DOI:10.16136/j.joel.2023.02.0146

基于注意力机制的多方向文本检测

徐 健[1*],郭湛澎[1],刘秀平[1],陈 博[1],闫焕营[2]

(1.西安工程大学 电子信息学院,陕西 西安 710048;2.深圳罗博泰尔机器人技术有限公司,广东 深圳 518109)

摘要:针对多方向排列的文本因其尺度变化大、复杂背景干扰而导致检测效果仍不甚理想的问题,本文提出了一种基于注意力机制的多方向文本检测方法。首先,考虑到自然场景下干扰信息多,构建文本特征提取网络(text feature information ResNet50,TF-ResNet),对图像中的文本特征信息进行提取;其次,在特征融合模型中加入文本注意模块(text attention module,TAM),抑制无关信息的同时突出显示文本信息,以增强文本特征之间的潜在联系;最后,采用渐进扩展模块,逐步融合扩展前部分得到的多个不同尺度的分割结果,以获得精确检测结果。本文方法在数据集CTW1500、ICDAR2015上进行实验验证和分析,其F值分别达到80.4%和83.0%,比次优方法分别提升了2.0%和2.4%,表明该方法在多方向文本检测上与其他方法相比具备一定的竞争力。

关键词:场景文本检测;注意力机制;文本特征提取网络(TF-ResNet);文本注意模块

中图分类号:TP391.41 **文献标识码**:A **文章编号**:1005-0086(2023)02-0166-08

Multi-directional text detection based on attention mechanism

XU Jian[1*], GUO Zhanpeng[1], LIU Xiuping[1], CHEN Bo[1], YAN Huanying[2]

(1. School of Electronics and Information, Xi'an Polytechnic University, Xi'an, Shaanxi 710048, China; 2. Municipal Robotel Robot Technology Co., LTD, Shenzhen, Guangdong 518109, China)

Abstract: Aiming at the problem that the detection effect of multi-directional arrangement text is still not ideal due to its large scale change and complex background interference, this paper proposes a multi-directional text detection method based on attention mechanism. Firstly, considering that there is a lot of interference information in natural scenes, a text feature extraction network is constructed to extract the text feature information in the image; Secondly, a text attention module (TAM) is added to the feature fusion model to suppress irrelevant information while highlighting textual information to enhance potential connections between text features; Finally, a progressive expansion module is used to gradually fuse the segmentation results obtained from the pre-expansion part at several different scales to obtain accurate detection results. The method is experimentally validated and analysed on datasets CTW1500 and ICDAR2015, and its F-values reach 80.4% and 83.0% respectively, which are 2.0% and 2.4% better than the next best method, indicating that the method is competitive with other methods in multi-directional text detection.

Key words: scene text detection; attention mechanism; text feature information ResNet50 (TF-ResNet); text attention module (TAM)

0 引 言

自然场景下的文本检测作为计算机视觉的重要分支,逐渐成为当下的研究热点,尤其是在目标定位、人机交互、图像搜索、机器导航和工业自动化等场景发挥着极其重要的作用[1]。同时,场景文本检测也面临着诸多挑战,包括复杂的背景、不同的形状和尺度等,因此,当下对多方向文本实例区域的检测定位仍然是一项极具挑战性的工作。

