智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

双分支跨级特征融合的自然场景文本检测

刘光辉, 张钰敏, 孟月波, 占华

引用本文:
刘光辉,张钰敏,孟月波,占华. 双分支跨级特征融合的自然场景文本检测[J]. 智能系统学报, 2023, 18(5): 1079–1089.
LIU Guanghui, ZHANG Yumin, MENG Yuebo, et al. Natural scene text detection based on double-branch cross-level feature fusion[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(5): 1079–1089.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202303005

DOI: 10.11992/tis.202303005
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20230809.1531.002

双分支跨级特征融合的自然场景文本检测

刘光辉, 张钰敏, 孟月波, 占华
(西安建筑科技大学 信息与控制工程学院, 陕西 西安 710055)

摘　要: 现有的场景文本检测方法在处理任意形状文本时, 由于复杂背景的影响会造成文本区域定位不准确、相邻文本漏检误检的问题, 基于此提出一种双分支跨级特征融合的自然场景文本检测方法。首先, 以 Resnet50 为主干网络提取初始特征, 设计跨级特征分布增强模块(cross-level feature distribution enhancement module, CFDEM), 增强跨级特征文本信息的交互性, 提高特征的表达能力; 然后, 为自适应地选择过滤非文本或冗余特征, 降低误检率和漏检率, 提出自适应融合策略(adaptive fusion strategy, AFS), 利用双分支结构加强不同维度特征之间的联系, 优化融合过程; 最后, 预测阶段采用可微分二值化的方法来生成文本检测结果。所提方法在 ICDAR2015、ICDAR2017、Total-Text、CTW1500 数据集上进行消融实验, 实验结果表明该方法能准确定位文本区域, 克服文本漏检误检影响。

关键词: 文本检测; 任意形状; 跨级特征分布增强; 自适应融合; 双分支; 空间维度; 通道维度; 可微分二值化
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)05−1079−11

中文引用格式: 刘光辉, 张钰敏, 孟月波, 等. 双分支跨级特征融合的自然场景文本检测 [J]. 智能系统学报, 2023, 18(5): 1079–1089.
英文引用格式: LIU Guanghui, ZHANG Yumin, MENG Yuebo, et al. Natural scene text detection based on double-branch cross-level feature fusion[J]. CAAI transactions on intelligent systems, 2023, 18(5): 1079–1089.

Natural scene text detection based on double-branch cross-level feature fusion

LIU Guanghui, ZHANG Yumin, MENG Yuebo, ZHAN Hua
(School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: Current scene text detection methods cause the inaccurate location of text regions and false detection of adjacent texts due to the influence of complex backgrounds in arbitrarily shaped texts. To solve this issue, a natural scene text detection method based on double-branch cross-level feature fusion is proposed. First, the initial features were extracted using Resnet50 as the backbone network, and then a cross-level feature distribution enhancement module was designed to improve the interaction of cross-level feature text information and the expression ability of features. Second, an adaptive fusion strategy was proposed to filter nontext or redundant features adaptively and reduce the false and missed detection rates using the double-branch structure to strengthen the relationship between different dimensional features and optimize the fusion process. Last, the differential binarization method was used to yield text detection results in the prediction phase. The proposed method was employed to perform ablation experiments on the ICDAR2015, ICDAR2017, Total-Text, and CTW1500 datasets. The findings revealed that this method can accurately locate the text area and overcome the impact of text miss and false detections.

Keywords: text detection; arbitrarily shaped; cross-level feature distribution enhancement; adaptive fusion; double branch; spatial dimension; channel dimension; differentiable binarization

由于图像文本中包含着大量的信息[1], 会出现在许多应用领域包括教育、物流、旅游等, 故场景文本检测在日常生活中发挥着重要的作用[1-2]。

收稿日期: 2023−03−02.　**网络出版日期:** 2023−08−09.
基金项目: 国家自然科学基金项目(52278125); 陕西省重点研发计划(2021SF-429).
通信作者: 刘光辉. E-mail: guanghuil@163.com.

在自然场景中图像文本形状任意，文本颜色和字体也不同，还存在不同程度的遮挡和文本行模糊不清的情况，其多样性和多变性给文本检测带来了巨大的挑战。

近年来，基于深度学习的场景文本检测因良好的检测效果而逐渐成为主流方向，其主要分为基于回归和基于分割的方法。基于回归的方法是根据文本特点对通用目标检测算法进行改进，使用回归文本框获取文本。Tian 等[3] 提出 CTPN 算法是基于 Faster RCNN(region-based convolutional neural networks) 框架，用一个固定宽度大小的锚框通过双向长短记忆神经网络进行文本检测，能准确定位水平文本，但无法处理多方向文本；Shi 等[4] 提出 Seglink 算法，将 CTPN 小尺度候选框和 SSD 算法融合，能处理多方向文本，但其合并算法采用线性回归方式，只能拟合直线无法拟合曲线，不能检测曲线文本；Wang 等[5] 提出 Contour-Net 算法，在文本轮廓点上建模提取特征信息，用临界点来表示文本区域，能对不同曲线文本进行检测，但其 pipeline 较长，计算成本较大，且效果过于依赖超参数。基于回归的方法需要设计锚框，易受文本边界坐标限制，在任意形状文本检测时具有局限性。

基于分割的方法是从图像分割中吸取经验，在像素级别预测分割出图像文本，能检测任意形状文本。Wang 等[6] 提出 PSENet 算法，利用渐近式扩展方法将文本进行像素级分类，检测多个不同尺度的文本区域，该方法能描述任意形状文本，但其后处理很复杂，模型预测速度慢；Wang 等[7] 设计了轻量级骨干网络和低成本 FPN 来预测文本区域，虽明显提高了推理速度，但仍依赖于遍历文本实例的所有像素，当图像文本中有较多文本实例时，会导致计算负担增大；Liao 等[8] 在 DB-Net 中提出一种可微分二值化模型，将其插入到分割网络中进行联合优化，降低了后处理算法的计算成本，但对网络中的语义信息利用不充分，其检测性能受到限制；文献 [9] 对多级通道注意力信息进行适当编码，构建判别特征图，从而提高文本检测的性能；文献 [10] 通过多维度卷积融合的方法来减少信息损失，提升检测性能，但对于高度弯曲形状的文本建模能力有限；文献 [11] 提出一种透视轮廓连接算法来生成区间区域轮廓，可以有效拟合高度弯曲文本轮廓；FARNet[12] 提出 TFS 和 DGAT 的联合模块来推断文本片段之间的链接关系，提高长弯曲文本的分组能力，进而提升检测性能；FCENet[13] 针对高度弯曲形状文本提出傅里叶轮廓嵌入方法对文本进行建模，能够灵活拟合各种不规则文本，但在常规文本检测上其效果略有下降；文献 [14] 利用 B-Spline 曲线直接预测轮廓点的方法，能检测任意形状文本；文献 [15] 采用 Sigmoid Alpha 函数来建模边界与内部像素之间的距离关系，减少噪声和缺陷，完整的重建出任意形状的文本区域；同时文献 [16] 利用空间自适应卷积来提升文本检测性能，并提高对任意形状文本检测在实际应用中的鲁棒性，但文本图像质量不同仍会造成文本区域定位不准确和文本漏检误检。

