计算机科学
COMPUTER SCIENCE

基于空间域和频率域特征融合的场景文本识别

霍华骑, 陆璐

引用本文

霍华骑, 陆璐. 基于空间域和频率域特征融合的场景文本识别[J]. 计算机科学, 2023, 50(11A): 230300101-8.

HUO Huaqi, LU Lu. Scene Text Recognition Based on Feature Fusion in Space Domain and Frequency Domain [J]. Computer Science, 2023, 50(11A): 230300101-8.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230300101

基于空间域和频率域特征融合的场景文本识别

霍华骑[1]　陆　璐[1,2]
1 华南理工大学计算机科学与工程学院　广州 510006
2 鹏城实验室　广东 深圳 518055
(hhq-yyq@qq.com)

摘　要　对于小样本语言无关场景的文本识别,现有的方法往往面临鲁棒性低和泛化能力差的问题。针对这一问题,一方面,在特征提取阶段,提出了基于空间域和频率域特征融合的双流网络结构,其包含一个提取空间域特征的深度残差卷积网络分支,以及提取频率域特征的一维快速傅里叶变换和浅层神经网络分支,接着使用通道注意力机制融合这两种特征。另一方面,在序列建模阶段,针对语言无关场景的特点,提出一种多尺度一维卷积模块用来代替双向长短期记忆网络。然后结合现有的TPS矫正模块和CTC解码器搭建完整模型。训练过程中采用了迁移学习的方法,先在大型英文数据集上进行预训练,后在目标数据集上进行微调。在文中整理的两个小样本语言无关数据集上的实验结果表明,所提模型在准确率上优于现有的模型,验证了其在该场景下的具有较高的鲁棒性和泛化能力;此外,在语言相关场景的5个基准数据集上的相关实验(不用微调)表明,使用文中所述特征提取模块的方法优于对比的基线方法,证明了所提出的双流特征融合网络的有效性和通用性。

关键词:深度学习;场景文本识别;双流网络;频率域分支;小样本
中图法分类号　TP391

Scene Text Recognition Based on Feature Fusion in Space Domain and Frequency Domain

HUO Huaqi[1] and LU Lu[1,2]
1 School of Computer Science and Engineering,South China University of Technology,Guangzhou 510006,China
2 PENGCHENG Laboratory,Shenzhen,Guangdong 518055,China

Abstract　Existing scene text recognition methods often face the problems of low robustness and poor generalization ability in the few-shot and language-independent scene. To solve this problem,on the one hand,a dual-stream network structure based on the fusion of space domain and frequency domain features is proposed in the feature extraction stage. It consists of a deep residual convolutional network branch for extracting spatial domain features,and a shallow neural network with one-dimensional fast fourier transform(FFT) branch for extracted frequency features. And then apply the channel attention mechanism to fuse the two features. On the other hand,in the sequence modeling stage,a multi-scale one-dimensional convolution module is proposed to replace the bidirectional long short-term memory(BiLSTM) according to the characteristics of the language-independent scene. Finally,a complete model is built by combining the existing TPS rectification module and CTC decoder. The transfer learning method is adopted in the training process. Pre-training is performed on the large English datasets first,and then fine-tuning is performed on the target datasets. Experimental results on two few-shot language-independent datasets compiled in the paper show that the method is superior to the existing methods in terms of accuracy,which verifies that it has high robustness and generalization ability in this scenario. Moreover,the method using the feature extraction module described in the paper is better than the baseline on the five benchmark datasets of language-dependent scene(no fine-tuning),which verifies the effectiveness and versatility of the dual-stream feature fusion network proposed in the paper.

Keywords　Deep learning,Scene text recognition,Dual-stream network,Frequency domain branch,Few-shot

1　引言

随着信息技术的发展,场景文本识别(Scene Text Recognition,STR)已成为实现自动化和智能化的重要技术之一。STR通常被认为是光学字符识别(Optical Character Recognition,OCR)的一种特殊形式,也被称为基于相机的OCR[1]。它的主要任务是将自然场景图片中的文本提取出来并转化成字符形式,而识别场景文本有助于场景理解,对工业自动化、自动驾驶、图片检索、票据识别等领域有着重要的研究意义。然而STR在不受约束的环境中会遇到诸多的难题,例如字体多样性、多尺度、光照干扰、复杂背景、图像模糊等,因此对算法的鲁棒性和泛化能力提出了更高的要求。

传统文本识别方法[2]依赖于手工制作的特征,以及随机森林和支持向量机等机器学习技术,这类传统方法通常只

基金项目:广东省重点领域研究计划(2022B0101070001)
This work was supported by the Research Plan of Key Fields of Guangdong Province(2022B0101070001).
通信作者:陆璐(lul@scut.edu.cn)

适用于背景清晰、布局简单的文档类文本识别，难以处理字体和背景复杂的情况，因此鲁棒性和泛化能力都较差。

而基于深度学习的场景文本识别方法在复杂场景下的准确率上限高于传统的方法。得益于 MJ[3] 和 ST[4] 等大规模英文合成数据集以及 CNN(Convolutional Neural Network)，LSTM(Long Short-Term Memory)，Transformer[5] 等基础网络架构的提出，对于复杂场景下的文本识别已经取得了巨大进展。基于深度学习的模型整体上都遵循编码器-解码器架构，根据解码方式的不同，大体上可以分为基于 CTC 解码器和基于注意力解码器的两种方法。例如 Shi 等[6] 创新性地提出了一种基于 CNN 和 CTC 解码器的 CRNN 网络，可以进行端到端的训练，并且可以支持变长的文本，但是 CRNN 较为简单的编码器网络结构使其在面对复杂的场景时存在鲁棒性差的问题。

