双向特征融合的快速精确任意形状文本检测

边 亮*① 屈亚东② 周 宇②
①(北京航空航天大学电子信息工程学院 北京 100191)
②(中国科学技术大学信息科学技术学院 合肥 230026)

摘 要：现有的基于分割的场景文本检测方法仍较难区分相邻文本区域，同时网络得到分割图后后处理阶段步骤复杂导致模型检测效率较低。为了解决此问题，该文提出一种新颖的基于全卷积网络的场景文本检测模型。首先，该文构造特征提取器对输入图像提取多尺度特征图。其次，使用双向特征融合模块融合两个平行分支特征的语义信息并促进两个分支共同优化。之后，该文通过并行地预测缩小的文本区域图和完整的文本区域图来有效地区分相邻文本。其中前者可以保证不同的文本实例之间具有区分性，而后者能有效地指导网络优化。最后，为了提升文本检测的速度，该文提出一个快速且有效的后处理算法来生成文本边界框。实验结果表明：在相关数据集上，该文所提出的方法均实现了最好的效果，且比目前最好的方法在F-measure指标上最多提升了1.0%，并且可以实现将近实时的速度，充分证明了该方法的有效性和高效性。

关键词：场景文本检测；双向特征融合；多尺度特征；后处理复杂度；任意形状文本
中图分类号：TN911.73 **文献标识码**：A **文章编号**：1009-5896(2021)04-0931-08
DOI: 10.11999/JEIT200880

Bi-directional Feature Fusion for Fast and Accurate Text Detection of Arbitrary Shapes

BIAN Liang① QU Yadong② ZHOU Yu②
①(*School of Aeronautic Science and Engineering, Beihang University, Beijing* 100191, *China*)
②(*School of Information Science and Technology, University of Science and Technology of China, Hefei* 230026, *China*)

Abstract: Existing segmentation based methods have problems, such as the difficulty in distinguishing adjacent text areas and the low efficiency of model detection caused by the complex steps in the post-processing stage. In order to solve this problem, this article proposes a novel scene text detection model based on fully convolutional network, which can solve the problem that adjacent texts are difficult to distinguish in existing methods and improve the detection speed of the model. First, it constructs a feature extractor to extract multi-scale feature map from the input image. Secondly, the bidirectional feature fusion module is used to fuse the semantic information of the two parallel branches and promote the joint optimization of the two branches. It then effectively differentiates adjacent texts by predicting both a reduced text area map and a full text area map in parallel. The former can guarantee the distinction between different text instances, while the latter can effectively guide the network optimization. Finally, in order to improve the speed of text detection, it proposes a fast and effective post-processing algorithm to generate text boundary boxes. The experimental results show that: on relative datasets, the method proposed in this article achieves the best performance, and improves the F-measure index by 1.0% at most compared with the current best method, and can achieve near-real-time speed, which proves fully the effectiveness and high efficiency of the method.

Key words: Scene text detection; Bi-directional feature fusion; Multi-scale feature; Post-processing complexity; Arbitrary-shaped texts

1 引言

自然场景文本检测任务旨在定位自然场景中文本区域的位置，其在智能交通系统、可穿戴翻译装备、视觉辅助系统等智能系统中有着广泛的应用，近年来成为研究者关注的热门话题之一。然而由于自然场景环境的复杂性，文本图像的检测可能受到

收稿日期：2020-10-16；改回日期：2021-01-29；网络出版：2021-02-05
*通信作者： 边亮 askquestionbl@163.com

形状模糊、光照不均等影响，使得自然场景文本检测技术的发展相较于传统文档检测技术更加困难。

传统文本检测技术[1–3]依赖于人工设计的文本特征以及一些文本的先验信息，如纹理、笔画宽度等。这些方法主要分为基于滑动窗口及基于联通组件的方法，它们的精度低、速度慢。近年来，基于深度学习的方法在文本检测中占据了主要地位。在深度学习发展的早期，文献[4]首先采用MSER方法获取候选文本组件，文献[5]通过训练字符分类器生成文本显著性图。这些算法仍一定程度依赖于人工设计的特征。随着目标检测和语义分割方法的发展，目前的文本检测方法得到了很大的改进，主要分为基于分割的方法和基于回归的方法。