基于以上分析，本文提出一种双分支跨级特征融合的自然场景文本检测方法。首先，设计跨级特征分布增强模块（cross-level feature distribution enhancement module, CFDEM），从初始特征中分别提取全局和局部信息进行编码，实现跨级特征交互，增强特征的表达能力；然后，提出自适应融合策略（adaptive fusion strategy, AFS），以双分支的结构分别构建空间维度和通道维度来计算相关权重，加强不同尺度特征之间的联系，能够自适应地选择过滤非文本或冗余特征，降低误检率和漏检率；最后，采用可微分二值化的方法对文本进行处理生成文本检测结果。

1　文本检测方法

1.1　网络结构

本文方法整体结构如图 1 所示，包括特征提取部分、自适应融合部分和可微分二值化部分。特征提取部分以 Resnet50 为骨干网络，输入图像以前向传播方式进行采样，得到初始特征 $\{F_2,F_3,F_4,F_5\}$，将 $\{F_3,F_4,F_5\}$ 输入到跨级特征分布增强模块中，增强跨级特征文本信息的交互性，得到增强后的特征 $\{M_3,M_4,M_5\}$；自适应融合部分为了帮助网络能自适应地过滤非文本或冗余特征，降低误检和漏检，通过双分子结构构建不同维度特征之间的权重关系，进行自适应的融合；可微分二值化部分利用融合的特征图预测概率图和阈值图并进行可微分二值化，得到最终的文本检测结果。

1.2　跨级特征分布增强模块

在网络训练中由于 Resnet50 的网络层数较浅，会造成提取的特征缺乏语义信息，导致文本检测效果差，且出现检测目标误检漏检的现象。基于此，本文设计如图 2 所示的跨级特征分布增强模块，提取不同层级特征信息，同时增强跨级特征文本信息的交互性。

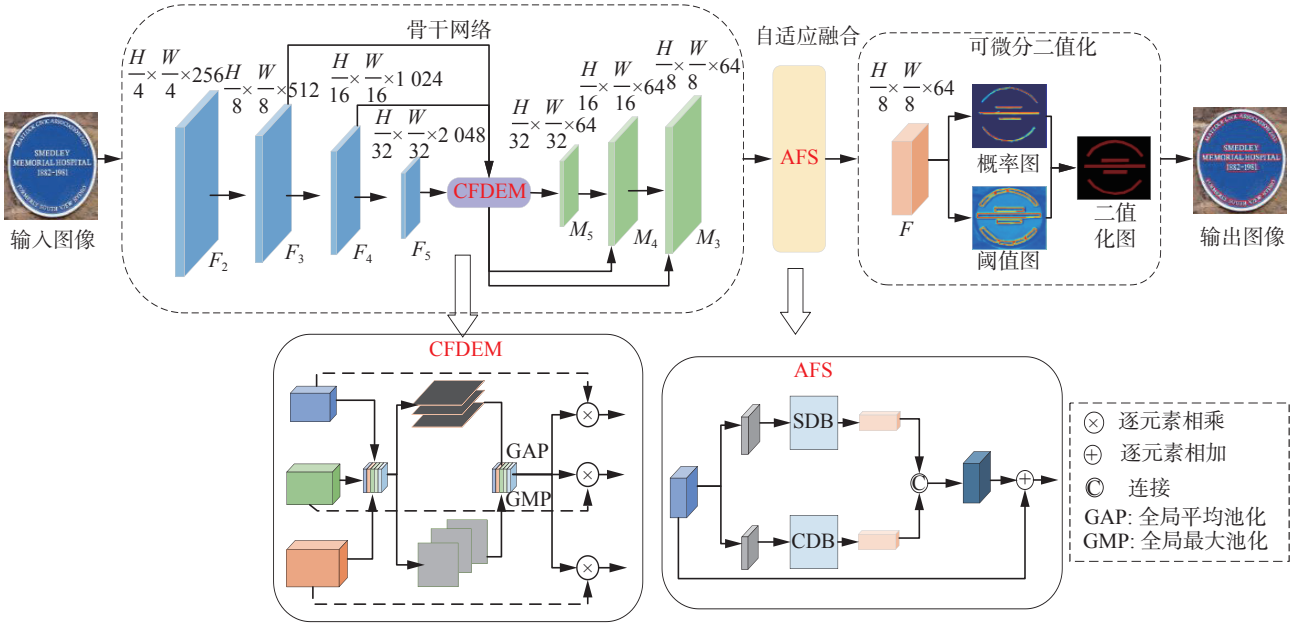


图 1　文本检测框架结构

Fig. 1　Text detection framework structure

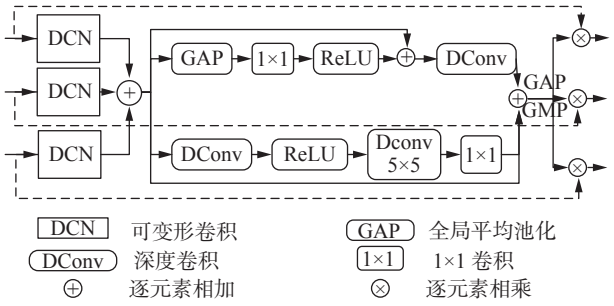


图 2　跨级特征分布增强模块

Fig. 2　Cross-level feature distribution enhancement

由于骨干网络提取的初始特征$\{F_3,F_4,F_5\}$分辨率为原始图像的$\{1/8,1/16,1/32\}$且包含较多语义信息，故将$\{F_3,F_4,F_5\}$作为跨级特征分布增强模块的输入。首先，为了增强任意形状文本的适应性，利用可变形卷积使卷积核的形状跟随不同文本形状而发生改变，将其特征进行逐元素相加得到特征和F^D，过程如下：

$$F^D=\sum_{i=3}^{5} f_{\mathrm{DCN}}(F_i) \tag{1}$$

式中：f_{DCN}为可变形卷积，F_i为初始特征。

然后，为了提取不同层级特征信息，增强跨级特征文本信息的交互性，通过双分支框架来捕获更多的特征信息。一个分支为全局分支，结构如图 2 所示，输入特征和F^D通过全局平均池化来捕获全局上下文信息，并且利用1×1卷积和 ReLU 激活函数来增强网络对特征信息的表达能力，其过程为

$$F_1=f_{\mathrm{ReLU}}\left(f_{\mathrm{Conv}_{1\times1}}\left(f_{\mathrm{GAP}}\left(F^D\right)\right)\right) \tag{2}$$

