结合感受野增强和全卷积网络的场景文字检测方法

李晓玉 [1] 宋永红 [2] 余 涛 [1]

摘 要 自然场景图像质量易受光照及采集设备的影响, 且其背景复杂, 图像中文字颜色、尺度、排列方向多变, 因此, 自然场景文字检测具有很大的挑战性. 本文提出一种基于全卷积网络的端对端文字检测器, 集中精力在网络结构和损失函数的设计, 通过设计感受野模块并引入 Focalloss、GIoUloss 进行像素点分类和文字包围框回归, 从而获得更加稳定且准确的多方向文字检测器. 实验结果表明本文方法与现有先进方法相比, 无论是在多方向场景文字数据集还是水平场景文字数据集均取得了具有可比性的成绩.

关键词 感受野增强, Focalloss, GIoUloss, 全卷积网络

引用格式 李晓玉, 宋永红, 余涛. 结合感受野增强和全卷积网络的场景文字检测方法. 自动化学报, 2022, **48**(3): 797−807

DOI 10.16383/j.aas.c190376

Text Detection in Natural Scene Images Based on Enhanced Receptive Field and Fully Convolution Network

LI Xiao-Yu[1] SONG Yong-Hong[2] YU Tao[1]

Abstract The quality of natural scene images is influenced easily by the shooting environment and conditions, and scene image background is relatively complex and has a strong interference for detection, besides, text in scene images may have different colors, fonts, sizes, directions, languages and so on, all these situations make natural scene text detection be still a challenging research topic. This paper proposes an end-to-end text detector based on fully convolution network. We focus on the design of the network structure and the loss function, through adding the enhanced receptive field module and introducing Focalloss, GIoUloss for pixels classification and text boxes regression respectively, we gain a more stable accurate multi-oriented text detector. Our method provides promising performance compared to the recent state-of-the art methods on both the multi-oriented scene text dataset and horizontal text dataset.

Key words Receptive field enhanced module, Focalloss, GIoUloss, full convolution network

Citation Li Xiao-Yu, Song Yong-Hong, Yu Tao. Text detection in natural scene images based on enhanced receptive field and fully convolution network. *Acta Automatica Sinica*, 2022, **48**(3): 797−807

场景图像文字中承载的高级语义信息可以帮助我们更好地理解周围的世界, 同时场景图像文字检测技术也可以广泛地应用于多媒体检索、视觉输入和访问, 以及工业自动化. 早期的文字检测技术都是使用传统的模式识别技术, 可以分为两大主流方法, 一种是以连通区域分析为核心技术的文字检测方法, 另一种则是以滑动窗为核心技术的文字检测方法. 传统的模式识别方法一般包含多个步骤: 字符候选区域生成、候选区域滤除、文本行构造和文本行验证, 繁琐的检测步骤致使文字检测结果过于依赖中间结果且非常耗时.

随着计算机视觉和模式识别领域的发展, 目标检测方法研究开始使用卷积神经网络 (Convolutional neural network, CNN), 研究者们开始借鉴基于深度学习的目标检测方法来检测文字, 因此产生了一系列基于回归的深度学习文字检测方法, 该类方法主要是基于目标检测框架 SSD (Single shot multibox detector)[1]、Faster-RCNN (Region CNN)[2] 等进行针对文字特性的改进得到. 这类方法的主要特点是通过回归水平矩形框、旋转矩形框以及四边形等形状来获得文字检测结果. 同时, 由于后续文字识别步骤需要精确的文字定位结果, 也诞生了一系列基于分割的深度学习文字检测方

收稿日期 2019-05-16 录用日期 2019-08-22
Manuscript received May 16, 2019; accepted August 22, 2019
陕西省自然科学基础研究计划 (2018JM6104), 国家重点研究开发计划 (017YFB1301101) 资助
Supported by Natural Science Basic Research Program of Shaanxi (2018JM6104) and National Key Research and Development Program of China (017YFB1301101)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 西安交通大学软件学院 西安 710049 2. 西安交通大学人工智能学院 西安 710049
1. School of Software Engineering, Xi′an Jiaotong University, Xi′an 710049 2. College of Artificial Inteligence, Xi′an Jiaotong University, Xi′an 710049

法[3−5], 该类方法主要借鉴语义分割的思路, 将文本像素分到不同的实例中, 并通过一些后处理方法获得文字像素级别的定位结果, 并且由于像素级检测的特点, 近年来该类方法逐渐开始用于解决曲线文本检测与识别问题[6]. 此外, 由于无论是基于目标检测还是基于分割都存在各自的局限, 因此也有学者尝试融合检测和分割的思想[7] 进行文字检测. 虽然近些年基于深度学习的文字检测方法已经取得巨大进步, 但是文字作为一种具有其独有特色的目标, 其字体、颜色、方向、大小等呈现多样化形态, 相比一般目标检测更加困难, 即便有许多的学者尝试根据文字的特点进行网络改进, 如使用旋转敏感的回归[8] 来适应任意方向文本, 亦或使用端到端的文字检测与识别方法联合优化检测和识别结果[9], 但在遇到多方向文字以及多尺度文字场景图像时, 检测准确性和有效性依旧差强人意. 另外, 现有检测方法有一阶端对端检测流程, 但当前一阶方法存在以下问题: 1) 一阶方法如果使用较小的网络结构进行检测, 速度快但精度不理想, 因此, 一阶方法一般会通过增加网络深度提高检测精度, 显然, 这种做法增大了计算开销, 检测速度无法得到满足; 2) 一阶检测方法存在严重的正负样本不均衡、对目标尺度不敏感等问题, 也导致检测器准确率不高.

本文提出一种可端对端训练的快速文本检测方法, 可以鲁棒地检测任意方向文本和多尺度文本. 为了提升网络的检测效果并尽量减少计算量, 受人类视觉系统感受野结构的启发, 在网络结构设计中加入手工设计的感受野增强模块, 从而在保持较快速度前提下提高检测精度, 克服了一阶检测方法速度快精度低的弊端. 在损失函数部分, 为了改善样本不均衡、文字尺度不敏感等问题, 引入 Focalloss[10] 和 GIoUloss[11] 训练网络, 进一步提升网络性能.

本文内容安排如下: 第 1 节介绍基于全卷积网络的检测框架的各部分结构设计; 第 2 节描述损失函数的设计; 第 3 节给出详细的实验结果与模型分析; 第 4 节对本文进行总结.

