利用可区分二值化技术进行实时场景文本检测

廖明辉1 , Zhaoyi Wan2 , Cong Yao2 ，Kai Chen3,4, Xiang Bai1 †
1华中科技大学、 2旷视科技、 3上海交通大学、 4昂立科技
{mhliao,xbai}@hust.edu.cn, i@wanzy.me, yaocong2010@gmail.com, kchen@sjtu.edu.cn

抽象的

最近,基于分割的方法在
场景文本检测,因为分割结果可以更准确地描述各种形状的场景文本,
例如曲线
文本。然而,二值化的后处理对于基于分割的检测至关重要,它将分割
方法生成的概率图转换为文本的边界框/区域。本文提出了一个名为可
微分二值化 (DB) 的模块,它可以在分割网络中执行二值化过程。与
DB 模块一起优化后,分割网络可以

自适应地设置二值化的阈值,这不仅
简化了后处理,同时也提升了文本检测的性能。基于一个简单的分割网
络,我们验证了DB在

五个基准数据集,在检测准确率和

速度。特别是,在轻量级主干上,DB 的性能提升非常显著,因此我们

可以在检测精度之间寻找理想的权衡
和效率。具体来说,以 ResNet-18 为骨干,
我们的检测器实现了 82.8 的 F 值,运行速度为 62
在 MSRA-TD500 数据集上进行 FPS 测试。代码可在此处获取:
https://github.com/MhLiao/DB。

介绍

近年来,阅读场景图像中的文本已成为
由于其广泛的实际应用,成为一个活跃的研究领域
例如图像/视频理解、视觉搜索、自动
驾驶和盲人辅助。
场景文本作为场景文本阅读的重要组成部分,
旨在定位每个文本实例的边界框或区域的检测仍然是一项具有挑战性的任
务,因为
场景文本通常具有各种比例和形状,包括
水平、多方向和弯曲的文本。基于分割的场景文本检测引起了很多关注

近年来,由于其像素级的预测结果,它可以描述各种形状的文本。然而,大多数
基于分割的方法需要复杂的

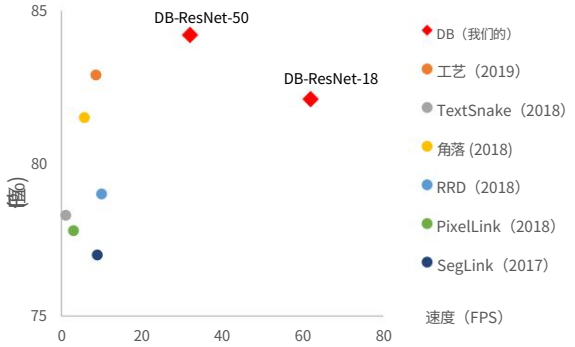

图 1:MSRA-TD500 数据集上几种近期场景文本检测方法的比较,包括

准确度和速度。我们的方法实现了理想的平衡
有效性和效率之间。

将像素级预测结果分组到检测到的文本实例中,从而产生大量

推理过程中的时间成本。以两种最新的场景文本检测方法为例:

PSENet (Wang et al. 2019a) 提出了后处理
逐步扩展规模以提高检测
准确度;使用 (Tian et al. 2019) 中的像素嵌入
根据分割结果对像素进行聚类,
需要计算像素之间的特征距离。

大多数现有的检测方法都使用类似的后处理流程,如图 2 所示（遵循蓝色

箭头）:首先,他们设定了转换的固定门槛
分割网络产生的概率图
变成二进制图像;然后,一些启发式技术,如
像素聚类用于将像素分组到文本实例中。或者,我们的管道（沿着红色箭头

图 2)旨在将二值化操作插入到
分割网络进行联合优化。这样,
图像中每个位置的阈值都可以自适应地预测,从而可以充分区分像素和

前景和背景。然而,标准二值化函数不可微,因此我们提出一个

称为可微分的二元化近似函数
二值化（DB）,在训练时完全可区分

作者贡献均等。
†通讯作者

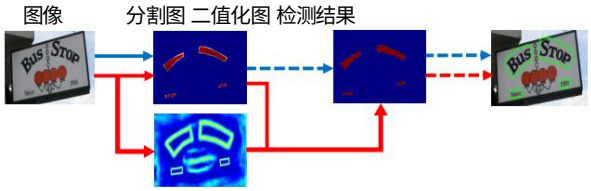

图 2:传统管道（蓝色流程)和我们的管道
（红色流程）。虚线箭头表示仅推理运算符；
实线箭头表示训练中的可微算子
和推理。

将其与分割网络一起进行。
本文的主要贡献是提出了 DB
可微分的模块,这使得
CNN 中端到端可训练的二值化。通过结合
一个简单的语义分割网络和提出的DB模块,我们提出了一个健壮且快速的场景文本
检测器。从使用DB模块的性能评估来看,我们发现我们的检测器比之前最先进的检测器有几个显著的优势

基于分割的方法。

1. 我们的方法在以下方面取得了持续更好的表现
五个场景文本基准数据集,包括水平、多方向和弯曲文本。

2. 我们的方法比以前的主要方法执行得更快,因为DB可以提供高度稳健的二值化图,从而大大简化了后处理。

3. 使用轻量级主干时,DB 运行良好,
这显著提高了检测性能
以 ResNet-18 为骨干。

4. 由于可以在推理阶段移除 DB 而不会牺牲性能,因此不需要额外的内存/时间
测试费用。

相关工作

最近的场景文本检测方法大致可以分为两类:基于回归的方法和

基于分割的方法。
基于回归的方法是一系列模型,
直接回归文本实例的边界框。
TextBoxes (Liao et al. 2017) 修改了锚点和
基于SSD的卷积核的尺度（Liu et al.
2016)用于文本检测。TextBoxes++（Liao、Shi 和 Bai
2018) 和 DMPNet (Liu 和 Jin 2017) 应用四边形回归来检测多方向文本。
SSTD (He et al.
2017a) 提出了一种注意力机制来粗略地识别文本区域。RRD (Liao et al.
2018) 通过使用旋转不变特征将分类和回归解耦。

