面向自然街景改进的文本检测

丁　泽,程艳云
(南京邮电大学 自动化学院、人工智能学院,江苏 南京 210023)

摘　要:近年来,随着深度学习的发展,在自然街景下的文本检测取得了巨大的进步,但在多方向和弯曲文本及对比度低的文本检测中的效果仍不理想。因此,针对弯曲文本和对比度低的文本的检测问题,提出了一种融合多尺度模块的文本检测方法,并通过检测效果的提升,提高端到端文本识别的识别效果。针对 RFB(Receptive Field Block)模块在下采样后局部信息丢失的问题,在 RFB 模块中嵌入极化自注意力(Polarized Self-Attention)机制以改进 RFB 来提取有效文本特征,提高特征图表征效果。针对特征金字塔(FPN)提取的特征不足、感受野小的问题,将改进的 RFB 模块嵌入特征金字塔(FPN)模块以增强特征提取融合。针对特征分布不确定性及远距离特征融合效果不佳的问题,引入条形池化(Strip Pooling)模块,进而提升检测方法的鲁棒性。在公开数据集 Total-Text 上的实验结果表明,该算法的 F-measure 值在端到端文本识别没有词汇表的情形下与目前高效的 MaskTextSpotterV3 相比高了 0.3 百分点,而在有词汇表的情形下则高出了 0.2 百分点;而在仅文本检测的情形下,该方法也有较为良好的表现。

关键词:文本检测;特征金字塔;极化自注意力;RFB 模块;条形池化模块
中图分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1673-629X(2023)04-0082-07
doi:10.3969/j.issn.1673-629X.2023.04.012

Improved Text Detection for Natural Streetscape

DING Ze,CHENG Yan-yun
(School of Automation,School of Artificial Intelligence,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)

Abstract:In recent years,with the development of deep learning,great progress has been made in text detection under natural streetscape,but the effect in multi-directional and curved text and text with low contrast is still unsatisfactory. Therefore,we propose a text detection method incorporating a multi-scale module for the detection of curved text and text with low contrast,and improve the recognition effect of end-to-end text recognition through the improvement of detection effect. To address the problem of local information loss in the RFB (Receptive Field Block) module after downsampling,a Polarized Self-Attention (PSA) mechanism is embedded in the RFB module to improve the RFB to extract effective text features and improve the feature map representation. To address the problem of insufficient features and small perceptual fields extracted by the feature pyramid (FPN),the improved RFB module is embedded in the feature pyramid (FPN) module to enhance feature extraction fusion. To address the problems of uncertain feature distribution and poor fusion of long-range features,a Strip Pooling module is introduced to improve the robustness of the detection method. Experimental results on the publicly available dataset Total-Text show that the F-measure value of the proposed algorithm is 0.3% higher than that of the current efficient MaskTextSpotterV3 in the case of end-to-end text recognition without vocabularies,and 0.2% higher in the case of vocabularies. In the case of text-only detection,it also has a better performance.

Key words:text detection;feature pyramid;Polarized Self-Attention (PSA);RFB module;strip pooling module

0　引　言

文本在人机交互中扮演着重要的角色,随着智能机器人、无人驾驶、医疗诊断的飞速发展,文本的检测与识别已经成为定位和理解物体信息的重要途径。

经典的文本检测方法可分为两大类:基于连通域分析的文本检测方法和基于滑动窗口的文本检测方法。然而,基于连通域的方法对噪声的包容性较差,而基于滑动检测窗的方法虽然可以避免该问题,但该方法却对滑窗依赖极大,通用性不强。近年来,出现了大量的基于深度学习的自然场景文本检测方法,这些方

收稿日期:2022-06-26　　**修回日期**:2022-10-27
基金项目:国家自然科学基金青年科学基金项目(61802204)
作者简介:丁　泽(1996-),男,硕士研究生,研究方向为图像处理与文本识别;通讯作者:程艳云,硕士,副教授,研究方向为模式识别、机器学习、文字识别。

法多采用 2 种深度学习图像处理策略:(1)目标检测算法中得到区域建议的策略;(2)图像语义分割策略。

基于区域建议的方法一般以通用目标检测网络作为基本模型,并在此基础上结合实际应用对算法进行改良。2017 年 Liao 等人[1]提出的 TextBoxs 网络可根据不同卷积层的多尺度特征有效检测出不同尺度文本。2018 年, Liao 等人[2]又在此基础上提出了 TextBoxs++文本检测模型,利用旋转角度的倾斜文本框实现不规则的文本检测窗。2019 年,Zhong 等人[3]提出一种无锚区域建议网络(AF-RPN)替代 Faster R-CNN 中的基于参考框的区域建议方法。该方法能够摆脱复杂的参考框设计,在水平和多方向文本检测任务中均取得了更高的召回率。2020 年,Wang 等人[4]提出了 ContourNet 文本检测模型,该模型设计了一种与尺度无关的自适应区域建议网络(Adaptive-RPN),该网络能有效地解决算法产生的伪召回及对尺度变化剧烈的文本检测不准确的问题。然而,上述方法在检测任意形状或极端纵横比的文本时效果依旧不理想。