1 网络整体框架

1.1 检测流程

图 1 是本文文字检测算法流程的一个高级概述. 可以看到图像送入全卷积网络 (Fully convolutional networks, FCN), 通过特征金字塔网络(Feature pyramid networks, FPN)[12] 随之产生多通道的像素级别的文本得分图和旋转矩形框预测图. 其中 1 通道的像素级别文本得分图的每一个像素值在 [0, 1] 之间, 代表该像素属于文本的置信度. 旋

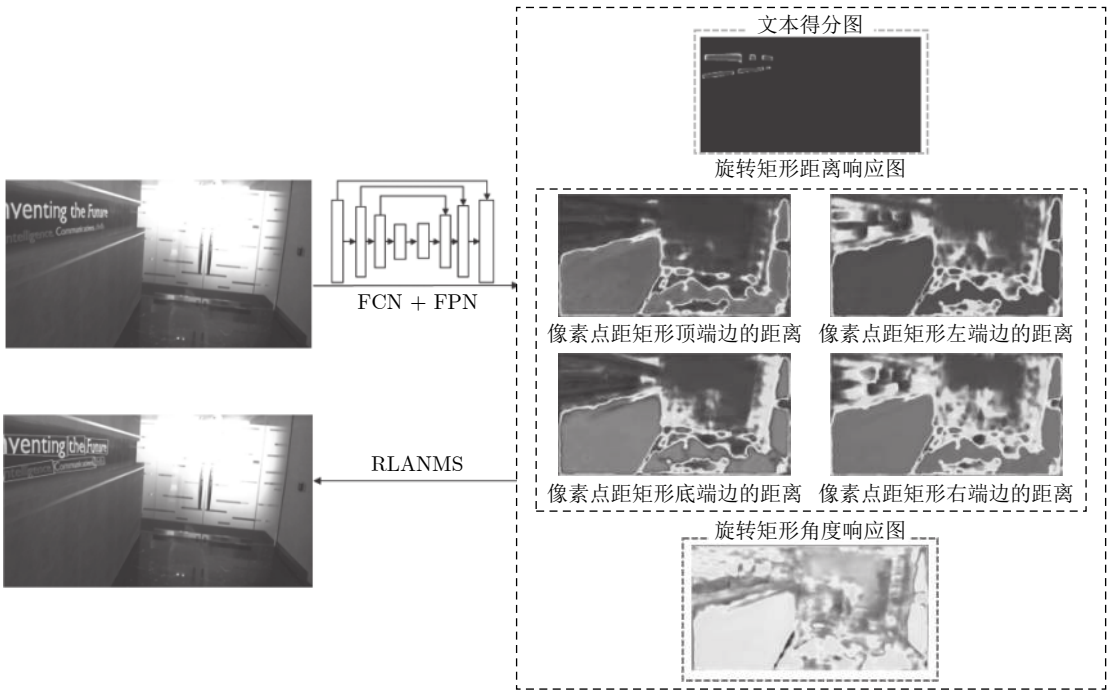


图 1 本文方法检测流程图

Fig. 1 Flow chart of our detection method

转矩形框预测图表示以当前像素点为中心, 可以包围文本区域的旋转矩形, 共包含 5 通道特征图, 分别代表以该像素点为中心, 预测的旋转矩形的 4 条边与该点的距离以及该矩形的旋转角度. 网络产生的旋转矩形框预测结果直接经过精细局部感 知非极大值抑制 (Refined locality aware non-maximum suppression, RLANMS) 产生最终的结果.

1.2　网络结构设计

图 2 展示了文字检测网络的详细结构图, 主要包括 4 部分: 特征提取主干、感受野增强模块、特征融合分支和输出层.

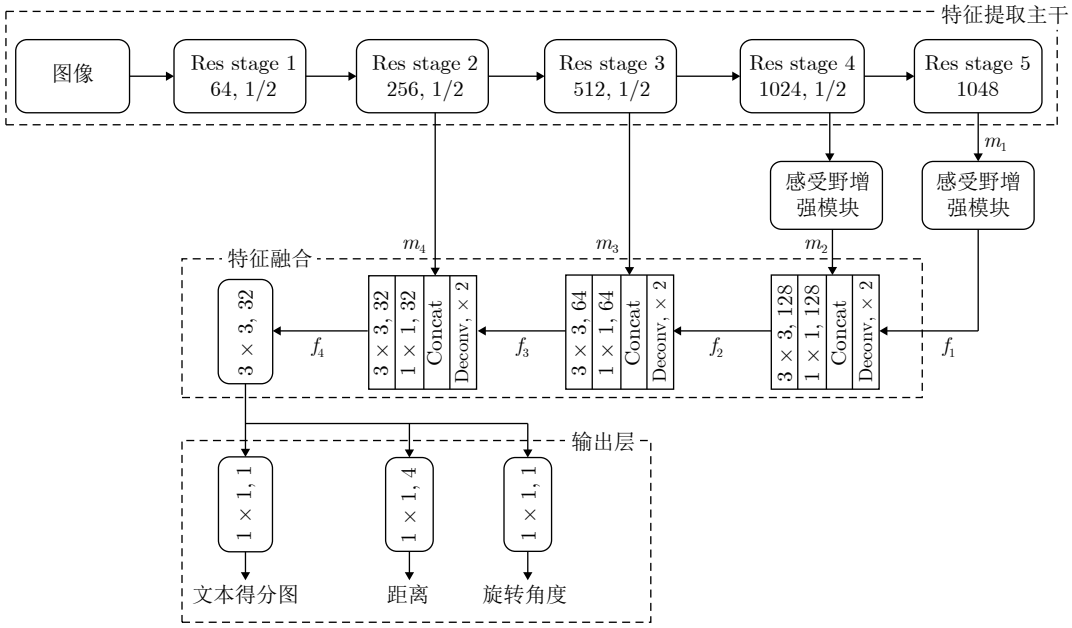


图 2　本文方法网络结构图

Fig. 2　Structure of our network

特征提取主干使用的是在 ImageNet[13] 数据集预训练的 50 层的残差网络 (ResNet50[14]). ResNet50 有 5 个级别特征图, 本文主要使用后四个级别的特征层, 如图 2 所示, 它们的尺寸分别是输入图像的 1/32, 1/16, 1/8, 1/4, 用 f_i 表示.

在特征融合分支中, 逐渐地合并从 ResNet50 中提取的 4 个级别特征图 m_i, 受 FPN[12] 启发, 具体融合方式如式 (1) 和式 (2) 所示.

$$b_i = \begin{cases} Deconv(f_i), & i \le 3 \\ conv_{3\times3}, & i = 4 \end{cases} \tag{1}$$

式 (1) 和式 (2) 中, b_i 是准备融合前的特征图; f_i 是融合后的特征图; $[\cdot;\cdot]$ 表示不同层次特征图进行通道方向上的拼接. 在每个融合阶段, 前一阶段的特征图 f_{i-1} 经过一个反卷积层放大两倍, 然后与当前特征图进行连接. 这里, 当 $i = 4, 5$ 时, m_i 会先经过一个感受野增强模块, 该模块的具体结构将在下一节详细介绍. 接着, 一个 $conv_{1\times1}$ 模块用于增加网络的非线性并降低特征图的通道数, 减少网络参数. 最后的融合阶段, 经过一个 $conv_{3\times3}$ 模块得到整个融合分支的最后输出 f_4, 作为输出层的输入.