为了提高文本识别的鲁棒性，后来提出的 RARE[7] 网络首次在该领域引入空间变换网络 STN 进行图片矫正，同时使用基于注意力机制的解码器进行特征向量到字符的解码工作。该方法提高了对倾斜文本的识别鲁棒性，同时基于注意力的解码器相比 CTC 解码器，在大规模语言相关的场景下可以更好地利用上下文依赖性，从而实现更高的准确率。然而该类方法只能提高倾斜文本的鲁棒性，对于字体多样性、模糊、复杂背景的识别仍然面临困难；而且在小样本语言无关场景下，上下文依赖的解码器也会产生过拟合以及泛化能力不足的问题。虽然 Yue 等[8] 也考虑到了该问题并且提出了 RobustScanner 模型，该模型在解码时融合特殊的位置编码信息降低了长距离的上下文依赖，提高了语言无关场景下的准确率，但他们未考虑小样本情况下泛化能力差的问题。

对于小样本语言无关场景下存在的上述问题，文中提出了基于空间域和频率域特征融合网络作为核心编码器，相比现有的 STR 研究，它通过引入额外的频率域信息提高了模型的鲁棒性和泛化能力。此外，针对语言无关场景的特点，提出了多尺度一维卷积模块用来代替双向长短期记忆网络，可以更好地避免上下文依赖导致的过拟合问题。然后结合现有的 TPS 矫正模块[9] 和 CTC 解码器搭建完整模型。对比实验和消融实验表明，在小样本语言无关场景下，本文提出的模型取得了最好的准确率，而且具有更高的鲁棒性；同时为了验证本文提出的特征提取方法的通用性，还在基准数据集上做了对比实验，相比基线方法，在非弯曲的规则数据集上有较大的准确率提升，验证了本文方法具有一定的通用性。

本文的主要贡献如下：

(1)首次关注了场景文本识别领域中的小样本语言无关场景，并且标注和生成了两个小样本语言无关数据集，弥补了这方面研究的空白。

(2)提出了基于空间域和频率域特征融合的特征提取模块，可以提取更加丰富的视觉特征信息，提高了模型的整体鲁棒性和泛化能力。

(3)为了更好地适应语言无关场景任务，提出一种只关注相邻局部信息的多尺度一维卷积神经网络，用来代替经典双向长短期记忆网络进行序列建模。

本文第 2 章回顾了相关工作；第 3 章介绍了本文提出的方法；第 4 章介绍了数据集和实验流程；第 5 章分析了实验结果；最后总结全文。

2 相关工作

2.1 基于 CTC 解码器

随着基于深度学习的识别方法成为主流，许多经典的 STR 模型被提出。例如，Shi 等[6] 在 2017 年提出了经典 3 阶段的卷积循环神经网络模型 CRNN。该网络先使用深度卷积网络进行特征提取，得到特征序列，然后堆叠多个双向长短期记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)进行序列建模，并将 CTC 解码器连接在 BiLSTM 网络的最后一层来实现端到端的文本识别。CRNN 大大提高了识别性能，是深度学习应用在场景文本识别领域的经典开山之作。为了进一步提高 CTC 解码器的准确率，Hu 等[10] 提出了使用图卷积网络指导 CTC 训练的 GTC 模型，其中 CTC 模型从一个更强大的注意引导中学习更好的对齐和特征表征；通过引导训练的协同训练，CTC 模型在保持快速推理速度的同时，对规则和不规则场景文本实现了鲁棒和准确的预测。

2.2 基于注意力解码器

为了改进对倾斜角度图片的识别效果，后来的研究者提出了 RARE[7] 和 ASTER[11] 等网络，他们将空间变换网络 STN 和薄板样条插值 TPS 引入模型中，将其变为更具鲁棒性的 4 阶段场景文本识别框架。空间变换网络可准确地识别透视变换过的文字以及弯曲的文字，可以将输入图片中的文字矫正成水平形状，从而达到了更好的识别效果。同时为了提高模型的解码能力，使用了基于注意力机制的编码器解码器架构来代替经典的 CTC 解码器，但是基于注意力解码器的方法通常训练时间更长。Li 等提出的 SAR[12] 在上述模型的基础上，考虑到现有的方法解码时只能根据一维信息进行预测，引入了二维注意力的解码器，并且达到了很高的识别准确率，但缺点是速度会变慢。

近年来，随着 Transformer 在视觉领域的快速发展，基于 Transformer 的各种方法也被提出。ViTSTR[13] 直接将视觉 Transformer(ViT) 引入文本识别领域，它将图片分块后添加位置编码，送入 Transformer 编码器和解码器中，直接得到字符序列，思想简单却有效。近两年来，为了进一步解决图片模糊、遮挡等问题，场景文本识别的热点方向转向引入语言模型进行联合训练，例如 SEED[14]，PARSeq[15] 等，这些方法虽然在大规模英文数据集上展现出了巨大的优势，但是还不能证明其对非英语的数据集的有效性，而且这种方法相当于使用了语言模型进行结果矫正，难以应用在语言无关的场景。此外，目前 SVTR[16] 算法是纯视觉模型里综合性能最优的，它并非使用卷积或者 ViT 网络，而是提出了一种包含混合-合并-组合等操作、可以提取丰富多粒度的特征成分的新型网络结构，在精度和速度方面都取得了很好的综合性能。

3 算法框架

本章将详细介绍本文提出的方法，首先从整体上介绍模型的结构，然后对每个阶段的模块进行描述。

模型整体遵循 Baek 等[9] 总结的 4 阶段场景文本识别框架，即图片矫正、特征提取、序列建模和序列预测。本文提出的整体方法如图 1 所示，橙色模块(1)和(4)代表使用现有的模块，二者中间的模块是本文新提出的。其中，输入的图片首先经过 TPS 矫正模块得到归一化的图片，该模块会将倾斜、弯曲的图片进行适当的矫正，有利于后续的识别。接下来将

矫正后的图片输入本文提出的融合空间域和频率域的神经网络中进行特征提取，其主要分为两个分支：以 ResNet[17] 为基础的空间域分支和经过纵向一维 FFT 之后再进行卷积的频率域分支，然后将二者提取的特征使用通道注意力机制进行融合后输出，改变形状后得到特征序列。在序列建模阶段，采用提出的多尺度一维卷积模块进行序列建模，提高了相邻特征的交互能力。最后使用经典的 CTC 解码器[6] 对特征序列进行序列解码，得到预测的最终结果。

