入图像是真实图像还是生成器生成的图像。二者对抗学习，不断优化。目标函数如公式2-1所示。其中x是真实样本，z是随机噪声,G和D分别是生成器和判别器。判别器D的优化目标是提升分类鉴别能力，能够正确区分真实样本和生成样本，因此从判别器D的角度出发，希望D(x)尽可能大，D(G(z))尽可能小。生成器G的优化目标是生成尽可能逼真的图像，可以干扰判别器D的判断,骗过判别器D,一次从生成器G的角度出发,希望D(G(z))尽可能大。这两个模型相互对抗进行学习，相互博弈，不断优化和提升。

$$min_G max_D \mathrm{V(D,G)} = \mathrm{E}_{x\sim p_{data}(x)}[logD(x)] + \mathrm{E}_{x\sim p_z(z)}[\log\left(1 - D\left(G(z)\right)\right)]$$

（公式2-1）

GAN 提出了一种很巧妙的生成数据的思想，利用未标注的数据学习数据样本分布，进行无监督学习，并且生成器和判别器的具体形式没有过多的限制，既可以选择卷积神经网络，也可以选择循环神经网络，或者其他网络结构。下图2-6展示了一个生成对抗网络的原理图，其中生成器和判别器都选择了卷积神经网络，设计该网络完成生成手写数字的任务。生成器由卷积神经网络构成，从随机噪声中学习生成样本，通过训练生成器学习从原始输入到生成对应的手写数字的映射关系。真实图像和生成器生成的样本图像一起作为判别器的输入，判别器进行分类鉴别，判断真伪，将真实样本和生成样本区分开来。生

成器和判别器互相促进，迭代优化，模拟手写数字数据的分布来生成数据。

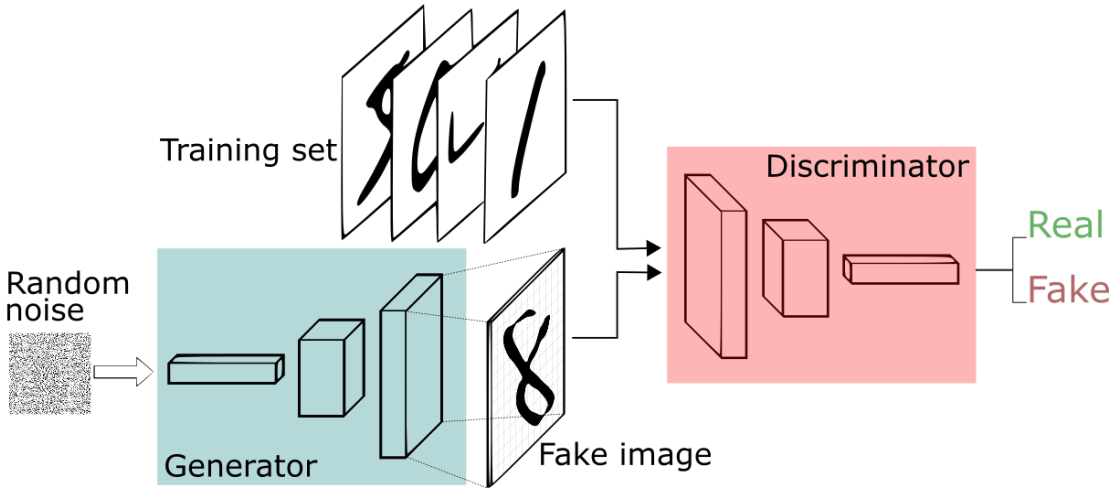

图 2-6 GAN 生成手写数字

GAN 是一种很有创造力的想法，借助这种思想，可以完成很多很有趣的任务。在图像和视频领域，可以使用生成对抗网络完成对图像着色、对图像去噪、进行图像补全、画画创作等任务，还可以改变图像的风格属性。在自然语言处理领域，使用生成对抗网络可以完成生成演讲稿、创作诗歌等任务。对抗博弈的思想不需要标注数据就可以捕捉数据的高阶相关性，所以可以完成很多有意思的创作。但在实际应用中，这种网络也有一些缺点：

（1）可控性差

如上文所说，GAN 对于 G 和 D 的具体网络结构没有要求，其精髓为相互竞争相互进步，对抗学习，不需要标注数据进行无监督学习。在网络开始训练前，没有假设的先验数据分布，而是从随机噪声或者隐变量中直接采样进行学习，从而捕捉数据分布，训练网络尽可能的生成逼真数据。因此，没有预先对数据进行建模的情况下，GAN 的生成方式过于自由，难以有效操控 GAN 的生成方向，尤其是对于大图像，图像中有很多像素，训练过程不稳定很容易出现一些生成效果很差的生成结果。下图 2-7 展示了一些生成效果较差的结果图，可以看到生成图像可能会出现模糊、低分辨率的现象或者只学习到了一部分数据特征，和原始图像相比还是有很大的差距。

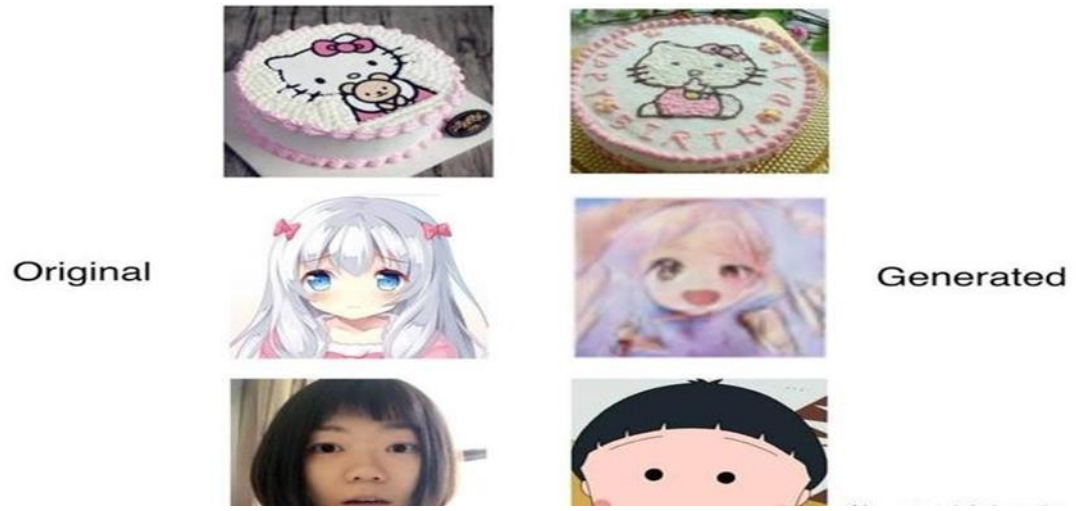

图 2-7 可控性差的现象

（2）模型崩塌问题

现实世界的数据分布是非常复杂的，具有很丰富的多样性。而生成对抗网络的生成器在最初的学习中质量非常差，此时判别器很容易就可以分辨真假，判别器从中能够学到的经验太少，而导致生成器的质量也提升不了，模型开始退化。这样就会使整个迭代训练的过程停滞不前，网络反复生成一些很相似的样本图像，这些生成样本只覆盖到真实数据的一小部分模式，并没有完全学习到数据中的特点，训练无法继续，所以这种现象叫做模型崩塌。

2.2.2 条件生成对抗网络

标准的生成对抗网络[14]从随机噪声开始学习，将数据由随机噪声变成符号目标数据分布的图像，学习数据映射的关系，完成无监督的生成过程。但是直接从随机噪声中采样建模去学习数据转换的过程，在实际训练过程中存在控制力差、生成过程过于自由，很难控制生成器的生成方向的问题。因此条件生成对抗网络[15]在标准生成对抗网络的思想中额外加入条件信息 y，作为生成器和判别器的输入来更好的引导网络的生成过程，约束生成数据的分布趋势，大大加快网络的收敛速度，解决生成网络过于自由的问题。

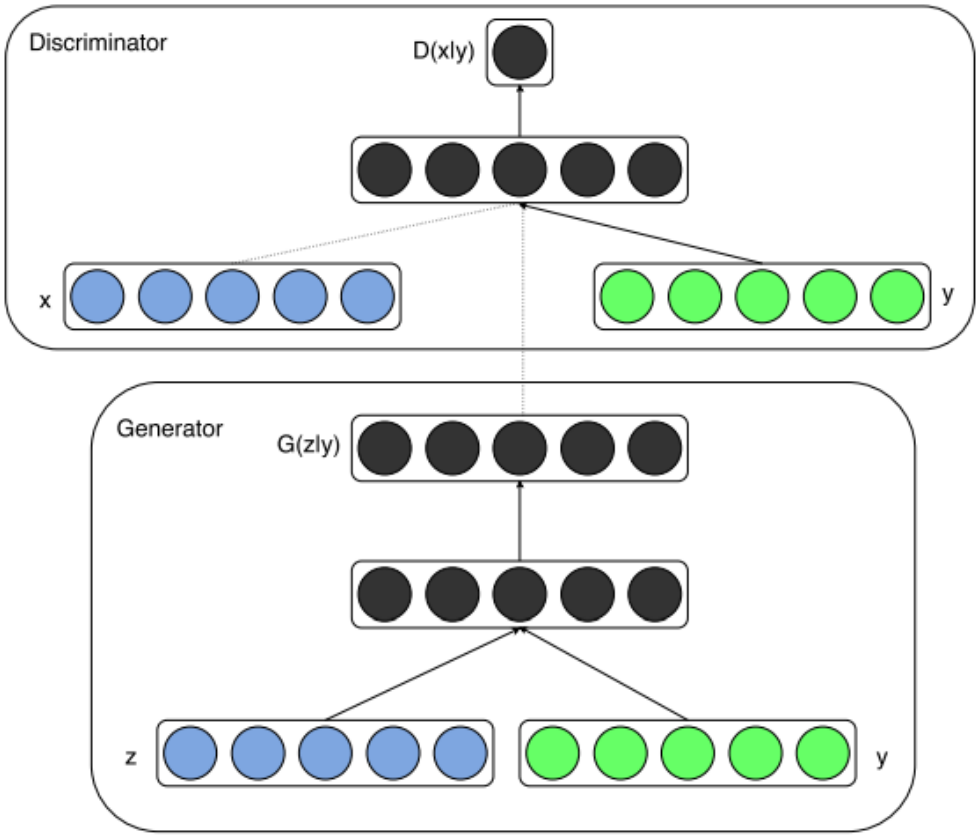

图 2-8 条件生成对抗网络[15]

条件生成对抗网络的目标函数如公式 2-2 所示。其中 x 表示真实样本，z 表示数据的初始状态，y 表示附加的条件信息，同时作为生成器和判别器的输入。判别器 D 的优化目标是能够将真实样本与生成器生成的样本区分开来，学习这两种图像的差异性，提升分类鉴别能力。生成器 G 的优化目标是在条件信息的辅助下，引导网络将初始数据转换为更加符合目标数据分布的图像。

$$min_G max_D \mathrm{V(D,G)} = \mathrm{E}_{x\sim p_{data}(x)}[logD(x|y)] + \mathrm{E}_{x\sim p_z(z)}[\log\left(1 - D\left(G(z|y)\right)\right)]$$

（公式 2-2）

下图 2-9 展示了条件生成对抗网络的训练过程，图中黑色的曲线表示真实样本。也就是训练时生成器的额外的条件信息 y 的概率分布，绿色的曲线表示生成器生成的样本的概率分布，蓝色曲线代表判别器对于输入图像的类别判断，判别器的输出值越大，表示输入图像是真实样本的概率越大。最下方还有两条平行线，其中 z 表示生成器输入的原始场景文本图像，x 表示真实图像。训练一开始，黑色曲线和绿色曲线的分布差异很大，生成器的生成性能较差，不能很好的拟合真实数据的分布，此时两种图像的差异很大，判别模型比较容易对图像是真实样本

还是生成样本进行判断。随着训练的不断推进，网络在目标函数的引导下不断迭代优化，生成器生成的样本逐渐接近真实样本的分布，生成器在与判别器的对抗博弈中学习到了真实样本的分布特点，掌握了将输入图像映射成为目标数据分布的能力，生成样本与真实样本的差异越来越小，图中黑色曲线与绿色曲线逐渐靠近。最后，生成器和判别器在对抗训练中不断提升自己的能力，生成样本的分布不断拟合真实样本分布，黑色曲线和绿色曲线重合，生成样本和真实样本的差异性太小，判别器不能够将这两种图像进行区分，模型达到了最优的状态，生成器可以完成对输入图像进行数据转换。