用于分类,以及用于回归的旋转敏感特征,在多方向和长文本实例中效果
更佳。EAST（Zhou et al. 2017)和DeepReg（He et al. 2019）

2017b) 是无锚方法,它应用了像素级
多方向文本实例的回归。SegLink（Shi,
Bai 和 Belongie 2017) 回归了片段边界
框并预测其链接,以处理长文本实例。DeRPN (Xie et al. 2019b) 提出了一个
维度分解区域提议网络来处理规模

场景文本检测中的问题。基于回归的方法
通常喜欢简单的后处理算法（例如非最大抑制）。然而,它们大多受到限制

表示不规则形状的精确边界框,
例如弯曲的形状。
基于分割的方法通常结合像素级
预测和后处理算法来获取边界框。(Zhang et al. 2016) 检测到多方向文本

通过语义分割和基于MSER的算法。
文本边框用于（Xue, Lu, and Zhan 2018)分割
文本实例,Mask TextSpotter（Lyu et al. 2018a；
Liao et al. 2019) 检测到任意形状的文本实例
一种基于Mask R-CNN的实例分割方式。
PSENet (Wang et al. 2019a) 提出了渐进尺度
通过使用不同的方法分割文本实例来扩展
尺度核。像素嵌入是在（Tian et al.
2019)对分割结果中的像素进行聚类。
PSENet（Wang 等人,2019a)和 SAE（Tian 等人,2019）
对分割结果提出了新的后处理算法,导致推理速度降低。相反,

我们的方法重点是通过以下方式改善分割结果
将二值化过程纳入训练阶段,
而不会损失推理速度。
快速场景文本检测方法注重准确率和推理速度。TextBoxes (Liao et al. 2017),

TextBoxes++ (Liao, Shi, and Bai 2018), SegLink (Shi, Bai,
和 Belongie 2017),以及 RRD (Liao et al. 2018) 实现了
通过遵循检测架构进行快速文本检测
SSD (Liu et al. 2016). EAST (Zhou et al. 2017) 提出
应用 PVANet (Kim et al. 2016) 来提高其速度。大多数
它们无法处理不规则形状的文本实例,
比如曲线形状。相比之前的快速场景
文本检测器,我们的方法不仅运行速度更快,而且可以
检测任意形状的文本实例。

方法论

我们提出的方法的架构如图 3 所示。
首先,将输入图像输入到特征金字塔主干中。其次,将金字塔特征上采样到

相同尺度并级联以生成特征 F。然后,特征 F 用于预测概率图（P)和

阈值图（T）。之后,
映射（B ）由P和F计算得出。在训练期间,
在概率图上施加监督,阈值
概率图和近似二值图共享相同的监督。在推理阶段,边界框可以

可以从近似二进制图或
通过盒子公式模块的概率图。

二值化

标准二值化给定概率图 P ∈
RH×W由分割网络生成,其中 H 和
W 表示地图的高度和宽度,必须
将其转换为二值映射 P ∈ RH×W ,其中像素

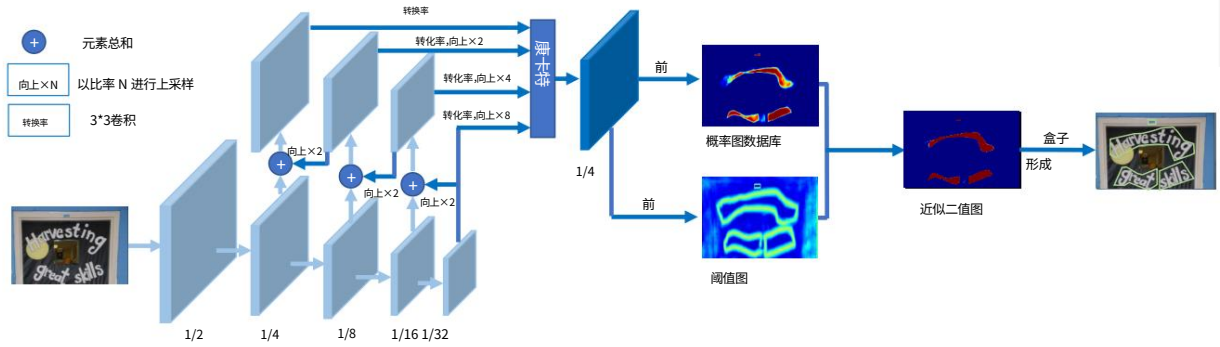

图 3:我们提出的方法的架构,其中“pred”由一个 3×3 卷积运算符和两个反卷积运算符组成
步幅为 2 的运算符。“1/2”、“1/4”、... 和“1/32”表示与输入图像相比的缩放比例。