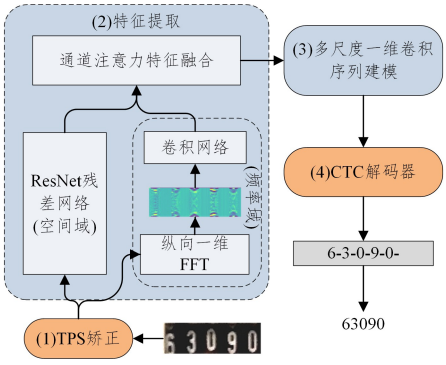


图 1　场景文本识别模型整体架构图(电子版为彩图)

Fig. 1　Architecture of scene text recognition model

3.1　图片矫正阶段

文本识别模块总是希望输入的图片尽可能是规则的文字行，因此这需要在识别之前做一个特殊的空间变换去矫正图片。在图片矫正阶段，本文直接采用了 ASTER 网络中[11] 成熟的 TPS 空间变换网络，TPS 通过学习一组参数，可以将倾斜、透视和弯曲文字图像转换为比较水平的形式。

具体来说，TPS 使用平滑样条线在基准点集合之间进行插值，识别上下包络点处的几个基准点，然后将字符区域归一化为预定义的矩形，如图 2 所示。此前的许多研究[9] 都证明了其对倾斜、弯曲文本识别的有效性。

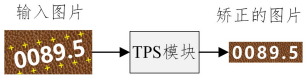


图 2　TPS 图片矫正

Fig. 2　Image rectification by TPS

3.2　特征提取阶段

现有的网络通常采用深度卷积神经网络进行特征提取，将输入的图片变为若干特征序列后预测字符。然而不少研究表明[18]，虽然卷积算子在学习高频细节方面有很好的能力，但是往往忽略了低频信息导致模型的泛化能力变差，这是因为卷积更加擅长从边缘捕获高频信息特征。在场景文本识别领域中的图片具有复杂性，而普通的 CNN 网络只能从单一空间域提取图像特征，从而限制了网络的整体性能。

受到 Li 等[19] 引入频率域信息用于提高人脸伪造检测能力，以及 Mao 等[20] 引入频率域信息可以实现更好的图像去模糊效果的启发，本文结合场景文本识别的特点，提出了融合空间域和频率域特征的双流骨干网络。上述两种方法都是对图片整体使用二维的 FFT 进行变换后提取整体特征，然而对于场景文本识别而言，文本行的方向性决定了无法直接套用这种方法。如果使用了横向 FFT 会对图片的每一行计算频谱图，相当于对横向的信息进行了某种混淆，这导致横向排布的文字无法区分每个字符。而如果采用纵向一维 FFT，则每个字符都有对应的频谱图，不会导致字符的混淆，其效果如图 1 中浅绿色图片所示。

频谱图来源于离散傅里叶变换 DFT 在图片领域的应用，它的一维形式可以表示为：

$$X[k]=\sum_{j=0}^{N-1}x[j]\mathrm{e}^{-i\frac{2\pi}{N}kj} \tag{1}$$

其中，$X[k]$ 表示频率为 $\omega_k=2\pi k/N$ 的频谱，i 是虚数单位，x 是 N 个复数的序列。经过 DFT 之后，任何频率的频谱都有全局信息。快速傅里叶变换 FFT 算法降低了复杂度，并更有效地计算了 DFT。在目前常用的深度学习框架中，可以直接调用 FFT 相关函数。

3.2.1　空间域和频率域特征提取

结合上述 FFT 的思想，本文设计的特征提取骨干网络详细结构如图 1 的第(2)部分所示。该骨干网络有两个分支，分别采用了 ResNet 作为空间域特征提取分支，同时使用纵向 FFT 变换之后拼接 CNN 作为频率域特征提取分支，最后将二者提取的特征使用通道注意力机制融合后输出。

其中空间域采用的 ResNet 与原始版本略有不同，为了保证将输入的图片高度降为 1 以及将宽度降为原来的 1/4，需要修改网络中池化(Pooling)层的步长大小，也就是把除了第一个和最后一个池化层外的中间所有原始步长为 2×2 的池化层都改为 2×1。这样修改之后中间的网络只会缩小高度而不会更改宽度，从而保证了得到足够长度的特征序列来预测每个字符。后续实验会讨论该采用何种深度的 ResNet。

在频率域方面，将输入的图片经过纵向一维 FFT 之后，将实部和虚部按通道维度进行合并得到频谱图，再经过卷积神经网络进行特征提取，其表达式如下：

$$f=FFT(x,-2) \tag{2}$$

$$x'=Conv(f\rightarrow real\oplus f\rightarrow imag) \tag{3}$$

其中，x 表示输入的图片，$FFT()$ 表示进行一维快速傅里叶变换，第二个参数 -2 表示对倒数第二个维度即高度进行变换，也就是纵向的 FFT。$\rightarrow read$ 和 $\rightarrow imag$ 分别表示取其中的实部和虚部；$\oplus$ 表示按照通道进行张量拼接；$Conv$ 表示卷积神经网络，其中网络的结构配置如表 1 所列。

表 1　频率域卷积神经网络结构

Table 1　Architecture of frequency domain convolution network

Name	Configuration	Output(CHW)
Sigmoid		3×32×100
Block1	$k=3,n=32,s=2\times2,p=1$ BatchNorm,ReLU	32×16×50
Block2	$k=3,n=128,s=2\times1,p=1$ BatchNorm,ReLU	128×8×50
Block3	$k=3,n=256,s=2\times1,p=1$ BatchNorm,ReLU	256×4×50
Block4	$k=3,n=512,s=2\times1,p=1$ BatchNorm,ReLU	512×2×50
MaxPool	$k=2\times2,s=2\times2$	512×1×25