输出层各部件的特征图通道数如图 2 所示, 输出层中, 输入的是 32 通道的融合特征, 目的是为了保证以少许的计算复杂度换取更高的检测精度. 最后的输出层包含 3 个 $conv_{1\times1}$ 模块, 分别将输入特征变换到 1 通道的文本得分图、4 通道的矩形距离响应图和 1 通道的旋转角度响应图, 文本得分图和旋转矩形框的标签制作具体可参照文献 [15].

$$f_i = \begin{cases} m_i, & i = 1 \\ conv_{3\times3}(conv_{1\times1}([b_{i-1}; m_i])), & i = 2 \\ conv_{3\times3}(conv_{1\times1}([m_{i-1}; ERFB(m_i)])), & \text{其他} \end{cases} \tag{2}$$

1.3 增强感受野模块

自然场景文字由于尺度大小和宽高比多变, 导致现有方法准确率欠佳. 本节通过加入感受野模块 (Receptive field block, RFB) 来提升不同尺度和宽高比文字检测准确率. 受目标检测领域中文献 [16] 方法的启发, 本文重新设计了这一模块, 并将其嵌入特征融合中. 图 3(a) 展示了在人类视觉系统中, 感受野的大小在人类视网膜图中是离心率的函数, 感受野随着离心率的增加不断增大; 在不同视觉系统中, 感受野也不同, 图 3(b) 展示了基于图 3(a) 中参数的感受野空间阵列, 显示了感受野的分布规律, 每个圆的半径表示在对应离心率下的感受野大小.

本节希望通过控制离心率来控制感受野大小, 因此设计了与人类视觉系统感受野结构有相似分布规律的感受野增强模块. 整个感受野增强模块用于在网络特征融合时, 主干网络中共 4 次特征融合, 为了保证此模块在发挥最大作用的同时, 尽量减少参数量以加快检测速度, 本文只将该模块用于高层语义 (实验时, 加在低层在 ICADAR2015 上仅有 0.1% 的提升), 即主干网络的 stage 4 和 stage 5. 该模块在参考 Inception-ResNet[17] 的基础上, 加入了空洞卷积, 使用不同尺度的卷积核作为不同视觉系统, 不同膨胀率的空洞卷积作为对应视觉系统中的离心率.

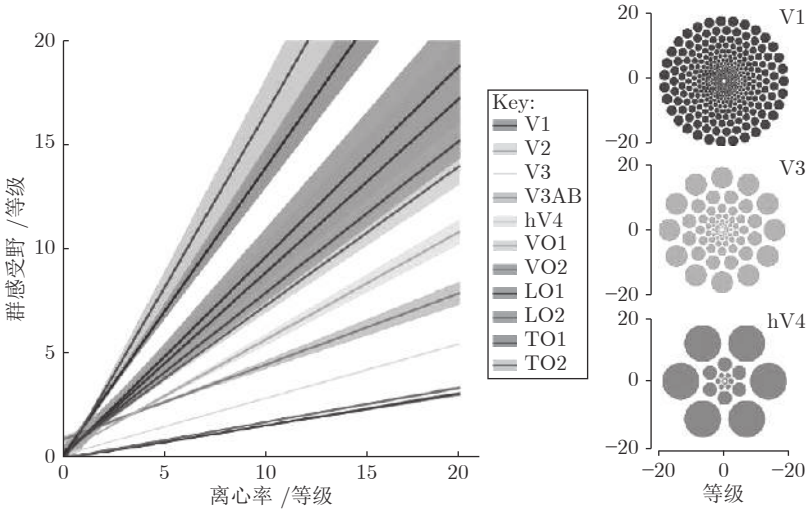
1.3.1 多分支卷积层

感受野增强模块是由多种尺度卷积核的卷积层构成的多分支结构[17]. 具体设计如图 4 所示, 从主干网络提取的特征图分别进入 6 个分支, 其中, 前五个分支都先经过一个 $conv_{1\times1}$ 模块以减少通道特征, 最后经过一层空洞卷积, 且其中间 4 个分支在空洞卷积前还要分别经过 $conv_{1\times3}$、$conv_{3\times1}$、$conv_{1\times5}$、$conv_{5\times1}$ 卷积, 最后一个分支为 shortcut. 使用 1、3、5 不同大小的卷积核相当于不同的视觉系统, 它们的基础感受野不同, 针对不同尺度的文字进行检测. 使用 $1\times n$ 和 $n\times1$ 代替 $n\times n$ 卷积是为了降低参数量, 使得提升网络性能的同时, 尽量减少计算成本的增加; 最后一个分支是直连, 该设计来自于 ResNet 和 Inception-ResNet. 5 个分支的输出进行通道上连接后与直连通道进行相加融合, 得到该模块的最终输出.

1.3.2 空洞卷积层

在图像分割领域, 为了保证在增大感受野的同时, 又不会因为池化操作而损失图像信息, 学者们提出空洞卷积[18]. 在文字检测中, 大的长文本需要比较大的感受野, 小的短文本检测需要保留尽量多的信息, 因此在本文的感受野增强模块中加入空洞卷积, 保证在感受野增大的同时, 避免信息损失. 在图 5 显示的结构中, 每个分支都是一个正常卷积后面加一个空洞卷积, 膨胀因子大小根据卷积核大小设计. 本文设计的感受野增强模块结构中, 分别在 $conv_{1\times1}$, $conv_{1\times3}$, $conv_{3\times1}$, $conv_{1\times5}$, $conv_{5\times1}$ 卷积后加膨胀因子大小为 1, 3, 3, 5, 5 的 $conv_{3\times3}$ 卷积. 图 5 展示了卷积核大小为 3×3 的卷积在膨胀因子分别为 1, 3, 5 情况下的感受野.

图 5(a) 表示当膨胀因子为 1 时, 与普通 3×3 的卷积相同, 感受野为 3; 图 5(b) 表示当膨胀因子为 3 时, 与普通 3×3 的卷积相比, 空洞卷积的感