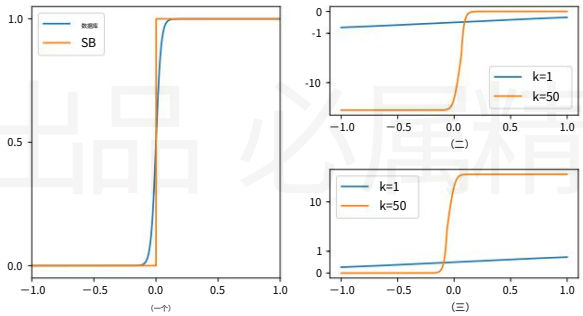

图 4:可微分二值化及其
导数。(a)标准二值化(SB)和可微二值化(DB)的数值比较。(b) l+的导数。 (c) l−的导数。

值 1 被视为有效文本区域。通常,此二值化过程可以描述如下:

$$B_{i,j} = \begin{cases} 1 \text{ 如果} P_{i,j} >= t, \\ \text{否则为 } 0。 \end{cases} \quad (1)$$

其中 t 是预定义阈值,(i, j) 表示
地图上的坐标点。

可微分二值化公式1中描述的标准二值化是不可微分的。因此,它不能在训练过程
中与分割网络一起进行优化。

周期。为了解决这个问题,我们建议使用近似阶跃函数进行二值化:

$$B_{i,j} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \quad (2)$$

其中 B 是近似二值映射;T 是自适应
从网络学习到的阈值图;k表示放大系数。经验上,k设置为50。这个近似

二值化函数的行为类似于标准二值化函数(见图 4),但可微,因此可以

与分割网络一起优化
训练期。具有自适应阈值的可区分二值化不仅可以帮助区分文本区域

从背景中,还有单独的文本实例
紧密接合。图7显示了一些示例。

DB 提高性能的原因可能是
通过梯度的反向传播来解释。让我们
以二分类交叉熵损失为例。定义 $f(x) = \frac{1}{1+e^{-x}}$ 作为我们的 DB 函数,其中 $x = P_{i,j} - T_{i,j}$。 然后
正标签的损失l+和负标签的损失l−分别为:

$$l_+ = -\log\left(1 + \frac{1}{e^{-kx}}\right)$$
$$l_- = -\log\left(1 - \frac{1}{1 + e^{-kx}}\right) \quad (3)$$

我们可以轻松计算损失与
链式法则:

$$\frac{\partial l_+}{\partial x} = -kf(x)e^{-kx}$$
$$\frac{\partial l_-}{\partial x} = kf(x) \quad (4)$$

图 4 还显示了l+和l−的导数。我们
可以从微分中看出(1)梯度是
乘以放大系数 k;(2)放大倍数
梯度对于大多数错误预测的
区域(L+ 为 x < 0; L−为 x > 0),从而有利于
优化,并有助于产生更独特的预测。此外,当 x = Pi,j − Ti,j 时,P 的梯度为

受到 T 的影响并在前景和背景之间重新调整。

自适应阈值

图1中的阈值图与文本边界图类似
(Xue, Lu, and Zhan 2018) 从外观上看。然而,
阈值图的动机和用法不同于
文本边界图。图 6 显示了有/无监督的阈值图。即使没有监督,阈值图也会突出显示文本边界区域。

阈值图。这表明边界阈值
阈值图对最终结果有利。因此,我们在阈值图上应用了边界监督,以获得更好的指导。
关于监督的消融研究在
实验部分。对于用法,文本边框图
(Xue, Lu, and Zhan 2018)用于分割文本实例
而我们的阈值图则作为二值化的阈值。

可变形卷积

可变形卷积 (Dai et al. 2017; Zhu et al. 2019) 可以为模型提供灵活的感受野,这对于极端长宽比的文本实例尤其有益。随后 (Zhu et al. 2019),在 ResNet-18 或 ResNet-50 主干网络的 conv3、conv4 和 conv5 阶段中,所有 3×3 卷积层都采用了调制可变形卷积 (He et al. 2016a)。

标签生成

图 5:标签生成。文本多边形的注释以红线显示。收缩和扩张的多边形分别以蓝线和绿线显示。

概率图的标签生成灵感来自 PSENet (Wang et al. 2019a)。给定一张文本图像,其文本区域的每个多边形由一组线段描述:

$$G = \{S_k\}_{k=1}^{n} \quad (5)$$

n 是顶点的数量,在不同数据集中可能不同,例如,ICDAR 2015 数据集为 4（Karatzas 等人,2015）,CTW1500 数据集为 16（Liu 等人,2019a）。

然后,使用 Vatti 裁剪算法 (Vati 1992)将多边形 G 收缩至Gs ,生成正区域。收缩的偏移量 D 由原始多边形的周长 L 和面积 A 计算得出:

$$D = \frac{A(1 - r^2)}{L} \quad (6)$$

其中 r 是收缩率,根据经验设置为 0.4。

通过类似的步骤,我们可以为阈值图生成标签。首先,将文本多边形 G 以相同的偏移量 D 膨胀到Gd。我们将Gs和Gd之间的间隙视为文本区域的边界,其中阈值图的标签可以通过计算到 G 中最近线段的距离来生成。