基于分割的方法针对输入图像生成像素级标签。文献[6]通过分辨属于不同文本实例的链接来定位文本区域位置。文献[2]提出了一种新的文本表征方式适用于曲形文本。文献[7]从最小分割图开始扩充像素直至最大的分割图结束，获取文本连通区域，但非常耗时。文献[8]将文本检测视为实例分割问题，并基于Mask R-CNN提出了文本上下文模块以及重评分机制来抑制假阳性文本案例。文献[3]提出了一种新思路，通过多维向量之间的距离约束文本像素之间的关系，采用聚类操作分离相近文本实例。文献[9]提出了一种可微二值化模块，提升了文本检测的性能。

基于回归的方法的主要思想是根据从预置锚框或像素点所预测的回归框偏置来生成文本检测框。借鉴目标检测领域相关算法，Textboxes++[10]直接预测回归预置框的四边形或多方向矩形框来检测任意方向的文本区域；文献[11]首先预测一系列小文本框，然后将属于同一大文本框的小文本框进行合并得到最终检测结果。与上述基于框偏置不同，文献[12]采用了另一种回归方式，直接预测文本实例内部像素各点到文本框4个顶点的偏置或到4个边界的距离。文献[13]通过直接回归模块生成文本的四边形预测框并进行迭代优化，从而逐渐感知到整个长文本。

上述方法的主要不同在于：传统的文本检测方法工作量大，后处理复杂，速度慢，效果差；早期的基于深度学习方法一定程度需要人为设计特征，相对简单，后处理仍比较复杂，速度中等；基于分割和回归的场景文本检测方法不需要手工设计特征，后处理简单，速度快，效果好。其中，基于分割的方法一般先将文本从图像中分割出来，然后进行阈值处理来得到文本区域的边界框。而基于回归的方法一般直接回归出文本区域的边界框，速度通常比基于分割的方法快，但是其对长文本及曲形文本的检测效果仍难以令人满意。

本文提出一种新的文本检测方案来解决上述两个问题以实现精确的任意场景文本检测。首先通过特征金字塔网络(Feature Pyramid Network, FPN)生成多尺度特征图，随后构建整体文本区域预测和缩小文本区域预测两个平行分支本，提出了双向特征融合模块对两个分支的特征进行融合处理，最终经过简单的后处理步骤对缩小文本区域分割图进行放大即可得到文本实例的检测结果。

本文的主要贡献有3个方面：(1)提出一种新颖的文本实例重建方式，可将中心分割图快速准确地重建出完整的文本区域，与目前的方法相比可以在后处理阶段节省很多计算消耗；(2)提出的双向特征融合模块可更好地融合文本语义特征，从而生成非常有效的文本特征；(3)提出的方法在SCUT-CTW1500, Total-Text以及MSRA-TD500上分别可以获得83.5%, 85.7%, 85.9%的F综合指标，检测速度可以分别达到15.2, 16, 17.1FPS，很好地实现了精度和速度的权衡。

2 模型构建

2.1 网络工作流程

如图1所示，文本图像经过ResNet进行特征提取，经过FPN特征融合及通道级联之后，生成特征图f；特征图f通过两个分支处理之后经过双向特征融合模块生成具有很强的特征表示能力的特征图p_5和p_6，分别用于预测完整和缩小的文本区域图。其中f_1到f_5是ResNet的5个阶段, h_1到h_4是FPN的4个特征图。缩小的文本区域图经过重建算法进行后处理，得到最终检测结果。