目前,用来检测自然场景中文本的方法主要

* E-mail:xujian@xpu.edu.cn

收稿日期:2022-03-09 修订日期:2022-04-29

基金项目:陕西省科技厅项目(2018GY-173)和西安市科技局项目(GXYD7.5)资助项目

分为两大类:基于回归的方法与基于分割的方法。基于回归的方法,文本对象往往由四边形框表示。横纵向序列检测网络(curved scene text detection via transverse and longitudinal sequence connection,CTD)[2]针对曲线文本形状复杂问题,提出多边形回归曲线文本方法,并设计多边形非极大值抑制与非多边形抑制两种后处理策略,但距离过近的曲线文本无法进行有效检测;段连接网络(detecting oriented text in natural images by linking segments,SegLink)[3]基于单发多框检测器(single shot multi-box detector,SSD)思想,将文本分解为两个局部可检测元素,检测多个元素并预测连接方式,通过直线拟合原则以处理多方向文本,但只通过直线拟合原则也致使曲线文本检测效果较差。文本连接建议网络(connectionist text proposal network,CTPN)[4]针对水平文本的检测问题,在 Faster-RCNN[5]架构上提出一种固定宽度的垂直锚回归机制,将 VGG16 与双向长短时记忆神经网络(long short term memory,LSTM)的连接模型串联后预测文本,但由于其锚框结构一般固定不变,因此这种方法难以处理多方向文本;由于四边形框的限制,使用上述方法对任意尺度、形状的文本实例往往难以处理。基于分割的方法,主要是基于像素级的分类来定位文本对象,可以描述各种形状的文本。文本蛇网络(flexible representation for detecting text of arbitrary shapes,Text-Snake)[6]通过设计文本中心线,然后通过线上多个堆叠但有序的圆盘来预测弯曲文本,以此处理弯曲文本检测的冗余问题,但对弯曲文本粘连问题仍无法处理。高效准确文本检测网络(efficient and accurate scene text detection,EAST)[7]提出一种基于 U-Net[8]结构的非极大值抑制(non-maximum suppression,NMS)算法与全卷积网络的架构,以此应对文本检测的复杂性,该架构可以利用预测像素与所归属文本边缘之间的距离来进行文本的检测,但在处理弯曲文本时会出现检测框冗余的不良现象;应用此类方法由于文本实例之间的距离较近,很难将其分开,并且大多需要通过繁琐的后处理过程,才可以把像素级的预测结果分组至检测到的文本实例中,这对于推理过程来说可能非常耗时且成本高昂。

针对以上问题,提出一种基于注意力机制的多方向文本检测方法。首先,通过应用改进的文本特征提取网络(text feature information ResNet50,TF-ResNet)作为主干网络,相比于 ResNet50 使用了分组卷积的方法,能提取文本特征的多样化信息;其次,在特征融合阶段加入文本注意力模块,使网络更加关注图像中的有用文本信息,抑制无用信息;最后,在后处理阶段,通过应用尺度扩展算法重建完整的文本区域。实验结果表明,该方法在多方向文本上的检测性能显著提升,而且在曲线、密集文本上也具备良好的检测效果。

1　基于深度学习的文本检测框架

1.1　残差模块

深层网络 ResNet 由残差模块(ResBlock)叠加而成,由于网络深化可能造成梯度爆炸现象,因此为残差模块中设计映射连接,以此保证文本信息能有效传输,防止发生网络退化现象[9]。ResBlock 由残差与映射两部分构成,$x^{[m]}$、$x^{[m+1]}$ 表示第 m 层 ResBlock 的输入、输出;conv1×1,64,256 表示卷积核大小为 1×1、输入通道数为 64、输出通道数为 256 的一个卷积层,conv1×1 用于增加和减少特征映射通道的维数,而 conv3×3 用于提取特征映射的文本特征信息,如图 1 所示。

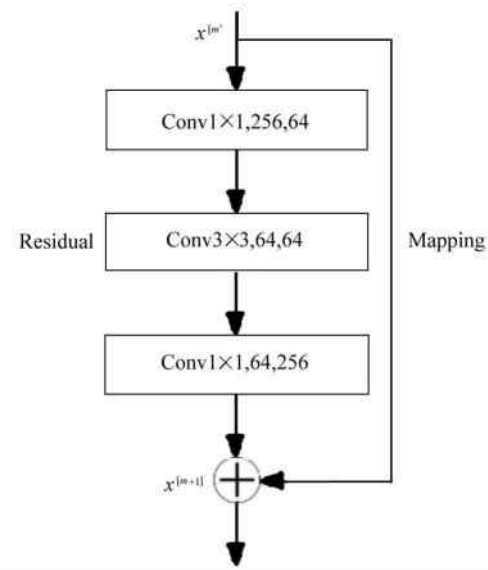

图 1　残差模块

Fig. 1　ResBlock

式(1)表示残差模块的基本原理,映射连接将网络学习 $x^{[m]}$ 到 $x^{[m+1]}$ 映射的过程变换成学习残差部分的映射过程。由于随着网络层数的不断深化,网络学习残差部分的映射比学习整个映射过程更简单,因此能够有效地提高网络的训练效果。