式中：F^D为经可变形卷积后进行逐元素相加的特征和，f_{GAP}为全局平均池化，$f_{\mathrm{Conv}_{1\times1}}$为$1\times1$卷积，$f_{\mathrm{ReLU}}$为 ReLU 激活函数。为了获取更多特征信息，将式(2)与F^D进行逐元素相加。同时空间位置信息对定位文本边界非常重要，可以有效提升文本检测能力，为了突出文本空间位置，将其进行深度可分离卷积处理，即

$$F_1'=f_{\mathrm{DConv}}\left(F_1+F^D\right) \tag{3}$$

式中：F_1为全局分支中增强后的特征，F_1'为全局分支输出的特征，f_{DConv}为深度可分离卷积。

另一个分支为局部分支，该分支先通过深度可分离卷积来捕获局部上下文信息，利用 ReLU 激活函数增强网络对局部特征信息的表达能力，再经过卷积核为5×5的深度可分离卷积层，深化网络的同时扩大感受野感知特征交互关系，过程为

$$F_2=f_{\mathrm{DConv}_{5\times5}}\left(f_{\mathrm{ReLU}}\left(f_{\mathrm{DConv}}\left(F^D\right)\right)\right) \tag{4}$$

式中：$f_{\mathrm{DConv}_{5\times5}}$为卷积核为$5\times5$的深度可分离卷积。同理该分支为了获取更多的特征信息，将式(4)与F^D进行逐元素相加，具体为

$$F_2'=f_{\mathrm{Conv}_{1\times1}}(F_2)\oplus F^D \tag{5}$$

式中：F_2为局部分支中增强后的特征，F_2'为局部分支输出的特征。

最后，为了收集不同层级特征信息，将双分支输出特征分别与输入特征进行逐元素相乘来促进

跨级特征的有效使用，提高任意形状文本检测能力。其中为了使双分支输出特征与输入顺利进行逐元素相乘，通过全局平均池化和全局最大池化来完成。跨级特征分布增强模块输出结果为

$$Y=(f_{\mathrm{GAP}}(F_1'\oplus F_2')f_{\mathrm{GMP}})\otimes F_i(i\in 3,4,5) \tag{6}$$

1.3　自适应融合策略

特征融合是提高检测性能的一个重要手段[17]，在融合过程中由于场景文本图像包含大量环境信息，会导致特征中包含冗余信息，降低场景文本检测效果。同时文本多尺度特征图感受野会造成特征信息之间存在差异，直接融合多尺度空间特征信息易形成噪声等，造成检测时出现混淆定位或漏检误检的问题。基于此，本文提出自适应融合策略，通过构建空间维度和通道维度来捕获像素的相关性和通道的依赖关系，优先将更多注意力专注于重要元素上，自适应地判断冗余部分，实现特征充分融合，生成更加丰富的特征图。

本文的自适应融合策略整体结构如图 3。

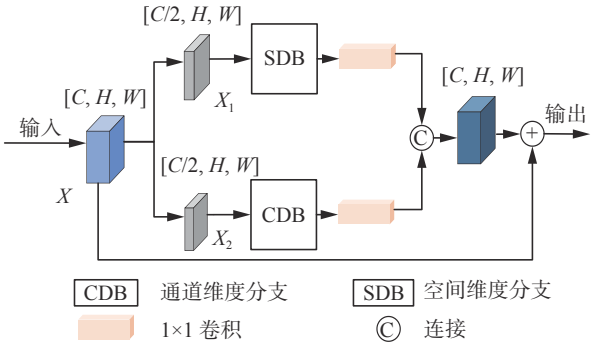


图 3　自适应融合策略

Fig. 3　Adaptive fusion strategy

给定特征图 X，维度大小为 $[C,H,W]$，其中通道数为 C、高度为 H、宽度为 W。首先，使用 1×1 卷积调整通道，从信道维度将 X 分成通道维度分支（channel dimension branch, CDB）和空间维度分支（spatial dimension branch, SDB），维度大小为 $X_1=X_2=[C/2,H,W]$，分别处理通道维度和空间维度。然后，将通道维度分支和空间维度分支两个不同的层级输出经卷积后连接，生成大小为 $[C,H,W]$ 的综合特征图；最后，为确保融合过程中网络不会过度关注感兴趣的区域，从而导致其他区域的权重不合理，增加了一条残差连接，来提高融合效果。具体过程如下：

$$\xi_1=f_{\mathrm{Conv}_{1\times1}}(f_{\mathrm{SDB}}(X_1)),\ \xi_1\in R^{[C/2,H,W]} \tag{7}$$

$$\xi_2=f_{\mathrm{Conv}_{1\times1}}(f_{\mathrm{CDB}}(X_2)),\ \xi_2\in R^{[C/2,H,W]} \tag{8}$$

$$F=(f_C(\xi_1,\xi_2))\oplus X,\ F\in R^{[C,H,W]} \tag{9}$$

式中：ξ_1 为空间维度分支，ξ_2 为通道维度分支，F 为融合网络输出的结果。

1.3.1　通道维度分支

通道维度分支中原始特征图各通道权重相同，会削弱文本重要特征，本文提出的通道维度分支专注于获取不同通道之间的重要性，赋予每个通道相应的权重，可以更加准确地关注重要文本信息，削减一些不必要的参数，使模型更加准确地预测文本位置，并降低计算负担。结构如图 4 所示。