2.2 特征提取阶段

本文使用ResNet和FPN进行特征提取和融合，并且ResNet使用了可变形卷积[14]。如图2所示，首先进行下采样，从而分别得到特征图f_1, f_2,

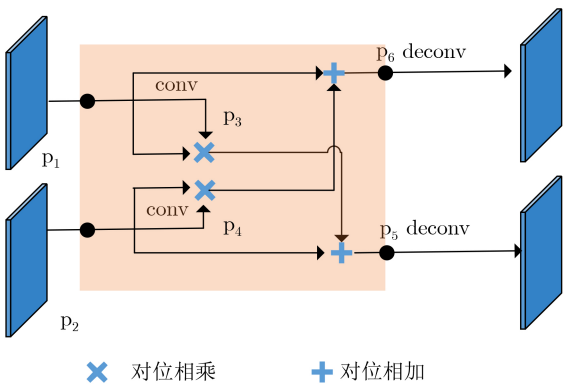


图 1 双向特征融合模块内部网络示意图

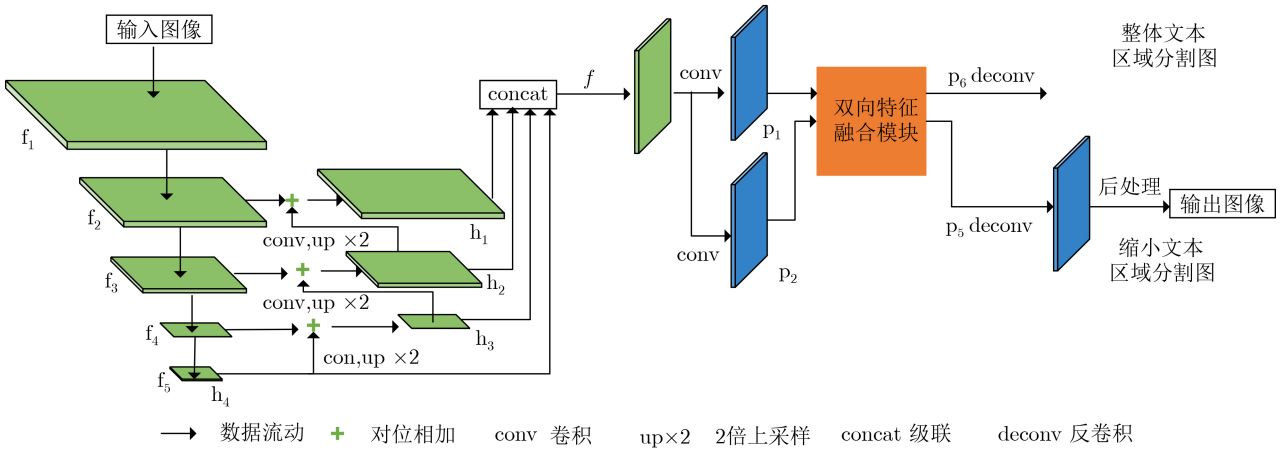


图 2 网络结构图

f_3, f_4, f_5。从f_1到f_5的分辨率依次降低为前者的1/2，包含的语义信息越来越丰富。小尺度的特征图如f_5有利于大文本的检测，大尺度的特征图如f_2对于小文本检测至关重要。然后，f_5, f_4, f_3依次经过上采样两倍之后分别与f_4, f_3, f_2进行特征融合，分别得到特征图h_3, h_2, h_1。最后特征图h_4, h_3, h_2分别上采样到h_1的大小，与h_1级联，得到特征图f。f融合了具有不同感受野和语义信息的特征图，从而对于不同尺度的文本具有更强的检测能力。

2.3 双向特征融合模块

本文提出双向特征融合模块有效地检测相邻文本。特征融合[15]能显著提升网络性能，广泛应用于深度学习的任务中。尽管特征提取阶段的FPN已经使用了特征融合，但是这对于检测背景复杂的文本还是不够的，因此本文提出了双向特征融合模块。如图1所示，p_1经过3×3的特征图处理之后得到p_3，p_1与p_3相乘之后与p_2相加得到p_5。类似地，p_2经过3×3的卷积处理之后得到p_4, p_4和p_2相乘之后与p_1相加得到p_6。双向特征融合模块能使两个分支的信息互相监督，彼此受益，从而促进网络优化。此外，还能融合上下文的尺寸信息，实现特征信息的互补，得到更准确的检测结果。

2.4 预测阶段

本文提出了一种新的预测分支模块，如图1所示，p_6和p_5经过步长为2的反卷积进行上采样到原图大小之后，分别用于预测完整和缩小的文本区域图。缩小的文本区域中不同的实例的间距变大，因此能有效地将相邻文本区分开。但缩小的文本区域并不是完整的文本实例，想要得到完整的文本边界框需要额外的后处理。因为完整的文本区域包括了很多边界信息，而网络对于边界信息是非常敏感的，因此本文预测了完整的文本区域图，进一步优化网络，降低了检测过程的漏检率。本文的实现中分别使用1个1×1的卷积来预测完整和缩小文本区域图。完整文本区域图和缩小文本区域图中每个像素点的预测值都在0～1范围内，表示该像素点属于文本的概率。