基于分割的方法以语义分割为基本技术手段,通过深度学习语义分割网络对自然场景图片进行处理,获取像素级别的标签预测。2018 年,Deng 等人[5]提出 PixelLink 模型,采用实例分割的方法,分割出文本行区域,然后直接找对应文本行的外接矩形框,但其需针对不同数据集调整 pixel 和 link 的阈值,并设计不同的后处理方法,且无法处理背景复杂的数据。2019 年,Xu 等人[6]提出 Text Field 来学习一个方向场来链接相邻像素,并使用一个简单的基于形态学的后处理来实现最终检测,但其后处理过程过于复杂,模型的检测速度很慢。2019 年,Wang 等人[7]提出了 PAN 模型,通过像素聚合的方式来让网络学习文本相似性矢量,有选择地聚合文本内核附近的像素,有效地提升了文本的检测速率但对任意形状的文本检测不够鲁棒。2021 年, Wang 等人[8]在 PAN 的基础上又提出了 PAN++网络,该网络展示了一种基于文本内核的任意形状文本的表示方法,不仅能够描述任意形状的文本,还能在保持精度的同时实现较高的推理速度,但该方法表征能力较弱,在应对极端纵横比和旋转文本的效果不佳。2020 年, Liao 等人[9]提出的 MaskTextSpotterV3 采用 ResNet50 作为主干网络,能有效地提取文本特征,并且该模型设计了一个无锚分割建议网络,可以提供对任意形状建议的准确描述,并且在检测旋转、极端长高比或不规则形状的文本实例时具有鲁棒性,但该方法因感受野较小且在特征融合阶段将不同尺度特征直接融合,故在处理极端纵横比、大尺度文本检测时容易出现漏检、误检的现象且易引入过多的噪声,影响模型对小尺度文本的检测效果。

为解决以上问题,该文在 MaskTextSpotterV3 的基础上提出了一种融合多尺度模块的文本检测方法(text detection method incorporating multi-scale modules,IMSM)。该检测方法采用改进的特征提取模块和改进的特征融合模块,在有效扩大感受野的同时抑制噪声信息,能有效地捕捉中长文本的特征信息,减少漏检、误检的现象且对极端纵横比的文本具有鲁棒性。

1　融合多尺度模块的文本检测网络

1.1　总体网络架构

该文提出的 IMSM 模块如图 1 所示,具体分为三个模块,分别是改进的特征提取模块、改进的特征融合模块和分割候选模块。主要内容如下:为了平衡模型的体积和检测效果,采用 Resnet50 作为主干网络,同时将 FPN 与改进的感受野模块(receptive field block for integrating attention,RFBIA)相融合以扩大感受野、捕捉中长文本的特征信息。针对 RFB 模块[10]下采样融合后与输入特征图相加引入过多的噪声信息,嵌入极化自注意力机制[11](polarized self attention,PSA)对特征进行处理,以提取有效的文本特征。针对特征分布不确定性及远距离特征融合效果不佳的问题,在特征融合模块中引入条形池化(strip pooling module,SPM)模块[12]来捕获更长距离之间的依赖关系,以此提升检测方法的鲁棒性。

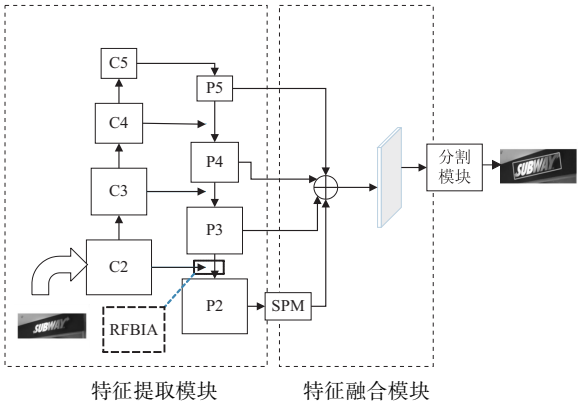


图 1　IMSM 结构

1.2　特征提取与多尺度模块

在特征金字塔网络对主干网络提取的高语义特征和高分辨率特征进行融合时,由于采用 3 * 3 的卷积,其对于极端纵横比、大尺度文本的融合效果较差,易造成漏检、误检的现象。为解决此问题,该文将融合后的高语义特征和高分辨率特征送入 RFBIA 模块,通过扩大感受野来对大尺度文本进行检测,同时 RFBIA 模块也能有效抑制因为扩大感受野而引入的噪声信息,提取有效特征,从而提高文本检测效果。