(a) 离心率与感受野的关系图
(a) Relationship between eccentricity and receptive field

(b) 感受野分布规律
(b) Distribution of receptive field

图 3 离心率与感受野的关系图

Fig. 3 Structure of the human visual system′s receptive field

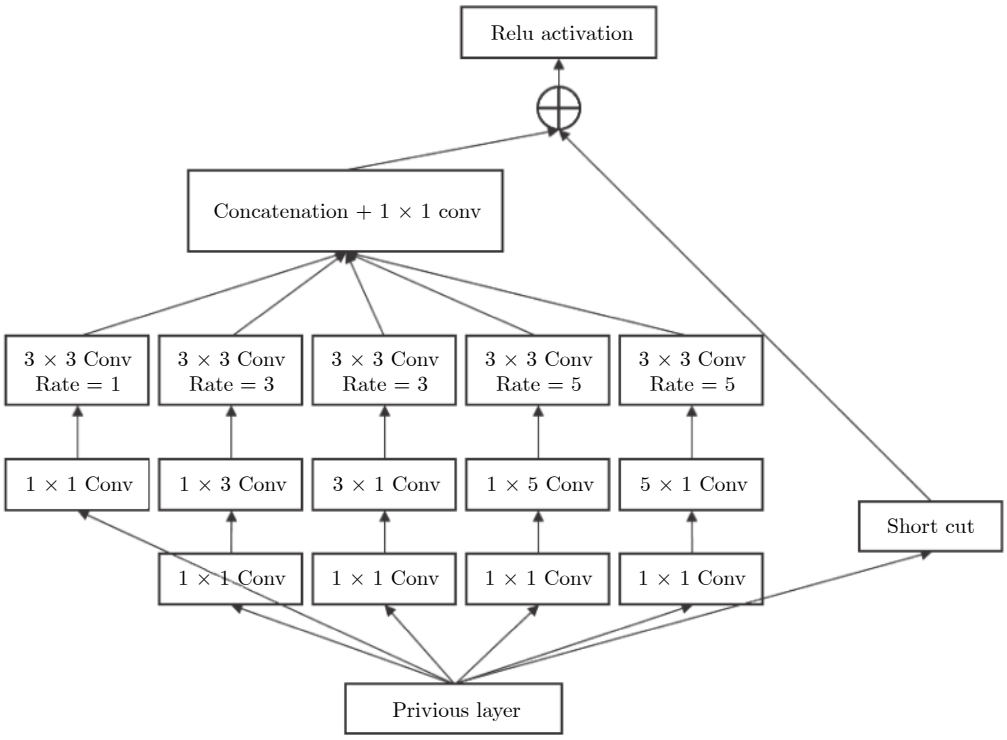
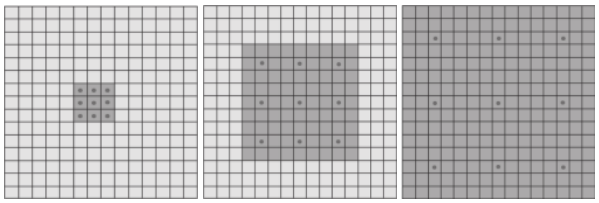


图 4　感受野增强模块

Fig. 4　Receptive field block



(a) 卷积核: 3 × 3 膨胀因子: 1 (a) Convolution kernel: 3 × 3 Dilation rate: 1

(b) 卷积核: 3 × 3 膨胀因子: 3 (b) Convolution kernel: 3 × 3 Dilation rate: 3

(c) 卷积核: 3 × 3 膨胀因子: 5 (c) Convolution kernel: 3 × 3 Dilation rate: 5

图 5　不同膨胀因子的空洞卷积

Fig. 5　Dilated convolution with different dilation rates

受野为 9; 图 5(c) 表示当膨胀因子为 5 时, 与普通 3×3 的卷积相比, 空洞卷积的感受野为 15. 图 5 直观展示了空洞卷积增大感受野的过程.

2　损失函数

本小节介绍本文模型的损失函数, 模型整体的损失函数表示为

$$L = L_{\text{conf}} + \lambda L_{\text{loc}} \tag{3}$$

式中, L_{conf} 和 L_{loc} 分别表示文本得分图和旋转矩形框损失; λ 是平衡因子, 用于均衡文本得分图损失和旋转矩形框损失; 在本文实验中, λ 设为 1.

2.1　文本得分图损失

在文字检测领域, 一幅图像可能生成成千上万的候选包围框, 但是一幅图像上真实目标包围框可能只是很少几个甚至没有, 这样就造成正负样本比例失衡的问题. 本文网络最后会得到大小为 $256\times256\times1$ 的文本得分图和 $256\times256\times5$ 的旋转矩形框几何特征图, 在每一个像素点位置都会预测一个候选包围框, 也即是 256×256 个候选框. 然而, 每幅图像上需要检测的文字数量只是很少几个甚至 0 个, 这样致使网络训练过程中文字区域与非文字区域样本比例严重失衡.

目前已有的很多检测方法也关注到了样本不均衡问题, 其一般做法是对样本进行数据增广或者训练过程中进行难样本挖掘. 这类做法确实在一定程度上改善了样本不均衡问题, 但是也在整个检测过程中引入额外的步骤, 这与本文 “简洁快速的端对端检测器”初衷是相违背的. 为了保持一个简单的训练过程, 同时又可以改善正负样本不均衡问题, 本文引入 Focalloss [10] 损失函数, 计算式为

$$\begin{aligned} FL(p_t) &= -\alpha_t(1-p_t)^{\gamma}\log(p_t) \\ pt &= \begin{cases} p, & y=1 \\ 1-p, & \text{其他} \end{cases} \end{aligned} \tag{4}$$

式中, α_t 用于控制正负样本的权重. 一般而言, 文

字检测任务中, 文字类的样本数量远远少于背景类的样本, 因此, α_t 取 [0, 0.5] 来增加文字类的样本的权重, 使网络关注文字类的学习. $(1-p_t)^\gamma$ 用于调控易分类样本和难分类样本的比重. 当一个样本越难分类, p_t 就越小, 那么其损失和反向梯度就会越大, 损失再乘以 $(1-p_t)^\gamma$ 则会更大; 易分类样本恰好相反, 损失和梯度会更小. 于是网络就会更多关注难分类样本的学习, 从而降低样本误检. 通过多次实验结果, 当 $\alpha_t = 0.5$, $\gamma = 0.5$ 时, 效果最好, 本文实验均在该参数设置下进行.

2.2 旋转矩形损失

2.2.1 矩形框损失

场景文字检测的一大难题是场景图像中文字的尺度、宽高比极其多变. 目标检测领域常用 L1、L2 损失来回归目标包围框, 这类损失的特点是对大数很敏感, 如果直接使用这类损失来回归文字区域, 那么大文字、长文字的损失就会相对更大, 不仅导致梯度难以控制, 也很可能指导网络学习出更大更长的文本包围框. 因此, 需要一个对文字尺度不敏感的函数进行文字区域回归.

EAST (Efficient and accurate scene text detector)[15] 中, 对于矩形框部分使用交并比 (Intersection over union, IoU) 损失, Zhou 等[15] 认为 IoU 的特性就是对尺度不敏感, 可以兼容文字的多种尺度, 但没有考虑 IoU 作为损失函数时存在以下问题: 1) 假设两个目标包围框没有发生重叠, 那么 IoU 值为零, 这种情况下, IoU 作为损失反向梯度也为 0, 网络得不到任何优化; 2) IoU 无法表达出两个目标矩形框的重合情况. 图 6 给出了两个目标包围框不同情况下的重合, 图 6(a) ~ 6(c) 三种情况下的 IoU 值相等, 但显然它们的重合情况完全不同. 这三种情况下, 图 6(a) 会得到一个很好的回归结果, 图 6(c)很难回归出理想的包围框. 因此, IoU 函数用作损失无法反映出两个目标包围框的重叠情况.