2.5 实例重建算法

一般基于分割的文本检测算法的后处理都非常复杂且耗时，本文提出了一个非常简单有效的文本重建算法，主要包括如下操作：(1)阈值处理。对预测的缩小的文本区域图使用固定的阈值t进行如式(1)的二值化操作，得到文本区域二值图；(2)去噪。得到每个文本实例的最小外接矩形，计算每个实例的面积与外接矩形框的比例作为文本实例的置信度，滤除置信度低的区域；(3)轮廓生成。得到二值图中每个文本实例的轮廓图；(4)扩张。使用放大比例系数K，计算出每个文本需要扩张的像素个数D，使用Vatti's clipping algorith对缩小的文本区域进行扩张。D的计算公式如式(2)所示

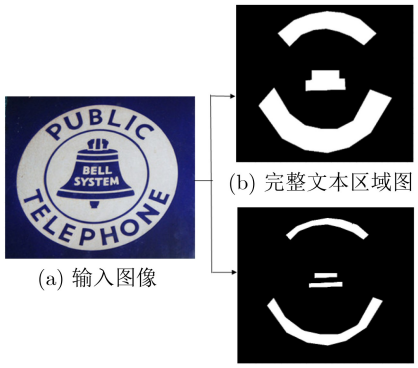
$$M = \begin{cases} 0, m \le t \\ 1, m > t \end{cases} \tag{1}$$

$$D = \frac{K \cdot S}{L} \tag{2}$$

其中，S和L分别表示缩小的文本区域的面积和周长。

2.6 标签生成

预测完整和缩小的文本区域图，需要对应的两种标签进行网络模型的训练。标签的生成过程如图3所示。对于完整的文本实例标签，本文直接将多边形标注的区域设置为1，将其他区域设置为0。对于缩小的文本区域的标签生成，与DB[9]一样，训练数据中的多边形标注框是由一系列2维坐标点表示的，因此可以采用 Vatti's clipping algorithm将多边形区域向内缩小D_1个像素。其中，D_1是由该多



(a) 输入图像 (b) 完整文本区域图 (c) 缩小文本区域图

图 3 标签生成示意图

边形框的周长L_1和面积S_1共同决定的，用公式表示则为

$$D_1 = \frac{(1-r^2)S_1}{L_1} \tag{3}$$

2.7 损失函数

整体的损失函数是预测阶段各个分支的加权和，用公式可以表达为

$$L = L_{\mathrm{g}} + \lambda L_{\mathrm{s}} \tag{4}$$

其中，L_{g}为完整文本区域分割图的损失函数，L_{s}为缩小文本区域分割图的损失函数，λ为平衡系数，本文设置为1。

本文使用二值交叉熵损失来计算L_{g}。此外，为避免正负样本严重不平衡，使得网络偏向于预测背景区域，我们在损失函数中加入了难样本挖掘处理，完整文本区域分割图的损失函数用公式表示为

$$L_{\mathrm{g}} = -\left[\sum_{i\notin S} y_i \lg \hat{y}_i + (1-y_i)\lg(1-\hat{y}_i)\right] \tag{5}$$

其中，$\hat{y}_i$表示预测结果，y_i是标签，S是对图像进行正负样本比例为1:3的采样样本。L_{s}采用的是Dice损失函数。Dice损失函数是用于度量集合相似度的一种函数，其公式为

$$L_{\mathrm{s}} = 1 - \frac{2|X\cap Y|}{|X|+|Y|} \tag{6}$$

其中，X和Y分别表示预测结果和标签，$|X\cap Y|$表示X和Y的交集，$|X|+|Y|$表示X和Y的并集。

3 实验结果及分析

3.1 数据集

本文在实验过程中共涉及了SynthText[16]、曲形文本数据集SCUT-CTW1500[17], Total-Text[18]以及四边形文本数据集MSRA-TD500[19]。

3.2 实验结果评价指标

本文选取场景文本检测常用的3个评价指标：准确率(Precision)、召回率(Recall)和F综合指标(F-measure)来评价模型性能。准确率为检测结果中正确检测的文本占实际被检测文本的总数的比例，召回率表示正确检测的文本数占正样本数的比例，F综合指标则为准确率与召回率的调和平均值，其计算方式为

$$\mathrm{F} = \frac{2\times \mathrm{Precision}\times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{7}$$