该网络主要包含：一个 Sigmoid 层将频谱图归一化；4 个包含批次归一化和 ReLU 激活函数的卷积层，进行特征提取；一个最大池化层将输出匹配到指定大小。其中 k 表示核大小，n 表示输出维度，s 表示步长，p 表示填充大小，BatchNorm 是批次归一化，ReLU 是非线性激活函数。该网络可以提取经过 FFT 变换之后的频域图的特征，输出维度大小与空间域相同。

3.2.2　基于通道注意力机制的特征融合

对于上述两个特征提取网络所得到的特征，需要进行

特征融合。一般特征融合可以采用直接拼接、直接相加、加权相加等方法，然而这些方法都具有一定的局限性，直接拼接或相加无法衡量每种特征的重要性，而加权相加则需要手动设置融合系数。为了实现更好的特征融合效果，本文结合 SE 注意力机制[21]提出了基于通道注意力机制的特征融合网络，如图 3 所示。

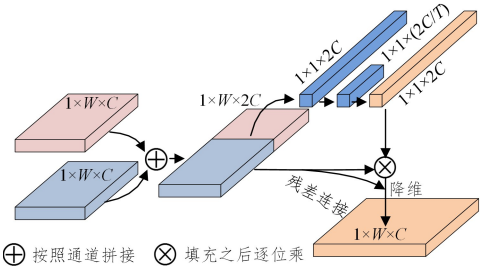


图 3 基于通道注意力机制的特征融合网络

Fig. 3 Feature fusion network based on channel attention mechanism

其工作流程为：将空间域和频率域的特征先按照通道维度进行拼接，然后使用全局平均池化得到 $1\times1\times2C$ 的向量，其中 $2C$ 是通道维度，即大小为 2 倍的输入通道数；使用 1×1 卷积核对通道维度降维到 $2C/T$ 后，再升维恢复原来的维度 $2C$，经过 Sigmoid 函数归一化得到通道注意力权重；将该权重平铺(Tile)回原来大小后与拼接的向量逐位相乘，将结果与拼接的向量相加形成残差连接，经过 Dropout 后，使用 1×1 卷积降维得到最后的特征融合向量。其中压缩比 T 取 SE 原论文中的最优值 16。

3.3 序列建模阶段

对于序列建模阶段，传统的 LSTM 结构擅长建立字符之间的全局联系，目的是利用字符之间的语义关系实现单词预判。但是对于没有语义关系的语言无关场景，字符出现的概率是独立的，LSTM 反而会因为建立了多余的依赖关系而导致错误的预测。

为此本文提出了多尺度一维卷积(Multi-Scale one-Dimensional，MS1D)模块，对骨干网络提取的特征进行序列建模，其内部细节如图 4 所示。

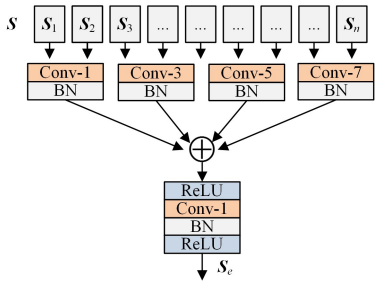


图 4 多尺度一维卷积序列建模模块

Fig. 4 Sequence modeling module using multi-scale 1D convolution

其中，BN 表示批归一化，ReLU 表示线性整流函数。对于输入维度是 $n\times d$ 的序列 $\boldsymbol{S}$，分别使用核大小为 1，3，5，7 的一维卷积来提取多尺度的特征，这样可以使得每个特征序列 $\boldsymbol{S}_i$ 可以与周围 $\boldsymbol{S}_{i-3}-\boldsymbol{S}_{i+3}$ 个共 6 个序列进行交互。同时为了保证序列 $\boldsymbol{S}$ 的长度不变，分别使用了大小为 0，1，2，3 的边缘填充。值得注意的是，为了保证大卷积核边缘节点不损失信息，卷积的填充模式采用了重复填充(Replicate)，即重复边缘的元素，而不是默认的零填充(Zero)。

更具体地，算法 1 描述了多尺度一维卷积序列建模算法的核心流程，其中输入为：特征图 $\boldsymbol{S}$，输入维度 i，隐藏维度 h。输出为：经过相邻交互建模之后的新特征图 $\boldsymbol{S}_e$。使用该算法之后，新的序列完成了与最多 6 个相邻序列的交互，而且避免了全局的交互，更加适合字符出现概率独立的语言无关场景。

算法 1 多尺度一维卷积序列建模算法

输入：特征图 **S**，输入维度 i，隐藏维度 h

输出：输出特征图 $\mathbf{S}_e$

1. 声明 4 个不同尺度特征图 $\mathbf{C}_1$，$\mathbf{C}_3$，$\mathbf{C}_5$，$\mathbf{C}_7$
2. 声明临时变量和最终输出变量 **H**，$\mathbf{S}_e$
3. for k in {1，3，5，7}：
4. 对 **S** 进行一维卷积得到 $\mathbf{C}_k$//其中核大小为 k，输入通道数为 i，输出通道数为 h
5. 对 $\mathbf{C}_k$ 进行批量归一化
6. end for
7. 按通道维度合并 $\mathbf{C}_1$，$\mathbf{C}_3$，$\mathbf{C}_5$，$\mathbf{C}_7$ 特征图得到 **H**
8. 对 **H** 进行 ReLU 激活
9. 对 **H** 进行一维卷积得到 $\mathbf{S}_e$//其中核大小为 1，输入通道数为 4×h，输出通道数为 i
10. 对 $\mathbf{S}_e$ 进行 ReLU 激活
11. return $\mathbf{S}_e$