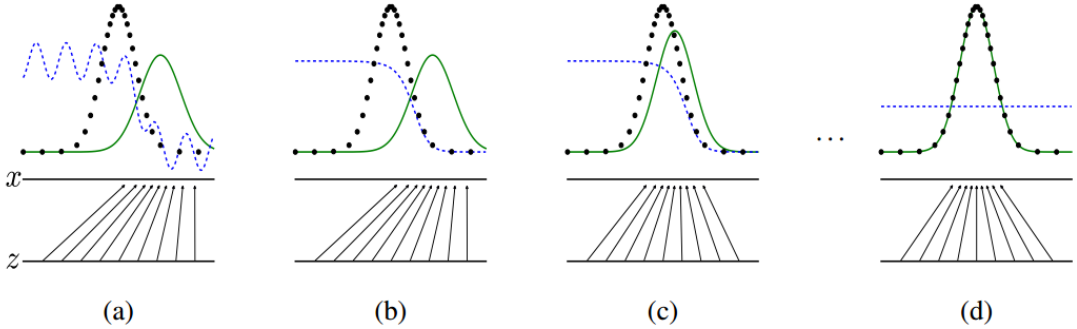

图 2-9 条件生成对抗网络训练过程

2.3 先进的卷积神经网络结构

卷积神经网络（Convolutional Neural Networks）是神经网络中的一个分支，通过模拟人类视觉系统的工作机制而搭建的层次模型，属于前馈神经网络。主要用于处理图像、视频等相关问题，有效提取数据中包含的视觉信息和语义特征。本文设计的文本识别算法借鉴了先进的卷积神经网络结构，可以更好的提取图像特征，扩大神经元的感受野，并且加入可变形卷积提升网络的空间变换建模能力。

• 压缩激励网络

典型的卷积层[16]通过矩形卷积核按照自上而下、自左向右的顺序与图像中相应位置的元素进行相乘求和，将图像映射成为对应的特征图来进行特征提取，这种方式在许多计算机视觉任务中证明是非常有效的。但是这种方式中不同通道得到的特征图有相同的权重，不加区分一起送到下一层，不能有效区分不同通道的重要性。

压缩激励网络（Squeeze-and-Excitation Networks，简称为 SENet[17]）使用了经过特殊设计的卷积模块来代替典型的卷积层。该模块的核心思想是为每一层卷积层不同通道分配相应的权重，通过训练去学习卷积层不同通道的重要性，增大为识别任务贡献程度大的通道的权重，减小对识别任务贡献程度小的通道的权

重。该模块主要通过压缩（Squeeze）和激励（Excitation）两个操作来学习不同通道的重要性，所以叫做压缩激励模块（Squeeze-and-Excitation block，简称为 SE 模块）。从图中可以看到该模块将卷积之后的得到的特征图进行了权重分配，不同通道的颜色不同，代表分配的权重大小不同，也意味着不同通道的有效性有差异，对任务的贡献程度也不同。接下来进行详细介绍。

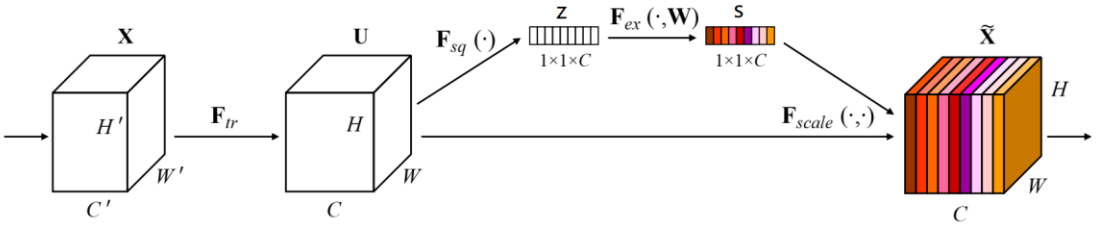

图 2-10 Squeeze-and-Excitation block[17]

图中最左边的 X 表示这一层的输入，用W'、H'、C'分别表示输入图像的图像宽度、图像高度和图像通道数，对 X 进行卷积操作F_{tr}后得到中间输出 U，用W、H、C表示中间输出 U 的图像宽度、图像高度和图像通道数。

$$F_{tr}:\mathrm{X}\to\mathrm{U},X\in R^{W'\times H'\times C'}\ \ U\in R^{W\times H\times C} \quad （公式 2-3）$$

F_{tr}是标准的卷积操作，v_c表示第 c 个卷积核，X 表示这一层的输入，总共有C'通道，不同通道的输入与对应的卷积核按照自上而下、从左到右的顺序依次进行对应位置的相乘求和操作，得到输出U_c，也就是 U 中第 C 个通道的特征图，将这些通道的输出组合起来组成中间输出 U。

$$\mathrm{U}_c=\ v_c*X=\sum_{s=1}^{C'}v_C^s*x^s \quad （公式 2-4）$$

对 U 的每个通道特征图进行全局平均池化，将每个通道的二维特征图转换为一个实数，表示该特征通道上响应的全局分布。将$W\times H\times C$的中间输出 U 转换为$1\times1\times C$的输出，对 U 的全局特征进行聚合统计，学习不同通道的数值分布情况得到输出 z。这一步叫做 Squeeze 操作。

$$\mathrm{z}_c=\ F_{sq}(\mathrm{u}_c)=\frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}u_c(i,j) \quad （公式 2-5）$$

接下来对 z 进行 Excitation 操作，学习不同通道的权重分布。首先 z 经过一层与所有通道相连的全连接层，乘以参数W_1，经过修正线性单元（Rectified Linear Unit，简称为 ReLu）学习非线性映射，得到中间结果$\delta(W_1z)$。再将中间结果经过一层全连接层，乘以参数W_2，经过 S 型激活函数 Sigmoid 层进行非线性学习，输出的结果$\sigma(W_2\delta(W_1z))$称为 s，s 的尺寸同样也是$1\times1\times C$，与 z 保持一致。这一操作相当于输入经过两个全连接层和两次非线性激活，对不同通道的特征图

的权重分布进行学习，刻画不同通道的重要性。

$$s = F_{ex}(\mathrm{z}, \mathrm{W}) = \sigma\big(g(z, W)\big) = \sigma(W_2\delta(W_1 z)) \quad （公式 2-6）$$

s 已经学习到了不同通道的权重分布，接下来将 s 学习到的权重数值赋予 U，将 U 的每个通道的二维特征图与对应的权重数值相乘，完成在通道维度上的特征重标定。将不同通道的结果拼接在一起就是最后的输出。

$$\tilde{x}_c = F_{scale}(u_c, s_c) = u_c \cdot s_c \quad （公式 2-7）$$

- **残差神经网络**

计算机视觉领域的多数任务会使用卷积神经网络进行特征提取，并且网络层数越深，模型参数就越多，模型具有更优异的表达能力。但是简单地堆叠网络会出现网络退化现象，信息在不同层之间进行传递会出现信息缺失的现象，网络训练速度也很慢。2015 年由微软研究院的 Kaiming He 等人提出的残差神经网络（Residual Neural Network，简称为 ResNet[18]），将计算机视觉常用的残差表示的思想应用在卷积神经网络的构建上，增加直连通道，保留原始输入信息，直接将输入信息通过旁路连接传输到这一层网络的输出，并且避免引入额外的参数和计算量。这样设计的网络层学习输入输出之间的残差，提升信息的完整性，简化学习目标和学习难度，避免卷积神经网络由于网络过深可能会出现的网络退化问题和梯度消失的问题。这样的模型设计，使得网络收敛速度大大提升，网络性能也更加优异,这样设计的网络叫做深度残差网络。

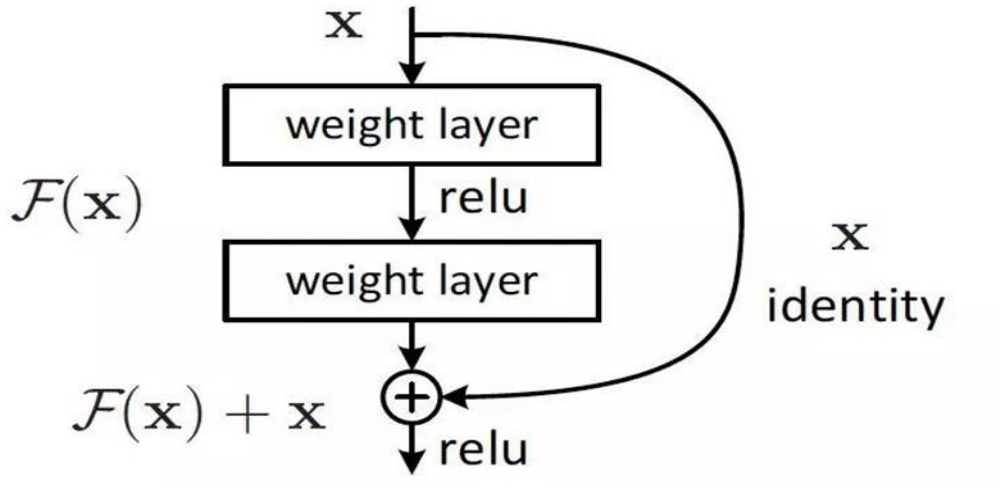

图 2-11 残差原理图[18]

- **可变形卷积**

卷积神经网络通过权值共享的卷积核矩阵对图像这类非结构化数据进行特征提取，典型的卷积网络使用的卷积核是固定的矩形结构，用矩阵的方式对图像

进行采样，这种采样方式对图像中的文本的几何建模能力有限。可变形卷积[19]在典型的卷积操作的基础上，增加了可以学习卷积核的位置变换的参数，对空间采样方式进行增强，自适应地根据图像中目标地位置分布调整空间采样的位置，使得网络对图像的平移、缩放、旋转等变换保持不变性，并且该结构不需要额外的监督信息，可以通过反向传播算法进行端到端的训练。

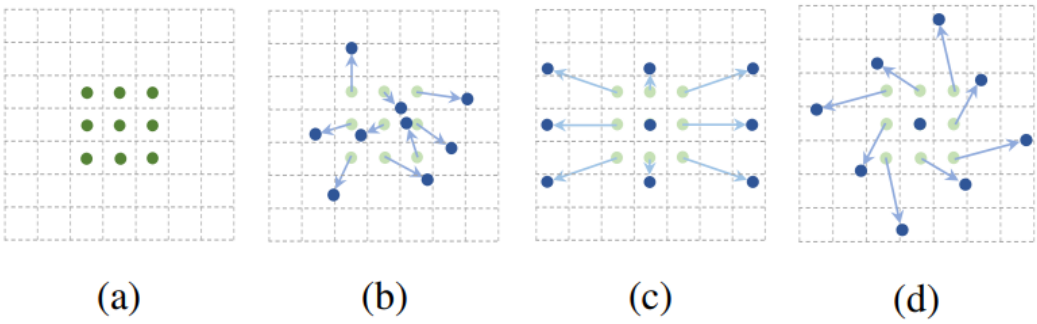

图 2-12 可变形卷积核[19]

上图 2-12 展示了可变形卷积的空间采样方式。(a)表示典型的矩形卷积核，图中的卷积核大小是 3×3，可以看到典型卷积操作在图像空间中进行采样的位置是固定的，(b)(c)(d)表示经过特殊设计的可变形卷积的卷积核，与典型的卷积核不同，可变形卷积在采样过程中会对每一个对应位置学习一个表示偏移量的参数，在图中用蓝色的箭头来表示这个偏移量，通过这种方式由传统的固定采样学习到了新的卷积采样点。(c)(d)表示(b)的特殊情况，卷积核的不同位置的偏移量进行组合可以表示各种各样的形变情况，(c)展示的是用可变形卷积学习图像中物体的尺度变换，(d)展示了可变形卷积学习图像中物体的旋转变换。