$$x^{[m+1]}=\sigma[F(x^{[m]})+x^{[m]}], \tag{1}$$

$$F(x^{[m]})=F_3\{F_2[F_1(x^{[m]})]\}, \tag{2}$$

式中,F_3,F_2,F_1 分别代表 conv1×1,conv3×3,conv1×1。

1.2　注意力机制

在计算机视觉领域,注意力机制是科研工作者

们关注研究的热点之一。在深度学习中，注意力机制类似于人类的视觉机制，通过多种操作为目标特征图分配权重，选择性接受和处理信息，使得网络在抑制无关信息干扰的情况下，更加关注目标图像的文本区域，从而筛选特征信息，以此达到对文本特征信息的挑选工作。目前，按照注意力应用于域的不同，主要分为通道注意机制与空间注意机制两种，能够分别从通道与空间维度筛选特征信息[10]。

空间注意力实质上是使用空间转换模块将图像中的空间信息转换至另一个空间，为其保存关键文本信息，并为每个文本信息的位置生成一个权重掩码，再对输出进行加权操作，通过此种方式增强对目标文本区域的感兴趣程度，而弱化图像中的无关背景区域；而通道注意力通过学习不断调整每个通道的权重，能够通过网络学习这种途径，自动获得每个特征通道的重要性，最后，为每个通道分配不同的权重因子，使网络更加关注重要的文本区域。

2　基于注意力机制的文本检测算法

本文提出的文本检测网络框架如图 2 所示，其中主要包括特征提取、特征融合和后处理 3 部分。在本文中，首先，使用平行堆叠多个卷积的 ResNet50，相比于传统的 VGG、ResNet 网络，网络参数更少，网络复杂度也更低。作为提取文本特征信息的基础网络 TF-ResNet，该部分主要将改进后的 ResNet50 的 conv2_x 到 conv5_x 的特征送进特征金字塔网络（feature pyramid networks，FPN）来得到基础特征；而特征融合部分由特征金字塔和文本注意模块（text attention module，TAM）组成，使用 FPN 融合基础网络提取特征后，额外添加一个 TAM，能够提升文本预测精度；在后处理阶段，使用 PSENet 的渐进尺度扩展模块，逐步整合前一部分

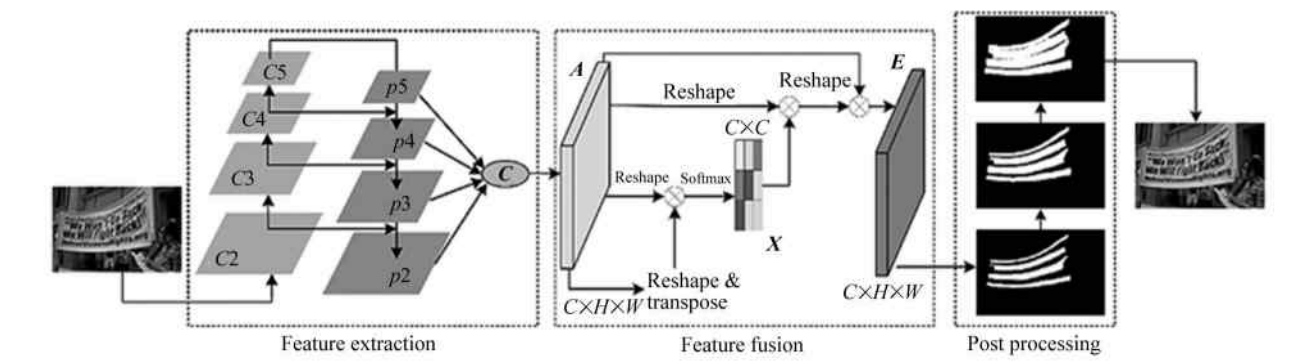

图 2　整体网络结构

Fig. 2　Overall network structure

生成的多个不同尺度的分割结果，以此获得完整的文本区域。

2.1　文本特征提取网络

在场景文本检测任务中，由于 ResNet、VGG 等网络的感受野大小有限且缺乏跨通道交互，无法较好进行特定场景文本检测任务。通常情况下，研究者会通过增加网络的宽度或者是深度（ResNet-101，ResNet-152）来提取文本目标的语义信息，但大量增加网络的参数会影响计算速度。而从 ResNext 中可以看出，增加网络基数比增加网络层数更好，可以在不增加网络参数量的情况下更好地提升网络特征的描述功能[11]。基于此情况，本文在网络层数较深的 ResNet50 上增加一定的基数，以提高网络的性能。