优化

损失函数L可以表示为概率图损失Ls、二值图损失Lb、阈值图损失Lt的加权和:

$$L = L_s + \alpha \times L_b + \beta \times L_t \quad (7)$$

其中Ls是概率图的损失， Lb是二值图的损失。根据损失的数值,α 和 β 分别设置为 1.0 和 10。

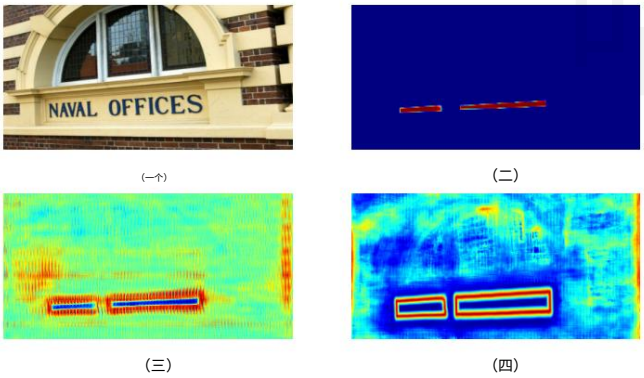

图 6:有/无监督的阈值图。(a)
输入图像。(b)概率图。(c)无监督的阈值图。(d)有监督的阈值图。

我们对Ls和Lb都应用了二元交叉熵 (BCE) 损失。为了克服正样本和负样本数量的不平衡,BCE 损失中采用了困难负样本挖掘 (Hard negative mining) 方法,即对困难负样本进行采样。yi log xi + (1 − yi) log (1 − xi)

$$L_s = L_b = \sum_{i \in S_l} \quad (8)$$

Sl为正负样本比为 1:3 的样本集。

Lt是将 L1 距离与
扩大文本多边形Gd 内的预测和标签:

$$L_t = \sum_{i \in R_d} \| y_i - x_i \| \quad (9)$$

其中Rd是扩张多边形Gd内部像素的索引集合； y
是阈值图的标签。

在推理阶段,我们可以使用概率图或近似二值图来生成文本边界框,其结果几乎相同。为了提高效率,我们使用概率图,以便消除阈值分支。边界框生成过程包含三个步骤：(1)首先将概率图/近似二值图以常数阈值（0.2)进行二值化,得到二值图；(2)从二值图中获取连通区域（收缩的文本区域）；(3)使用 Vatti 裁剪算法（Vati 1992)对收缩区域进行偏移量 D 的扩张。D 的计算公式为

$$D = \frac{A \times r}{L} \quad (10)$$

式中 ,A为收缩多边形的面积;L为收缩多边形的周长 ;r根据经验设置为1.5。

实验

数据集

SynthText (Gupta、Vedaldi 和 Zisserman 2016) 是一个合成数据集,包含 80 万张图片。这些图片由 8000 张背景图片合成。该数据集仅用于预训练我们的模型。1 MLT -2017 数据集

是一个多语言数据集。它包括

[1] https://rrc.cvc.uab.es/?ch=8

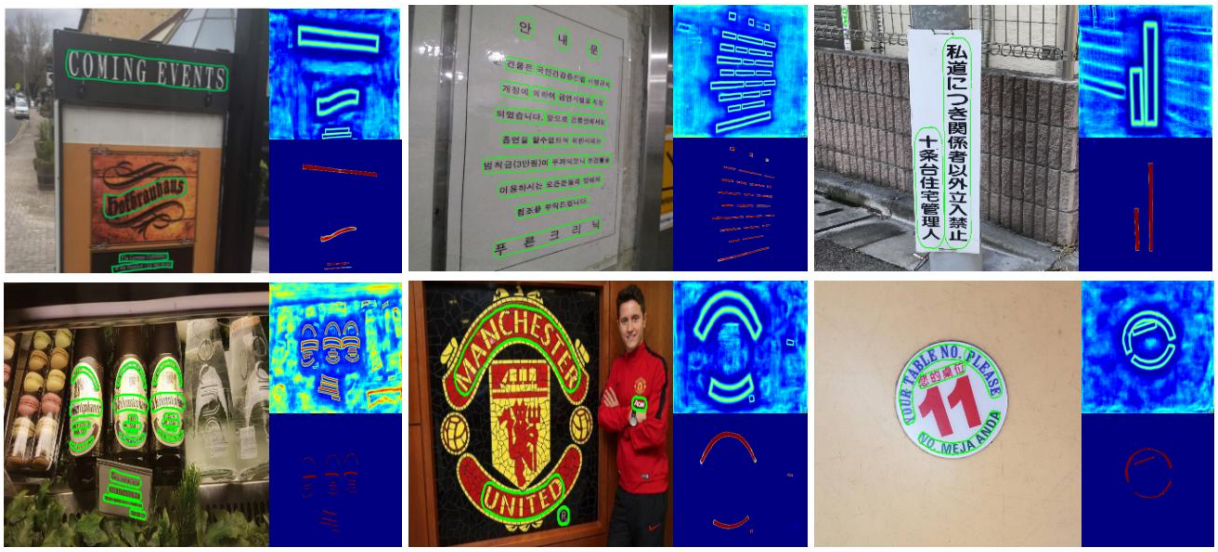

图 7:各种形状的文本实例的一些可视化结果,包括弯曲文本、多方向文本、垂直文本、
和长文本行。对于每个单元,右上角是阈值图;右下角是概率图。

表 1:不同设置的检测结果。“DConv”表示可变形卷积。“P”、“R”和“F”分别表示准确率、召回率和 f 值。

骨干 DConv DB		MSRA-TD500				CTW1500			
		磷	R	F	第一人称射击游戏	磷	R	F	第一人称射击游戏
ResNet-18 ×	×	85.5	70.8	77.4	66	76.3	72.8	74.5	59
ResNet-18	×	86.8	72.3	78.9	62	80.9	75.4	78.1	55
ResNet-18 ×		87.3	75.8	81.1	66	82.4	76.6	79.4	59
ResNet-18		90.4	76.3	82.8	62	84.8	77.5	81.0	55
ResNet-50 ×	×	84.6	73.5	78.7	40	81.6	72.9	77.0	二十七
ResNet-50	×	90.5	77.9	83.7	32	86.2	78.0	81.9	22
ResNet-50 ×		86.6	77.7	81.9	40	84.3	79.1	81.6	二十七
ResNet-50		91.5	79.2	84.9	32	86.9	80.2	83.4	22