RFBIA 模块如图 2 所示,RFB 模块由多分支卷积

层和膨胀卷积层组成,图中用大小不同的圆形表示不同尺寸卷积核构成的卷积层;膨胀卷积层的作用在于增加感受野,图中用不同的 rate 表示膨胀卷积层的参数。其中,多分支卷积层使用多种尺寸的卷积核来实现,相比于固定尺寸的卷积核而言,多尺寸的卷积核提取的信息更加丰富,从而能尽量避免信息的丢失。每个分支的卷积层后面会级联一个膨胀卷积层,膨胀卷积层在保持参数量的同时能扩大感受野,用来获取更高分辨率的特征。

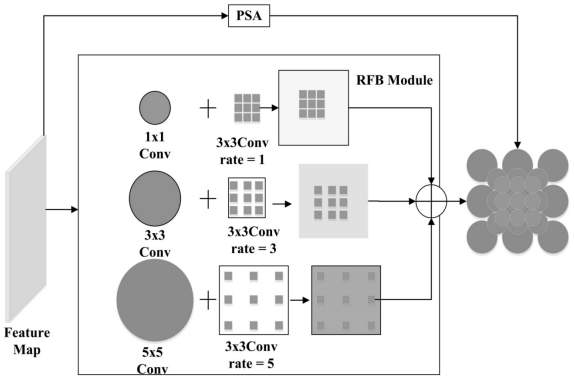


图 2 RFBIA 结构

然而,在 RFB 模块下采样得到高语义信息并扩大感受野的同时,由于分辨率的降低会丢失输入图像的部分特征信息。为了精确地从特征图像中分割出文本信息,需要底层的特征图提供重要的细节信息和边缘信息,所以该文设计将输入特征图通过一个极化自注意力机制(PSA)来提供所需的细节信息和边缘信息。输入特征图经过 PSA 模块后提取出丰富的局部信息和边缘信息;而 RFB 模块扩大感受野后,提取出不同尺度的空间信息,得到包含高语义、抽象化的特征信息的输出,将两者提取出的信息相融合以进行联合预测,从而提高检测效果。

在 RFBIA 模块中,为有效地提取重要的细节信息和边缘信息,并联了一个精细的双重注意力机制(PSA)。PSA 采用了一种极化滤波(polarized filtering)的机制,类似于光学透镜过滤光一样,每个自注意力的作用都是用于增强或抑制特征,该机制在通道和空间维度能保持较高的分辨率,这能够减少降维所造成的信息损失。该模块还在通道和空间分支中采用了 Softmax 和 Sigmoid 相结合的非线性函数,从而能够拟合出细粒度回归结果的输出分布,如图 3 所示。PSA 分为两个分支,一个分支做通道维度的自注意力机制,另一个分支做空间维度的自注意力机制。两分支采用并行的方式来获取注意力权重,这充分利用了自注意力结构的建模能力,在保证计算量的情况下,实现了一种非常有效的长距离建模。输入的特征再对分别经过这两个分支后产生的结果进行融合就得到了极化自注意力结构的输出。