本文将改进的 ResNet50 定义为 TF-ResNet，参考分组－转换－融合理念，因为文本特征信息的主要提取部分为 ResBlock 中的 conv3×3，所以，本文把 conv3×3 改换为多个具有相同结构的卷积层组，并对其进行并行叠加，如图3所示。图4（a）显示了普通卷积的卷积过程，输出特征图的单条通道却需

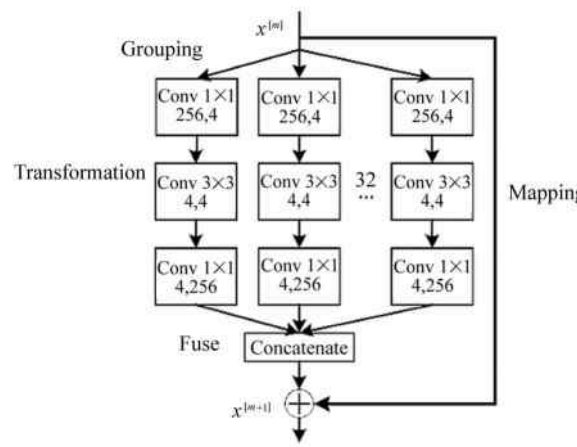

图 3　TF-ResNet 的 ResBlock

Fig. 3　The ResBlock of TF-ResNet

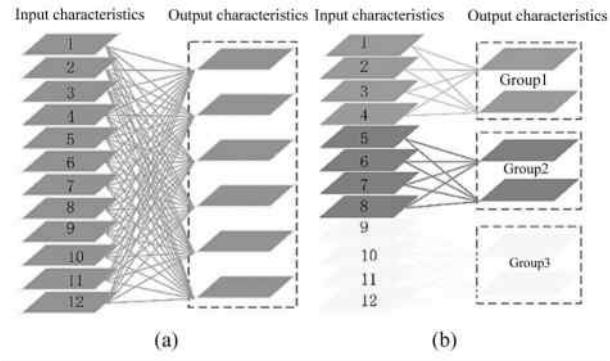

图 4　卷积过程:(a) 普通卷积; (b) 分组卷积

Fig. 4　Convolution process: (a) Ordinary convolution; (b) Grouping convolution

要输入特征图全部通道来进行计算。图 4(b)显示了并行堆叠操作的卷积过程。通过分组卷积(grouping convolution),将 64 个通道的 conv3×3 平均分成 32 组 4 个通道的 conv3×3。其中,不同的卷积层组能够表示为不同的子空间,而不同的子空间能够学习到的特征信息也具有不同的侧重点,即提取了文本特征的多样化信息,以此能够有效地应对检测过程中相似形态干扰、复杂的背景等挑战。

2.2　TAM

由于注意力机制在文本检测中能够突出显示重点信息、减少无关信息对检测结果造成的干扰,基于此特性,本文通过引入注意力机制来提高文本检测的准确率。

注意力机制已被证实在 CNN 中提取鲁棒特征方面有巨大的潜力,但是,现有的大多数办法专注于利用复杂的注意力机制来达到更好的特征提取效果,这不可避免地增加了计算量。DANet[12] 通过加入通道注意模块和空间注意模块两种模块,分别捕获通道与空间维度之间的全局依赖关系,用来捕获远程上下文信息,以此达到分割结果的准确性。

事实上,场景文本检测任务既要求高性能,也要求高效率,但是特征图中每个通道往往具有不同的重要性,因此平等地使用每个通道的特征信息,不可避免地会导致计算资源的浪费并限制网络特征的表达能力。因此,本文的文本检测模型分离出通道注意模块,用于以不同方式处理特征图中不同通道的特征信息,在不增加计算量和大幅降低推理速度的情况下用来提高文本预测精度。