可变形卷积通过对卷积采样的方式进行调整，在卷积核中加入可以学习采样位置偏移情况的方向参数来学习图像中密集的空间变换，卷积核在训练过程中可以自适应调整采样点的位置，扩大采样范围，通过视觉任务的监督信息自适应图像中不同目标的几何空间变换，更好的学习到对目标任务有意义的特征和信息，对于复杂场景下的视觉任务非常有效。下图 2-13 展示了可变形卷积的网络结构，图中以尺寸为 3×3 的卷积核为例，可变形卷积层的输入是上一层卷积学习得到的特征图，先对特征图进行卷积操作，得到通道数为输入图的双倍的特征图，分别表示空间位置 x, y 在不同方向上的偏移情况，根据学习到的不同像素的偏移方向和偏移量将特征图中的像素重新进行整合，自适应学习图像中目标的位置分布和空间几何变换。训练过程中，网络从特征图中学习到的偏移情况通过双线性插值来计算，使用反向传播算法对网络参数进行端到端训练，引导网络从训练数据中学习图像中不同目标文本的分布特征，自适应图像中不同目标的空间变换，显

著提升网络的识别性能。

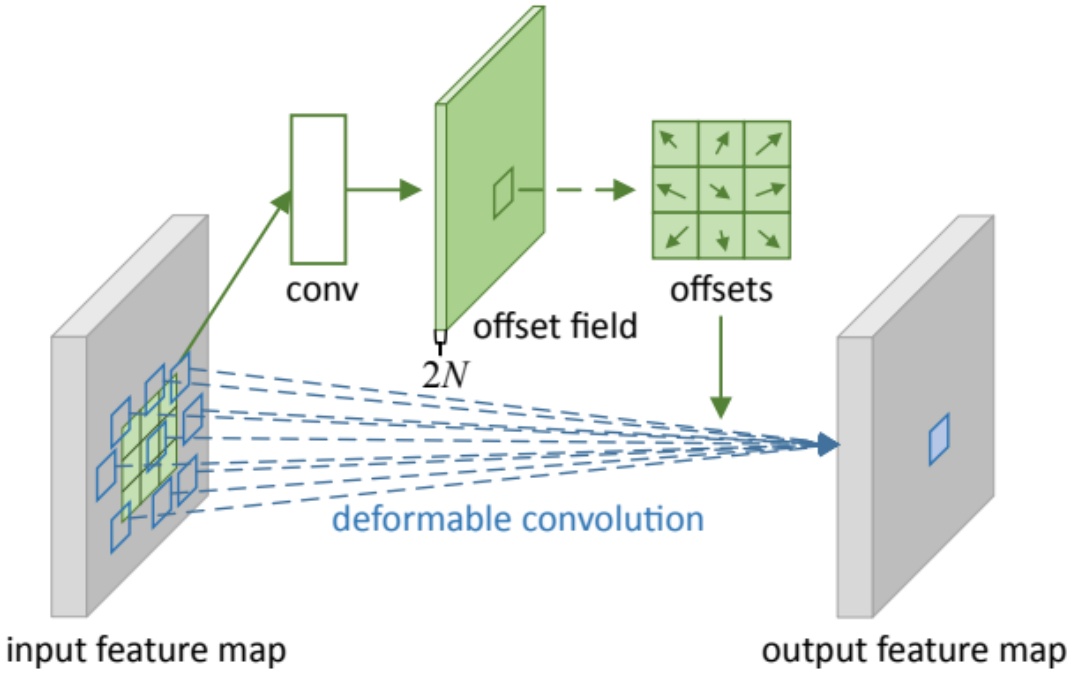

图 2-13 可变形卷积网络[19]

2.4 长短期记忆网络

循环神经网络（Recurrent Neural Network，简称为 RNN）[20]可以处理序列化数据，在隐藏层的同一层的神经元之间存在连接，相当于网络学习到了一定的记忆能力，循环神经网络可以理解为将同一个网络进行多次复制，理论上可以处理任意长度的序列数据。但是循环神经网络如果处理长序列的依赖关系的话，当前时刻隐藏层的状态计算需要将之前多个时刻的状态计算挂钩，计算量会呈现指数式增长，导致网络训练速度大大降低，训练进度缓慢。而且对于长序数据，目标任务的完成当前时刻的隐藏层的状态想要依赖于相距较远的时刻的隐藏层输出的话，数据传输中经过多次参数相乘，梯度经过多阶段的传输计算，非常容易出现梯度消失的问题，少数情况下也会梯度爆炸的现象，这也是循环神经网络很难解决的长期依赖问题，长序列数据的处理对循环神经网络是个很大的挑战。

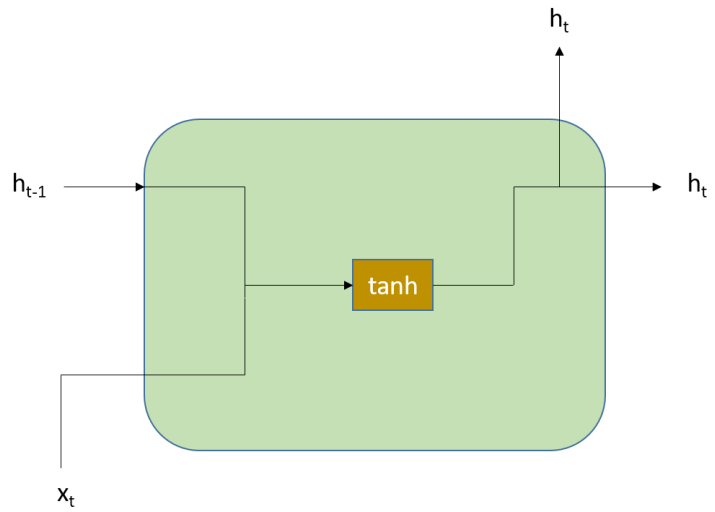

图 2-14 RNN

长短期记忆网络（Long Short-Term Memory，简称为 LSTM）[21]是循环神经网络的一种特殊类型，对网络进行修改设计，可以避免出现上述提到的典型循环神经网络处理长序数据时面临的长期依赖问题，训练网络学习到序列数据中的长期依赖信息。典型的循环神经网络的重复网络模块的形式是一个简单的 tanh 非线性映射层。长短期记忆网络对重复的网络模块进行重新设计，变成增强版的小模块，通过设计“门”的结构来平衡信息的传输状态，让信息有选择的按权重比例进行传输。其中丢弃门来决定是否将传输的某些信息进行丢弃，只保留有意义的信息，输入门对传输进的信息进行状态更新，保留重要信息，最后输出门会对更新后的信息就行筛选，将有价值的信息进行输出。

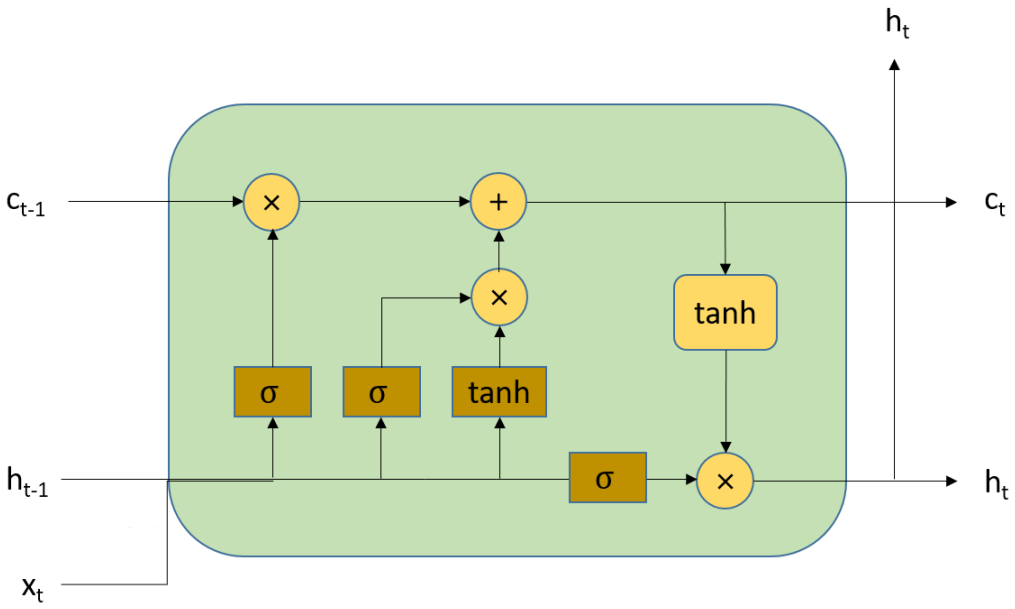

图 2-15 LSTM

2.5 本章小结

本章回顾了场景文本识别的算法的发展历程，展示了神经网络应用在场景文本识别任务之后，网络由浅到深，任务由字符分类到任意长度的文本序列识别到不规则文本的识别任务的递进发展过程，介绍了代表性算法并对这些算法进行了分析。还介绍了本文设计的场景文本识别算法会借鉴的先进神经网络，包括条件生成对抗网络，深度残差网络，可变形卷积和长短期记忆网络等。

目前的主流算法在自然场景的文本识别中还存在一些不足，在接下来的章节中，我们会去对算法模型进行创新和设计，解决本文的主要问题即自然场景下的文本识别。

第三章 场景文本识别算法设计

场景文本识别要面临复杂的自然场景，应对不同光线、不同气候、不同拍摄角度带来的干扰，图像经常会有模糊、分辨率低的问题[28]，给识别任务带来很大的挑战，而且自然场景中的文本有一部分比例是存在弯曲、倾斜、仿射等变换，这些不规则文本的识别难度更大，对识别算法的要求非常高。目前主流的文本识别算法在复杂场景和不规则文本的情况下表现一般，本章在对主流场景文本识别算法进行分析之后，针对复杂场景和不规则文本的情况，在识别模型前加入数据预处理模块，用生成对抗网络[14]将复杂场景的文本图像转换为背景干净、易于识别的图像，并且在识别模块的特征提取器中加入可变形卷积[19]，应对不规则文本难以识别的问题，可变形卷积会对图像中复杂多样的形变情况进行自适应，更好的提取图像特征并加以识别。

3.1 场景文本识别算法模型设计

3.1.1 算法模型设计

本文提出的基于生成对抗网络改进的场景文本识别算法由两个模块组成：

第一个模块对输入图像进行数据预处理，去除复杂背景的干扰，采用条件生成对抗网络[15]来实现。该网络由生成器和判别器组成，生成器将输入图像转换为背景较为干净的尺寸大小相同的图像。判别器对输入图像的真假进行判断。生成器和判别器的输入中都包括额外的条件信息，也就是输入的场景图像对应的背景干净的图像。生成器和判别器二者的优化目标不同，生成器倾向于生成符合条件数据分布，并且可以欺骗判别器的图像，而判别器倾向于对每一张输入图像都做出正确的判断。生成器和判别器进行对抗训练，迭代优化，不断进行改进和提升，在目标函数的约束下最终达到一种平衡，在这个过程中，生成器和判别器的性能都得到了提升，经过训练后的生成器可以将复杂场景的文本图像转换为背景较为干净的图像，更加易于识别。

第二个模块要对第一个模块输出的图像进行文本识别，这一模块以上文提到的主流的 CTC 模型为基础[11]，网络模型由特征提取、序列建模和转录层组成。本文对主流模型做了如下改进：首先对特征提取器进行改进，将普通的七层卷积神经网络替换成可以对通道进行权重调整的 SENet[17]网络，可以更好的提取图像特征，其次针对弯曲、缩放、旋转、仿射变换等形变文本的问题，在特征提取器内

加入可变形卷积，可以自动学习图像中的空间几何变换，自适应图像中各式各样的形变问题，提升识别精度。

综上所述，本文在对主流场景文本识别算法进行分析之后，总结了主流算法的不足之处，即主流场景文本识别算法不能很好的应对复杂场景带来的干扰，在不规则文本的识别过程中难以应对图像中复杂多样的形变问题。

3.1.2 算法创新点

为此本文提出的算法模型做了如下改进：

（1）加入数据预处理模块对复杂场景下的文本图像进行处理，将原始输入图像转换为背景干净、易于识别的图像。预处理模块选取条件生成对抗网络来实现，生成器学习从原始输入图像到原图对应的背景较为干净、尺寸保持一致的图像的映射关系，判别器去学习将真实图像判定为真，将生成器生成的图像判定为假的图像分类鉴别能力。通过设计恰当的目标函数，二者进行对抗训练，逐渐提升自身能力，生成出逼真清晰的图像。

（2）文本识别模块以主流算法为基础，对主流场景文本识别算法做出改进。为了更好的从图像中提取到重要信息和有意义的特征，将特征提取器的网络替换成性能更加强大的 SENet[17]，可以主动学习卷积不同通道的权重分布，在通道层次使用注意力机制，更好的提取图像特征。为了应对复杂场景下，图像中的文本存在各式各样的形变情况的问题，加入可变形卷积[19]，改变卷积核的采样方式，主动学习图像中的空间几何变换，提升网络识别性能。