9 种语言,代表 6 种不同的文字。共有 7,200 种
训练图像、1,800 张验证图像和 9,000 张测试图像
数据集中的图像。我们使用训练集和
微调期间的验证集。

ICDAR 2015 数据集（Karatzas et al. 2015)包含 1000
训练图像和 500 张测试图像,这些图像是
通过谷歌眼镜,分辨率为 720 × 1280。文本
实例在单词级别进行标记。

MSRA-TD500 数据集（Yao et al. 2012)是一个多语言
包含英语和中文的数据集。共有 300
训练图像和 200 张测试图像。文本实例
在文本行级别进行标记。继上一个
方法（Zhou et al. 2017;Lyu et al. 2018b;Long et al. 2017）
2018）,我们从 HUST-TR400（Yao、Bai 和 Liu 2014)中加入了额外的 400 张
训练图像。

CTW1500 数据集CTW1500 (Liu et al. 2019a) 是一个数据集
该数据集专注于弯曲文本。数据集包含 1000 张训练图像和 500 张测试图像。
文本实例如下:
在文本行级别进行注释。

Total-Text 数据集Total-Text (Chng and Chan 2017) 是一个
包含各种形状文本的数据集,包括
水平、多方向和弯曲。它们是1255张训练图像和300张测试图像。文本实例是

在单词级别进行标记。

实现细节

对于所有模型,我们首先使用 SynthText 进行预训练
数据集进行 10 万次迭代。然后,我们对模型进行微调
对应的真实世界数据集,共 1200 个 epoch。
训练批次大小设置为 16。我们遵循多项学习率
当前迭代的学习率等于
初始学习率乘以（1 −
$\frac{\text{伊特}}{\text{最大迭代次数}}$
)幂,其中
初始学习率设为 0.007,功率设为 0.9。我们
使用 0.0001 的权重衰减和 0.9 的动量。
max iter 表示最大迭代次数,取决于
最大时期数。

训练数据的数据增强包括:(1)
随机旋转,角度范围为（-10 ◦ ,10 ◦ ）;(2)
随机裁剪;(3)随机翻转。所有处理过的
为了提高训练效率,图像被重新调整大小为 640×640。

在推理期间,我们保持
测试图像并通过设置合适的
每个数据集的高度。推理速度通过
批量大小为 1,在单线程中使用单个 1080ti GPU。
推理时间成本包括模型前进时间
成本和后处理时间成本。后处理
时间成本约为推理时间的30%。

消融研究

我们对 MSRA-TD500 数据集进行了消融研究
和 CTW1500 数据集来展示我们的
提出了可微分二值化、可变形卷积和不同的主干模型。详细的实验

结果如表1所示。

可微分二值化在表 1 中,我们可以看到我们的
建议的数据库显著提高了
在两个数据集上都进行了 ResNet-18 和 ResNet-50 的测试。对于
ResNet-18 主干,DB 在 MSRA-TD500 上的 F 值性能提升了 3.7% 和 4.9%

数据集和 CTW1500 数据集。对于 ResNet-50 主干,DB 带来了 3.2%（在 MSRA-TD500 数据集上)和
4.6%（在 CTW1500 数据集上)的提升。此外，
由于可以在推理阶段移除 DB,因此速度
与没有DB的相同。

可变形卷积如表1所示,可变形卷积也能带来1.5-5.0的性能

因为它为骨干网络提供了一个灵活的接收场,并且额外时间成本很小,因此获得了更高的收益。对于 MSRA-TD500
数据集,可变形卷积增加了F-measure
分别提升 1.5%（使用 ResNet-18)和 5.0%（使用 ResNet-50）。对于
CTW1500 数据集,3.6%（使用 ResNet-18)和 4.9%
（使用 ResNet-50)通过可变形卷积实现了改进。

表 2:在 MLT-2017 数据集上监督阈值图的效果。“Thr-Sup”表示对

阈值图。

主干Thr-Sup	P	R	F	
ResNet-18 ×	81.3	63.1	71.0	41
ResNet-18	81.9	63.8	71.7	41
ResNet-50 ×	81.5	64.6	72.1	19
ResNet-50	83.1	67.9	74.7	19

监管门槛图虽然门槛
有/无监督的地图外观相似，
监督可以带来性能增益。如图所示
表 2,监督提高了 0.7% (ResNet-18) 和
MLT-2017 数据集上为 2.6%（ResNet-50）。
主干:建议使用 ResNet-50 主干的检测器
取得了比 ResNet-18 更好的性能,但运行
速度较慢。具体来说,最好的 ResNet-50 模型表现优于
比最好的 ResNet-18 模型快 2.1%（在 MSRA-TD500 上）
在 CTW1500 数据集上,平均速度提高了 1.7 倍,而 CTW2000 数据集上的速度提高了 2.4 倍,时间成本大约翻倍。

与以前的方法的比较

我们将我们提出的方法与以前的方法进行了比较
基于五项标准基准,包括两项
弯曲文本、多方向文本的一个基准,以及两个
针对长文本行的多语言基准测试。一些定性结果如图 7 所示。

表3:Total-Text数据集上的检测结果。括号中的值表示输入图像的高度。“*”