本文将引入的通道注意模块定义为 TAM,其结构如图 5 所示。网络直接从原始特征图 $\boldsymbol{A} \in \boldsymbol{R}^{C\times H\times W}$ 计算文本注意图 $\boldsymbol{X} \in \boldsymbol{R}^{C\times C}$,其中,$C$ 为特征图的通道数,H 和 W 分别代表此特征图的高度与宽度。将 $\boldsymbol{A}$ 重塑为 $\boldsymbol{R}^{C\times N}$,接下来,将 $\boldsymbol{A}$ 与 $\boldsymbol{A}$ 的转置矩阵执行乘法操作,最后,通过应用 softmax 层得到文本注意图,可以表达为:

$$x_{ij} = \frac{\exp(\boldsymbol{A}_i \cdot \boldsymbol{A}_j)}{\sum_{i=1}^{C}\exp(\boldsymbol{A}_i \cdot \boldsymbol{A}_j)}, \tag{3}$$

式中,x_{ij} 表示网络第 i 个信道对第 j 个信道的影响。此外,通过在文本注意图 $\boldsymbol{X}$ 和原特征图 $\boldsymbol{A}$ 的转置之间进行一个矩阵乘法操作,再把结果乘以比例参数 α,并对 $\boldsymbol{A}$ 进行元素求和得到最终输出 $\boldsymbol{E} \in \boldsymbol{R}^{C\times H\times W}$:

$$\boldsymbol{E}_j = \alpha\sum_{i=1}^{C}\exp(x_{ij}\boldsymbol{A}_i) + \boldsymbol{A}_j, \tag{4}$$

式中,α 从 0 逐步增加权重,由式(4)可以看出,所有通道的特征与原始特征的加权和构成了每个通道的最终特征,模拟了特征图之间的长期语义依赖关系。

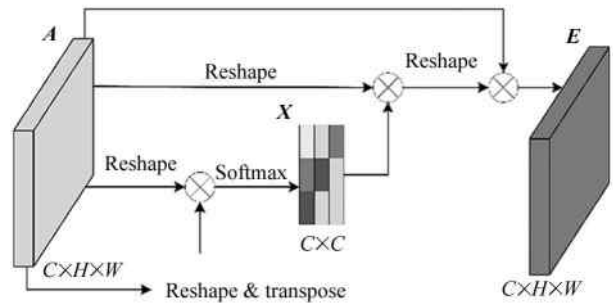

图 5　文本注意模块

Fig. 5　Text attention module

如图 5 所示,本文的 TAM 对特征融合的输出进行处理,用以增强特征的表达能力,通过此种方法,有助于提高文本特征的可识别性。

2.3　后处理模块

由于基于分割的办法难以将相邻的文本实例分离出来,因此,从广度优先搜索(breadth first search, BFS)中得到灵感,网络能够从一个顶点开始,沿径向优先遍历其四周的更大区域空间,渐进尺度扩展模块就是采用此思想来优化此问题。

其中,渐进尺度扩展的实现过程如图 6 所示,首先分别采用不同的色彩将不同文本实例的内核标记,然后使每个连通域逐像素进行尺度扩展,当出现冲突像素时内核按照先到先得原则进行合并。

图 7 为本文算法的后处理过程,其中 EX 表示尺度扩展算法,图(a)、(e)和(f)分别指 $p1$、$p2$ 和 $p3$,(b)为初始连通分量,图(c)和(d)是逐步扩张的结果。首先,生成 3 个分割结果 $p = \{p1, p2, p3\}$,$p1$ 为最原始的分割结果(最小内核),其内部生成了 4 个

不同的连通域 $D=\{d1, d2, d3, d4\}$，其中，不同的灰度区域表示网络生成了不同的连通分量，此时，能够检测到所有文本实例的最小内核。图7(b)表示最初的检测结果，再将属于 $p2$ 但不属于 $p1$ 中内核的像素点进行分配。在 $p2$ 的内核范围内，将从图7(b)中所找到连通域的每一个像素点以BFS的方式，逐一向四周扩展，以此逐步扩展 $p1$ 中预测的文本行区域，对于 $p2$ 中的内核重复上述操作，以此得到两种比例尺扩展结果，如图7(c)和图7(d)所示。最终，网络提取较好的连接分量，作为文本实例的最终预测，如图7(d)所示。