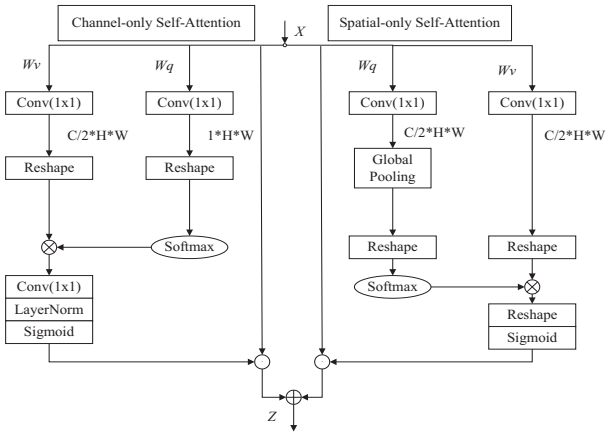


图 3 PSA 网络

通道维度的自注意力机制中,输入的特征会经过一个 1 * 1 的卷积将特征 X 转换成 $Q(C/2\times H\times W)$ 和 $V(1\times H\times W)$,其中 Q 通道被完全压缩,而 V 的通道维度依旧保持在 $C/2$ 的水平,由于 Q 的通道维度被完全压缩,故而采用 Softmax 对 Q 通道的信息进行增强。然后将 Q 和 V 进行矩阵乘法,特征图大小变为 $C/2\times1\times1$,然后特征图再经过一个 1×1 的卷积和 LayerNorm 层将通道维度从 $C/2$ 上升为 C。最后使用 Sigmoid 函数使得所有的参数都保持在[0,1]的范围内。通道维度的注意力权重如下:

$$A^{\mathrm{ch}}(X)=F_{\mathrm{SG}}[W_{Z|\theta_1}(\sigma_1(W_v(X)))\times F_{\mathrm{SM}}(\sigma_2(W_q(X))))]$$

其中,W_q、W_v、W_z 均为 1×1 的卷积层,σ_1、σ_2 是两个张量 reshape 操作,而 $F_{\mathrm{SM}}(\cdot)$ 代表 Softmax 运算,×则代表矩阵乘法运算。通道分支的输出结果则为通道权重与输入特征的逐通道相乘。

与通道维度的自注意力机制相似,空间自注意力机制中输入的特征图也是先经过一个 1×1 的卷积,将特征转换为 $Q(C/2\times H\times W)$ 和 $V(C/2\times H\times W)$,其中特征 Q 采用了全局池化来对空间维度进行压缩转换成 1×1 的大小,而特征 V 的空间维度则保持在 $H\times W$ 的水平。由于特征 Q 的空间维度被完全压缩,故而在全局池化后使用 Softmax 函数对 Q 的信息进行增强。然后再将 Q 和 V 进行矩阵乘法,将输出结果进行 reshape 和 Sigmoid 操作后,使得所有的参数都保持在[0,1]之间。空间维度的注意力权重如下:

$$A^{\mathrm{sp}}(X)=F_{\mathrm{SG}}[\sigma_3(F_{\mathrm{SM}}(\sigma_1(F_{\mathrm{GP}}(W_q(X)))))\times\sigma_2(W_v(X)))]$$

其中,W_q 和 W_v 是 1×1 的卷积,σ_1、σ_2 和 σ_3 表示三个张量 reshape 操作,$F_{\mathrm{SM}}(\cdot)$ 表示 Softmax 操作,$F_{\mathrm{GP}}(\cdot)$ 表示全局池化函数,×表示矩阵点积运算。

以上两个分支并联运算输出的结果为 $\mathrm{PSA}(X)=A^{\mathrm{ch}}(X)\odot X_{\mathrm{ch}}+A^{\mathrm{sp}}(X)\odot X_{\mathrm{sp}}$,其中+代表逐元素相加。

1.3　特征融合模块

在特征金字塔(FPN)融合高层信息和底层信息后,融合的特征图将送到后续分割模块中进行分割以进行文本的检测与后续的识别,这就需要对输出的特征进行融合,将多尺度的特征融合到一张特征图中。由于自然场景中的文本信息大多呈长条形,或离散分布,为解决特征分布不确定性及远距离特征融合效果不佳的问题,该文在特征融合中引入 SPM 来捕获更长距离之间的依赖关系,以此提升检测方法的鲁棒性。该模块与 RFBIA 模块相互补充,提升了整个网络的性能。

SPM 是一个新的池化策略,该策略采用了一个长而窄的核即 $1\times N$ 或 $N\times 1$,以此来捕获场景像素级预测任务的远程上下文信息,输入的特征图大小为 $C\times H\times W$,图 4 所示为一个通道的处理过程。