3.3 实验平台及模型参数

实验所用到的软硬件环境为：1块NVIDIA 1080Ti显卡。操作系统为Ubuntu16.04，使用Pytorch框架实现。

本文使用在SynthText[16]上预训练1轮的ResNet50模型作为基础网络，随后使用各目标数据集的官方训练数据进行相应的训练。数据增广方式包括训练图像的随机水平翻转、旋转、缩放以及裁剪。训练批量大小设定为16，参数优化采用随机梯度下降法，且权重衰减设为0.0001，梯度冲量设为0.9，初始学习率为0.007。

3.4 模型消融实验

为了验证本文所提出方法的有效性，本文进行了相关模块的消融实验，同时将更换ResNet50基础网络为ResNet18进行相关实验。

双向特征融合模块：本文首先验证双向特征融合模块对于模型整体性能的影响，该对比实验固定模型其余部分网络结构以及学习率等影响因素，仅针对双向特征融合模块进行消融。

表1对比第2, 3行及5, 6行为双向特征融合模块的消融实验结果，可以看出，双向特征融合模块分别在准确率、召回率及F综合指标方面获得了0.2%, 0.4%及0.3%的提升，在检测速度方面只有0.8FPS的差距，ResNet18基础网络也在添加了双向特征融合模块之后，F综合指标有0.5%的提升。并且检测速度方面仅低0.9FPS。

整体文本区域图：网络通过缩小文本区域预测分支来分辨相邻的文本实例，增加整体文本区域预测分支来学习整个文本实例分割图，让网络对整体文本区域敏感。表1对比第1, 2行及4, 5行可以看到，基于ResNet50基础网络整体文本区域预测分支可以将网络F综合指标提升0.4%，且网络检测速度相当。而基于ResNet18基础网络，整体文本区域预测分支的加入也提升了0.2%的F综合指标，且检测速度仅有0.5FPS的下降。

基础网络：对比表1中1, 4行或2, 5行可以看出基于基础网络ResNet50的模型的表现相较于ResNet18分别在F综合指标方面有1.9%, 2.1%的提

表 1　双向特征融合模块及整体文本框分支在不同基础网络下的性能增益及检测效率

基础网络	双向特征融合模块	整体文本区域预测分支	评价指标(%) 准确率	召回率	F综合指标	FPS
ResNet-50	×	×	87.4	82.7	85.0	17.4
ResNet-50	×	√	87.8	83.1	85.4	16.8
ResNet-50	√	√	88.0	83.5	85.7	16.0
ResNet-18	×	×	86.6	79.8	83.1	31.0
ResNet-18	×	√	85.9	80.8	83.3	30.5
ResNet-18	√	√	86.5	81.2	83.8	29.6

升，而基于ResNet18基础网络的模型也因为网络层数较浅，在检测速度方面表现更好。

3.5　本文方法与先进方法的比较

3.5.1　Total-Text数据集测评

如表2所示，本文方法在Total-Text[18]数据集上取得了准确率88.0%，召回率83.5%及F综合指标85.7%的结果，相较于表中的方法在F综合指标上取得了最好的结果，且超出最好方法1%(85.7% vs. 84.7%)。相较于针对曲形文本提出的场景文本检测方法TextSnake[2], PSENet[7]等，本文方法相比于其中最高性能方法在F综合指标上提升了3.8%，并可以达到检测速度16FPS，与PSENet[7]的检测速度相比是4倍。尽管CRAFT[20]使用了字符级的图像标注信息来指导网络的学习，本文方法在准确率、召回率及F综合指标分别提升0.4%, 3.6%, 2.1%。与DB[9]相比虽速度方面有些差距，但在精度方面本文分别在准确率、召回率及F综合指标上超出其0.9%, 1%, 1%。

3.5.2　MSRA-TD500数据集测评

不同方法在MSRA-TD500[19]数据集上的定量检测结果如表3所示。从表3可以看出，本文方法相比于目前最好方法DB在召回率及F综合指标上分别超出了2.1%, 1%，且在检测速度方面也达到了较为可观的数值。

3.5.3　SCUT-CTW1500数据集测评

如表4所示，本方法取得了准确率84.7%，召回率82.3%，F综合指标83.5%的结果，不仅F综合指标相较于其他方法有提升，检测速度也达到了15.2FPS，优于PSENet[7]。尽管LOMO[13]中的迭代矫正模块对长文本的检测效果有不错的提升，本文方法仍在F综合指标上比其高出了5.1%(83.5% vs. 78.4%)。