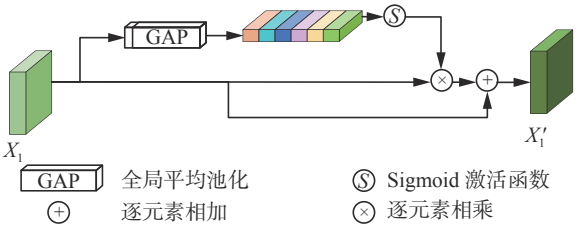


图 4　通道维度分支

Fig. 4　Channel dimension branch

首先使用全局平均池化获取通道重要性 u，即

$$u=f_{\mathrm{GAP}}(X_1)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}X_1(i,j) \tag{10}$$

式中 f_{GAP} 为全局平均池化。为防止网络过度关注感兴趣区域，添加一个残差连接来防止通道被过度放大或抑制，避免其他区域权重不合理，更好地对通道重要性进行建模。然后利用激活函数为每个通道生成权重集合，进而反应通道相关性，最后输出为

$$X_1'=f_{\mathrm{sig}(u)}\otimes X_1\oplus X_1,\ X_1'\in R^{[C/2,H,W]} \tag{11}$$

式中：$f_{\mathrm{sig}(u)}$ 为 Sigmoid 函数，X_1' 为通道维度分支输出。

1.3.2　空间维度分支

通道维度分支主要关注特征图中的通道权重，缺乏空间相关性。空间维度分支可以通过捕获像素之间的相关性来重新分配文本区域中字符和非字符区域之间的权重，使模型更加关注有用的特征区域，减少无用信息的干扰，提高文本区域的检测精度。空间维度分支通过对文本的空间信息编码，能适应多尺度文本，既能满足大尺度文本的需求，又能提高小尺度文本的检测效果。结构如图 5 所示。

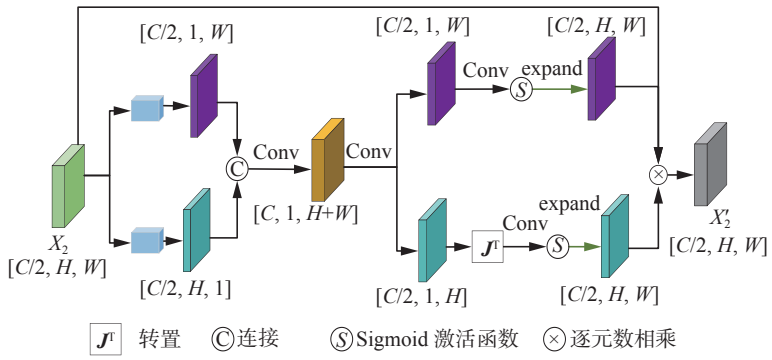


图 5　空间维度分支

Fig. 5　Spatial dimension branch

空间维度分支在水平和垂直方向上对X_2进行编码, 分别使用大小为$(H,1)$和$(1,W)$卷积核进行卷积和全局平均池化, 此时高度 h 和宽度 w 处输出为

$$X_2^h(h)=\frac{1}{W}\sum_{0<i<W}X_2(h,i),\quad X_2^h\in R^{[C/2,H,1]} \tag{12}$$

$$X_2^w(w)=\frac{1}{H}\sum_{0<j<H}X_2(j,w),\quad X_2^w\in R^{[C/2,1,W]} \tag{13}$$

式中: X_2^h为高度方向的变换, X_2^w为宽度方向的变换。这两个变换被聚合成沿空间的两个方向, 即高度和宽度方向的单独的方向感知特征。由于两个变换的方向不同, 当 AFS 在一个方向上捕获要素的更长依赖性时, 不会影响另一个方向上的位置信息。

将X_2^h和X_2^w连接, 经过 1×1 卷积层后发送到激活功能层, 生成维度大小为$[C,1,H+W]$的特征, 将其沿宽度维度分成两个子张量, 其维度大小分别为$x_h^c=[C/2,1,H]$、$x_w^c=[C/2,1,W]$。将X_h^c经$\boldsymbol{J}^{\mathrm{T}}$转置为$[C/2,H,1]$, 让$X_h^c$和$X_w^c$分别进入 1×1 卷积层, 进一步建立空间映射, 此时输出特征上加入 sigmoid 函数, 以获得水平和垂直方向上的权重特征图, 即

$$X_2^c=f_{\mathrm{ReLU}}\left(f_{\mathrm{Conv}_{1\times1}}\left(f_C\left(X_2^h,X_2^w\right)\right)\right),\quad X_2^c\in R^{[C,1,H+W]} \tag{14}$$

$$Q_h=f_{\mathrm{sig}}\left(f_{\mathrm{Conv}_{1\times1}}\left(X_h^c\right)\right),\quad Q_h\in R^{[C/2,H,1]} \tag{15}$$

$$Q_w=f_{\mathrm{sig}}\left(f_{\mathrm{Conv}_{1\times1}}\left(X_w^c\right)\right),\quad Q_w\in R^{[C/2,1,W]} \tag{16}$$

式中: f_C为串联操作, Q_h为水平方向权重特征图, Q_w为垂直方向权重特征图。

最后为了调整水平和垂直方向的权重分布, 更好地捕获空间维度中的重要区域, 将Q_h和Q_w扩展到和X_2一样大, 然后将两个权重特征图和X_2相乘, 输出X_2'的结果为

$$X_2'=X_2(i,j)\times Q_h(i)\times Q_w(j),\quad X'\in R^{[C/2,H,W]} \tag{17}$$

1.4　可微分二值化

可微分二值化如图 1 右部分所示, 采用 DB-net[8] 中的二值化方法。利用融合后的特征图预测概率图和阈值图, 再进行处理生成近似二值化图, 即

$$B_{i,j}=\frac{1}{1+\mathrm{e}^{-k\left(P_{i,j}-T_{i,j}\right)}} \tag{18}$$

式中 (i,j) 表示像素点, $B_{i,j}$、$P_{i,j}$、$T_{i,j}$分别为近似二值化图、概率图、阈值图上 (i,j) 点的值; k 为放大因子, 根据经验设置为 50。