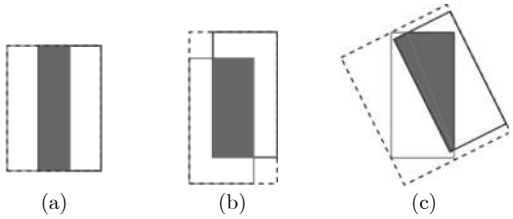


图 6 三种 IoU 相等的情况[11]

Fig. 6 Three situations with the same IoU[11]

针对上述 IoU 的缺点, 本文参考文献 [11] 引入 GIoU (Generalized IoU) 作为文字回归的损失, GIoU 计算过程很简单, 详细计算步骤如下:

1) 对于两个任意形状凸边形, $q_1, q_2 \subseteq Q \in \mathbf{R}^n$, 求出可以封闭两者的最小凸边形 q_3.这里 $q_3 \subseteq Q \in \mathbf{R}^n$. 例如, 图 6(c) 中的虚线部分即两个矩形框的最小凸边形.

2) 计算 q_1, q_2 的 IoU 值, $IOU = \frac{|q_1 \cap q_2|}{|q_1 \cup q_2|}$.

3) 计算 GIoU 值, $GIOU = IOU - \frac{|q_3/(q_1 \cup q_2)|}{|q_3|}$.

综上所述, 对于矩形框部分, 本文模型使用 GIoU 损失的表达式为

$$L_{AABB} = 1 - GIOU = 1 - \frac{|\hat{R} \cap R^*|}{|\hat{R} \cup R^*|} + \frac{|R/(\hat{R} \cup R^*)|}{|R|} \tag{5}$$

式中, $\hat{R}$ 表示网络预测的矩形形状; R^* 表示其对应的真实包围框; R 表示可以封闭 R^* 和 $\hat{R}$ 的最小矩形. $|\hat{R} \cup R^*|$ 的宽高可以简单地表示为

$$\begin{aligned} w_i &= \min(\hat{d}_2, {d_2}^*) + \min(\hat{d}_4, {d_4}^*) \\ h_i &= \min(\hat{d}_1, {d_2}^*) + \min(\hat{d}_3, {d_3}^*) \end{aligned} \tag{6}$$

式中, d_1, d_2, d_3 和 d_4 分别代表一个像素位置到其对应的矩形框上、右、下、左边的距离. R 的宽高计算式为

$$\begin{aligned} w_R &= \min(\hat{d}_2, {d_2}^*) + \max(\hat{d}_4, {d_4}^*) \\ h_R &= \min(\hat{d}_1, {d_2}^*) + \max(\hat{d}_3, {d_3}^*) \end{aligned} \tag{7}$$

重叠区域计算式为

$$|\hat{R} \cup R^*| = |\hat{R}| + |R^*| - |\hat{R} \cap R^*| \tag{8}$$

因此, 根据上述计算式, GIoU 可以很容易地计算出来.

2.2.2 角度损失

角度损失简单地使用余弦损失, 计算式为

$$L_\theta(\hat{\theta}, \theta^*) = 1 - \cos(\hat{\theta} - \theta^*) \tag{9}$$

式中, $\hat{\theta}$ 是预测的旋转角度; θ^* 代表真值.

综上, 几何形状损失可以整合表示为

$$L_{\text{loc}} = L_\alpha + \lambda_\theta L_\theta \tag{10}$$

式中, λ_θ 在实验时设置为 20. 值得一提的是, 本文在计算 L_α 时假设两个目标包围框角度相同, 即忽略了角度差异. 虽然网络在训练过程中, 两个包围框的角度有较大差异, 但是这样的近似 GIoU 值依然可以反映两个包围框的重合情况.

3 实验和结果分析

为了证明本文模型的有效性, 分别在 ICDAR-2013, ICDAR2015, 以及 MSRATD-500 数据集上

进行测试. 并对实验结果进行了详细的对比和分析.

3.1　实验数据

ICDAR2013[19]: 高分辨率的自然场景图像, 包含训练图片 229 幅, 测试图片 233 幅.

ICDAR2015[20]: 该数据集来自 ICDAR2015 鲁棒阅读竞赛中的任务 4: 偶然场景文字检测. 该数据集包含的图片是随机拍摄的生活场景, 不是刻意针对文字拍摄的.训练集包含1 000 幅图片, 测试集包含 500 幅图片, 文本的标注是以单词为单位.

MSRATD-500[21]: 该数据集是多方向自然场景文字数据集, 训练集包含 300 幅图片, 测试集包含 200 幅图片. 该数据集不仅包含英文文本也包含中文文本, 并且中英文标注都是以行为单位. 因为该数据集数据量太少, 所以在使用该数据集时, 加入 HUSTTR400[22] 数据集共同作为训练数据.

3.2　模型训练

本文方法利用 ADAM 优化器进行网络训练. 为了加速训练, 统一地从原始图片上随机采样 512 × 512 像素大小的图片块作为每一批次的训练样本, 训练的批次大小设置为 12. ADAM 的初始学习率为 0.0001, 每迭 代 10 000 次下降为原来的 0.94 倍, 训练均在一块 TITAN X GPU 上进行, 一共迭代 100 000 次.

3.3　实验结果和分析

3.3.1　精度性能

首先在两个比较流行的多方向偶然场景文字数据集 ICDAR2015 和 MSRA-TD500 上进行实验, 以此验证本文模型解决偶然场景下多方向文本检测的能力. 并且, 为了验证本文方法的多功能性, 又在比较流行的水平自然场景文字数据集 ICDAR2013 进行训练与测试, 并与现有方法的性能进行了详细对比.

1) 多方向偶然场景文字数据集

本节实验首先在广泛使用的多方向偶然场景数据集 ICDAR2105 上实施, 与其他方法的部分检测结果列举在图 7 中. 从图 7 中列举的检测结果可以看到, Zhang 等[23] 和 Shi 等[24]. 对多方向文本和多尺度文本出现了大量的漏检现象, 而本文方法在所列举的这几幅图像上表现出了对多尺度文本和多方文本鲁棒的检测性能.