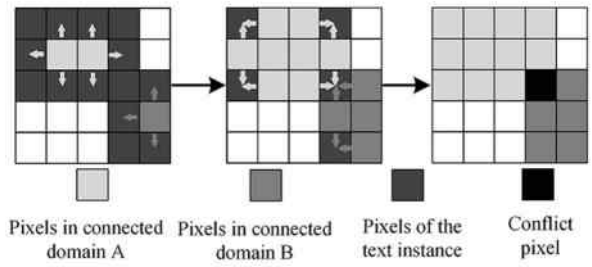

图6　渐进尺度扩展的实现过程

Fig. 6　The realization process of progressive scale expansion

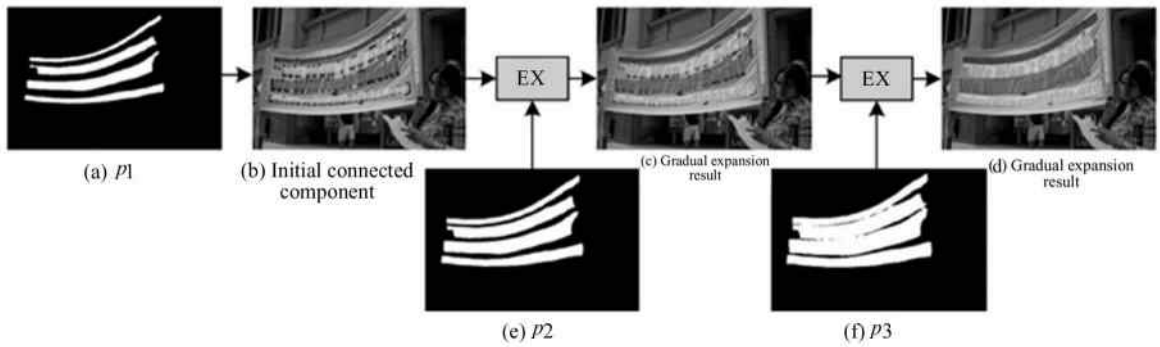

图7　后处理过程

Fig. 7　Post processing process

3　实验结果及分析

本文算法实验环境主要基于Pytorch深度学习框架，使用Linux操作系统，编程语言为Python3.7，CPU为i7-7800X，GPU为NVIDIA GTX 2080Ti×2，内存大小为64 G。所有网络采用Adam优化器进行优化，初始学习率、权重衰减系数分别设置为0.001、5×10^{-4}，批量大小设置为12，迭代36 000次，第12 000次迭代和第24 000次迭代时学习率衰减为前次学习率的1/10。

3.1　数据集

为验证本文方法的有效性，在公开数据集ICDAR2015[13]、CTW1 500[14]上进行实验，并与其他先进的文本检测方法进行比较。

ICDAR2015是由1 500张图片组成的多方向文本数据集，其中训练与测试图像数比值为2∶1，由于图像中的背景环境都是随机的自然场景，所以场景中的文本方向是任意的。

CTW1500是由1 500张图片组成的曲线文本数据集，其中训练与测试图像数比值为2∶1，每幅图像至少有一条曲线文本线。数据集中的文本实例由一个多边形标记，该多边形有多达14个点位，因此能够表述任意曲线文本的形状。

3.2　性能评价指标

场景文本检测算法的性能评价指标包括：准确率(precision，P)、召回率(recall，R)以及综合评价指标 F 值(F-measure，F)。其中 R 是正确检测文本数与正样本数的比值，P 为检测结果中正确检测的文本与实际被检测文本的总数的比值，F 值是 P 和 R 的调和平均值，当 F 值越高时，表明算法对文本的综合检测性能越好。这3个评价指标的计算式分别为：

$$P=\frac{TP}{TP+FP}, \tag{5}$$

$$R=\frac{TP}{TP+FN}, \tag{6}$$

$$F=\frac{2\times P\times R}{P+R}, \tag{7}$$

式中，TP 为正确检测数，FP 为错误检测数，$TP+FP$ 代表真实框的集合，FN 是网络未能检测到的正样本，$TP+FN$ 是所有标注的真实样本。

3.3　消融实验

为了验证TF-ResNet与TAM的有效性，本文在学习率、训练批次等参数相同的情况下在IC-

DAR2015 数据集上进行消融实验的对比,其中 ResNet50 为本次实验对 PSENet 算法的复现结果,并以此作为基准线,TF-ResNet 为仅加入文本特征提取网络的网络结构,ResNet50＋TAM 为仅加入 TAM 的网络结构,TF-ResNet＋TAM 为本文提出的网络结构,实验的结果如表 1 所示。