2　实验结果分析

本文所有实验均在 Ubuntu 系统下进行, GPU 型号为 RTX2080Ti, 环境配置为 CUDA9.0+anacond−a3+Python3+Tensorflow1.8.0。主干网络 ResNet50 选择 ImageNet 预训练结果作为初始化参数, 其余模块的初始化参数采用随机生成方式。使用带动量的随机梯度下降算法 (SGD) 来优化模型, 能够处理大量文本数据, 提高模型的性能和训练效率。初始学习率能加速优化算法的收敛速度, 其设置过小如 0.001 时, loss 会无法收敛, 设置过大会导致训练不稳定, 无法收敛, 故本文将初始学习率设置为 0.005。动量为 0.9, 减少训练过程中的梯度震荡现象, 权重衰减系数为 0.000 1, 防止模型过拟合。损失函数采用交叉熵损失。Batch-Size 设置为 8, 迭代次数为 5×10^4 时, 能获得最优的检测结果。采用准确率 P(Precision)、召回率 R(Recall) 和 F 值 (F-score) 来评价模型的精度, 使用参数量 (parameters, Param) 和浮点运算次数 (floating point operations, FLOPs) 对模型的复杂度进行评估。

2.1　对比实验

本文在 ICDAR2015、Total-Text、CTW1500 和 ICDAR2017 等 4 个数据集下进行实验, 结果如表 1~4 所示, 部分检测效果对比如图 6 所示。

表 1　ICDAR2015 多算法性能指标结果对比

Table 1　Comparisons of performance index results of multiple algorithms in the ICDAR2015

方法	主干网络	P/%	R/%	F/%	Param/10^6	FLOPs/10^9
SegLink[4]	VGG	73.1	76.8	75.0	27.3	82.1
CTPN[3]	VGG	74.2	51.6	60.9	17.7	29.2
EAST[18]	VGG	83.6	73.5	78.2	26.2	36.4
PSENet[5]	Resnet50	81.5	79.7	80.6	—	—
LOMO[19]	Resnet50	83.5	89.3	87.2	—	—
TextSnake[20]	VGG	84.9	80.0	82.6	19.1	37.3
PAN[7]	Resnet	85.5	81.9	82.9	12.5	7.0
DBnet[8]	Resnet50	88.2	82.7	85.4	—	—
SPCNet[21]	Resnet	88.7	85.8	87.2	—	—
CRAFT[22]	VGG16	89.8	84.3	86.9	—	—
Boundary[23]	Resnet50	88.1	82.2	85.0	—	
DRRG[24]	VGG16	88.5	84.7	86.6	—	—
ContourNet[5]	Resnet50	87.6	86.1	86.9	96.6	184.3
RSCA[25]	Resnet50	87.2	82.7	84.9	—	—
STKM[26]	Resnet	88.7	84.8	86.7	—	—
TextMountain[27]	Resnet50	88.5	84.2	86.3	59.8	98.5
本文	Resnet50	**91.9**	**88.5**	**90.1**	**82.2**	**96.1**

注: 表中“—”表示原文献中未给出, 无法对其进行衡量

表 2　Total-Text 多算法性能指标结果对比

Table 2　Comparisons of performance index results of multiple algorithms in the Total-Text　%

方法	主干网络	P	R	F
SegLink[4]	VGG	30.3	23.8	26.7
EAST[18]	VGG16	36.2	50.0	42.0
TextSnake[20]	VGG16	82.7	74.5	78.4
LOMO[19]	Resnet50	87.6	79.3	83.3
PSENet[6]	Resnet50	84.0	78.0	80.9
PAN[7]	Resnet50	89.3	81.09	85.0
DBnet[8]	Resnet50	87.1	82.5	84.7
Boundary[23]	Resnet50	85.2	83.5	84.3
TextRay[28]	Resnet50	83.5	77.9	80.6
CRAFTS[29]	Resnet50	89.5	85.4	87.4
ABCNet[30]	Resnet50	87.9	81.3	84.5
RSCA[25]	Resnet50	86.6	83.3	85.0
STKM[26]	Resnet	86.3	78.3	82.2
本文	Resnet50	**90.3**	**83.1**	**86.8**

表 3　CTW1500 多算法性能指标结果对比

Table 3　Comparisons of performance index results of multiple algorithms in the CTW1500　%

方法	主干网络	P	R	F
SegLink[4]	VGG	42.3	40.0	40.8
TextSnake[20]	VGG16	67.9	85.3	75.6
EAST[18]	VGG16	78.7	49.1	60.4
PSENet[6]	Resnet50	80.6	75.6	78.0

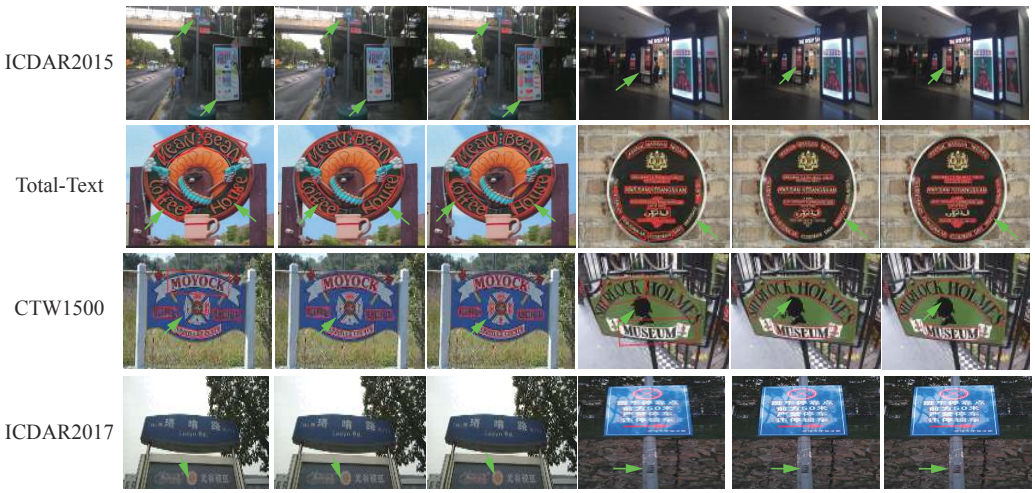
续表 3

方法	主干网络	*P*	*R*	*F*
CTPN[3]	VGG16	60.4	53.8	56.9
PAN[7]	Resnet50	86.4	81.2	83.7
DBnet[8]	Resnet50	86.9	80.2	83.4
TextRay[28]	Resnet50	82.8	80.4	81.6
ABCNet[30]	Resnet50	83.8	79.1	81.4
STKM[26]	Resnet	85.1	78.2	81.5
TextFuseNet[31]	Resnet50	85.0	85.8	85.4
TextMountain[27]	Resnet50	82.9	83.4	83.2
本文	Resnet50	**87.3**	**82.0**	**84.1**