3.4 序列解码阶段

对于解码阶段，本文采用了经典的 CTC 解码器[6]，而不是基于注意力的序列解码器 Att[11]。因为在一些综述[1]以及实际测试中发现：Att 有注意力漂移的缺点并且不适用于语言无关的场景，而且在小样本数据集下，Att 又是训练困难和泛化能力较差的。这是因为这种解码器默认每个字符都是上下文相关的，其严重依赖上一个字符的解码。因此考虑到小样本语言无关场景的因素，本文采用了之前的序列识别研究中经典的 CTC 作为解码模块。CTC 解码器是一个特殊的损失函数，其条件概率的公式可以描述如下：

输入特征用 $\boldsymbol{y}=(y_1,y_2,\cdots,y_K)$ 表示，其 K 为序列长度。$\boldsymbol{y}$ 的每个元素都是字符集 L 上的概率分布。具体来说，L 表示一个字符集合包括所有有效字符和一个表示无效输出的额外空白符号。CTC 路径 π 是一个长度为 K 的序列，它包括有效字符和空白字符。定义了一个特殊映射函数 G 来删除路径 π 中重复的字符和空白符号，得到真实标签 l。然后，将映射到 l 的所有路径的概率相加，计算条件概率。

$$p(\pi|y)=\prod_{k=1}^{K}y_{\pi_k}^{k} \tag{4}$$

$$p(l|y)=\sum_{\pi:G(\pi)=l}p(\pi|y) \tag{5}$$

其中，$p(\pi|y)$ 表示每条路径 π 对输入特征 y 下的条件概率，$y_{\pi_k}^{k}$ 表示在序列 k 处有标签 π_k 的概率。对所有满足 $G(\pi)=l$ 的路径 π，将其所有的条件概率相加，就可以得到在网络输出的 y 下得到标签 l 的条件概率 $p(\pi|y)$。CTC 损失函数就是最大化条件概率 $p(\pi|y)$。

4 实验设置

4.1 实验数据集

4.1.1 提出的小样本语言无关数据集

EMeterSTR：为了弥补小样本语言无关场景文本识别方面的空白，结合网上公开的脱敏之后的电表图片，使用

PPOCRLabel 工具标注之后，将读数和设备编号等关键识别区域裁剪下来，得到了 593 张可以用于场景文本识别任务的图片，命名为 EMeterSTR，含义是电表场景文本识别。另外还划分了 200 张作为测试集，数据集的字典只包含“0123456789.-A”这 13 个字符。其中部分样例图如图 5(a)所示。

(a) (b)

图 5 EMeterSTR 和 RandGen1200 数据集部分样例图

Fig. 5 Partial samples of EMeterSTR and RandGen1200 datasets

RandGen1200：由于 EMeterSTR 中只包含 13 个字符，为了进一步验证模型在更多字符集下的能力，使用开源的 Text Recognition Data Generator 工具，亦即 Python 中的 trdg 包，生成了 1 200 张包含数字和大写字母共 36 字符的图片，其中 200 张作为测试集。为了模拟在真实场景下的复杂文本识别，添加了随机的弯曲、高斯模糊、倾斜、不同字体等变换，其部分样本如图 5(b)所示。

4.1.2 基准数据集

ICDAR2013(IC13)[22]：从 561 张图像中裁剪、删除具有非字母数字字符的单词后，包含 857 张图片。

IIIT5K[23]：来自于谷歌地图街景图像，包含 5 000 张文本实例图像，其中 3 000 张用于评估。

SVT[24]：也是来自于街景的图片，但是具有更严重的噪声、模糊、低分辨率，裁剪后有 647 张图片用于评估。

ICDAR2015(IC15)[25]：这是谷歌眼镜在没有保证图像质量的情况下拍摄的，而且包含完全文本。IC15 数据集裁剪和过滤一些几何扭曲严重的图像后包含 1 811 张图片。

CUTE80[26]：该数据集包含来自于 80 张高分辨图片裁剪得到的 288 张实例图，它更加关注弯曲文本，同时图片拥有复杂的背景以及透视失真等情况。

4.1.3 大规模人工合成数据集

MJ[3]：这是场景文本识别常用的大型训练集，包含了人工合成的约 900 万张复杂场景图片，同时它是语言相关的，文本内容来自 9 万个单词。

ST[4]：同样是常用的大型合成复杂场景文本检测数据集，裁剪之后可用于场景文本识别的约有 600 万张图像。

4.1.4 数据集的分析说明

其中提出的小样本语言无关数据集用来验证本文提出的整体模型的有效性，而基准数据集则是为了进一步验证所提模型在语言相关场景下是否也具有通用性。大规模合成在小样本语言无关实验中因为存在数据分布和字典数量等不一致情况，只作为预训练集，而在基准数据集上则作为训练集。

4.2 评价指标

为了更好地评价模型性能，实验中主要是使用了文本识别领域常用的 1-NED 和 ACC。

1-NED：NED 即归一化编辑距离(Normalized Edit Distance，NED)[27]，常用于比赛等高精度文本识别场景中。NED 越接近 0 表示两个字符串越相似，识别准确率越高，因此 1-NED 也可以被看作字符级的准确率。1-NED 的公式为：

$$1\text{-}NED=1-\frac{1}{N}\sum_{i=1}^{N}\frac{D(s_i,\hat{s}_i)}{\max(l_i,\hat{l}_i)} \tag{6}$$

其中，N 为文本行的样本数量，D 代表莱文斯坦(Levenshtein) 距离，是一种常用的编辑距离。s_i 和 $\hat{s}_i$ 分别表示真实标签和预测的文本，而 l_i 和 $\hat{l}_i$ 是它们的文本长度，max 函数表示取最大值。

ACC：即准确率，如果预测字符串和标签字符串完全相同，则可以视为准确。因此 ACC 的定义是测试时完全识别正确的图片数量 C 除以被测的图片总数量 A，如式(7)所示：

$$ACC=\frac{C}{A} \tag{7}$$

4.3 实验环境与训练方法

实验环境如下，电脑 CPU 是 AMD 5950X，内存 64GB，显卡是 AMD MI100 32GB；编程语言为 Python 3，使用 PaddlePaddle 深度学习框架搭建模型。评估方法采用场景文本识别领域通用的训练期间多次评估取最优值的方法。默认输入图片宽高为 100×32。