根据文献 [20] 定义的召回率 (R)、精确率 (P)、F 值三个指标, 将本文方法与其他方法的定量比较结果列举在表 1 中. 本文模型单尺度测试的结果已经达到与现有先进方法相当的水平. 更重要的是,

(a) Zhang 等[23] 检测结果 (a) Results of Zhang et al.[23]　(b) Shi 等[24] 检测结果 (b) Results of Shi et al.[24]　(c) 本文方法检测结果 (c) Our results

图 7　各种方法在 ICDAR2015 测试集检测结果比较

Fig. 7　Qualitative comparison on ICDAR2015 dataset

表 1　ICDAR2015 测试集检测结果对比

Table 1　Qualitative comparison on ICDAR2015 dataset

方法	召回率 (R)	精确度 (P)	F 值
CNN MSER[22]	0.34	0.35	0.35
Islam 等[25]	0.64	0.78	0.70
AJOU[26]	0.47	0.47	0.47
NJU[22]	0.36	0.70	0.48
StradVision2[22]	0.37	0.77	0.50
Zhang 等[23]	0.43	0.71	0.54
Tian 等[27]	0.52	0.74	0.61
Yao 等[28]	0.59	0.72	0.65
Liu 等[29]	0.682	0.732	0.706
Shi 等[24]	0.768	0.731	0.750
East PVANET[15]	0.7135	0.8086	0.7571
East PVANET2x[15]	0.735	0.836	0.782
EAST PVANET2x MS[15]	0.783	0.833	0.807
TextBoxes++[30]	0.767	0.872	0.817
RRD[8]	0.79	0.8569	0.822
TextSnake[6]	0.804	0.849	0.826
TextBoxes++ MS[30]	0.785	0.878	0.829
Lv 等[7]	0.895	0.797	0.843
本文方法	0.789	0.854	0.82

本文模型与以 PVANET 作为基网络的 EAST 相比, 在都使用单尺度测试的情况下, F 值相比 EAST 高出 6.29%, 当 EAST 基网络 PVANET 通道增加为原来两倍时, 本文方法 F 值高出 3.8%, 更进一步, 本文方法在单尺度测试的情况下, 依然高出 EAST 多尺度测试版本 1.3%. 与方法 TextBoxes++ 相比, 本文方法的单尺度测试结果高出 TextBoxes++ 的单尺度测试结果.

本文方法在 MSRA-TD500 数据集上实验成绩

与现有方法相比也达到了相当的水准, 如表 2 所示.

表 2 MSRA-TD500 测试集检测结果对比
Table 2 Qualitative comparison on MSRA-TD500 dataset

方法	召回率 (R)	精确度 (P)	F 值
Epshtein 等[31]	0.25	0.25	0.25
TD-ICDAR[21]	0.52	0.53	0.50
Zhang 等[23]	0.43	0.71	0.54
TD-Mixture[21]	0.63	0.63	0.60
Yao 等[28]	0.59	0.72	0.65
Kang 等[32]	0.62	0.71	0.66
Yin 等[33]	0.62	0.81	0.71
East PVANET[15]	0.6713	0.8356	0.7445
EAST PVANET2x[15]	0.6743	0.8728	0.7608
TextSnake[6]	0.739	0.832	0.783
本文方法	0.689	0.925	0.79

从表 2 可知, 本文方法与以 PVANET 作为基网络的 EAST 相比, 在 R 值和 F 值上分别高出 1.77%, 8.94% 和 4.45%, 当 EAST 基网络 PVANET 通道增加为原来两倍时, 本文方法 F 值依旧高出 2.8%. Zhang 等[23] 的方法是之前发表的先进多方向文字检测方法, 与其相比, 本文方法在 R 值、P 值、F 值三个指标上分别提升了 25.9%, 21.5%, 25%.

2) 水平自然场景文字数据集

除了在多方向数据集上进行实验, 本文也在水平文本数据集 ICDAR2013 上进行实验, 该数据集是目前最为广泛使用的水平文本数据集. 表 3 展示了本文方法与其他先进文字检测方法的成绩对比情况.

由表 3 可以观察到, 除了 Tang 等[42] 的方法, 本文方法成绩全方面超越表 3 中所列举的先进文字检测方法. 然而, Tang 等[42] 的方法使用的是包含了两个网络的级联结构, 检测一幅图片平均耗时 1.36 s, 更进一步, 该方法只可以检测水平文本数据集, 对多方向文本失效. 与相似网络[43] 结构 EAST 相比, 表 3 中列出以 2 倍通道 PVANET 作为基网络的 EAST 的检测成绩, 召回率、精确率和 F 值分别为 0.8267, 0.9264, 0.8737, 本文方法在三个指标上分别超出 EAST 3.13%, 0.46%, 1.93%.

表 3 ICDAR2013 测试集检测结果对比
Table 3 Qualitative comparison on ICDAR2013 dataset

方法	召回率 (R)	精确度 (P)	F 值
Fasttext[34]	0.69	0.84	0.77
MMser[35]	0.70	0.86	0.77
Lu 等[36]	0.70	0.89	0.78
TextFlow[37]	0.76	0.85	0.80
TextBoxes [38]	0.74	0.86	0.80
TextBoxes++[30]	0.74	0.86	0.80
RRD[8]	0.75	0.88	0.81
He 等[39]	0.73	0.93	0.82
FCN[23]	0.78	0.88	0.83
Qin 等[40]	0.79	0.89	0.83
Tian 等[41]	0.84	0.84	0.84
TextBoxes MS[38]	0.83	0.88	0.85
Lv 等[7]	0.933	0.794	0.858
TextBoxes++ MS[30]	0.84	0.91	0.88
EAST PVANET2x[15]	0.8267	0.9264	0.8737
Tang 等[42]	0.87	0.92	0.90
本文方法	0.858	0.931	0.893

表 4 多种文字检测方法在 ICDAR2015 上的精度和速度对比结果
Table 4 Comparison of accuracy and speed on ICDAR2015 dataset

方法	测试图片尺寸 (像素)	设备	帧率 (帧/s)	F 值
Zhang 等[23]	MS	TitanX	0.476	0.54
Tian 等[27]	ss-600	GPU	7.14	0.61
Yao 等[28]	480 p	K40m	1.61	0.65
Shi 等[24]	768 × 768	TitanX	8.9	0.750
EAST PVANET[15]	720 p	TitanX	16.8	0.757
EAST PVANET2x[15]	720 p	TitanX	13.2	0.782
TextBoxes++[30]	1024 × 1024	TitanX	11.6	0.817
RRD[8]	1024 × 1024	TitanX	6.5	0.822
TextSnake[6]	1280 × 768	TitanX	1.1	0.826
TextBoxes++ MS[30]	MS	TitanX	2.3	0.829
Lv 等[7]	512 × 512	TitanX	1	0.843
本文方法	720 p	TitanX	12.5	0.82

3.3.2 时间性能

本文方法不仅检测准确, 而且检测快速. 在 ICDAR2015 数据集上对本文方法和部分先进检测算法[44−45] 的运行速度进行比较, 结果如表 4 所示.

由表 4 可知, 本文方法在取得 82% 的 F 值的情况下, 检测速度为 12.5 帧/s. 相较其他方法, 这样的结果在性能和速度上达到了相对均衡. 观察表 4, 可以看到 Tian 等[27] 提出的 ss-600 方法, 训练时图片的最短边缩放到 600, 其在 ICDAR2015 数据集上的最优结果是在将图片最短边放大到 2 000 时得到的, 这种情况下, 该方法的时间相对表 4 中显示的时间会更慢. 对于 Zhang 等.[23] 的方法, MS 表示使用三个尺度测试 (如 200, 500, 1 000). EAST 方法在以 PVANet 为基网络时, 可以达到 16.8 帧/s 的速度, 虽然 EAST 方法比本文方法略快, 但是在

ICDAR2015 数据集上 F 值低于本文方法 6.3%. EAST 方法为了提高检测成绩, 将 PVANet 的通道数增加为原来的两倍, 速度增为 13.2 帧/s, 与本文方法速度相近, 但检测的 F 值依然比本文方法低 4% 左右.