图4(b)、图4(c)分别为对图4(a)的整体和缩小文本区域分割图的预测结果。从图4(c)可以看出，缩小文本区域分割图可以将整体分割图中因距离较近互相粘连的文本区域分开，最终对文本实例精确定位。图4(b)、图4(c)也验证了本文的网络结构可以适应于任意形状文本实例。

图5是不同方法在3个数据集上的速度和精度比较，FPS表示每秒运行几帧。图6是3个测试数据集上的检测结果，从图中可以看出，本文的方法可以精确定位文本，并对于距离较近的文本实例也能正确分割。

表 2　TotalText数据集模型性能对比

方法	评价指标(%) 准确率	召回率	F综合指标	FPS
EAST*[12]	36.2	50.0	42.0	–
TextSnake[2]	74.5	82.7	78.4	–
MSR[21]	74.8	83.8	79.0	4.3
PSENet-1s[7]	78.0	84.0	80.9	3.9
Textfield[22]	81.2	79.9	80.6	6
LOMO[13]	87.6	79.3	83.3	–
CRAFT[20]	87.6	79.9	83.6	–
DB[9]	87.1	82.5	84.7	32
本文方法	88.0	83.5	85.7	16

表 3　MSRA-TD500数据集模型性能对比

方法	评价指标(%) 准确率	召回率	F综合指标	FPS
RRPN[23]	82.0	68.0	74.0	–
MCN[24]	88.0	79.0	83.0	–
PixelLink[6]	83.0	73.2	77.8	3.0
TextSnake[2]	83.2	73.9	78.3	1.1
CRAFT[20]	88.2	78.2	82.9	8.6
Tian等人[32]	84.2	81.7	82.9	–
DB[9]	91.5	79.2	84.9	32.0
本文方法	91.1	81.3	85.9	17.1

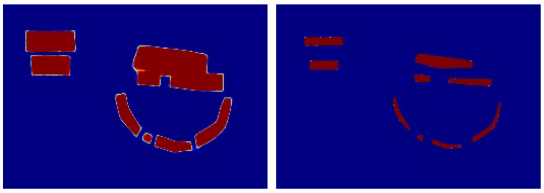
表 4　CTW1500数据集模型性能对比

方法	评价指标(%) 准确率	召回率	F综合指标	FPS
CTPN[25]	60.4	53.8	56.9	7.14
EAST[12]	78.7	49.1	60.4	21.2
Seglink[11]	42.3	40.0	40.8	10.7
TextSnake[2]	67.9	85.3	75.6	1.1
PSENet-1s[7]	84.8	79.7	82.2	3.9
Tian等人[3]	77.8	82.7	80.1	3
LOMO[13]	69.6	89.2	78.4	4.4
DB[9]	86.9	80.2	83.4	22
本文方法	84.7	82.3	83.5	15.2

3.6 缺点与不足

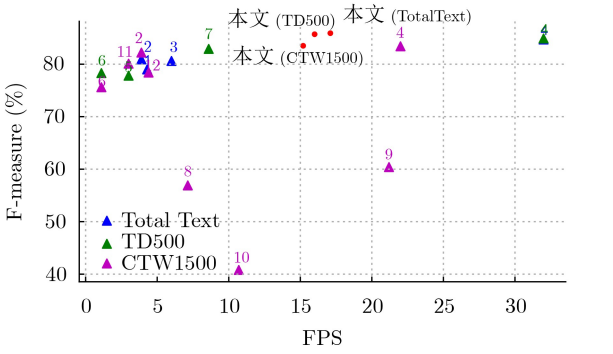
尽管本文的方法能在多个公开数据集上取得最好的效果，但是由于场景文本背景的复杂性，字体大小、形状、颜色等的多样性，本文的方法对于一些极端的文本仍然存在检测困难的问题。

(a) 检测对象



(b) 预测的整体文本区域图 (c) 预测的缩小文本区域图

图 4 检测的最终结果



1: MSR[22] 4: Liao等人[9] 7: CRAFT[20] 10: Seglink*[11]
2: PSENet-1s[7] 5: PixelLink[6] 8: CTPN*[25] 11: Tian等人*[3]
3: Textfield[22] 6: TextSnake[2] 9: EAST*[12] 12: LOMO[13]