3.2 生成式对抗网络设计细节

第一个算法模块借助生成对抗网络完成对输入图像进行预处理转换的任务，通过训练，网络学习从输入图像到对应的背景干净的图像的映射关系，并且学习用于训练和约束这种映射关系的损失函数。训练生成器保证能够由输入图像生成去除复杂背景干扰的图像，将图像转换为背景更加干净且易于识别的图像，训练判别器能够判别输入的图像是生成的图像还是真实图像，二者对抗博弈，相互学习，最终达到一个平衡。

数据预处理模块采用条件生成对抗网络，生成模型的输入不仅包括原始输入图像，还包括与原始图像相同尺寸、背景干净的文本图像，在网络中加入了额外的条件约束[15]，这二者联合起来作为隐层表征。条件生成对抗网络中加入额外的条件信息在数据维度上作为潜在的约束范围，可以更好的引导生成器的生成过程。

同样，判别模型的输入既包括生成器的生成样本，也包括条件信息，判别器学习两种图像的差异性，将这两种图像类型区分开来。

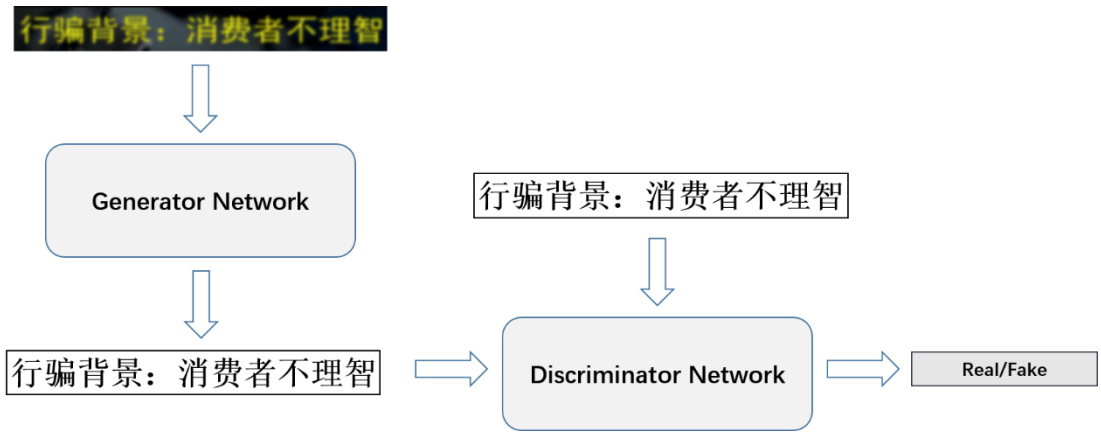

图 3-1 预处理模块

3.2.1 生成模型

生成模型使用采用编码器-解码器的结构，编码器的输入是原始图像，这一部分主要由卷积神经网络组成，浅层通过卷积操作对图像进行映射，提取输入图像的视觉信息和特征，之后加入一层注意力模块，采用注意力机制更好的学习图像不同区域的权重分布，深层采用残差模块，加入残差连接[18]，在加深网络层数的同时，提高训练速度，更好的提取到图像中有意义的特征。随着网络层数的加深，卷积通道数逐渐增加，同时通过最大池化层逐渐减少特征图的尺寸，对图像进行下采样，尽可能地保留空间位置信息。解码器的输入是卷积层输出得到的特征图，解码器模块对输入图进行上采样，逐层恢复特征图的尺寸，与原始输入图像的分辨率保持一致，并减少卷积通道数，最后一层是卷积层。这样生成器通过训练和学习，去模拟数据的分布特征，学习输入图像的转换，将输入图像映射成为背景干净的图像。

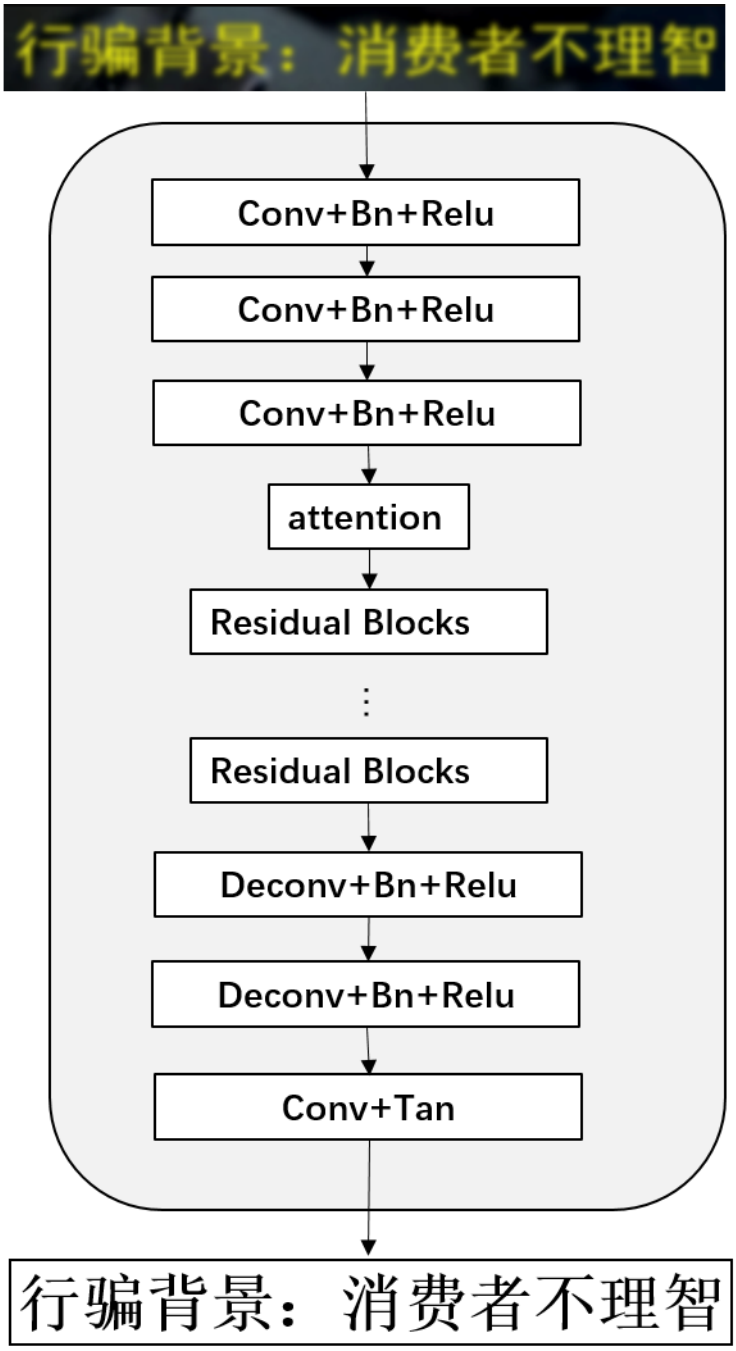

图 3-2 生成模型

3.2.2 判别模型

判别模型使用全卷积神经网络对生成器的生成图像和条件信息进行分类，网络输入既包括生成器生成的背景较为干净的图像，也包括真实的背景干净的图像。判别器的分类能力一定程度上反映了生成器的生成水平，如果判别器轻易将图像区分开，那么说明生成器生成的图像跟真实图像的差距很大，没有很好的学习到

数据分布的特征。当生成器生成的图像不能欺骗判别器的时候，那么生成器生成的图像与真实图像的差异性就很小，可以很好的模拟数据分布，生成高水平高质量的输出图像。

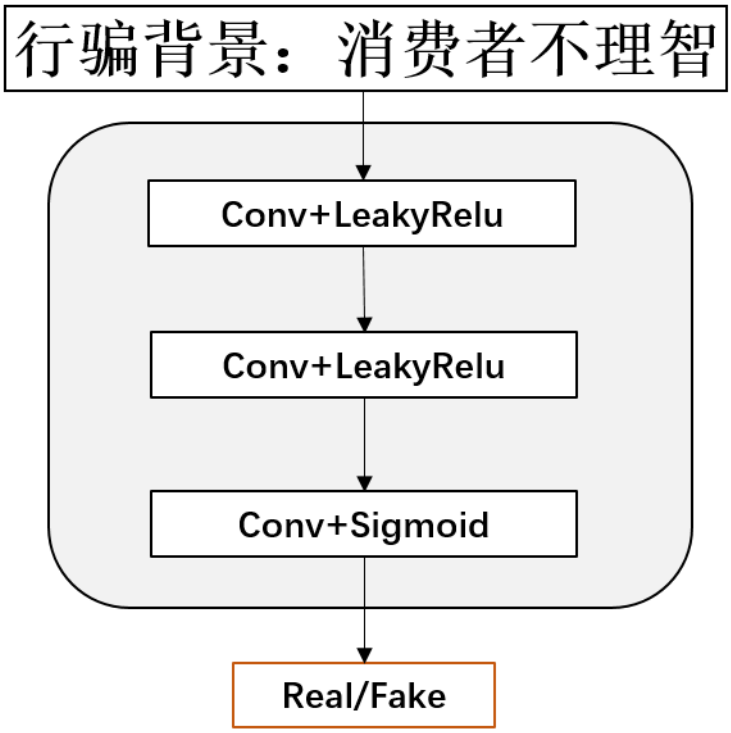

图 3-3 判别模型

生成模型和判别模型一起组成条件生成对抗网络[15]完成数据预处理的任务。不同于原始的生成对抗网络，条件生成对抗网络在生成模型和判别模型的输入中都加入了额外的条件信息 x，这里的条件信息 x 指的是与原始图像相对应的背景干净并且尺寸大小一致的真实图像。条件信息作为监督，约束和引导生成器的生成过程，增强条件对抗网络训练的控制力，这样生成过程就变成了带条件概率的极小极大的二元博弈问题，

3.2.1 目标函数

目标函数如公式 3-1 所示。其中 G 代表生成模型，D 表示判别模型，z 表示生成器的输入，x 表示条件信息，y 表示真实图像。

$$\mathcal{L}_{cGAN}(G,D)=E_{x,y}[\log D(x,y)]+E_{x,z}[\log(1-D(x,G(x,z))] \quad \text{（公式 3-1）}$$

原始的生成对抗网络并没有额外的条件约束信息 x，生成器输入随机噪声 z 生成相应的样本 G(z)，判别器对真实图像 y 和生成器生成的样本 G(z)进行分类判别，目标函数如公式 3-2 所示。

$$\mathcal{L}_{GAN}(G,D)=E_y[\log D(y)]+E_z[\log(1-D(G(z))] \quad （公式 3-2）$$

为了更好的约束生成器的生成效果，额外加入 L1 损失函数，约束生成器生成样本和真实样本数据的差异，目标函数如公式 3-3 所示。

$$\mathcal{L}_{L1}(G)=E_{x,y,z}[||y-G(x,z)||_1] \quad （公式 3-3）$$

将条件生成对抗网络的目标函数和 L1 损失函数按比例相加，就得到了数据预处理模块的目标函数，如公式 3-4 所示。

$$G^* = argmin_G max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G) \quad （公式 3-4）$$

3.3 文本识别算法设计细节

本节介绍第二个算法模块-文本识别模块，这一模块会对第一个数据预处理模块的生成器输出的图像中的文本序列进行识别。文本识别算法选取主流的基于 CTC 的方法[11]，将识别图像中字符串的任务当成序列任务来解决。网络模块由三部分组成：特征提取层、序列建模层、转录层。特征提取层选用卷积神经网络从输入图像中提取关键特征，序列建模层选用循环神经网络对文本序列进行建模，转录层采用基于 CTC[7]的损失函数来训练整个网络，对输出结果进行矫正，解决输入序列和输出序列不能对齐的问题。