3.4　模型分析

3.4.1　模型各组件作用

为了直观地观察模型中各组件的作用, 本节进行控制变量实验来观察各组件如何影响模型的最终效果. 由于 ICDAR2015 数据集为自然场景图, 在该数据集上的结果更能体现方法的实用性, 因而整个实验在该数据集上进行. 本节的所有实验除了控制变量, 其他条件均相同, 实验结果如表 5 所示.

从表 5 中可以看出: 1) 本文模型通过使用基网络 ResNet50, F 值得到提升, 在 ICDAR2015 数据集上达到 79.7%. 2) 在本文网络结构中, 对 ResNet50 的第 4 阶段和第 5 阶段特征图之后嵌入感受野增强模块, F 值得到 0.5% 的提升. 这样的实验结果说明增大网络的感受野对网络性能确实有所提升, 提升不是很明显的主要原因是 ICDAR2015 数据集主要特点在于自然场景背景的复杂, 而不在于长、大文本, 因此基础模型由于感受野不足而误检的情况并不常见. 3) 当对像素点进行文本/非文本分类时, 引入 Focalloss 作为分类损失, F 值提高到 81.3%. 这组实验一定程度上说明正负样本不均衡问题确实影响网络性能, 并且 Focalloss 确实改善了网络性能. 4) 使用 GIoU 作为网络回归矩形的损失, 使得 F 值再次得到提升. 最终, 本文方法在 ICDAR2015 数据集上的召回率、精确率、F 值分别为 78.9%, 85.4% 和 82%.

3.4.2　模型优缺点

图 8 给出了本文模型在 ICDAR2013、ICDAR2015 以及 MSRA-TD500 数据集测试集上的部分图像检测结果. 从这些检测结果图可以看出, 本文方法在多方向数据集、水平数据集上都表现出了优异的检测结果, 并且对于一幅图像上出现文字尺度多变、宽高比多边的情况, 本文方法检测依然具有一定的鲁棒性. 另外, 从图 8(c) 可看出本文方法不仅可以检测英文文本, 中文文本同样可以检测. 但是本文方法也存在检测效果不理想的情况, 如图 9(a) 所示, 对于过长文本和特大文字, 本文方法会出现检测不全甚至漏检的情况. 考虑到长文本和特大文字需要更大的感受野, 虽然添加了感受野增强模块, 但感受野依然受限, 导致长文本和特大文字检测失败. 图 9(b) 显示了本文方法对曲线文本检测[46]的效果差强人意, 主要原因是一方面旋转矩形框无法准确地表示出曲线文本的形状, 另一方面可能是因为所使用的三个数据集的训练集中包含曲线文本的图像样本几乎没有. 图 9(b) 也显示出本文方法在垂直文本检测方面效果欠佳, 这个问题出现的一个主要原因可能是在训练集中包含垂直文本的样本图片数量较少, 导致网络对垂直文本的学习程度不够.

4　结束语

本文提出并介绍了一种结合感受野增强和全卷积网络的多方向文本检测方法. 该方法基于以 ResNet50 为基网络的全卷积网络 (FCN), 不仅可以鲁棒地检测任意方向文本和多尺度文本, 而且消除了冗余且耗时的中间步骤, 可端对端训练. 首先, 为了提升不同尺度和宽高比文字检测准确率, 受人类视觉的感受野结构的启发, 使用多层卷积和空洞卷积设计了感受野增强模块, 使得网络对尺度、宽高比多变的文字检测更加鲁棒, 然后, 针对文字检测中样本不均衡问题, 引入 Focalloss 对像素点进行文本/非文本预测, 从而一定程度上提升了网络的检测性能; 其次, 针对以往 IoUloss 使用存在的几个弊端问题, 引入 GIoU 作为包围框回归损失, 改善文本定位精确性; 最后, 在多方向文本数据集 ICDAR2015 和 MSRA-TD500 以及水平文本数据集 ICDAR2013 上与现有的顶级方法进行对比实验和模型分析, 最后结果显示本文方法达到了现有先进水平, 并且也验证了本文各部件的作用.

表 5　本文方法各组件在 ICDAR2015 数据集上的作用效果
Table 5　Effectiveness of various designs on ICDAR2015 dataset

ResNet50	感受野增强模块	Focalloss	GIoUloss	召回率 (R)	精确度 (P)	F 值
×	×	×	×	0.735	0.836	0.782
√	×	×	×	0.764	0.833	0.797
√	√	×	×	0.766	0.845	0.802
√	√	√	×	0.776	0.853	0.813
√	√	√	√	0.789	0.854	0.82

(a) ICDAR2013 检测结果
(a) Results on ICDAR2013 dataset

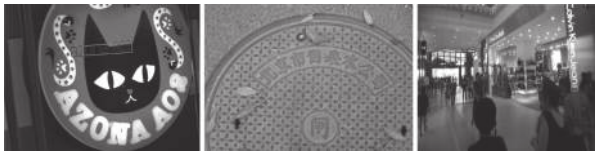
(b) ICDAR2015 检测结果
(b) Results on ICDAR2015 dataset

(c) MSRA-TD500 检测结果
(c) Results on MSRA-TD500 dataset

图 8 本文方法在各个数据集上检测结果比较

Fig. 8 Comparison of detection results on different datasets

(a) 长文本和特大文字出现漏检或检测不全
(a) Failure of detecting long text and extra text



(b) 曲线文本和垂直文本检测失败
(b) Failure of detecting curve text and vertical text

图 9 本文方法检测失败的一些场景图像

Fig. 9 Some scene image of detect failure

References

1 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot multibox detector. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 21–37

2 Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the 2015 Advances in Neural Information Processing Systems. NIPS, 2015. 91–99

3 He W H, Zhang X Y, Yin F, Liu C L. Deep direct regression for multi-oriented scene text detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 745–753

4 Deng D, Liu H F, Li X L, Cai D. Pixellink: Detecting scene text via instance segmentation. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, LA, USA: AAAI, 2018. 6773–6780

5 Yan S, Feng W, Zhao P, Liu C L. Progressive scale expansion network with octave convolution for arbitrary shape scene text detection. In: Proceedings of the 2019 Asian Conference on Pattern Recognition. Springer, Cham, 2019. 663–676

6 Long S B, Ruan J Q, Zhang W J, He X, Wu W H, Yao C. TextSnake: A flexible representation for detecting text of arbitrary shapes. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 19–35

7 Lv P Y, Yao C, Wu W H, Yan S C, Bai X. Multi-oriented scene text detection via corner localization and region segmentation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7553–7563

8 Liao M H, Zhu Z, Shi B G, Xia G S, Bai X. Rotation-sensitive regression for oriented scene text detection. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 5909–5918

9 Lyu P Y, Liao M H, Yao C, Wu W H, Bai X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 71–88

10 Lin T Y, Goyal P, Girshick R, He K M, Dollar P. Focal loss for dense object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2999–3007

11 Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019. 658–666

12 Lin T Y, Dollar P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 936–944

13 Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 2017, **60**(6): 84–90

14 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016. 770–778

15 Zhou X Y, Yao C, Wen H, Wang Y Z, Zhou S C, He W R, et al. EAST: An efficient and accurate scene text detector. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 2642–2651

16 Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection. In: Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018. 404–419

17 Szegedy C, Ioffe S, Vanhoucke V, Alemi A A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 4278–4284

18 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv: 1511.07122, 2015.