图 5 不同方法在3个数据集上的速度-精度对比

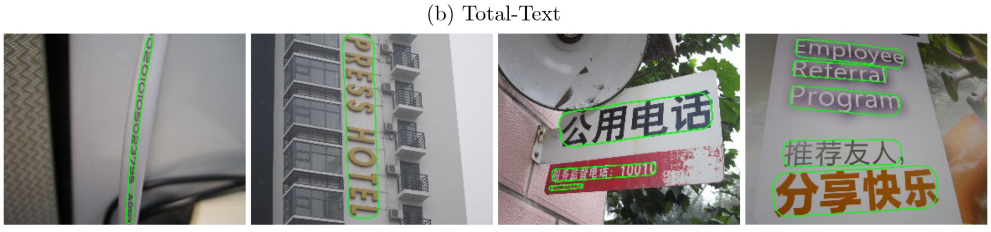
如图7所示，图7(a)显示的是本文方法对于特别长的文本存在检测不完整的问题，这是网络的感受野不够大导致的，可以通过使用空洞卷积来解决这个问题，但是这将会降低网络的速度。图7(b)显示的是对于一些字符与字符之间距离很大的文本，模型会将它检测为多个文本实例，这是缺乏足够的上下文信息和语言模型导致的，即使是人也很难分辨出来。图7(c)显示的是，在背景非常复杂的情况下，一些外形类似于文本的目标很容易被检测为文本，主要是因为这些目标的外形与文本的笔画很相似。图7(d)揭示的是对于一些在水平方向和垂直方向都存在靠得很近的相邻文本的时候，在只基于图像的视觉信息而没有加入语意信息进行检测的时候，模型很难辨别出文本应该是从水平方向读取还是从垂直方向读取。图7(e)和图7(f)显示的是模型对于一些非常小的文本可能存在漏检的问题，这是因为经过多次下采样之后，这些小的文本区域已经变得非常小了，会被模型当成是噪声过滤掉。上述的问题是文本检测存在的普遍问题，它们共同的最主要的原因是训练集中缺乏相应的文本，使得模型无法学习到相关知识。

4 结束语

首先，为了有效地区分相邻文本实例，本文设计了缩小和整体两个并行的文本区域预测分支。其中前者可以准确地分离粘连的文本实例，而后者能够通过预测图像中文本实例的整体文本区域分割图来辅助整体网络优化以及缩小文本区域预测分支的

(a) SCUT-CTW1500

(b) Total-Text



(c) MSRA-TD500

图 6 不同数据集模型的测试结果可视化图

(a) 检测不完整　(b) 检测为多个实例　(c) 检测出非文本

(d) 水平、垂直识别问题　(e) 微小文本漏检1　(f) 微小文本漏检2

图 7 模型检测错误的一些例子

学习。其次，本文提出了双向特征融合模块用于充分融合两个分支所获取的文本语义信息以及提升文本检测的精度。最后，为了解决现有方法存在的检测速度过慢的问题，本文设计了一个简单高效的文本重建算法。在3个公开数据集上的实验证明，本文的方法在F综合指标上超越了现有方法，并且达到了与现有最快的方法相当的速度，充分证明了本文方法的有效性和高效性。后续工作计划将网络进行进一步扩展，构成端到端文本检测识别系统。

参 考 文 献

[1] 黄剑华, 承恒达, 吴锐, 等. 基于模糊同质性映射的文本检测方法[J]. 电子与信息学报, 2008, 30(6): 1376–1380.
HUANG Jianhua, CHENG Hengda, WU Rui, *et al*. A new approach for text detection using fuzzy homogeneity[J]. *Journal of Electronics & Information Technology*, 2008, 30(6): 1376–1380.

[2] LONG Shangbang, RUAN Jiaqiang, ZHANG Wenjie, *et al*. Textsnake: A flexible representation for detecting text of arbitrary shapes[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 19–35.

[3] TIAN Zhuotao, SHU M, LYU P, *et al*. Learning shape-aware embedding for scene text detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 4229–4238.

[4] HUANG Weilin, QIAO Yu, and TANG Xiaoou. Robust scene text detection with convolution neural network induced MSER trees[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 497–511.

[5] JADERBERG M, VEDALDI A, and ZISSERMAN A. Deep features for text spotting[C]. The 13th European Conference on Computer Vision, Zurich, Switzerland, 2014: 512–528.