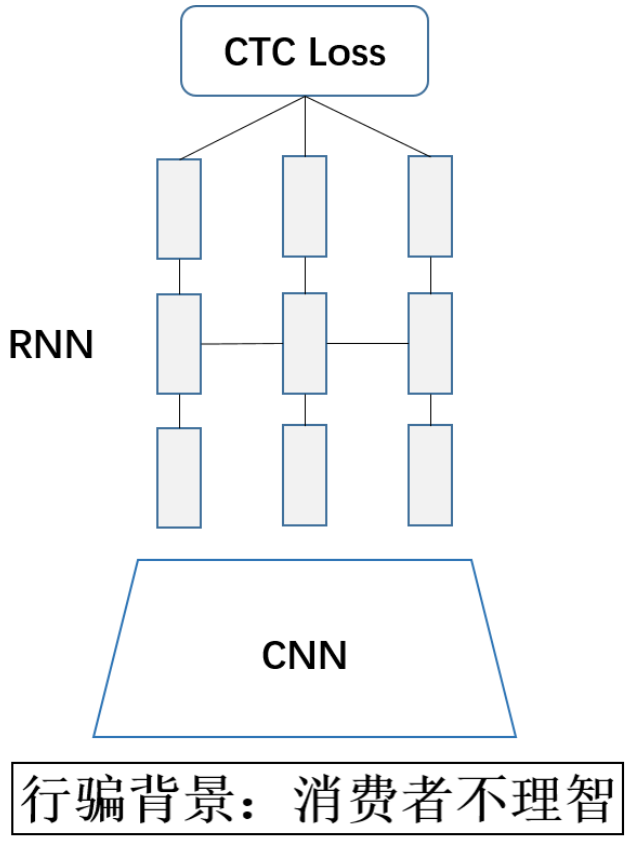

图 3-4 文本识别算法

3.3.1 特征提取层

第一层为特征提取层，从输入图像中提取关键特征，通过卷积、池化等操作将输入图像映射成为特征图。这一层将前文介绍过的 SENet[17]和 ResNet[18]网络相结合，在卷积神经网络中加入残差模块，每一层学习上层输入输出的残差，解决梯度消失的问题。在此基础上将典型卷积层替换成 SE 模块，卷积层在对图像进行特征提取之后可以学习到不同通道的权重分布，相当于引用了注意力机制。将学习输入输出之间的残差的旁路连接和学习卷积通道的权重分布模块组合在一起，对典型的卷积层进行改进，增强网络对输入数据的特征表达和特征映射能力，大大加快网络的收敛速度，提升网络对数据特征的抽取能力，提升特征提取器的性能表现，以及整个识别网络的识别效果。

下图 3-5 展示了将这二者组合之后的网络层，组合之后的模块称为 SE-ResNet 模块[17]。文本识别算法的特征提取器就选用这个模块进行堆叠，从文本图像中进行特征提取，抽取有价值的图像信息，搜索对文本识别任务有帮助的相关特征，学习场景文本数据的数据特点和分布，对数据中的视觉特征进行表达和建模。特征提取器输入图像，输出输入图像对应的特征图，展示了数据中的视觉特征和关键信息。

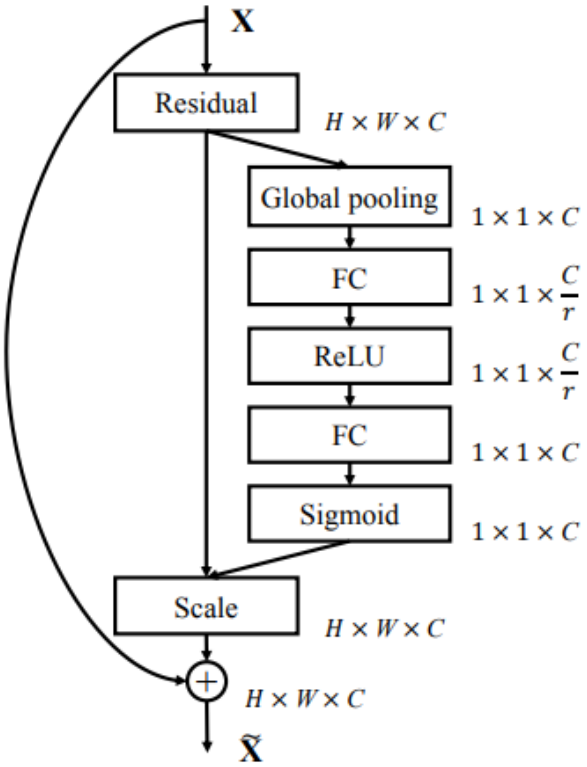

图 3-5 SE-ResNet Module[17]

场景文本识别的一大难点是自然场景下有相当一部分比例的文本序列属于不规则文本，文本序列中的字符大小不一，呈现弧形弯曲排列分布，或者由于拍摄角度的不同，存在一定的仿射变换问题。这些不规则文本序列存在一定的空间几何变换和形变问题，给文本识别任务带来很大的挑战。本文设计的文本识别算法的在浅层网络中加入前文介绍过的可变形卷积[19]，在卷积核参数中额外加入自适应图像中不同字符变换的偏移参数，改变典型卷积层使用矩形框进行采样的方式，扩大卷积核的采样范围。

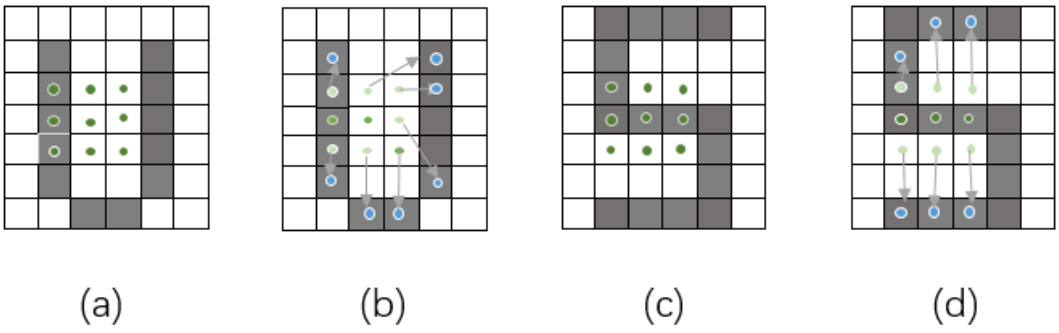

图 3-6 可变形卷积核

上图 3-6 展示了可变形卷积在文本识别任务中的空间采样方式。(a)(b)展示的是一张包含大写字母 U 的字符图像，(a)展示了普通卷积核在这张图像上的采样方式，采用标准的 3×3 大小的卷积核按照从左到右、自上而下的顺序对图像进行特征提取。(b)展示了可变形卷积的采样方式，额外加入参数 w 学习卷积核中不同位置的偏移情况，经过训练后，卷积核自适应图像中的字符分布区域，改变卷积核的采样方式，对图像中包含文本字符的区域进行采样，更好的适应字符的分布情况。(c)(d)同样是一张包含数字 5 的图像，(c)展示了标准卷积核的采样方式，(d)展示了可变形卷积的自适应采样方法，通过调整卷积核对应位置的偏移参数，对图像中的字符区域进行采样，扩大了卷积核在图像中的采样范围，更好的进行特征建模。

可变形卷积在卷积核参数中额外加入偏移参数，自适应改变卷积层的采样位置，可以更好的应对不规则文本的识别问题。偏移参数的加入使网络可以灵活适应文本序列的不同分布，从图像中学习到局部密集的形变情况。

文本识别算法的特征提取器由卷积神经网络组成，整体结构选择 SE-ResNet 来进行特征提取，并且为了更好的解决不规则文本的识别问题，在浅层加入可变形卷积，提升网络的空间几何建模能力。特征提取器通过训练，从输入图像中抽取出对文本识别任务有价值的特征信息，将特征提取器从输入图像中提取到的特征转换为特征序列作为序列建模层的输入。

3.3.2 序列建模层

序列建模层的输入是特征提取层从输入的图像中提取出的特征序列，输出是特征序列对应的文本序列，每一个时间步的特征对应一个字符类别，完成对特征序列建模的任务。

长短期记忆网络（LSTM）[21]通过设计精巧的门控机制，对传递的信息进行有选择的更新和传递，解决了循环神经网络的长期依赖问题，可以学习到任意长度的序列数据的依赖关系。本文设计的文本识别算法的序列建模层选择两层的双向长短期记忆网络结构来对输入图像的特征序列进行建模，继续从图形特征中提取文本序列特征，学习文本序列的上下文依赖关系，序列建模层对特征序列的标签分布进行预测，输出该图像的文本预测结果。

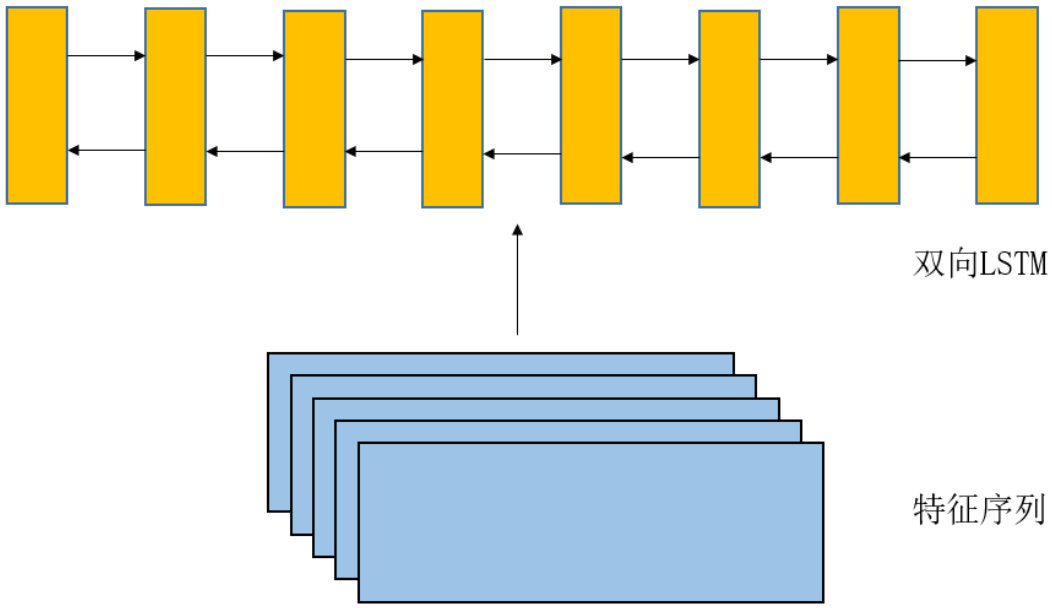

图 3-7 序列建模层

3.3.3 转录层

序列建模层的输出是对特征序列的每一个时间步都预测相应的字符类别，预测结果存在冗余的现象，输入序列和输出序列不能一一对齐，CTC[7]提出一种目标函数的计算方式，并以此对文本识别模型进行端到端训练，可以不用对序列进行标注对齐，可以直接训练模型输出正确的字符串结果，选择正确的类别路径。目标函数的计算如公式 3-5 所示，其中 x 表示序列建模层输入的特征序列，z 表示序列建模层的预测结果，$\alpha(t,u)\beta(t,u)$表示在 t 时刻对应的数据选择节点 u 的所

有路径的概率和。网络训练过程中，通过目标函数的约束，对序列建模层的输出进行去重，找到概率最大的标签路径组合，得到最后的预测序列，并且不需要对训练数据进行额外的标注对齐，可以对任意长度的序列进行预测。

$$\mathrm{L(x,z)} = -ln\sum_{u=1}^{|z'|}\alpha(t,u)\beta(t,u) \quad \text{（公式 3-5）}$$

综上所述，本文设计的文本识别算法采用主流的 CTC 架构，将文本识别的问题当作序列问题来解决，并加以改进。网络模型由特征提取器、序列建模层和转录层组成，整个模型可以端到端进行训练，训练数据是有文本序列标注的图像。首先，特征提取器对输入图像进行特征提取，网络选用加入残差模块[18]学习输入输出之间残差映射以及加入 SE 操作学习卷积通道的重要性分布的 SE-ResNet[17]，为了更好的识别不规则文本，浅层的卷积选用可变形卷积[19]，卷积核中额外加入学习卷积核的不同元素的偏移情况的参数，学习卷积核对应的采样方式和位置，自适应图像中不同字符的形变情况，从图像中抽取对文本识别有意义的特征和信息。紧接着，特征提取器提取到的特征图转换为特征序列作为序列建模层的输入，序列建模层选取两层的双向长短期记忆网络[21]，学习特征序列的上下文依赖关系，可以对任意长度的序列数据进行处理，输出预测结果。整个网络使用 CTC Loss 作为目标函数进行端到端训练，引入空白字符表示没有对应的字符类别输出，对循环神经网络的预测序列进行去重，并去掉空白字符，解决输出序列无法与输入序列对齐的问题。如此完成对输入图像进行文本识别的任务。