19 Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda L G I, Mestre S R, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, USA: IEEE, 2013. 1484–1493

20 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. ICDAR 2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE, 2015. 1156–1160

21 Yao C, Bai X, Liu W Y, Ma Yi, Tu Z W. Detecting texts of arbitrary orientations in natural images. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 1083–1090

22 Yao C, Bai X, Liu W Y. A unified framework for multioriented text detection and recognition. *IEEE Transactions on Image Processing*, 2014, **23**(11): 4737–4749

23 Zhang Z, Zhang C Q, Shen W, Yao C, Liu W Y, Bai X. Multi-oriented text detection with fully convolutional networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision

and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016. 4159−4167

24 Shi B G, Bai X, Belongie S. Detecting oriented text in natural images by linking segments. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE 2017. 3482−3490

25 Islam M R, Mondal C, Azam M K, Islam A S M. Text detection and recognition using enhanced MSER detection and a novel OCR technique. In: Proceedings of the 5th International Conference on Informatics, Electronics and Vision (ICIEV). Dhaka, Bangladesh: IEEE, 2016. 15−20

26 Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 2315−2324

27 Tian Z, Huang W L, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer, 2016. 56−72

28 Yao C, Bai X, Sang N, Zhou X Y, Zhou S C, Cao Z M. Scene text detection via holistic, multi-channel prediction. arXiv: 1606.09002, 2016.

29 Liu Y L, Jin L W. Deep matching prior network: Toward tighter multi-oriented text detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: 2017. 3454−3461

30 Liao M H, Shi B G, Bai X. TextBoxes++: A single-shot oriented scene text detector. *IEEE Transactions on Image Processing*, 2018, **27**(8): 3676−3690

31 Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE, 2010. 2963−2970

32 Kang L, Li Y, Doermann D. Orientation robust text line detection in natural images. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 4034−4041

33 Yin X C, Pei W Y, Zhang J, Hao H W. Multi-orientation scene text detection with adaptive clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(9): 1930−1937

34 Buta M, Neumann L, Matas J. FASText: Efficient unconstrained scene text detector. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015. 1206−1214

35 Zamberletti A, Noce L, Gallo I. Text localization based on fast feature pyramids and multi-resolution maximally stable extremal regions. In: Proceedings of the 2015 Asian Conference on Computer Vision. Singapore, Singapore: Springer, 2014. 91−105

36 Lu S J, Chen T, Tian S X, Lim J H, Tan C L. Scene text extraction based on edges and support vector regression. *International Journal on Document Analysis and Recognition (IJDAR)*, 2015, **18**(2): 125−135

37 Tian S X, Pan Y F, Huang C, Lu S J, Yu Kai, Tan C L. Text flow: A unified text detection system in natural scene images. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: 2015. 4651−4659

38 Liao M H, Shi B G, Bai X, Wang X G, Liu W Y. Textboxes: A fast text detector with a single deep neural network. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI, 2017. 4161−4167

39 He T, Huang W L, Qiao Y, Yao J. Text-attentional convolutional neural network for scene text detection. *IEEE Transactions on Image Processing*, 2016, **25**(6): 2529−2541

40 Qin S Y, Manduchi R. A fast and robust text spotter. In: Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision. Lake Placid, USA: IEEE, 2016. 1−8

41 Tian C N, Xia Y, Zhang X N, Gao X B. Natural scene text detection with MC-MR candidate extraction and coarse-to-fine filtering. *Neurocomputing*, 2017, **260**: 112−122

42 Tang Y B, Wu X Q. Scene text detection and segmentation based on cascaded convolution neural networks. *IEEE Transactions on Image Processing*, 2017, **26**(3): 1509−1520

43 Li Wen-Ying, Cao Bin, Cao Chun-Shui, Huang Yong-Zhen. A deep learning based method for bronze inscription recognition. *Acta Automatica Sinica*, 2018, **44**(11): 2023−2030
(李文英, 曹斌, 曹春水, 黄永祯. 一种基于深度学习的青铜器铭文识别方法. 自动化学报, 2018, **44**(11): 2023−2030)

44 Wang Run-Min, Sang Nong, Ding Ding, Chen Jie, Ye Qi-Xiang, Gao Chang-Xin, et al. Text detection in natural scene image: A survey. *Acta Automatica Sinica*, 2018, **44**(12): 2113−2141
(王润民, 桑农, 丁丁, 陈杰, 叶齐祥, 高常鑫, 等. 自然场景图像中的文本检测综述. 自动化学报, 2018, **44**(12): 2113−2141)

45 Jin Lian-Wen, Zhong Zhuo-Yao, Yang Zhao, Yang Wei-Xin, Xie Ze-Cheng, Sun Jun. Applications of deep learning for handwritten Chinese character recognition: A review. *Acta Automatica Sinica*, 2016, **42**(8): 1125−1141
(金连文, 钟卓耀, 杨钊, 杨维信, 谢泽澄, 孙俊. 深度学习在手写汉字识别中的应用综述. 自动化学报, 2016, **42**(8): 1125−1141)

46 Wang W H, Xie E Z, Li X, Hou W B, Lu T, Yu G, Shao S. Shape robust text detection with progressive scale expansion network. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE, 2018. 9328−9337

李晓玉 西安交通大学软件学院硕士研究生. 主要研究方向为自然场景文字检测技术.
E-mail: 18155760591@163.com
(**LI Xiao-Yu** Master student at the School of Software Engineering, Xi′an Jiaotong University. Her research interest covers text detection in natural scenes.)

宋永红 西安交通大学人工智能学院研究员. 主要研究方向为图像与视频内容理解, 智能软件开发. 本文通信作者. E-mail: songyh@xjtu.edu.cn
(**SONG Yong-Hong** Professor at the College of Artificial Intelligence, Xi′an Jiaotong University. Her research interest covers image and video content understanding, and intelligent software development. Corresponding author of this paper.)

余 涛 西安交通大学软件学院硕士研究生. 2018 年获得西安交通大学软件学院学士学位. 主要研究方向为自然场景文字检测技术.
E-mail: yyttmonster@outlook.com
(**YU Tao** Master student at the School of Software Engineering, Xi′an Jiaotong University. He received his bachelor degree from Xi′an Jiaotong University in 2018. His research interest covers text detection in natural scenes.)