表示使用多个尺度进行测试。“MTS”和“PSE”是
Mask TextSpotter 和 PSENet 的缩写。

方法	磷	射频	
TextSnake（Long等人,2018）	82.7	74.5	78.4 -
ATRR（Wang 等人,2019b)	80.9	76.2	78.5 -
MTS（Lyu 等人,2018a)	82.5	75.6	78.6 -
TextField（Xu 等人,2019）	81.2	79.9	80.6 -
LOMO（Zhang 等人,2019)*	87.6	79.3	83.3 -
CRAFT（Baek等人,2019）	87.6	79.9	83.6 -
CSE（Liu 等人,2019b）	81.4	79.1	80.2 -
PSE-1s (Wang 等人 2019a)	84.0	78.0	80.9 3.9
DB-ResNet-18 (800)	88.3	77.9	82.8 50
DB-ResNet-50 (800)	87.1	82.5	84.7 32

表 4:CTW1500 的检测结果。
“*” 收集自 (Liu et al. 2019a)。
括号表示输入图像的高度。

方法	磷	射频	
CTPN*	60.4	53.8	56.9 7.14
EAST*	78.7	49.1	60.4 21.2
SegLink*	42.3	40.0	40.8 10.7
TextSnake (Long 等人 2018)	67.9	85.3	75.6 1.1
TLOC（Liu等人,2019a）	77.4	69.8	73.4 13.3
PSE-1s（Wang等人,2019a）	84.8	79.7	82.2 3.9
SAE（Tian等人,2019）	82.7	77.8	80.1 3
Ours-ResNet18（1024）	84.8	77.5	81.0 55
Ours-ResNet50（1024）	86.9	80.2	83.4 22

弯曲文本检测我们证明了
我们在两个曲线文本基准（Total-Text 和
CTW1500）。如表3和表4所示,我们的方法
在准确性和
速度。具体来说，“DB-ResNet-50”在Total-Text和CTW1500数据集上的表现分别比之前的最佳方法高出1.1%和1.2%。“DB-ResNet-50”的运行速度更快

比所有以前的方法都快,并且通过使用 ResNet-18 主干,速度可以进一步提高,性能下降较小。与最近基于分割的

检测器（Wang et al. 2019a)在 Total-Text 上的运行速度为 3.9 FPS，“DB-ResNet-50 (800)” 快 8.2 倍，“DB-ResNet-18 (800)” 快 12.8 倍。
多方向文本检测ICDAR 2015 数据集是
一个多方向的文本数据集,包含大量小而
低分辨率文本实例。在表 5 中,我们可以看到

“DB-ResNet-50 (1152)” 在准确率方面达到了最高水平。与之前最快的

方法 (Zhou et al. 2017) 中，“DB-ResNet-50 (736)” 的准确率比后者高出 7.2%，运行速度也快了一倍。对于“DB-ResNet-18 (736)”，当 ResNet-18 达到 48 fps 时，速度可达
应用于主干，f 值为 82.3。

表 5：ICDAR 2015 数据集上的检测结果。

括号中的值表示输入图像的高度。
“TB”和“PSE”分别是TextBoxes++和PSENet的缩写。

方法	磷	射频	
CTPN（Tian等人，2016）	74.2 51.6 60.9 7.1		
EAST（周等人，2017）	83.6 73.5 78.2 13.2		
SSTD（He 等人，2017a)	80.2 73.9 76.9 7.7		
WordSup（Hu等人，2017）	79.3 77 78.2 -		
角球（Lyu 等人，2018b）	94.1 70.7 80.7 3.6		
TB (Liao, Shi, and Bai 2018)	87.2 76.7 81.7 11.6		
RRD（Liao等人，2018）	85.6 79 72 80	82.2 6.5	
MCN（刘等人，2018）		76	-
TextSnake（Long等人，2018）	84.9 80.4 82.6 1.1		
PSE-1s (Wang 等人 2019a)	86.9 84.5 85.7 1.6		
SPCNet (Xie et al. 2019a)	88.7 85.8 87.2 -		
LOMO（Zhang 等人，2019)	91.3 83.5 87.2 -		
CRAFT（Baek等人，2019)	89.8 84.3 86.9 -		
SAE(720)（Tian 等人，2019）	85.1 84.5 84.8 3		
SAE(990)（田等人，2019)	88.3 85.0 86.6 -		
DB-ResNet-18 (736)	86.8 78.4 82.3 48		
DB-ResNet-50 (736)	88.2 82.7 85.4 26		
DB-ResNet-50（1152）	91.8 83.2 87.3 12		

多语言文本检测我们的方法具有鲁棒性
多语言文本检测。如表6和表7所示，
“DB-ResNet-50” 在准确率和速度上均优于以往的方法。准确率方面，“DB-ResNet-50” 比之前的 SOTA 方法高出 1.9%，

在 MSRA-TD500 和 MLT-2017 数据集上分别提高了 3.8%。在速度方面，“DB-ResNet-50” 的速度提高了 3.2 倍
比之前最快的方法（Liao et al. 2018）
MSRA-TD500 数据集。采用轻量级主干网络的“DB-ResNet-18 (736)”实现了与之前最先进的方法 (Liu et al.) 相当的准确率。

2018 年) (82.8 vs 83.0) 并以 62 FPS 运行，是
比之前最快的方法（Liao et al. 2018)更快，
MSRA-TD500。速度可进一步加快至
通过减小输入大小，实现 82 FPS（“ResNet-18 (512)”）。