3.4 本章小结

本章具体介绍本文设计和提出的基于生成对抗网络改进的场景文本识别算法。

3.1 节介绍了本文设计的场景文本识别算法的思路和创新点，算法包括两大模块，训练生成对抗网络做数据预处理模块，以及场景文本识别模块对转换后的图像中的文本进行识别和分类，输出预测结果。

3.2 节介绍了第一个模块-数据处理模块，这个模块采用生成对抗网络对输入图像进行预处理，训练生成器保证能够由输入的复杂场景图像转换成为去除复杂背景干扰的背景干净的图像，训练判别器能够判别输入的图像是生成的图像还是真实图像，二者对抗博弈，最终达到一个平衡。

3.3 节介绍了第二个模块-场景文本识别模块，这一模块选用了主流的 CTC 模型，卷积神经网络提取输入图像的特征，将图像特征转化为特征序列送入循环神经网络，学习特征序列的关系，并输出每一个时间步对应的字符类别，输出该

图像的文本识别序列，并使用 CTC 作为转录层对预测结果进行矫正处理，得到最终的识别结果。本文对主流模型做了如下改进：首先对改进了特征提取器，替换成性能更好的 SENet，可以更好的提取图像特征，其次加入可变形卷积，自适应图像中各式各样的形变问题。

第四章 实验与分析

前一章介绍了本文设计的场景文本识别算法，实现从输入图像中识别文本序列的任务。算法分为两个模块，第一个模块选取条件生成对抗网络对场景图像进行数据预处理，第二个模块的输入是第一个模块输出的图像，对图像中的文本进行识别并输出，得到最后的识别结果。接下来介绍本文设计的算法的具体实现，以及对比实验相应的实验环境、实验设置以及实验结果与分析。

4.1 数据集与数据增强

深度学习与神经网络的快速发展离不开信息时代海量数据的支持，这些数据一方面展示了各个不同的深度学习任务的概况，从数据中可以观察到很多数据分布特点和规律，另一方面，数据集也为评估深度学习模型性能提供了基准测试。本节将对实验过程中使用到的公开数据集及其预处理过程进行介绍。

4.1.1 数据集

深度学习的快速发展一定程度上得益于信息时代不同行业应用产生的海量数据，训练深度神经网络从大量数据中学习数据分布和模式，在检测、识别、分割等视觉任务中表现更好。为了验证不同算法的性能，需要在相同的数据集上对这些算法进行基准测试，来比较不同算法的优劣。因此，计算机视觉领域有很多大规模开源数据集，这些数据集适用于不同的计算机视觉任务和不同的场景，为该领域的算法评测提供便于测试和比较的评测标准。

本文研究场景文本识别任务，选取了 ICPR MTWI、ICDAR 等大规模数据集来进行评测，这些数据集的共同特点是均为场景文本识别的公开数据集，并且数据以中文为主，文本序列长度不固定，有较多的背景干扰。下面对实验部分用到的数据集分别进行介绍。

• MTWI 数据集：该数据集是由阿里巴巴集团举办的 MTWI 2018 挑战赛中公开的数据集，数据集中总共包括一万张图像，每张图像对应一个文本文件，文本文件中包含了该图像文本区域的矩形框的坐标和相应的文本内容。用于文本识别任务的数据集需要从原图中对文本区域图像进行截取，总共截取出十四万张文本图像。

该挑战赛比赛任务是网络图像的文本识别，数据集是基于网络图像的、以中

文为主的大规模数据集，数据集中的图像来源于电商平台，文本主要由产品描述、广告语组成。数据难点在于图像中有很多商品作为背景，干扰较多，而且图像中的文本存在旋转、仿射变换、缩放等形变问题。

图 4-1 MTWI 数据集

• ArT 数据集：该数据集是 ICDAR 2019 国际挑战赛中 Robust Reading Challenge on Arbitrary-Shaped Text 任务公开的数据集。数据集中总共包括五万张图像，每张图像有对应的文本内容的标注信息。

这一分赛着重研究任意形变文本序列的识别任务，数据集是场景图像，以街景图像为主的中英文数据集，数据难点在于该数据集的图像中弯曲形变文本的比例更大，以任意朝向的文本为主。

图 4-2 ArT 数据集

• LSVT 数据集：该数据集是 ICDAR 2019 国际挑战赛中 Robust Reading Challenge on Large-scale Street View Text with Partial Labeling 任务公开的数据集，图像来源是街景图像，有复杂背景的干扰。数据集里面包含五万张街景图像，每张图像都有对应的文本区域定位标注信息和文本内容的标注。同样，对于文本识别任务，需要将文本区域从图片中截取出来，该数据集总共截取出二十一万张文本图像。

这三个数据集都是公开的场景文本数据集，对数据进行观察可以发现数据集中的图像可以分为几种类型。如下图 4-3 所示，第一行的图像是场景中的规则文本，这类文本排列整齐，没有太多的遮挡和干扰，识别难度较低。第二行展示是场景中的弯曲文本，图像中的文本呈现扇形排布，第三行是透视变换文本，是由拍摄角度不同引起的变换，第四行是数据集中其他类型的不规则文本。

图 4-3 文本图像类别

4.1.2 数据增强

本文实验部分使用的是场景文本识别任务的开源数据集，实验过程中会对读取的图像数据进行数据增强。数据增强通过对同一张图像进行多种变换，由此生成一些新的训练样本，进行数据扩充，丰富数据分布的多样性，而不需要重新采集新的数据样本，同时避免神经网络在数据集上过拟合，提升网络泛化性，使训练出的神经网络有更好的鲁棒性[48]。数据增强的变换多种多样，下面介绍计算机视觉任务的实验中常用的增强策略。

首先，常用的有空间几何变换，比如对训练图像进行水平翻转或者垂直翻转，增强网络对图像物体朝向的识别不变性，该方法在图像分类任务中对模型性能有所提升，但不适用于文本识别任务；对训练图像进行随机裁剪，保留图像中感兴趣的部分区域进行训练，这种方法可以引导网络关注感兴趣的区域，提取到更好的图像特征，但是在文本识别中裁剪会丢失掉一些字符，造成语义不全和部分字符遮挡的情况，不太适合文本识别任务；对训练图像进行缩放，改变图像比例，一般做法是从原始图像上随机裁剪出区域图像，然后统一到原图像的尺寸，保持图像大小不变，缩放后的图像会有一定程度的变形；对训练图像进行不同角度的旋转变换，改变物体朝向，对于文本识别任务可以进行小幅度的旋转，不改变文本语义的前提下进行旋转丰富图像多样性；对图像进行仿射变换，一般仿射是由旋转、缩放、裁剪等多种操作混合完成的。

其次，常见的还有改变图像像素颜色的变换，比如在原有图像的基础上随机添加噪声，并且可以选择在图像上添加噪声或者在图像的每个通道上都添加随机噪声，移动像素点来改变像素颜色；改变原有图像的对比度，改变图像的色调和饱和度；改变图像 RGB 通道的颜色参数，进行颜色扰动；对原有图像进行模糊处理，对图像像素点进行像素平滑，减小不同像素的差异性，实现模糊效果；在图像上选取面积不一，位置随机的小块矩形区域进行丢弃，如果选定区域丢失掉所有的通道信息，那么选定区域就变成黑色的矩形噪声区域，如果选定区域丢失掉的是部分通道的信息，那么选定区域变成彩色的噪声；改变原有图像的对比度，改变图像的色调和饱和度；从图像中随机选取矩形区域进行擦除，去掉该区域的视觉信息。这些方法对图像的像素进行改变，对图像进行不同类别的增强变换，比较适合于文本识别任务。不同的增强策略还可以改变变换的程度，设置增强策略作用在训练数据中的概率，将这些策略进行组合，丰富数据多样性，更好的在神经网络的训练过程中对数据分布进行拟合，提升网络性能。

图 4-4 数据增强

经过观察可以发现，并不是所有的视觉任务中常用的数据增强方法都适用于文本识别的任务。不同于常见的图像分类任务，文本识别的图像样本中要识别的字符的区域相对较小，有一些数据增强策略会丢失掉一些文本区域，改变序列文本的语义。所以在场景文本识别算法的对比实验中，对训练样本主要进行第二类增强，在训练图像的基础上添加随机噪声、模糊、对比度变换、颜色扰动等变换，再辅助小幅度的旋转变换，将这些策略进行组合，用多式多样的数据增强方法对训练数据进行数据扩充，防止神经网络过拟合。下图 4-5 是实验部分选取的增强策略示例。

图 4-5 增强后的图像

4.2 场景文本识别算法训练流程

本节对本文设计的基于生成对抗网络改进的场景文本识别算法进行实现，将数据预处理的生成模块和识别网络通过目标函数进行端到端的联合训练，完成对场景文本进行识别的任务。

4.2.1 生成式对抗网络训练细节

数据预处理模块采用生成式对抗网络[15]实现对原始输入图像进行预处理，达到去除复杂背景，尽量让图像背景较为干净单一的目的。本文设计的数据预处理模块选取条件生成对抗网络来进行学习。其中，生成器由编码器-解码器结构组成，编码器对原始输入图像进行下采样，解码器对特征图进行上采样，判别器采用全卷积神经网络进行分类鉴别。数据预处理模块使用两种数据进行训练，一种是作为输入的复杂场景下的文本图像，另一种是与之相对应的背景干净的文本图像，作为条件信息加快网络的学习过程，这两种图像组成图像对对网络进行训练。数据预处理模块的设计目的是对复杂场景下的图像进行数据转换，将复杂场景的文本图像映射成为背景干净的图像，尽量避免复杂背景带来的干扰，提升后续的识别模型的识别精度。生成器和判别器在目标函数的约束下进行学习，迭代训练，逐步提升各自网络模型的性能。生成器从训练数据中学习数据映射关系，经过训练之后的生成器可以将输入图像转换为背景干净的图像，判别器在对真实图像和生成图像的差异性的学习中，增强自己的鉴别能力。二者在博弈中完成强化模型性能的目的。

图 4-6 用于 GAN 训练的数据示例

数据预处理模块在目标函数的引导下进行训练，生成器和判别器进行对抗博弈，迭代优化。首先，从输入数据中进行采样送到生成器网络，生成器网络的输入不是随机噪声，而是训练时指定的复杂场景下的文本图像，生成器网络对输入图像进行转换完成数据生成工作，然后将生成的图像送入判别器网络，判别器网

络对图像进行分类。判别器网络的数据来源既包括真实数据集，也包括生成网络生成的图像，判别器网络不断地从这两种数据来源中获得图像，学习如何正确区分真实和虚假的图像。生成器可以完成对输入图像进行数据转换，将复杂场景文本图像映射成为背景干净的图像，尽量减少复杂背景给图像特征提取以及文本识别带来的干扰，提升网络的识别精度。

4.2.2 文本识别网络训练细节

文本识别算法由三个模块组成，包括特征提取层、序列建模层和转录层，其中，特征提取层选用卷积神经网络对输入图像进行特征提取，卷积神经网络主体选择可以学习残差映射以及卷积通道权重分布的 SE-ResNet[17]，浅层加入可变形卷积[19]来应对复杂场景下不规则文本的识别问题，增强网络对输入图像提取重要特征的能力，从图像中学习对场景文本识别任务有意义的特征和信息。序列建模层选用深层双向长短期记忆网络[21]对特征序列进行学习和建模，学习特征序列的上下文依赖关系，以及文本序列的语义信息，对文本序列做出预测。转录层选用 CTC loss[7]作为目标函数来引导和约束文本识别模型的训练和学习，对循环神经网络的预测结果进行矫正，解决输入序列和输出序列不能一一对齐的问题，三个模块在目标函数的约束下端到端进行训练，从训练数据中学习对图像中的文本进行分类识别的能力。

文本识别算法将卷积神经网络和循环神经网络连接在一起，特征提取层输出的特征序列作为序列建模层的输入，使用目标函数来对网络参数进行训练和优化。实验部分的数据集使用之前介绍过的三个数据集，按照八比二的比例将数据集拆分为训练集和测试集进行训练，数据集的图像包含对图像中文本序列的字符串标注，数据集中的文本序列包含中英文两种语言以及特殊符号，从标注信息中进行字符类别的提取作为学习的字典，也就是网络分类的类别总和。在训练过程中依据网络性能表现对学习率进行调整，引导网络的学习过程。经过训练之后的网络可以对场景图像中的文本进行识别。