表 4　ICDAR2017 多算法性能指标结果对比

Table 4　Comparisons of performance index results of multiple algorithms in the ICDAR2017　　%

方法	主干网络	*P*	*R*	*F*
文献[32]	Resnet	64.6	53.8	58.7
PixelLink[33]	VGG16	69.0	64.8	66.8
SPCNet[21]	Resnet50	73.4	66.9	70.0
PSENet[6]	Resnet50	73.5	67.8	70.5
LOMO[19]	Resnet50	78.8	60.6	68.5
CTPN[3]	VGG16	76.3	61.6	68.1
CRAFT[22]	VGG16	80.6	68.2	73.9
FOTS[34]	Resnet50	79.5	57.5	66.7
CharNet[35]	Resnet50	77.1	70.0	73.4
本文	Resnet50	**81.2**	**73.2**	**74.9**



(a) DBnet 效果 1　(b) PSEnet 效果 1　(c) 本文 效果 1　(d) DBnet 效果 2　(e) PSEnet 效果 2　(f) 本文 效果 2

图 6　部分检测效果图

Fig. 6　Partial detection effect diagram

ICDAR2015 数据集是文本检测任务中最常用的数据集, 共包含 1 500 张图像。根据表 1 分析可知, 本文提出的方法准确率达到 91.9%, 在主干网络一致的情况下, 与 TextMountain 相比, 准确率提高了 3.4%, 且与其他主干网络一致的算法相比, 本文算法的准确率均有一定的提升, 在主干网络不一致的情况下, 本文算法的准确率仍优于其他算法; 召回率比 DBnet、PSENet 等经典算法提高 5.8%、8.8%, 优于其他算法; *F* 值分别超过 DBnet、PSENet 等算法 4.7%、9.5%, 且本文方法优于大多

数先进的方法，说明本文模型更有效果。同时本文模型的总参数量达到 82.2×10^6，FLOPs 为 96.1×10^9，在提高文本区域定位准确率的过程中，牺牲了一定的模型复杂度性能。由于 Param 和 FLOPs 主要和模型本身有关，因此只在一个数据集上进行实验结果说明。

Total-Text 数据集是 2017 年提出的用于任意形状场景文本的数据集，该数据集从各种场景中采集，包含文本场景复杂度和低对比度背景，总共包含 1 555 张图像。由表 2 数据分析可知，本文算法准确率达到 90.3%，召回率达到 83.1%，F 值达到 86.8%，与较近的 STKM 算法相比，准确率提高了 4%，召回率提高了 4.8%，F 值提高了 4.6%，相较于其他算法均有提升，验证了本文方法对 Total-Text 数据集中任意形状场景文本图像都具有较好的检测能力，且对于 Total-Text 数据集中复杂度高和低对比度背景的文本图像具有稳定的检测效果。

CTW1500 数据集是用于任意形状场景文本的数据集，文本包含多方向和任意形状，主要侧重于弯曲文本，包含 1 500 张图像。由表 3 分析可知，本文方法在 CTW1500 数据集上准确率达到 87.3%，比 DBnet 算法高了 0.4%，高于表 3 中的所有算法；F 值达到了 84.1%，比 PAN 算法高 0.4%。验证了本文方法对弯曲文本检测效果较好，充分表明本文方法具有很强的检测能力和较强的鲁棒性。

ICDAR 2017 数据集是任意形状文本检测数据集，包含水平、倾斜、垂直、弯曲和长文本，由 12 263 张图片组成。根据表 4 分析可知，本文方法在 ICDAR 2017 数据集上准确率达到 81.2%，比 FOTS 算法高 1.7%；召回率达到 73.2%，比 CharNet 高出 3.2%；F 值达到 74.9%，比 CharNet 高出 1.5%。通过比较表明本文方法在准确率、召回率和 F 值上都优于表 4 中的其他算法，验证了本文方法检测文本得有效性。

由图 6 可看出，与其他检测效果对比。在 ICDAR 2015 数据集中，图 6(a)存在误检漏检现象，其误检现象是由于其外形类似文本目标，在特征映射过程没有明确重点特征，导致信息损失，造成非文本目标的误判。而本文方法优先将更多注意力专注于重要元素上，赋予文本区域更多的权重，使模型更加准确地关注重要文本信息，减少文本误检现象。图 6(b)存在漏检现象，图 6(d)存在文本区域定位不准确的现象，而图 6(c)、(f)能够准确检测出文本区域的具体位置，表明本文方法能够提高文本定位的准确性，准确捕获漏检误检文本。在 Total-Text 数据集中，图 6(a)存在定位不准确的现象，这是由于网络感受野不够大导致的，本文模型在特征提取部分使用可变形卷积，能增强任意形状文本的适应性。图 6(d)、(e)存在漏检的现象，而图 6(c)、(f)文本检测效果较好，表明本文方法能够准确定位文本区域，改善漏检现象，且在任意形状文本检测中具有一定的竞争力。在 CTW1500 数据集中，图 6(a)、(b)存在漏检现象，这是由于小目标文本经多次采样后会变得模糊和失真，被一些模型当成噪声给过滤掉。而本文更注重捕获像素之间的相关性，重新分配字符与非字符区域之间的权重，使模型更加关注有用的文本区域，提高文本检测精度，减少漏检现象。图 6(d)存在文本区域定位不准确的现象，而图 6(c)、(f)能够准确检测弯曲文本图像，表明本文方法能够准确检测曲线文本，提高任意形状文本的检测效果。在 ICDAR2017 数据集中，图 6(a)、(b)、(d)、(e)存在漏检现象，而图 6(c)、(f)能够准确检测文本区域，表明本文方法能够准确定位文本区域，改善任意形状文本检测漏检现象，具有一定的竞争力。

2.2　消融实验

1) 为验证本文方法模块的有效性，在 ICDAR 2015、ICDAR2017、Total-Text 和 CTW1500 等 4 个数据集上进行消融实验，结果如表 5 所示。

表 5　ICDAR2015、ICDAR2017、Total-Text 和 CTW1500 消融实验结果

Table 5　Comparison of performance index results of multiple algorithms in the ICDAR2015、ICDAR2017、Total-Text and CTW1500