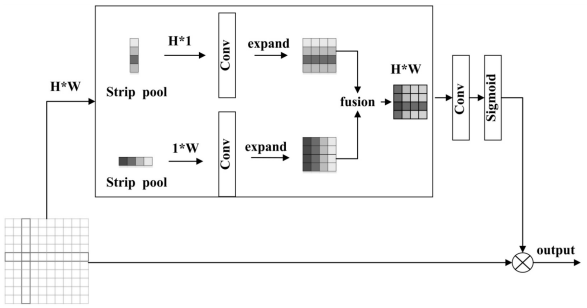


图 4　条形池化网络

SPM 由两条路径组成,它们分别侧重于沿着水平和垂直空间两个维度捕获远程上下文,这样的好处是在一个方向上可以得到 non-local 的信息,而在另一个方向上则可以得到细节的信息。输入的特征图经过水平和竖直条纹池化后变为 $H\times 1$ 和 $1\times W$,之后对池化核内的元素值求平均,并以此平均值作为池化输出值,其中水平方向上的运算公式为 $y_i^h=\frac{1}{w}\sum_{0\leqslant j\leqslant W}x_{i,j}$,竖直方向的运算公式为 $y_i^v=\frac{1}{H}\sum_{0\leqslant i\leqslant H}x_{i,j}$,随后经过卷积核为 3 的 1D 卷积对两个输出的特征图分别沿着左右和上下的方向进行扩容,扩容后的两个特征图尺寸相同,最后将两个尺寸相同的特征图按元素进行相加得到一个 $H\times W$ 的特征图。该特征图在经过一个 1×1 的卷积层和 Sigmoid 处理后与原输入图像进行元素乘法以得到最终的输出结果 $z=\mathrm{Scale}(x,\sigma(f(y)))$。其中 $\mathrm{Scale}(\cdot,\cdot)$ 表示逐元素相乘,σ 表示 Sigmoid 函数,f 表示 1×1 的卷积,y 表示每个通道中的水平和竖直方向上的输出值之和,其公式为 $y_{c,i,j}=y_{c,j}^h+y_{c,j}^v$。在上述过程中,输出张量中的每个位置都与输入张量中的各个位置建立了关系。因此,通过多次重复上述聚合过程,可以在整个场景中构建长期依赖关系。

1.4　分割候选模块

分割候选模块采用了 U-net 结构,该结构沿用了全卷积网络(FCN)进行图像语义分割的思想,包括收缩路径和扩张路径,其中收缩路径用于捕获上下文,扩张路径用于精确定位。相较于 FCN 而言,U-net 在扩张路径上采样的过程中拥有更多的通道数,这使得 U-net 网络能进行多尺度的图像特征识别,将上下文的信息向更高层分辨率传播。同时,U-net 结构在上采样融合特征提取部分的输出时采用了拼接的特征融合方式,将特征在通道维度拼接在一起形成更厚的特征,这也提高了其对于尺度的鲁棒性。

与基于特征金字塔结构的区域候选网络在多个尺度的特征图上产生候选框不同,分割候选网络从分割图中生成候选区域,其中分割图由上文中融合后的特征图映射预测得到。融合后的特征图连接了不同感受野的特征映射,其大小为 $H/4\times W/4$,其中 H 和 W 分别是输入图像的高度和宽度。预测的文本分割图的大小为 $1\times H\times W$,其值在[0,1]的范围内。

1.4.1　分割标签生成

为了分离相邻的文本实例,通常采用的方法是将文字实例区域进行收缩,故而该文的分割候选网络采用了 Vatti clipping 算法来收缩文字区域,通过裁剪 d 个像素来缩小文本区域,偏移像素 d 可由公式 $d=\frac{A(1-r^2)}{L}$ 计算,其中 A 和 L 分别表示多边形文本区域得到面积和周长,r 为收缩率,根据经验将其设置为 0.4,分割标签的生成示例如图 5 所示。

图 5　分割标签生成

左图中外多边形和内多边形分别是原始注释和收缩区域,右图为分段标签,黑色和白色分别代表 0 和 1 的值。

1.4.2　候选区域生成

给定一个值在[0,1]范围内的文本分割图 S,将 S 二值化为二值图 B。如上文所述,文本分割标签被收缩,然后分割候选网络在二值图中搜索出连通的区域,这些连通区域可以被视为收缩的文本区域,之后再通过 Vatti clipping 算法取消裁剪 d 像素,以此膨胀回文字区域。如上所述,分割候选网络能够精确地产生多边形候选区域。因此,它能够为极端长宽比的文字行和密集多方向、不规则形状的文字生成合适的候选区域,同时也为后续模块提供了精确的多边形位置信息。

1.4.3 损失函数

分割候选网络采用的损失函数为 dice loss，dice 系数是一种集合相似度度量函数，通常用于计算两个样本点的相似度，其公式为 $S=\frac{2|x\cap y|}{|x|+|y|}$，$|x\cap y|$ 表示 x、y 之间的交集，$|x|$、$|y|$ 分别表示 x 和 y 的元素个数，而分子中的系数 2 是因为分母存在重复计算 x 和 y 之间的共同元素的原因。而 dice loss 用公式表达为 $L=1-\frac{2|x\cap y|}{|x|+|y|}$，其中 dice 系数越大，表明集合越相似，损失越小；反之亦然。