网络训练使用的数据包括场景文本图像以及对应的序列标注，训练过程中使用之前介绍过的数据增强策略对训练数据进行数据扩充，增加训练数据的多样性，丰富数据特征的分布，使网络有更好的特征表达能力，并且避免出现过拟合的现象，使网络具有更好的泛化性。数据预处理模块进行训练之后，生成器可以对输入图像进行转换，将场景文本图像转换为对应的背景干净的图像，对输入数据进行处理，更加有利于网络对图像中文本序列的识别。将生成器生成的图像送入文本识别模块进行训练和识别，特征提取层对图像中有价值的视觉特征和信息进行

提取，并传输到序列建模层作为输入，序列建模层对序列进行预测，得到图像对应的预测结果。

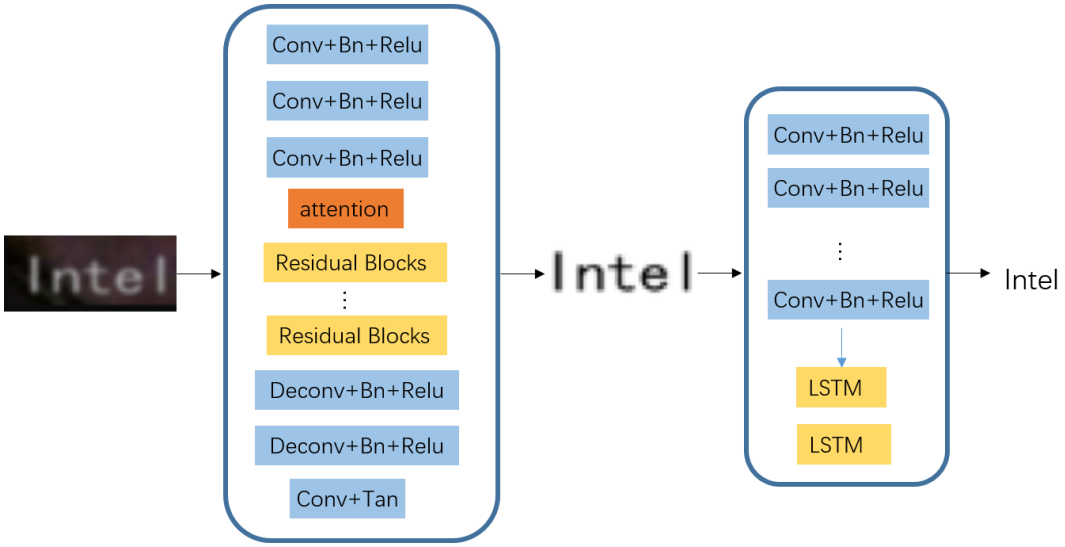

图 4-7 模型架构

4.3 算法实验与分析

本文设计的场景文本识别算法基于主流的算法进行了改进，在对主流算法进行分析之后，发现主流算法在复杂场景和不规则文本的识别过程中性能表现一般，有待于进一步的改进。因此在主流算法的基础上加入预处理模块，使用条件生成对抗网络训练生成器学习从输入图像到对应的背景干净的图像的映射过程，完成图像转换，尽量减少复杂背景给识别任务带来的干扰。对于不规则文本的识别，则在识别算法的特征提取器中加入可变形卷积[19]，改变典型卷积网络使用矩形卷积核对图像进行固定采样的方式，额外加入学习卷积核对应元素偏移情况的参数，自适应学习图像中目标文本的形变情况，增强网络对场景图像的空间几何建模能力。

4.3.1 实验设置

下面将设计对比实验，使用之前介绍过的三个中英文场景文本识别数据集来证明算法改进点的有效性，实验过程中不同算法模型的训练数据和测试数据保持一致。

MTWI 数据集[53]包括一万张网络图像，将随机抽取出的九千张图像进行文本区域的截取用于模型训练，对剩余的一千张图像中的文本区域进行截取用于模型测试，总共截取出十二万六千张训练图像和一万四千张测试图像。

ArT 数据集[54]的图像不需要进行截取，总共五万张文本图像按照八比二的比例进行划分，随机抽取四万五千张图像作为训练数据，其余的五千张图像作为测试数据。

LSVT 数据集[55]中有五万张街景图像和对应的标注信息，从数据集中随机挑选出四万五千张图像，对图像中的文本区域进行截取，得到十九万张文本图片作为训练数据。另外的五千张街景图像对文本区域进行截取，截取出两万张文本图像作为测试数据。

设计对比实验时，不同算法模型的实验设置和参数选择保持一致。优化算法选择自适应矩估计（adaptive moment estimation，简称 Adam）[52]方法，初始学习率设置为 0.0001，并且在完成所有训练数据的十次训练之后对学习率进行调整，将学习率调整为原学习率的十分之一。在实验设置保持一致的条件下，比较不同算法模型在场景文本识别任务上的性能表现，验证算法改进点的有效性。

4.3.2 评测规则

实验评判规则包括词准确率和归一化编辑距离。其中词准确率表示测试数据集的图像预测结果与图像序列标注完全匹配的占比，数值范围在零到一之间，准确率越高表明网络模型能够正确识别对的图像占比更高，网络模型的性能也就更加优异。

归一化编辑距离对测试集图像的预测序列与图像对应的标注序列的差异程度进行量化评测，并进行归一化，数值范围也是在零到一之间，归一化编辑距离的数值越小，表明图像预测结果与数据标注的差异越小，网络性能也就更加优异。

Norm 表示归一化编辑距离，D（:）表示编辑距离，s_i表示文本序列的预测结果，$\hat{s}_i$表示对应的文本序列的标注。

$$Norm = \frac{1}{N}\sum_{i=1}^{N} D(s_i, \hat{s}_i) \quad （公式 4-1）$$

4.3.3 实验结果与分析

下面三个表格分别展示了三个数据集的对比实验的实验结果，对比实验的数据集以及实验参数设置均保持不变，只改变算法模型，使用全匹配准确率和归一化编辑距离作为评判标准。表格中展示了不同网络结构下模型的性能表现。

三个表格中的 Aug 表示是否有数据增强，SEResNet 表示是否将特征提取器进行替换，DeformConv 表示是否加入可变形卷积，cGAN 表示是否加入条件生成对抗网络对图像进行预处理。表格对应位置为空表示没有做对应的改进，“√”表

示在算法中加入了相应的改进。评测规则包括准确率（Accuracy）和归一化编辑距离（Norm）。

表 4-1 MTWI 实验结果

MTWI dataset					
Aug	**SEResNet**	**DeformConv**	**cGAN**	**Accuracy(%)**	**Norm**
				58.00	**0.0943**
√				**58.31**	**0.0914**
√	√			**59.17**	**0.0859**
√	√	√		**60.21**	**0.0788**
√	√	√	√	**61.33**	**0.0632**

从 MTWI 数据集的实验结果可以看出，初始模型在该数据集上的实验结果不够理想，在对特征提取器进行改造，网络替换成 SE-ResNet 后模型性能有所提升，特征提取更加科学。在此基础上加入可变形卷积，增强网络对字符变换的空间建模能力，识别性能有所提升。之后再加入使用条件生成对抗网络训练好的生成器对输入图像进行转换，再次提升识别精度。

表 4-2 ArT 实验结果

ArT dataset					
Aug	**SEResNet**	**DeformConv**	**cGAN**	**Accuracy(%)**	**Norm**
				44.92	**0.3436**
√				**45.16**	**0.3312**
√	√			**46.71**	**0.3158**
√	√	√		**50.63**	**0.2546**
√	√	√	√	**53.47**	**0.2014**

ArT 数据集中包含大量自然场景下的任意形状的文本，识别难度增大，所以整体实验精度低于 MTWI 数据集。从 ArT 数据集的实验结果可以看出，不同算法模型下的实验结果变化趋势与 MTWI 数据集的相似，在原始模型下实验结果的精度最低，在初始模型的基础上，依次改进模型的特征提取器，将卷积神经网络不断优化，在典型的卷积网络中加入残差模块学习上一层输入输出的残差，加入 SE 模块学习卷积通道的权重分布都可以帮助网络更好的从训练数据中提取有效特征。ArT 数据集中不规则文本所占比例很大，针对性的加入可以自适应文本的弯曲、旋转、仿射等变换的可变形卷积，学习图像中局部密集的形变情况，模型性

能大幅提升。加入生成器对图像进行转换后进一步提升了模型性能。

表 4-3 LSVT 实验结果

LSVT dataset					
Aug	**SEResNet**	**DeformConv**	**cGAN**	**Accuracy(%)**	**Norm**
				48.95	**0.2814**
√				**49.21**	**0.2796**
√	√			**50.67**	**0.2671**
√	√	√		**52.46**	**0.2146**
√	√	√	√	**55.29**	**0.1678**

LSVT 数据集主要由街景图像组成，相较于 MTWI 这种以网络图像为主的数据集，LSVT 数据集中的图像背景更加复杂，给识别带来干扰，因此同一个算法模型在 LSVT 数据集上的实验精度要低于 MTWI 数据集。LSVT 中不规则文本所占的比例低于 ArT 数据集，因此整体实验精度高于 ArT 数据集。从实验结果可以看出，对特征提取器的改进和预处理模块的加入都提升了网络性能，实验精度逐步提升，模型性能更加优异。

从上述实验结果可以看出，在主流模型的基础上进行改进之后模型性能都有所提升，预处理模块可以减少复杂背景的干扰，改进之后的特征提取器具有更好的特征表达能力，并且可以自适应图像中的形变情况，尤其对于 ArT 这种任意形状的文本的比例较高的数据集。实验结果证明了算法改进点的有效性，改进之后的模型在复杂场景的文本识别任务中有更好的性能表现。

4.4 本章小结

本章对本文设计和提出的基于生成对抗网络改进的场景文本识别算法进行具体实现，并设计对比实验在多个真实场景的文本识别数据集上证明算法改进点的有效性。

其中，4.1 节介绍了实验中使用到的三个真实场景下的场景文本识别数据集，这些数据集的图像中存在复杂背景的干扰，数据集的文本序列中的字符包括中英文字符和特殊符号，有一定比例的图像中存在不规则文本的情况，识别难度较大。在实验过程中会对训练数据进行数据增强来丰富数据的多样性，提升网络的泛化性能，避免出现过拟合的情况。4.2 节介绍了场景文本识别算法的实现细节，介绍了生成对抗网络以及文本识别网络的训练过程。4.3 节设计了对比实验，使用三个真实场景的数据集来对模型性能进行评测，验证本文提出的算法改进点的有

效性。实验证明算法模型中加入生成器对图片进行转换，转换后的图片中减少了复杂背景带来的干扰，更易于识别模型对文本序列进行预测，识别准确率更高。另外将特征提取器替换成 SE-ResNet，可以更好的从图像中提取出有效特征，实验证明这种做法可以提升模型的识别准确率。在特征提取器中加入可变形卷积来解决文字的形变问题，使卷积核在图片中进行自适应采样，扩大卷积核的感受野，实验证明可变形卷积的加入提升了模型在场景文本识别，尤其是不规则文本识别任务上的识别准确率。

第五章 总结与展望

本文在对场景文本识别任务的国内外研究现状进行调研之后，分析了现有算法的不足，并重点解决复杂背景干扰和不规则文本识别难度大的问题，并设计了基于生成对抗网络改进的场景文本识别算法，对主流算法做出了改进，并设计对比实验证明了算法创新的可行性和有效性。本章对全文进行总结，并在此基础上展望未来的工作。

5.1 全文总结

本文关注于场景文本识别任务的研究，借助深度学习和神经网络来解决这一经典的计算机视觉任务。主要完成了以下的工作：

（1）明确场景文本识别研究的课题意义和背景，对场景文本识别的国内外研究现状进行调研，对现有的场景文本识别算法进行分析，找出现有算法的不足和局限性，明确本文的研究内容，针对性的对现有算法的不足之处进行改进，并提升模型性能。

（2）对深度学习和神经网络的相关基础进行介绍，着重学习并介绍卷积神经网络、循环神经网络和生成对抗网络的基本概念和技术原理，为接下来的场景文本识别算法设计做原理指导。

（3）针对主流算法在复杂背景和不规则文本识别上性能表现差的问题，针对性的对网络模型进行改进。加入生成对抗网络对数据进行预处理，将原始输入图像转换为背景干净的图像，减少复杂背景的干扰。特征提取器中加入残差模块、SE 模块和可变形卷积，提升网络的特征抽取能力。

（4）借助多个深度学习工具对本文设计的场景文本识别算法进行代码实现，用目标函数约束网络的学习和训练，并使用多个真实场景的文本识别数据集设计对比实验，在保持实验设置一致的前提下，通过全匹配准确率和归一化编辑距离两个评测指标对不同算法的模型性能进行量化评测，证明本文提出的算法创新点的有效性。

5.2 未来展望

本文设计的场景文本识别算法针对图像中过多的复杂背景干扰和不规则文本给文本识别带来挑战的问题进行特殊设计，在多个场景文本识别任务的公开数据集上表现优异，与主流算法相比模型性能有所提升，证明了算法创新点的有效

性。

但是目前的场景文本识别算法还存在一些不足，未来可以做进一步的算法创新和改进。首先，自然场景下的文本存在遮挡的现象，被遮挡的字符存在结构缺失的现象，语义信息不够明确，非常容易被识别成错误的字符类别，给文本识别任务带来很大的挑战。其次目前出现了很多在移动设备上可以使用的文本识别的功能，本文设计的场景文本识别算法模型参数较多，不适合用于移动设备这类存储空间和计算能力都有所限制的使用场景，未来可以对模型进行剪枝量化，在保持模型识别性能的同时，精简网络设计，减少模型存储空间和计算资源的消耗。最后在实际业务场景中，经常会出现网络训练数据中没有出现过的数据类型，给识别任务带来挑战，未来可以通过算法设计，使得网络可以使用在场景中拥有自我学习和自我纠正的能力，让网络更加智能。

参考文献

[1] Karaoglu S, Tao R, Gevers T, et al. Words matter: Scene text for image classification and retrieval[J]. IEEE Transactions on Multimedia, 2016, 19(5): 1063-1076.