表 1　不同网络框架对比实验　(%)

Tab. 1　Comparative experiment of different network frameworks　(%)

Backbone	*P*	*R*	*F*
ResNet50	81.5	79.7	80.6
TF-ResNet	81.7	80.2	80.9
ResNet50＋TAM	83.2	80.8	82.0
TF-ResNet＋TAM	84.8	81.3	83.0

从表 1 结果可以看出,与 PSENet 算法相比,仅引入 TF-ResNet 模块,准确率、召回率、*F* 值分别有了 0.2%、0.5%、0.3%的提升;当仅采用 TAM 模块,本文模型的准确率、召回率、*F* 值分别有 1.7%、1.1%、1.4%的提升;而两者都采用时,本文方法的召回率、准确率、*F* 值比 PSENet 算法分别提升了 3.3%、1.6%、2.4%。消融实验结果表明,TF-ResNet 模块与 TAM 的引入,有效地提升了网络模型的检测性能。

3.4　对比实验

为更好验证本文网络框架的有效性,分别在数据集 ICDAR2015、CTW1500 上将本文所提出的方法与目前文本检测的其他多种方法(SegLink[3]、CTPN[4]、TextSnake[6]、EAST[7]、WordSup[15]、PSENet[16])进行了比较,结果如表 2、表 3 所示,表中对比方法的数据均来自于其对应的论文。

表 2 为不同方法在 ICDAR2015 数据集上的文本检测性能对比。根据表 2 数据可知,本文方法的召回率、准确率、*F* 值达到 81.3%、84.8%、83.0%,比 PSENet 算法分别提升了 1.6%、3.3%、2.4%。此外,与先前其他方法相比,本文所提出方法在准确率、召回率、*F* 值上均优于之前的算法,从而证明本文算法对于倾斜文本和常规文本检测的有效性。如图 8 为本文算法在 ICDAR2015 数据集上测试的部分可视化结果。

表 3 为不同方法在 CTW1500 数据集上的文本检测性能对比。通过表 3 数据可知,本文方法的召回率、准确率、*F* 值达到 81.3%、84.8%、83.0%,比 PSENet 算法分别提升了 2.6%、2.1%、2.4%。此外,与其他方法相比,本文方法在准确率、*F* 值上均优于之前的先进算法,这是因为在数据集中许多文本太近,甚至重叠,很难将其分开,在使用渐近尺度扩展算法的情况下,算法改进了骨干网并加入 TAM 从而增强模型的特征提取能力,证明本文算法对于曲线文本检测的有效性。图 9 为本文算法在 CTW1500 数据集上测试的部分可视化结果。

表 2　不同算法在 ICDAR15 数据集上的性能对比　(%)

Tab. 2　Performance comparison of different algorithms on the ICDAR15 dataset　(%)

Methods	*P*	*R*	*F*
CTPN	74.2	51.6	60.9
SegLink	73.1	76.8	75.0
EAST	83.6	73.5	78.2
WordSup	77.0	79.3	78.2
PSENet	81.5	79.7	80.6
Ours	84.8	81.3	83.0

表 3　不同算法在 CTW1500 数据集上的性能对比　(%)

Tab. 3　Performance comparison of different algorithms on the CTW1500 dataset　(%)

Methods	*P*	*R*	*F*
CTPN	60.4	53.8	56.9
SegLink	42.3	40.0	40.8
EAST	78.7	49.1	60.4
TextSnake	67.9	85.3	75.6
PSENet	80.6	75.6	78.0
Ours	82.7	78.2	80.4