局限性

我们的方法的一个限制是它不能处理
案例“文本中的文本”，这意味着文本实例
在另一个文本实例内。虽然缩小文本区域对于文本实例不在

另一个文本实例的中心区域，当
文本实例恰好位于另一个文本实例的中心区域
文本实例。这是基于分割的场景文本检测器的常见限制。

表 6：MSRA-TD500 数据集上的检测结果。
括号中的值表示输入图像的高度。

方法	磷	射频	
(He et al. 2016b) 71 61 69 DeepReg (He et al. 2017b) 77 70 74 RRPN (Ma et al. 2018) 82 68 74 RRD (Liao et al. 2018) 87 73 79 MCN (Liu et al. 2018) 88 79 83 PixelLink （邓等人，2018)83 73.2 77.8 3			- 1.1 - 10 -
角落（Lyu 等人，2018b）	87.6 76.2 81.5 5.7		
TextSnake（Long等人，2018）	83.2 73.9 78.3 1.1		
(Xue, Lu, and Zhan 2018)	83.0 77.4 80.1 -		
(Xue, Lu, and Zhang 2019)	87.4 76.7 81.7 -		
CRAFT（Baek等人，2019)	88.2 78.2 82.9 8.6		
SAE（田等人，2019)	84.2 81.7 82.9 -		
DB-ResNet-18 (512)	85.7 73.2 79.0 82		
DB-ResNet-18 (736)	90.4 76.3 82.8 62		
DB-ResNet-50 (736)	91.5 79.2 84.9 32		

表 7：MLT-2017 数据集上的检测结果。标有“*”的方法摘自 (Lyu et al. 2018b)。MLT-2017 数据集中的图像尺寸调整为 768 × 1024，

我们的方法。“PSE”是PSENet的缩写。

方法	磷	R	F	
SARI FDU RRPN V1* 71.2 55.5 62.4 商汤 OCR* 56.9 69.4 62.6 SCUT DLVlab1* 80.3 54.5 65.0 e2e ctc01 多尺度* 79.8 61.2 69.3 Corner (Lyu et al. 2018b) 83.8 55.6 66.8 PSE (Wang et al. 2019a) 73.8 68.2 70.9				- - - - - -
DB-ResNet-18	81.9	63.8 71.7 41		
DB-ResNet-50	83.1	67.9 74.7 19		

结论

本文提出了一种用于检测任意形状场景文本的新型框架，该框架在分割网络中引入了我们提出的可微分二值化过程（DB）。实验验证了我们的

方法（ResNet-50 主干）始终优于
在速度和准确率方面，它在五个标准场景文本基准上达到了最先进的水平。特别是，即使

借助轻量级主干（ResNet-18），我们的方法可以
在所有测试数据集上取得有竞争力的表现
具有实时推理速度。未来，我们有兴趣扩展我们的端到端文本识别方法。

致谢

该工作得到国家重点研发计划的支持
中国（编号 2018YFB1004600），由
国家拔尖青年人才支持计划和华中科技大学青年学术前沿计划

团队 2017QYTD08。

参考

Baek, Y.;Lee, B.;Han, D.;Yun, S.;以及 Lee, H.,2019 年。基于字符区域感知的文本检测。在 Proc. CVPR,9365–9374 中。

Chng, CK 和 Chan, CS 2017. Total-text:用于场景文本检测和识别的综合数据集。在 Proc. ICDAR,935–942 中。

Dai, J.;Qi, H.;Xiong, Y.;Li, Y.;Zhang, G.;Hu, H.;和 Wei, Y.
2017. 可变形卷积网络。In Proc. ICCV, 764–773。

Deng, D.;Liu, H.;Li, X.;以及 Cai, D.,2018 年。Pixellink:通过实例分割检测场景文本。在 Proc. AAAI 上。

Gupta, A.;Vedaldi, A.;以及 Zisserman, A.,2016 年。《自然图像中文本定位的合成数据》。刊登于 Proc. CVPR。

何凯;张晓玲;任胜;孙建军,2016a. 深度残差学习在图像识别中的应用。CVPR 论文集,770–778。

He, T.; Huang, W.; Qiao, Y.; 和 Yao, J. 2016b. 用于场景文本检测的文本注意卷积神经网络。IEEE Trans.
图像处理 25(6): 2529–2541。

何平;黄伟;何婷;朱倩;乔燕;李晓玲,2017a。
具有区域注意机制的单次文本检测器。In Proc. ICCV, 3047–3055。

He, W.; Zhang, X.; Yin, F.; 和 Liu, C. 2017b. 深度直接回归在多方向场景文本检测中的应用。在 Proc. ICCV 上发表。

胡,H。张,C.;罗,Y。王,Y。韩,J。和丁,E.2017。
Wordsup:利用词语标注进行基于字符的文本检测。In Proc. ICCV,4940–4949。

卡拉察斯,D.;戈麦斯-比戈达,L.;尼古拉,A.;戈什,S.K.;巴格丹诺夫,AD;岩村,M.;麦塔斯,J.;诺依曼,L.;钱德拉塞卡 (Chan-drasekhar),V.R.;卢,S。沙菲特,F.;内田,S.;和瓦尔尼,E.

2015. ICDAR 2015稳健读取竞赛。刊于 Proc. IC-DAR。

Kim, K.;Cheon, Y.;Hong, S.;Roh, B.;以及 Park, M.,2016 年。
PVANET:用于实时物体检测的深度轻量级神经网络。CoRR abs/1608.08021。

Liao, M.; Shi, B.; Bai, X.; Wang, X.; 和 Liu, W. 2017. 文本框:基于单个深度神经网络的快速文本检测器。在 Proc.
AAAI。

廖 M.;朱 Z.;石 B.;夏 G.;白 X.,2018。用于定向场景文本检测的旋转敏感回归。在 Proc.