对于小样本语言无关数据集的实验方法：由于样本太少导致模型无法收敛，必须采用迁移学习的方法进行训练。具体地，先在 MJ 和 ST 两个大规模合成的英文数据集上训练到收敛，再将输出层改为目标数据集的字典对应大小，然后进行微调。微调时在目标数据集上训练若干个批次(在 EMeterSTR 数据集上为 200，RandGen1200 为 60)，批次大小设置为 18，采用 Adadelta 自适应学习率优化器，优化器的参数 $rho=0.95$，$clip\_norm=5.0$。训练 100 step 后每隔 32 step 计算一次测试集的 NED，选取最优值作为结果。由于数据较小，训练默认使用了几种常用的数据增强方法，包括随机裁剪、颜色反转、高斯模糊、颜色抖动，所有数据增强的概率都设置为 0.4。为了排除随机误差，重复上述过程 3 次(预训练过程不重复)，取中位数作为最终结果，以 1-NED 作为主要指标。

对于基准数据集的实验方法：采用与之前的研究(如 BEAK[9]，ASTER[11]等)类似的方法，即在 MJ 和 ST 两个大规模合成的英文数据集上训练，每间隔一段时间后在基准数据集上评估，直到结果收敛。训练批次大小设置为 100，采用 Adadelta 自适应学习率优化器，优化器的参数 $rho=0.95$，$clip\_norm=5.0$。训练 600 step 后每隔 600 step 计算一次测试集的 ACC，当 ACC 连续 5 次不再上升后停止训练，取最优值作为结果。

5 实验结果和分析

5.1 参数设置和网络结构实验

5.1.1 参数设置

在特征融合部分，随机失活层(Dropout) 的不同值会对

结果有一定的影响。为了进一步探究其对结果的影响以及设置最佳的随机失活值，首先进行了不同的 Dropout 概率对特征融合影响的实验。实验模型基于第 3 章提出的架构，使用 ResNet18 作为空间域特征网络，使用第 4 章对小样本数据集的评估方法，在 EMeterSTR 和 RandGen1200 数据集上评估 Dropout 概率对字符级准确率(1-NED)的影响。

结果如图 6 所示。可以看出，使用 Dropout 对特征融合整体上是有益的，在特征融合的过程中提高了不同特征之间的联合适应性，使得模型可以更好地融合空间域和频率域提取的特征，增强了泛化能力；但是 dropout 也不能太大，否则会导致训练难度增大，以及特征损失过多。综合上述实验结果，在后续实验中特征融合模块的 Dropout 设置为 0.2。

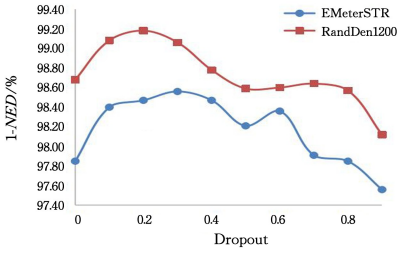


图 6 不同的 Dropout 概率对特征融合的影响

Fig. 6 Effects of different dropout rate on feature fusion

5.1.2 空间域网络结构选择

对于特征提取网络的空间域的卷积网络，不同深度的 ResNet 网络的效果是不一样的。为了探究在小样本语言无关场景下不同的 ResNet 效果，基于第 3 章提出的模型框架，使用了 18，34 和 50 这 3 种不同深度的网络，分别命名为 ReFT18，ReFT34，ReFT50。通过第 4 章介绍的实验步骤，评估其在两个小样本数据集上的效果。

实验结果如表 2 所列，可以看到，随着深度的增加，模型的 1-*NED* 和 *ACC* 都有提升，但是 ReFT34 和 ReFT50 的性能接近。为了平衡速度和精度，后续研究采用 ReFT34 作为基础。而且还注意到，为达到近似的单词级准确率(ACC)，RandGen1200 数据集需要更高的字符级准确率(1-NED)，这主要是因为 RandGen1200 数据集的字符平均长度更长，平均长了约 2 个字符。

表 2 不同深度 ResNet 对模型的影响

Table 2 Effects of ResNet at different depths on models

Name	EMeterSTR		RandGen1200	
	1-*NED*	*ACC*	1-*NED*	*ACC*
ReFT18	98.56	90.5	99.18	93.0
ReFT34	99.08	93.5	99.75	97.0
ReFT50	**99.12**	**94.0**	**99.78**	**97.5**

5.2 小样本语言无关数据集的对比实验

为了验证本文的核心论点，即提出的模型对小样本语言无关场景下提高鲁棒性和泛化能力的有效性，将第 3 章提出算法与基于深度学习的场景文本识别算法进行对比分析，同时进行了消融实验分析。表 3 列出了本文方法与其他 5 种常见方法的对比结果。从表中的结果可知，本文提出的方法远优于上述几种常用的方法，其中次优的 SVTR[16] 是 2022 年的最优场景文本识别方法之一，而本文方法在 EMeterSTR 和 RandGen1200 这两个小样本语言无关数据集上的 ACC 比 SVTR 分别高出了 9.4% 和 4.3%。这是因为现有的算法是针对大规模的合成数据集设计的，而且主要关注的是语言相关的场景，造成模型在样本较少的情况下，即使使用了迁移学习的方法，但仍面临特征提取能力差、泛化能力弱的问题。而本文通过引入频率域信息构建了双流网络，加强了特征提取能力，同时设计了多尺度的一维卷积模块来避免长距离语义干扰，因此在小样本语言无关数据集下也具有很强的泛化能力和较高的鲁棒性。