图 8　ICDAR 2015 数据集上测试的部分可视化结果

Fig. 8　Partial visualization results of tests on ICDAR 2015 dataset

图 9 CTW1500 数据集上测试的部分可视化结果

Fig. 9 Partial visualization results of tests on CTW1500 dataset

3.5 缺点与不足

尽管本文方法在多个数据集上取得较好效果，但是由于场景文本背景的复杂性以及字体的多样性，对部分极端文本（椭圆标出），如小文本（见图 10(a)）和大字符间距（见图 10(b)），仍然会出现漏检失败例子。对于非常小的文本这是因为在经过多次的下采样后，小的区域文本变得更小，会被当成噪声过滤掉。对与字符间距过大的文本，这是由于缺乏足够的上下文信息以及语言模型导致的，即使是人也较难正确分辨。对于以上存在的问题，主要原因是缺乏大量的训练样本，使得模型无法学习到相关的知识导致的，当有足够训练样本时，这些问题会得到缓解。

图 10 失败例子

Fig. 10 Failure examples

4 结 论

本文提出了一种高效的基于注意力机制的多方向文本检测算法，所用模型采用 TF-ResNet 作为骨干网络，能够更准确地提取语义信息，在特征融合阶段引入 TAM，增强模型对文本信息的可识别性，在后处理阶段，使用了渐进尺度扩展模块，增强对稀疏排列任意形状文本的检测后处理能力。最终实验结果表明，本文所设计的方法在当前场景文本检测算法中表现良好，在多方向和弯曲的文本检测中均有较好效果。由于本文算法在追求精度的同时，丧失了一定的速度，因此，在未来的研究中，将进一步地优化算法，并与文本识别方法相结合，从而设计一种完整的端到端的文本检测与识别算法。

参考文献：

[1] LI Y X, MA J W. The developments and challenges of text detection algorithms[J]. Journal of Signal Processing, 2017, 33(4): 558-571.
李翌昕, 马尽文. 文本检测算法的发展与挑战[J]. 信号处理, 2017, 33(4): 558-571.

[2] LIU Y, JIN L, ZHANG S, et al. Curved scene text detection via transverse and longitudinal sequence connection [J]. Pattern Recognition, 2019, 90: 337-345.

[3] SHI B, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 2550-2558.

[4] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C]// European Conference on Computer Vision, October 11-14, 2016, Amsterdam, The Netherlands. Cham: Springer, 2016: 56-72.

[5] SUN X, WU P, HOI S C H. Face detection using deep learning: An improved faster RCNN approach[J]. Neurocomputing, 2018, 299: 42-50.

[6] LONG S, RUAN J, ZHANG W, et al. Textsnake: A flexible representation for detecting text of arbitrary shapes [C]//Proceedings of the European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Cham: Springer, 2018: 20-36.

[7] ZHOU X, YAO C, WEN H, et al. EAST: An efficient and accurate scene text detector[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 5551-5560.

[8] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-assisted Intervention, October 5-9, 2015, Munich, Germany. Cham: Springer, 2015: 234-241.

[9] HAO J, ZHENG Z W, SUN Z A, et al. Improved moving

target detection based on Yolo and residual network algorithm [J]. Journal of Optoelectronics · Laser, 2020, 31(1): 81-88.
郝骏,郑紫微,孙滋昂,等. 基于 Yolo 与残差网络算法改进的运动目标检测[J]. 光电子 · 激光, 2020, 31(1): 81-88.

[10] XU G Y, YIN M Y. Improved ssd object detection algorithm based on space-channel attention[J]. Journal of Optoelectronics · Laser, 2021, 32(9): 970-978.
许光宇, 尹孟园. 基于空间-通道注意力的改进 SSD 目标检测算法[J]. 光电子 · 激光, 2021, 32(9): 970-978.

[11] XIE S, GIRSHICK R, DOLLÁR P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 1492-1500.

[12] FU J, LIU J, TIAN H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 3146-3154.

[13] LIU Y, JIN L, ZHANG S, et al. Curved scene text detection via transverse and longitudinal sequence connection [J]. Pattern Recognition, 2019, 90(10): 337-345.

[14] MA J, SHAO W Y, YE H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[15] HU H, ZHANG C, LUO Y, et al. WordSup: Exploiting word annotations for character based text detection [C]//Proceedings of the IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4940-4949.

[16] WANG W, XIE E, LI X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2019: 9336-9345.

作者简介:

徐　健　(1963—),男,教授,硕士生导师,主要从事图像处理、机器视觉方面的研究.