[2] Bai X, Yang M, Lyu P, et al. Integrating scene text and visual appearance for fine-grained image classification[J]. IEEE Access, 2018, 6: 66322-66335.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In TPAMI, 2017. 10

[4] A. Shrivastava, A. Gupta, and R. Girshick. Training regionbased object detectors with online hard example mining. In CVPR, 2016. 2, 5

[5] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In CVPR, 2017. 3, 9

[6] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In CVPR, 2017. 3, 9

[7] Graves et al. Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks.

[8] Bissacco A, Cummins M, Netzer Y, et al. Photoocr: Reading text in uncontrolled conditions[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 785-792.

[9] Jaderberg M, Simonyan K, Vedaldi A, et al. Reading text in the wild with convolutional neural networks[J]. International Journal of Computer Vision, 2016, 116(1): 1-20.

[10] He P, Huang W, Qiao Y, et al. Reading scene text in deep convolutional sequences[C]//Thirtieth AAAI Conference on Artificial Intelligence. 2016.

[11] B. Shi, X. Bai, C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition", CoRR, vol. abs/1507.05717, 2015.

[12] Baoguang Shi, Xinggang Wang, Pengyuan Lv, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. arXiv preprint arXiv:1603.03915, 2016.

[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. CoRR, abs/1506.02025, 2015.

[14]Goodfellow Ian, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.

[15]Mirza M, Osindero S. Conditional Generative Adversarial Nets[J]. Computer Science, 2014:2672-2680.

[16]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[17]J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507, 2017. 5, 7

[18]K.He, X.Zhang, S.Ren, J.Sun, "Deep residual learning for image recognition", 2015.

[19]J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. arXiv:1703.06211, 2017.

[20]Graves, Alex. Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850, 2013.

[21]Hochreiter, Sepp and Schmidhuber, Jurgen. Long short-term memory. ¨ Neural computation, 9(8): 1735–1780, 1997.

[22]刘保.基于神经网络深度学习的车牌识别算法[J].中国交通信息化,2019(08):122-125+128.

[23]赵汉理,刘俊如,姜磊,沈建冰,胡明晓.基于卷积神经网络的双行车牌分割算法[J].计算机辅助设计与图形学学报,2019,31(08):1320-1329.

[24]魏娇.基于机器视觉的车牌字符自动识别系统设计[J].自动化与仪器仪表,2019(08):49-53.

[25]丁蒙,戴曙光,于恒.卷积神经网络在手写字符识别中的应用研究[J/OL].软件导刊:1-5[2019-09-07]

[26]王瑞,李海峰,马琳.超低分辨率视频中的字符识别技术研究[J].智能计算机与应用,2019,9(01):203-207.

[27]邓立,张祺,张伟新.基于神经网络的准考证号码识别研究[J].工业控制计算机,2019,32(01):104-106.

[28]沈臻,韩震宇.基于机器视觉的 OCR 自动识别系统的研发[J].科技与创新,2019(08):144-145+147.

[29]胡泽枫,张学习,黎贤钊.基于卷积神经网络的批量发票识别系统研究[J].工业控制计算机,2019,32(05):104-105+107.

[30]洪洋. 基于深度卷积神经网络的验证码识别[A]. 中国自动化学会系统仿真专业委员会、中国系统仿真学会仿真技术应用专业委员会.第 19 届中国系统

仿真技术及其应用学术年会论文集（19th CCSSTA 2018）[C].中国自动化学会系统仿真专业委员会、中国系统仿真学会仿真技术应用专业委员会:中国自动化学会系统仿真专业委员会,2018:4.

[31]孟彩霞,王腾飞,王鑫.基于深度残差网络的文字识别算法研究[J].计算机与数字工程,2019,47(06):1487-1490+1501.

[32]杜泽炎,任明武.一种在低质量图像上提高字符识别率的深度学习框架[J].计算机与数字工程,2019,47(06):1491-1496.

[33]徐英杰,李国勇,洪文焕.基于多粒度级联多层梯度提升树的选票手写字符识别算法[J].计算机应用,2019,39(S1):26-30.

[34]阎晨阳. 小字符喷码机的字符识别方法研究[D].广西师范大学,2019.

[35]邵文良. 基于深度学习的医疗单据图文识别关键技术研究与实现[D].北京邮电大学,2019.

[36]Tyrrell Christopher D. A method to implement continuous characters in digital identification keys that estimates the probability of an annotation.[J]. Applications in plant sciences,2019,7(5).

[37]J E M Adriano,K A S Calma,N T Lopez,J A Parado,L W Rabago,J M Cabardo. Digital conversion model for hand-filled forms using optical character recognition (OCR)[J]. IOP Conference Series: Materials Science and Engineering,2019,482(1).

[38]E Arrieta-Rodríguez,L F Murillo,M Arnedo,A Caicedo,M A Fuentes. Prototype for identification of vehicle plates and character recognition implemented in Raspberry pi[J]. IOP Conference Series: Materials Science and Engineering,2019,519(1).

[39]Zongjhe Yang,Keisuke Doman,Masashi Yamada,Yoshito Mekada. Character recognition of modern Japanese official documents using CNN for imbalanced learning data[P]. Other Conferences,2019.

[40]Abhinav Kaushal Keshari,Rajat Sharma,Madhav J. Nigam. Digitizing physical documents using optical character recognition[P]. International Conference on Signal Processing Systems,2019.

[41]肖坚.基于学习的 OCR 字符识别[J].计算机时代,2018(07):48-51.

[42]王泽天.基于神经网络对手写字符的研究[J].科技经济导刊,2018,26(29):14-16.

[43]张展睿.基于 SVM 算法的英文识别工具[J].通讯世界,2018(09):250-251.

[44]Ali Farhat,Omar Hommos,Ali Al-Zawqari,Abdulhadi Al-Qahtani,Faycal Bensaali,Abbes Amira,Xiaojun Zhai. Optical character recognition on heterogeneous SoC for HD automatic number plate recognition system[J]. EURASIP Journal on Image and Video Processing,2018,2018(1).

[45] Zhishan Hu,Juan Zhang,Tania Alexandra Couto,Shiyang Xu,Ping Luan,Zhen Yuan. Optical Mapping of Brain Activation and Connectivity in Occipitotemporal Cortex During Chinese Character Recognition[J]. Brain Topography,2018,31(6).

[46] Named Entity Recognition; Study Results from Q.P. Tan et al Provide New Insights into Named Entity Recognition (Character Feature Learning for Named Entity Recognition)[J]. Computers, Networks & Communications,2018.

[47] Aro Taye Oladele,Musa Abdullahi Yola,Abdulkadir Ikeola Suhurat,Adeoye Latifat Bukola. Recognition of Alphabet Characters and Arabic Numerals Using Back Propagation Neural Network[J]. Journal of Computer Science and Control Systems,2018,11(2).

[48] Xiwen Qu,Weiqiang Wang,Ke Lu,Jianshe Zhou. Data augmentation and directional feature maps extraction for in-air handwritten Chinese character recognition based on convolutional neural network[J]. Pattern Recognition Letters,2018,111.

[49] Ondrej Bostik,Jan Klecka. Recognition of CAPTCHA Characters by Supervised Machine Learning Algorithms[J]. IFAC PapersOnLine,2018,51(6).

[50] Yang Jia,Wang Fan,Chen Zhao,Jungang Han. An Approach for Chinese Character Captcha Recognition Using CNN[J]. Journal of Physics: Conference Series,2018,1087(2).

[51] Bušta M, Patel Y, Matas J. E2E-MLT-an unconstrained end-to-end method for multi-language scene text[J]. arXiv preprint arXiv:1801.09919, 2018.

[52] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[53] Mengchao He, Zhibo Yang,, et al. ICPR MTWI 2018 text recognition challenge on network images.

[54] Yipeng Sun, Chee-Kheng, Chng, et al. ICDAR 2019 robust reading challenge on arbitrary-shaped text [C]//2019 15th International Conference on Document Analysis and Recognition (ICDAR).

[55] Yipeng Sun, Chee-Kheng, Chng, et al. ICDAR 2019 robust reading challenge on large-scale street view text with partial [C]//2019 15th International Conference on Document Analysis and Recognition (ICDAR).

致谢

时光如梭，不知不觉中研究生生活已经要接近尾声了，回想在复旦大学计算机科学技术学院度过的两年半的学习时光，我不仅收获了研究生期间研究方向的专业知识，还学到了科学的研究方法和思路，也懂得了很多为人处事的道理。这段经历会是我人生道路上的财富，激励我在以后的前进过程中更加勇敢、更加自信。在此我要感谢研究生期间给予我很多辅导和帮助的导师和学院各位老师，感谢同窗的同学和实验室的小伙伴们，感谢父母亲人和朋友们的陪伴。

首先，衷心感谢研究生期间对我影响最大的我敬爱的导师薛向阳教授，感谢薛老师对我孜孜不倦的教诲。生活上，薛老师是一位和蔼可亲的长辈，品德高尚，待人和善，关心我的生活，给予无微不至的关心和帮助，从老师身上我学到了很多做人做事的道理。学业上，薛老师治学严谨，学识渊博，不断追求创新，对我严格要求。本次毕业设计从论文开题到中期报告到最后的论文撰写，薛老师提出了很多建设性的意见，倾注了大量心血。再次对我的导师薛向阳教授表示最衷心的感谢，祝老师身体健康，桃李满天下！

其次，我要感谢李斌老师在我攻读研究生期间对我的教育和指导，李老师对学术孜孜不倦的追求，以及勇于攀爬科研高峰的精神给我留下了很深刻的印象，激励我树立远大的学术目标，永不放弃对学术创新的追求。李老师不仅及时解答我学业上的疑惑，还悉心指导我的学习和研究，提出了很多改进性意见。然后要要感谢实验室师兄师姐们在实验过程中帮助我解决问题，分享了很多经验，感谢同窗好友们在撰写论文中给予大量支持和帮助，感谢实验室师弟师妹们在学习和生活上的关心和帮助。感谢大家的陪伴让我收获了珍贵的友谊，也让我的研究生生活变得更加丰富多彩。

此外，还要感谢无论在物质上还是精神上都给予我很多支持的父母，感谢父母的养育之恩，感谢他们作为我最坚定的后盾，让我有积极的生活态度，可以对知识不懈追求，是父母的鼓励支持着我不断克服困难，是父母的爱激励着我坚持奋斗，让我在人生道路上坚定追求理想的信念，谢谢你们！

最后，感谢百忙之中抽出时间参与评审本论文的各位专家和老师，感谢你们的辛苦工作，向你们表示诚挚的谢意。

张培尧

2019 年 9 月

复旦大学
学位论文独创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：_____ 日期：_____

复旦大学
学位论文使用授权声明

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：_____