文中将分割图设为 S，目标图设为 G。损失函数表示为：

$$I=\sum(S*G)\ ;\ U=\sum S+\sum G\ ;\ L=\frac{2\times I}{U}$$

其中，I 和 U 分别表示分割图与目标图的交集和并集，$*$ 则代表逐元素相乘。

2 实验与分析

该文评估了所提出的改进的文本检测方法，并在不同标准场景文本基准上测试了对旋转、纵横比、小尺度文字的鲁棒性，并对提出的方法进行了消融实验。

2.1 数据集

SynthText 是一个包含 800k 文本图像的合成数据集，它为单词/字符边界框和文本序列提供了注释。

Rotated ICDAR 2013 dataset（RoIC13）是由 ICDAR2013 数据集生成的，该数据集的图像集中在文本内容周围，文本实例在水平方向上并且由轴对齐的矩形框标记，且该数据集提供了字符级的分割注释。该数据集包含 229 张训练图片和 233 张测试图片，为了测试旋转的鲁棒性，该文还创建了旋转的 ICDAR2013 数据集，方法是将 ICDAR 测试集中的图像和注释旋转到一些特定的角度。

S-CUT 是一个具有挑战性的曲线文本数据集，由 1 000 张训练图像和 500 张测试图像组成。不同于传统的文本数据集，SCUT 中的文本实例由 14 个点的多边形标记，因此它可以描述一个任意曲线文本的形状。

Total-Text 数据集包含 1 255 张训练图片和 300 张测试图片。它提供各种形状的文本实例，包括水平的、定向的和弯曲的形状。尽管 Total-Text 数据集提供了字符级的注释，但该文并未使用。

ICADR2015 数据集包含 1 000 张训练图像和 500 张测试图像，这些图像都用矩形边界框标注。该数据集中的大多数图像的分辨率较低，并且包含小文本实例。

2.2 实验细节

该文使用 SGD 来优化模型，权重衰减为 0.001，动量为 0.9。在消融实验和对比实验中使用 SynthText 预训练的 ResNet 50 模型作为主干网络，然后使用 SynthText、ICDAR2013 数据集、ICDAR2015 数据集、S-CUT 数据集和 Total-Text 数据集进行 300 000 次迭代的混合微调，这些数据集之间的采样率设置为 2∶2∶2∶1∶1。

在微调期间，初始学习率为 0.01，然后分别在 100 000次迭代和 200 000 次迭代时降低 10 倍。在推理期间，输入图像的短边在 RoIC13 数据集上调整为 1 000，在 ICDAR2015 数据集上调整为 1 440，以保持纵横比。

2.3 实验环境

实验使用 python3.7 作为编程语言，pytorch 版本为 1.4.0。所有的实验都是在 Linux18.04 操作系统进行，显卡配置为两张 NVIDIA RTX2080TI。

2.4 评价指标

利用准确率 P、召回率 R 和 F-measure 值这三个指标来衡量算法性能，其数值越大表示性能越好，即：$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$，$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$，$F=\frac{2PR}{P+R}$。其中 TP 表示正确预测的正样本数量；FP 表示本来为负样本，预测为正样本的数量；FN 表示本来是正样本，预测为负样本的数量。文中算法的效率由 FPS 来衡量，其数值表示在推理阶段每秒处理的图片数，故数值越大表示效率越高。

2.5 消融实验

为了验证 RFBIA 和 SPM 的有效性，在 ICDAR2015 和 RoIC13 数据集上分别进行了消融实验。

如表 1 所示，在添加提出的 RFBIA 后，原始网络的准确率下降了 0.7 百分点，检测速率下降了 0.3 fps，而召回率、F1 值均有所提升。在添加 SPM 后，原始网络的准确率、召回率、F1 值均有提升，其中召回率提升了 4 百分点，但检测速率下降了 0.1 fps。而在 RFBIA 和 SPM 的联合使用下，原始网络的准确率上升了 1.7 百分点，召回率上升了 4.2 百分点，F1 指标上升了 3.3 百分点，与此同时，检测速率也下降了 0.4 fps。

表 1 ICDAR2015 消融实验

骨干网络	RFBIA	SPM	P/%	R/%	F/%	FPS
ResNet50	×	×	80.2	62.6	70.3	2.9
ResNet50	√	×	79.5	63.5	70.6	2.6
ResNet50	×	√	80.9	66.6	73.1	2.8
ResNet50	√	√	81.9	66.8	73.6	2.5