方法	ICDAR2015			ICDAR2017			Total-Text			CTW1500			Param/10^6	FLOPs/10^9
	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%	P/%	R/%	F/%		
本文	91.9	87.5	90.1	81.2	73.2	74.9	90.3	83.1	86.8	87.3	82.0	84.1	82.2	96.1
本文(-CFDEM)	91.1	84.6	87.7	80.9	71.8	73.6	88.6	82.2	85.2	87.0	80.4	83.7	78.5	94.4
本文(-AFS)	89.4	82.5	85.9	79.2	70.1	72.4	87.8	79.9	83.7	85.9	78.7	81.3	78.9	93.0
本文(-CFDEM-AFS)	88.2	82.7	85.4	79.0	67.9	71.7	87.1	82.5	84.7	86.9	80.2	83.4	76.2	91.3

根据表 5 可以看出, 采用 CFDEM 时, ICDAR 2015 的准确率、召回率和 F 值分别提升了 1.2%、0.2% 和 0.5%; ICDAR2017 的准确率、召回率和 F 值分别提升了 0.2%、2.2% 和 0.7%; Total-Text 的准确率提升了 0.7%, 验证了 CFDEM 的有效性。采用 AFS 时, ICDAR2015 的准确率、召回率和 F 值分别提升了 2.9%、1.9% 和 2.3%; ICDAR2017 的准确率、召回率和 F 值分别提升了 1.9%、3.9% 和 1.9%; Total-Text 的准确率和 F 值分别提升了 1.5% 和 0.5%; CTW1500 的准确率、召回率和 F 值分别提升了 0.1%、0.2% 和 0.3%, 验证了 AFS 的有效性。同时采用 CFDEM 和 AFS 时, ICDAR2015 的准确率、召回率和 F 值分别提升了 3.7%、4.8% 和 4.7%; ICDAR2017 的准确率、召回率和 F 值分别提升了 2.2%、5.3% 和 3.2%; Total-Text 的准确率、召回率和 F 值分别提升了 3.2%、0.6% 和 2.1%; CTW1500 的准确率、召回率和 F 值分别提升了 0.4%、1.8% 和 0.7%, 证明了本文方法能够显著提高任意形状文本检测性能。

2) 为研究本文方法中各模块的空间复杂性和时间复杂性, 对各模块的参数量 (Param) 和浮点运算次数 (FLOPs) 进行度量, 如表 5 右侧所示。可以看出单独引入 CFDEM 时, Param 增加 2.7 M, FLOPs 升高 1.7 G; 单独引入 AFS 时, Param 增加 3.3 M, FLOPs 升高 3.1 G; CFDEM 和 Param 同时采用时, Param 增加 6.0 M, FLOPs 升高 4.8 G。

3) 为验证本文方法的性能, 分别以 Resnet18、Resnet50、Resnet101 为骨架网络进行测试, 以 Total-Text 数据集为例, 测试结果如表 6 所示。根据表 6 可以看出本文方法在不同深度的骨架网络上都具有良好的性能, Resnet18 网络层数较浅, 检测速度表现良好, 而 Resnet50 的模型相比 Resnet18 模型的评价指标具有较大的提升, 而 Resnet101 由于数据提升较小, 检测速度相对慢, 考虑到服务器设备的情况, 本文方法主要以 Resnet50 为主。

表 6　本文方法在 Total-Text 上的性能对比

Table 6　The performance comparison of this method on Total-Text

主干网络	P/%	R/%	F/%	检测速度/(f/s)
Resnet18	88.2	79.8	83.7	43.7
Resnet50	90.3	83.1	86.8	29.1
Resnet101	90.8	83.5	87.1	26.9

3　结束语

本文提出了一种双分支跨级特征融合的自然场景文本检测方法, 该方法通过设计跨级特征分布增强模块能够提取不同层级特征信息, 从而增强了跨级特征文本信息的交互性; 提出的自适应融合策略, 通过双分支结构来加强不同尺度特征之间的联系, 自适应地选择过滤非文本或冗余特征, 能够降低误检率和漏检率; 在不同数据集下的实验结果表明本文方法能够准确的定位文本区域, 改善漏检误检, 提高检测精度。

参考文献:

[1] 黄剑华, 唐降龙, 刘家锋, 等. 一种基于 Homogeneity 的文本检测新方法 [J]. 智能系统学报, 2007, 2(1): 69–73.
HUANG Jianhua, TANG Xianglong, LIU Jiafeng, et al. A new method for text detection based on Homogeneity[J]. CAAI Transactions on Intelligent Systems, 2007, 2(1): 69–73.

[2] 吕国宁, 高敏. 视觉感知式场景文字检测定位方法 [J]. 智能系统学报, 2017, 12(4): 569.
LYU Guoning, GAO Min. Scene text detection and localization scheme with visual perception mechanism[J]. CAAI transactions on intelligent systems, 2017, 12(4): 569.

[3] TIAN Z, HUANG W, HE T, et al. Detecting text in natural image with connectionist text proposal network[C] // Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands: Springer International Publishing, 2016: 56−72.

[4] 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. Honolulu: IEEE, 2017: 2550−2558.

[5] WANG Y, XIE H, ZHA Z J, et al. ContourNet: taking a further step toward accurate arbitrary-shaped scene text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11753−11762.

[6] 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. Long Beach: IEEE, 2019: 9336−9345.

[7] WANG W, XIE E, SONG X, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 8439−8448.

[8] LIAO M, WAN Z, YAO C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI conference on artificial intelligence. Washington, D.C.: AAAI, 2020, 34(07): 11474−11481.

[9] WU Y, ZHANG W, WAN S. CE-text: a context-aware and embedded text detector in natural scene images[J]. Pattern recognition letters, 2022, 159: 77–83.

[10] 孟月波, 石德旺, 刘光辉, 等. 多维度卷积融合的密集不

规则文本检测 [J]. 光学精密工程, 2021, 29(9): 2210–2221.
MENG Yuebo, SHI Dewang, LIU Guanghui, et al. Dense irregular text detection based on multi-dimensional convolution fusion[J]. Optics and precision engineering, 2021, 29(9): 2210–2221.

[11] YANG C, CHEN M, YUAN Y, et al. BiP-net: bidirectional perspective strategy based arbitrary-shaped text detection network[C]//ICASSP 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2022: 2255−2259.

[12] CHEN H, CHEN P, QIU Y, et al. FARNet: fragmented affinity reasoning network of text instances for arbitrary shape text detection[J]. IET image processing, 2023, 17(6): 1959–1977.