[6] DENG Dan, LIU Haifeng, LI Xuelong, *et al*. Pixellink: Detecting scene text via instance segmentation[C]. The 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 6773–6780.

[7] WANG Wenhai, XIE Enze, LI Xiang, *et al*. Shape robust text detection with progressive scale expansion network[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 9328–9337.

[8] XIE Enze, ZANG Yuhang, SHAO Shuai, *et al*. Scene text detection with supervised pyramid context network[C]. The 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 9038–9045.

[9] LIAO Minghui, WAN Zhaoyi, YAO Cong, *et al*. Real-time scene text detection with differentiable binarization[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 11474–11481. doi: 10.1609/aaai.v34i07.6812.

[10] LIAO Minghui, SHI Baoguang, and BAI Xiang. Textboxes++: A single-shot oriented scene text detector[J]. *IEEE Transactions on Image Processing*, 2018, 27(8): 3676–3690. doi: 10.1109/TIP.2018.2825107.

[11] SHI Baoguang, BAI Xiang, and BELONGIE S. Detecting oriented text in natural images by linking segments[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 3482–3490.

[12] ZHOU Xinyu, YAO Cong, WEN He, *et al*. EAST: An efficient and accurate scene text detector[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 2642–2651.

[13] ZHANG Chengquan, LIANG Borong, HUANG Zuming, *et al*. Look more than once: An accurate detector for text of arbitrary shapes[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 10544–10553.

[14] DAI Jifeng, QI Haozhi, XIONG Yuwen, *et al*. Deformable convolutional networks[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 764–773.

[15] 谢金宝, 侯永进, 康守强, 等. 基于语义理解注意力神经网络的多元特征融合中文文本分类[J]. 电子与信息学报, 2018, 40(5): 1258–1265. doi: 10.11999/JEIT170815.
XIE Jinbao, HOU Yongjin, KANG Shouqiang, *et al*. Multi-feature fusion based on semantic understanding attention neural network for Chinese text categorization[J]. *Journal of Electronics & Information Technology*, 2018, 40(5): 1258–1265. doi: 10.11999/JEIT170815.

[16] GUPTA A, VEDALDI A, and ZISSERMAN A. Synthetic data for text localisation in natural images[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2315–2324.

[17] LIU Yuliang, JIN Lianwen, ZHANG Shuaitao, *et al*. Curved scene text detection via transverse and longitudinal sequence connection[J]. *Pattern Recognition*, 2019, 90: 337–345.

[18] CH'NG C K and CHAN C S. Total-text: A comprehensive dataset for scene text detection and recognition[C]. The 2017 14th IAPR International Conference on Document Analysis and Recognition, Kyoto, Japan, 2017: 935–942.

[19] YAO Cong, BAI Xiang, LIU Wenyu, *et al*. Detecting texts of arbitrary orientations in natural images[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012: 1083–1090.

[20] BAEK Y, LEE B, HAN D, *et al*. Character region awareness for text detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 9357–9366.

[21] XUE Chuhui, LU Shijian, ZHANG Wei. MSR: Multiscale shape regression for scene text detection[C]. KRAUS S. The 28th International Joint Conference on Artificial Intelligence, Macao, China, 2019: 989–995.

[22] XU Yongchao, WANG Yukang, ZHOU Wei, *et al*. Textfield: Learning a deep direction field for irregular scene text detection[J]. *IEEE Transactions on Image Processing*, 2019, 28(11): 5566–5579.

[23] MA Jianqi, SHAO Weiyuan, YE Hao, *et al*. Arbitraryoriented scene text detection via rotation proposals[J]. *IEEE Transactions on Multimedia*, 2018, 20(11): 3111–3122.

[24] LIU Zichuan, LIN Guosheng, YANG Sheng, *et al*. Learning markov clustering networks for scene text detection[C]. 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 6936–6944.

[25] TIAN Zhi, HUANG Weilin, HE Tong, *et al*. Detecting text in natural image with connectionist text proposal network[C]. The 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 56–72.

边 亮：男，1982年生，博士生，研究方向为图像获取与处理.
屈亚东：男，1998年生，硕士生，研究方向为场景图像文字合成、检测与识别.
周 宇：男，1992年生，博士生，研究方向为场景图像文字合成、检测与识别.

责任编辑：马秀强