如表 2 所示，在添加提出的 RFBIA 后，在旋转 45°、60°时，原始网络的准确率、召回率、F1 指标、检测

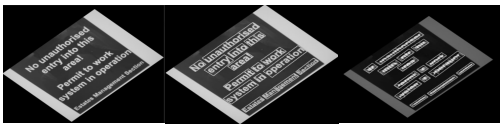
速率均有上升,其中在旋转 60°时,各项指标提升较多。在添加 SPM 后,在旋转 45°时,原始网络各项指标均有所下降,仅检测速率上升了 0.6 fps;而在旋转 60°时,原始网络的各项指标均有提升。在 RFBIA+SPM 联合使用下,在旋转 45°时,原始网络的召回率、F1 指标、检测速率有所提升,但准确率降低了 1.6 百分点;而在旋转 60°时,原始网络的准确率、F1 指标、召回率提升较大,其中召回率上升了 5.8 百分点,检测速率无变化。

表 2　RoIC13 消融实验

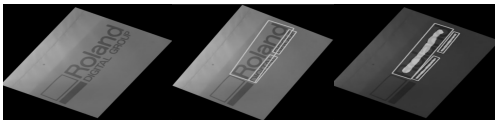
ANGLES(°)	RFBIA	SPM	*P*/%	*R*/%	*F*/%	FPS
45°	×	×	86.9	65.5	74.7	4.4
	√	×	87.5	66.4	75.5	4.8
	×	√	86.0	64.0	73.4	5.0
	√	√	85.3	68.1	75.8	4.2
60°	×	×	84.8	65.8	74.1	4.5
	√	×	87.2	67.8	76.3	4.8
	×	√	86.5	66.6	75.3	4.8
	√	√	88.0	71.6	79.0	4.5

提出的算法在 ICDAR2015 数据集和 RoIC13 数据集上消融实验的可视化测试结果如图 6 所示,其中按列从左到右分别为测试图、原始网络检测结果图和文中算法检测结果图。

(a)小尺度密集文本检测结果(ICDAR2015)



(b)旋转文本检测结果(RoIC13 45°)



(c)旋转文本检测结果(RoIC13 60°)

图 6　对比结果展示

将以上结果进行分析可得,在原始算法框架中添加了 RFBIA 模块后,由于 RFBIA 弥补了 FPN 提取特征时感受野较小的缺点,增强了模型检测大尺度弯曲文本的能力,模型的检测准确率在各数据集上均有提升,但该模块对于小尺度密集文本较多的 ICDAR2015 数据集的检测效果并不明显。在原始网络中添加 SPM 模块后,模型的各项检测指标在 ICDAR2015 数据集上有明显提升,但在其他数据集上则表现效果一般,这是因为 SPM 模块能有效捕获长距离的依赖关系,其条纹池化操作也可以认为是一种注意力机制,能有效地挖掘小尺度信息,对特征进行提取。而 RFBIA 和 SPM 的联合使用不仅增强了模型检测大尺度文本的能力,降低了特征图分辨率的损失,而且对有效文本特征信息的提取也有所增强。同时,该算法在 RoIC13 数据集旋转 60°实验中的表现也证明了所提出的算法对于旋转的鲁棒性。

2.6　对比实验

该方法与其他方法在 Total-Text 数据集上对综合评价指标 *F* 值的对比结果如表 3 所示,表 3 展示了文中方法在检测(Detection)和端到端(End-to-End)识别的情况下与其他模型的对比分析,为了使对比分析更加直观、公平,端到端识别的情况又分为无词汇表识别(None)和有词汇表识别(Full)两种情况。提出的方法在检测效果方面相比于针对处理多方向和曲线文本的 CharNet[13] 高了 0.1 百分点,相较于 MaskTextSpotter 高出了 0.5 百分点,这是因为 CharNet 和 MaskTextSpotter 采用的传统的区域建议网络对极端纵横比的文本识别效果不佳。而在进行端到端的文本定位时,文中方法在没有词汇表的情况下相较于 PAN++高出了 2.9 百分点,这是由于 PAN++采用的轻量级网络的表征能力较弱,虽然其推理速度较高,但识别精度还有待提高;相较于 MaskTextSpotter V3 高出了 0.3 百分点,体现了文中方法在无监督的情况下对文本识别效果的提升。而在有词汇表的情况下相较于 ABCNet[14] 高出了 1.2 百分点,持平于 PAN++;相较于 MaskTextSpotter V3 则高出了 0.2 百分点。MaskTextSpotter V3 在提取特征时感受野较小,而在特征融合阶段则将不同尺度的特征直接相加,容易造成漏检、误检,从而导致学习到的特征较为分散,这也体现了文中方法在弯曲文本上的有效性及对多方向文本检测有较强的鲁棒性。在提升整体网络对文本识别性能的同时,所提出的算法由于后处理过程较为复杂,在识别效率方面仅比 CharNet 高出 1.9,相较于其他网络模型还有待提升。