[13] ZHU Y, CHEN J, LIANG L, et al. Fourier contour embedding for arbitrary-shaped text detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3123−3131.

[14] YOU Y, LEI Y, ZHANG Z, et al. Arbitrary-shaped text detection with B-spline curve network[J]. Sensors, 2023, 23(5): 2418.

[15] ZHANG S X, ZHU X, CHEN L, et al. Arbitrary shape text detection via segmentation with probability maps[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(3): 2736–2750.

[16] TANG Q, FENG X, ZHANG X. A spatial feature adaptive network for text detection[J]. Multimedia tools and applications, 2022, 81(11): 15285–15302.

[17] 赵文清, 杨盼盼. 双向特征融合与注意力机制结合的目标检测 [J]. 智能系统学报, 2021, 16(6): 1098–1105.
ZHAO Wenqing, YANG Panpan. Target detection based on bidirectional feature fusion and an attention mechanism[J]. CAAI transactions on intelligent systems, 2021, 16(6): 1098–1105.

[18] ZHOU X, YAO C, WEN H, et al. EAST: An efficient and accurate scene text detector[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2642−2651.

[19] ZHANG C, LIANG B, HUANG Z, et al. Look more than once: an accurate detector for text of arbitrary shapes[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 10552−10561

[20] LONG S, RUAN J, ZHANG We, et al. TextSnake: A flexible representation for detecting text of arbitrary shapes[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 20−36.

[21] XIE E, ZANG Y, SHAO S, et al. Scene text detection with supervised pyramid context network[C]// Proceedings of the AAAI conference on artificial intelligence. Honolulu: AAAI, 2019, 33(1): 9038−9045.

[22] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9365−9374.

[23] WANG Hao, LU Pu, ZHANG Hui, et al. All You need is boundary: toward arbitrary-shaped text spotting[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020, 34(7): 12160−12167.

[24] ZHANG S X, ZHU X, HOU J B, et al. Deep relational reasoning graph network for arbitrary shape text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2020: 9699−9708.

[25] LI J, LIN Y, LIU R, et al. RSCA: real-time segmentation-based context-aware scene text detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 2349−2358.

[26] WAN Q, JI H, SHEN L. Self-attention based text knowledge mining for text detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio. Nashville: IEEE, 2021: 5983−5992.

[27] ZHU Y, DU J. TextMountain: accurate scene text detection via instance segmentation[J]. Pattern recognition, 2021, 110: 107336.

[28] WANG F, CHEN Y, WU F, et al. TextRay: contour-based geometric modeling for arbitrary-shaped scene text detection[C]//Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM, 2020: 111−119.

[29] BAEK Y, SHIN S, BAEK J, et al. Character region attention for text spotting[C]//Computer Vision–ECCV 2020: 16th European Conference. Glasgow: Springer International Publishing, 2020: 504−521.

[30] LIU Y, CHEN H, SHEN C, et al. ABCNet: real-time scene text spotting with adaptive bezier-curve network[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2020: 9809−9818.

[31] YE J, CHEN Z, LIU J, et al. TextFuseNet: Scene Text Detection with Richer Fused Features[C]//International Joint Conference on Artificial Intelligence. Yokohama: IJCAI, 2021: 512−518.

[32] BUŠTA M, PATEL Y, MATAS J. E2E-MLT - an unconstrained end-to-end method for multi-language scene text[C]//Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision. Perth: Springer International Publishing, 2019: 127−143.

[33] DENG D, LIU H, LI X, et al. PixelLink: detecting scene text via instance segmentation[C]// Proceedings of the AAAI conference on artificial intelligence. New Orleans: AAAI, 2018, 32(1): 6773-6780.

[34] LIU X, LIANG D, YAN S, et al. FOTS: fast oriented text spotting with a unified network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-

nition. Salt Lake City: IEEE, 2018: 5676−5685.

[35] XING L, TIAN Z, HUANG W, et al. Convolutional character networks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach: IEEE, 2019: 9126−9136.

作者简介：

刘光辉，副教授，主要研究方向为计算机视觉理解、建筑智能化技术。近年来主持/参与多项国家自然科学基金项目、陕西省重点研发计划项目、陕西省基础研究项目，获陕西高等学校科学技术优秀成果奖。

张钰敏，硕士研究生，主要研究方向为图像处理、场景文本检测与识别。

孟月波，教授，博士生导师，博士，主要研究方向为机器视觉信息处理与分析、建筑智能化。

[责任编辑：刘冰洁]

2023 第九届中国智能技术与大数据会议

中国智能技术与大数据会议是由中国人工智能学会智能服务专委会发起的系列会议，每年举办一次。继 2015 年召开第一届中国智能技术与大数据会议（北京）、2016 年召开第二届中国智能技术与大数据会议（河南焦作）、2017 年召开第三届中国智能技术与大数据会议（广东广州）、2018 年召开第四届中国智能技术与大数据会议（重庆）、2019 年召开第五届中国智能技术与大数据会议（江苏常州）、2020 年召开第六届中国智能技术与大数据会议（北京）、2021 年召开第七届中国智能技术与大数据会议（北京）、2022 年召开第八届中国智能技术与大数据会议（北京）后，2023 年第九届中国智能技术与大数据会议（CITBD2023）将于 2023 年 10 月 21 -22 日在山东烟台举行。本届会议由中国人工智能学会（CAAI）主办，CAAI 智能服务专委会、烟台大学、北京邮电大学计算机学院联合承办。

本届会议将就智能技术与大数据相关的科学基础理论、关键技术方法与系统进行探讨和交流，旨在加强相关方向的基础理论研究，掌握最新和实用技术、了解前沿发展趋势，从而推动我国智能技术与大数据领域的学术繁荣及其在智能服务领域的应用推广。会议将邀请中国工程院院士、欧洲科学院院士、国家高层次人才做大会特邀报告，同时，举办青年论坛、优秀论文评奖论坛、前沿技术和应用论坛。报告专家将介绍智能服务与大数据相关技术的最新学术成果和发展趋势，并就其关键技术和主要战略发展方向进行深入地交流和研讨。

会议期间还将召开 CAAI 智能服务专委会会议并发展新委员。诚挚欢迎全国各高等院校、科研院所和企事业单位的科技工作者参加本届会议。

会议的主题包括但不限于以下方面：智能大数据、大模型、认知计算、数据挖掘、机器学习、人工智能理论与前沿发展等。