表 3 6 种算法在小样本语言无关数据集上的对比

Table 3 Comparison of 6 algorithms on few-shot language-independent datasets

名称	EMeterSTR		RandGen1200	
	1-*NED*	*ACC*	1-*NED*	*ACC*
CRNN-Res34[14]	96.48	75.5	99.12	92.0
ASTER[8]	95.58	73.5	99.01	91.0
RobustSc[26]	97.13	81.5	99.26	92.5
ViTSTR-tiny[10]	96.71	79.5	98.74	89.5
SVTR-base[13]	97.63	85.5	99.32	93.0
ReFT34	**99.08**	**93.5**	**99.75**	**97.0**

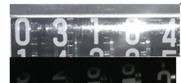
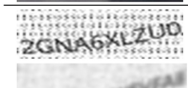
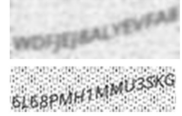
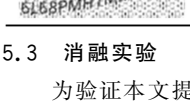
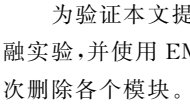
注：加粗表示最优；下划线表示次优。

此外，实验结果还表明，迁移学习中的目标数据集与源数据集的差异对结果也会有较大的影响：EMeterSTR 数据集的字符是真实的电表读数、编号等难以合成的字符，在预训练数据集 MJ 和 ST 中没有这种字符，因此许多算法在该数据集上效果较差；而另一个目标数据集 RandGen1200 是使用 Python trdg 包随机生成的，因此所有算法都取得了较高的准确率。

表 4 列出了在 EMeterSTR 和 RandGen1200 数据集上的部分识别结果，其中斜体文字表示全部正确，则其余表示出现错误。可以发现本文提出的方法在高光、部分遮挡、严重模糊等情况下仍具有很高的鲁棒性。但是当同一列包含两个字符时，可能会出现无法命中最中间字符的情况(如第三行所示)，主要原因是本文方法是以一维信息为主，对于二维信息利用较少，后续的改进方向可以向这方面考虑。而在该场景下，虽然 SVTR-base 和 RobustSc 的整体识别准确率不如本文方法，但是在某些实例上取得了更好的结果(如第 3 和第 6 行)。

表 4 在语言无关数据集上的部分识别结果

Table 4 Partial recognition results on language-independent datasets

	REFT34	RobustSC	SVTR
	***03184***	**03104**	**03104**
	***43772***	**46779**	**48772**
	**030130**	**020131**	***030131***
	***2GNA6***	**2GNA6**	**2GNA6**
	XLZUO	**XLZUD**	**XLZUD**
	***WDFJEJ8A***	**WDJJJ4A**	**WOFJEJ8A**
	LYEVFA8	**LYEVFAMI**	**LYRVFA5**
	**6L6BPMH1**	**6L68PMH1**	***5L68PMH1***
	MMU3SKG	**MMU3SKG**	***MMU3SKG***

5.3 消融实验

为验证本文提出的各个模块的有效性，需要对其进行消融实验，并使用 EMeterSTR 数据集作评估，在实验过程中，依次删除各个模块。当删除空间域或者频率域的网络时，直接返回一个全零的张量；当删除特征融合模块时，使用向量拼接后降维的方法代替；当删除多尺度一维卷积序列(MS1D)建模时，用传统的 BiLSTM 代替。

实验结果如表 5 所列，可以发现本文提出的基于空间域

和频率域特征融合的网络结构整体上是有效的，也可以看出各个部分的重要性不一样。其中空间域的 ResNet 对结果的影响最大，将其移除之后模型准确率相比完整模型下降了 20.5%，这主要可能是本文设计的频率域特征提取网络层数较少，特征提取能力仍有提升空间。而另外 3 个模块的重要性比较接近，都一定程度上辅助提高了模型对小样本语言无关场景下的文本识别准确率。

表 5　消融实验结果

Table 5　Ablation experimental results

频率域	空间域	融合	MS1D	1-*NED*	*ACC*
×				98.42	89.5
	×			95.61	73.0
		×		98.74	91.5
			×	98.34	89.0

5.4　基准数据集的对比实验

为了进一步验证本文提出的方法在非小样本语言无关的基准数据集上的有效性和通用性，本文设计了如下实验：首先将本文提出的模型整体使用第 4 章提出的基准测试方法进行训练和评估，目的是验证整体模型的效果；其次选取了一个四阶段的方法 ASTER[11] 作为基线方法（baseline），仅使用本文提出的双流特征提取网络代替其中的特征提取网络并记为 ASTER(our)，对比效果是否提升，目的是验证本文提出的双流特征提取网络的有效性和通用性。所选的对比的方法都是不包含语言模型的纯视觉模型，而且除了 CRNN 之外，选择的都是近年来提出的方法，ASTER[11]，RobustSc[8]，ABINet[28]，SVTR[16] 分别是 2019－2022 本年度内综合性能最好的方法，而且都是基于注意力解码器的，因此更加适用于的语言相关场景。

实验结果如表 6 所列，可以看到，使用本文提出的模型 ReFT34 在基准数据集上的 ACC 指标优于经典的 CRNN，但是劣于近几年提出的新方法。这是因为本文提出的模型主要是为语言无关场景设计的，专门设计了 MS1D 模块代替 BiLSTM，这导致了其会忽略语言相关场景中文本的字符依赖关系，因此不适用于语言相关的场景是可以预见的。

表 6　在基准数据集上的 ACC 指标对比结果

Table 6　Comparison results of ACC on benchmark datasets

Models	IC13 857	IIIT5K 3000	SVT 647	IC15 1811	CUTE80 288
	Regular Text			Irregular Text	
CRNN[6]	91.1	82.9	81.6	69.4	65.5
RobustSc[8]	94.8	95.3	88.1	77.1	90.3
ABINet-vision[28]	95.3	94.9	89.3	83.2	86.5
AoA[29]	95.0	88.4	89.7	79.1	75.3
SVTR-base[16]	**97.1**	**96.0**	91.5	**85.2**	**91.7**
ReFT34(our)	91.6	91.1	86.7	72.2	78.5
ASTER(baseline)[11]	91.8	93.4	89.5	76.1	79.5
ASTER(our)	96.0	94.9	**92.6**	79.2	82.6