表 3　Total-Text 数据集上模型性能对比

Method	Detection	End-to-End		FPS
	F-measure/%	None /%	Full /%	
CharNet	85.6	66.6	-	1.2
ABCNet	-	64.2	75.7	17.9
PGNet[15]	86.1	63.1	-	-
MaskTextSpotter[16]	85.2	65.3	77.4	4.8
Qin et al.[17]	-	67.8	-	4.8
MaskTextSpotter V3	-	71.2	78.4	-
PAN++	-	68.6	78.6	21.1
Ours	85.7	71.5	78.6	3.1

3 结束语

文中的研究具有一定的应用前景,例如检测路牌文字、辅助自动驾驶的导航、机器人送货上门等。但是,目前街景的文本检测中仍然存在一些问题,因此,提出了一个融合多尺度模块的文本检测方法(IMSM)。其中 RFBIA 和 SPM 能将有效特征精准地覆盖到目标文本区域,在突出特征的同时能有效抑制噪声影响。实验结果表明,文中算法在弯曲文本、旋转文本、密集小尺度文本的检测上有着优异的表现。后续工作将对提升文本对极端纵横比的鲁棒性、提高模型检测效率以及模型的轻量化展开深入研究,进一步提高检测效果。

参考文献:

[1] LIAO M,SHI B,BAI X,et al. TextBoxes:a fast text detector with a single deep neural network[C]//The AAAI conference on artificial intelligence. San Francisco: AAAI, 2017: 4161-4167.

[2] LIAO M,SHI B,BAI X. TextBoxes++:a single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing,2018,27(8):3676-3690.

[3] ZHONG Z,SUN L,HUO Q. An anchor-free region proposal network for faster R-CNN based text detection approaches [J]. Document Analysis and Recognition,2019,22(3):315-327.

[4] WANG Y,XIE H,ZHA Z,et al. ContourNet:taking a further step toward accurate arbitrary-shaped scene text detection [C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE, 2020: 11750-11759.

[5] DENG D,LIU H,LI X,et al. PixelLink:detecting scene text via instance segmentation[C]//Thirty-second AAAI conference on artificial intelligence. New Orleans: AAAI, 2018: 6773-6780.

[6] XU Y,WANG Y,ZHOU W,et al. TextField:learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5566-5579.

[7] WANG W,XIE E,SONG X,et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network [C]//IEEE conference on computer vision and pattern recognition. California:IEEE,2019:8440-8449.

[8] WANG W,XIE E,LI X,et al. PAN++:towards efficient and accurate end-to-end spotting of arbitrarily-shaped text[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,44(9):5349-5367.

[9] LIAO M,PANG G,HUANG J,et al. Mask TextSpotter v3: segmentation proposal network for robust scene text spotting [C]//European conference on computer vision. Piscataway: IEEE,2020:706-722.

[10] LIU S,DI H,WANG Y. Receptive field block net for accurate and fast object detection[C]//European conference on computer vision. Munich:IEEE,2018:404-419.

[11] LIU H,LIU F,FAN X,et al. Polarized self-attention:towards high-quality pixel-wise regression[J]. arXiv: 210700782,2021.

[12] HOU Q,ZHANG L,CHENG M M,et al. Strip pooling:rethinking spatial pooling for scene parsing[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle:IEEE,2020:4001-4002.

[13] XING L,TIAN Z,HUANG W,et al. Convolutional character networks[C]//IEEE/CVF international conference on computer vision (ICCV). Seoul:IEEE,2019:9126-9136.

[14] LIU Y,CHEN H,SHEN C,et al. ABCNet:real-time scene text spotting with adaptive bezier-curve network[C]//Proceedings of IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE, 2020: 9809-9818.

[15] WANG P,ZHANG C,QI F,et al. PGNet:real-time arbitrarily-shaped text spotting with point gathering network[J]. arXiv:210405458,2021.

[16] LYU P,LIAO M,YAO C,et al. Mask TextSpotter:an end-to-end trainable neural network for spotting text with arbitrary shapes[C]//Proceedings of the European conference on computer vision. Munich:Springer,2018:67-83.

[17] QIN S,BISSACCO A,RAPTIS M,et al. Towards unconstrained end-to-end text spotting[C]//2019 IEEE/CVF international conference on computer vision (ICCV). Seoul: IEEE,2019:4703-4713.