CVPR, 5909–5918.

廖敏;吕平;何敏;姚晨;吴伟;白雪,2019 年。
蒙版文本识别器:一种端到端可训练神经网络,用于识别任意形状的文本。IEEE 模式分析与机器翻译汇刊,2017 年,第 157-165 页。
智力。

廖明;石斌;白晓玲,2018. Textboxes++:面向单样本场景文本检测器。IEEE 图像处理学报 27(8):3676–3690。

Liu, Y. 和 Jin, L. 2017. 深度匹配先验网络:迈向更紧密的多方向文本检测。在 Proc. CVPR 中。

刘,W.;安格洛夫,D.;埃尔汗,D.;塞格迪,J.;和东南部里德
2016. SSD:单次多框检测器。在 Proc. ECCV 中。

刘,Z.;林,G。杨,S。冯,J。林,W。和 Goh,WL 2018。
学习马尔可夫聚类网络进行场景文本检测。在 Proc. CVPR,6936–6944 中。

刘英;金琳;张淑英;罗晨;张淑英,2019a。通过横向和纵向序列连接检测弯曲场景文本。模式识别 90:337–345。

Liu, Z.;Lin, G.;Yang, S.;Liu, F.;Lin, W.;以及 Goh, WL 2019b。
基于条件空间扩展的稳健曲线文本检测。Proc. CVPR,7269–7278。

Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; 和 Yao, C. 2018 年。
Textsnake:一种用于检测任意形状文本的灵活表示方法。ECCV 论文集,第 20-36 页。

Lyu, P.;Liao, M.;Yao, C.;Wu, W.;Bai, X. 2018a. Mask textspotter:一种用于识别任意形状文本的端到端可训练神经网络。载于 Proc. ECCV,67–83。

Lyu, P.; Yao, C.; Wu, W.; Yan, S.; 以及 Bai, X. 2018b. 通过角点定位和区域分割实现多方向场景文本检测。在 Proc. CVPR,7553–7563 中。

马建军; 邵伟; 叶华; 王琳; 王辉; 郑颖; 薛晓玲; 2018 年。通过旋转提议实现任意场景文本检测。IEEE 多媒体学报 20(11):3111–3122。

Shi, B.;Bai, X.;以及 Belongie, SJ,2017 年。通过链接片段检测自然图像中的定向文本。在 Proc. CVPR 上。

Tian, Z.; Huang, W.; He, T.; He, P.; and Qiao, Y. 2016. 使用联结文本提议网络检测自然图像中的文本。在 Proc. ECCV 中。

Tian, Z.; Shu, M.; Lyu, P.; Li, R.; Zhou, C.; Shen, X.; and Jia, J.
2019.学习形状感知嵌入以进行场景文本检测。
在 Proc. CVPR,4234–4243 中。

Vati, BR 1992. 多边形裁剪的通用解决方案。ACM通讯 35(7):56–64。

王伟;谢英;李晓燕;侯伟;陆婷;余刚;邵胜。2019a。基于渐进尺度扩展网络的形状稳健文本检测。CVPR 论文集,9336–9345。

Wang, X.;Jiang, Y.;Luo, Z.;Liu, C.-L.;Choi, H.;以及 Kim, S. 2019b。基于自适应文本区域表示的任意形状场景文本检测。在 Proc. CVPR,6449–6458 中。

Xie, E.;Zang, Y.;Shao, S.;Yu, G.;Yao, C.;和 Li, G. 2019a。
基于监督金字塔上下文网络的场景文本检测。载于 Proc. AAAI,第 33 卷,9038–9045。

Xie, L.;Liu, Y.;Jin, L.;以及 Xie, Z. 2019b. Derpn:迈向更通用的物体检测。刊于 Proc. AAAI,第 33 卷,9046–9053。

徐勇;王勇;周伟;王勇;杨哲;白雪,2019 年。
文本场:学习深度方向场用于不规则场景文本检测。IEEE 图像处理学报 28(11):5566–5579。

Xue, C.;Lu, S.;以及 Zhan, F.,2018 年。通过边界语义感知和引导实现准确的场景文本检测。在 Proc.

ECCV, 355–372.

Xue, C.;Lu, S.;以及 Zhang, W. 2019. MSR:用于场景文本检测的多尺度形状回归。IJCAI Pro,989–995。

Yao, C.; Bai, X.; and Liu, W. 2014. 多方向文本检测与识别的统一框架。IEEE 图像处理学报 23(11):4737–4749。

Yao, C.; Bai, X.; Liu, W.; Ma, Y.; 和 Tu, Z. 2012. 检测自然图像中任意方向的文本。在 Proc. CVPR 中。

Zhang, Z.;Zhang, C.;Shen, W.;Yao, C.;Liu, W.;和 Bai, X. 2016。
基于全卷积网络的多方向文本检测。在 Proc. CVPR 上。

Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; 和 Ding, X. 2019. 多次查看:一种用于任意形状文本的精确检测器。在 Proc. CVPR 中。

Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; and Liang, J. 2017. EAST:一种高效准确的场景文本检测器。在 Proc. CVPR 上。

Zhu, X.;Hu, H.;Lin, S.;以及 Dai, J. 2019. 可变形卷积网络 v2:更佳可变形性,更佳结果。刊于 Proc. CVPR,9308–9316。