改进的 ASTER(our)使用了注意力解码器，因此可以更好地验证本文提出的双流特征提取网络的有效性和通用性，相比基线方法，其在各个数据集上 ACC 都得到了一定的提升，其中在 IC13 数据集上提升幅度最大，提升了 4.2 个百分点。改进后的 ASTER 相比最先进的方法的 SVTR 在水平文本为主的数据集 SVT，IC13，IIIT5K 上取得了具有竞争力的结果，这说明本文提出的双流特征提取网络不仅适用于小样本语言无关数据集，而且在语言相关场景下也具有通用性。但是也可以看到，在 CUTE80 和 IC15 这两个含弯曲文本的数据集上，所提方法与最先进的方法相比差距较大。这主要是本文特征提取的过程主要是一维的，不能很好地关注二维信息，对于排列不整齐的文本识别，准确率还有提升空间。

结束语　为了进一步提高模型在小样本语言无关场景下文本识别的准确率，文中提出了一种基于双流网络和多尺度一维卷积的新模型。其中双流网络分别提取空间域和频率域两种信息后使用通道注意力机制进行特征融合，提高了模型的泛化能力；而多尺度一维卷积作为序列建模的模块，代替了经典的 LSTM 顺序建模，使其更适合语言无关的场景。对比实验和消融实验结果表明，本文提出的模型在小样本语言无关的场景下远远优于现有的方法，具有较高的鲁棒性和泛化能力。即使在大规模语言相关场景下，双流特征提取也可以加强现有模型的特征提取能力，取得了具有竞争力的结果。在下一步的工作中，有两个改进方向：一是在频率域使用更深的网络实现更好的性能，但也要平衡速度和精度；二是本文提出的方法整体上是一维的，对弯曲文本图片的识别能力存在不足，后续可以研究是否能够实现二维的双流网络而不会导致水平方向的字符发生混淆。

参 考 文 献

[1] CHEN X, JIN L, ZHU Y, et al. Text recognition in the wild: A survey[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-35.

[2] YAO C, BAI X, LIU W. A unified framework for multioriented text detection and recognition[J]. IEEE Transactions on Image Processing, 2014, 23(11): 4737-4749.

[3] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Synthetic data and artificial neural networks for natural scene text recognition[J]. arXiv: 1406.2227, 2014.

[4] GUPTA A, VEDALDI A, ZISSERMAN A. Synthetic data for text localisation in natural images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 2315-2324.

[5] HAN K, WANG Y, CHEN H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[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 and Machine Intelligence, 2016, 39(11): 2298-2304.

[7] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 4168-4176.

[8] YUE X, KUANG Z, LIN C, et al. Robustscanner: Dynamically enhancing positional clues for robust text recognition[C]// European Conference on Computer Vision. Springer, 2020: 135-151.

[9] BAEK J, KIM G, LEE J, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis [C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE, 2019: 4715-4723.

[10] HU W, CAI X, HOU J, et al. Gtc: Guided training of ctc towards

efficient and accurate scene text recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2020, 34(7): 11005-11012.
[11] SHI B, YANG M, WANG X, et al. Aster: An attentional scene text recognizer with flexible rectification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2035-2048.
[12] LI H, WANG P, SHEN C, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, 2019, 33(1): 8610-8617.
[13] ATIENZA R. Vision transformer for fast and efficient scene text recognition[C]// International Conference on Document Analysis and Recognition. Springer, 2021: 319-334.
[14] QIAO Z, ZHOU Y, YANG D, et al. Seed: Semantics enhanced encoder-decoder framework for scene text recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 13528-13537.
[15] BAUTISTA D, ATIENZA R. Scene Text Recognition with Permuted Autoregressive Sequence Models[C]// European Conference on Computer Vision. Springer, 2022: 178-196.
[16] DU Y, CHEN Z, JIA C, et al. Svtr: Scene text recognition with a single visual model[J]. arXiv: 2205. 00159, 2022.
[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.
[18] WANG H, WU X, HUANG Z, et al. High-frequency component helps explain the generalization of convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 8684-8694.
[19] LI Y, BIAN S, WANG C, et al. Detection of Deepfakes Based on Dual-stream Network[J]. Computer Science, 2022, 49(S2): 558-566.
[20] MAO X, LIU Y, SHEN W, et al. Deep residual fourier transformation for single image deblurring[J]. arXiv: 2111. 11745, 2021.
[21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2018: 7132-7141.
[22] KARATZAS D, SHAFAIT F, UCHIDA S, et al. ICDAR 2013 robust reading competition[C]// 2013 12th International Conference on Document Analysis and Recognition. IEEE, 2013: 1484-1493.
[23] MISHRA A, ALAHARI K, JAWAHAR C. Scene text recognition using higher order language priors[C]// BMVC-British Machine Vision Conference. BMVA, 2012: 1-11.
[24] WANG K, BABENKO B, BELONGIE S. End-to-end scene text recognition[C]// 2011 International Conference on Computer Vision. IEEE, 2011: 1457-1464.
[25] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]// 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015: 1156-1160.
[26] RISNUMAWAN A, SHIVAKUMARA P, CHAN C, et al. A robust arbitrary text detection system for natural scene images [J]. Expert Systems with Applications, 2014, 41(18): 8027-8048.
[27] HE M, LIU Y, YANG Z, et al. ICPR2018 contest on robust reading for multi-type web images[C]// 2018 24th International Conference on Pattern Recognition (ICPR). Elsevier, 2018: 7-12.
[28] FANG S, XIE H, WANG Y, et al. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 7098-7107.
[29] XIAO Z, NIE Z, SONG C, et al. An extended attention mechanism for scene text recognition[J]. Expert Systems with Applications, 2022, 203: 117377.

HUO Huaqi, born in 1998, postgraduate. His main research interests include deep learning and scene text recognition.

LU Lu, born in 1971, Ph.D, professor, Ph.D supervisor. His main research interests include deep learning, software reliability and high performance computing.