DOI:10.11918/201908104

短边顶点回归网络:新型自然场景文本检测器

游洋彪,石繁槐

(同济大学 电子与信息工程学院,上海 201804)

摘 要:近年来许多基于通用目标检测框架的文本检测方法相继被提出,这些方法往往是直接预测文本的整个边界框,受网络感受野的限制而难以有效检测长文本。为改进长文本难以有效检测的问题,提出了基于短边顶点回归网络的文本检测方法。该方法将文本区域划分为3类区域,即两条短边附近的区域及中间区域,采用分离再组合的方式检测文本,不再直接预测文本的整个边界框。首先,在一个融合多层特征的残差网络上预测分割3类文本区域,同时还将在每个短边区域的像素点处预测与之邻近的一条短边的两个顶点。然后,在后处理过程中,利用文本中间区域与短边区域相邻的关系将文本两类短边区域进行组合,两类短边区域预测的短边顶点将随之结合,便能产生完整精确的文本检测结果。在一个长文本检测数据集和公开的MSRA-TD 500,ICDAR 2015及ICDAR 2013自然场景文本检测数据集上进行测试比较,该方法的精度与速度超过目前绝大部分方法。实验结果表明,该方法在文本检测,尤其是长文本检测,具有一定的优越性。

关键词:自然场景;文本检测;卷积神经网络;感受野;长文本

中图分类号:TP391.4 **文献标志码**:A **文章编号**:0367-6234(2021)12-0089-09

Short edge vertices regression network: A new natural scene text detector

YOU Yangbiao, SHI Fanhuai

(College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China)

Abstract: In recent years, many scene text detection methods based on generic object detection framework have been proposed. These methods usually predict the entire bounding box of the text directly, while it is difficult for them to detect long text effectively due to the limit of receptive field. To solve such problem, a scene text detection method based on short edge vertices regression network was proposed. The method divides the text region into three kinds of regions, namely regions near two short edges and middle region, where separate text regions are firstly predicted and then combined, whereas the entire bounding box of the text is not predicted directly. Specifically, three kinds of regions were segmented on a residual network combined with multi-scale features, and two vertices of a short edge were predicted at each pixel in the region near the short edge. Then the regions near the two short edges were combined on the basis of the adjacent relationship between middle region and short edge regions in the post process, and vertices of short edges predicted by the two regions near short edges were combined to generate complete and accurate detection results. Finally, experiments were performed on a long text detection dataset and several public scene text detection datasets such as MSRA-TD 500, ICDAR 2015, and ICDAR 2013. The proposed method outperformed most of existing methods in accuracy and speed. Experimental results demonstrate that the method has advantages in text detection, especially for long text.

Keywords: natural scene; text detection; convolutional neural network; receptive field; long text

近年来,自然场景图像中的文本检测成为了计算机视觉领域的一个研究热点。自然场景图像文本检测在图像检索、定位导航、盲人辅助、数据录入等领域具有重要的实用价值。自然场景图像背景千变万化,并且自然场景图像文本具有大小和长宽比变化剧烈、多方向等特点;此外,与一般目标检测不同的是,文本目标框可能使用水平矩形、四边形、旋转矩形,甚至是多边形等形式进行精确表达,所以自然场景图像文本检测一直是一个难点问题。

通用目标检测(generic object detection)[1],定位图像上预先定义类别的目标实例的位置,检测结果通常以外接矩形框的形式呈现,不同于专用目标检测只适用于一种或一些类别,通用目标检测适用于广泛的类别,典型方法有Faster R-CNN[2]、SSD[3]等。随着深度学习技术的迅速发展,通用目标检测性能得到了显著的提升。很多学者在通用目标检测方法的基础上提出了许多自然场景文本检测方法。这些方法可以大致分为两类:间接回归方法与直接回归方法。间接回归方法,通常借鉴Faster R-CNN[2]、SSD[3]等目标检测方法,预先设定一些锚(anchor)或先验框(default box),利用卷积神经网络判定它们是否与文本区域高度重叠并调整它们的大小和位置以准确定位文本。然而自然场景中的文本方向多

收稿日期:2019-08-19
基金项目:上海市科技兴农重点攻关项目(沪农科创字(2018)第3-6号)
作者简介:游洋彪(1995—),男,硕士研究生
通信作者:石繁槐,fhshi@tongji.edu.cn

变,大小、长宽比变化剧烈,为了使预设的锚或先验框能与文本区域高度重叠,很多方法增加了多种方向、多种大小及多种长宽比的预设框,但这无疑增加了方法的复杂度与计算量。

为适应文本的特性,Liao 等[4]、Zhong 等[5]分别在 Faster R-CNN、SSD 的基础上增加了不同尺寸与长宽比的预设框。Ma 等[6]、Liu 等[7]为检测多方向的文本设置了多方向的预设框。为了能够输出四边形的检测结果,Liu 等[7]、Liao 等[8]直接预测预设框与文本边界四边形 4 个顶点的坐标差;Jiang 等[9]通过预测两个顶点坐标与一条边长得到旋转矩形的检测结果;Zhu 等[10]通过预测多个文本边界上的点得到多边形的检测结果。

直接回归方法不需要预设框,相比于间接回归方法,它更加灵活简便。直接回归方法借鉴了 DenseBox[11]的思想,这类方法通常采用全卷积网络[12](fully convolutional network,FCN)的架构,在分割出文本区域中的候选点的同时在每个点处预测对应的文本区域边界。为了得到四边形的检测结果,He 等[13]在分割出文本区域中像素点的同时预测该点与四边形边界顶点的坐标偏差。为了降低复杂度,Zhou 等[14]预测文本区域中的点到文本外接旋转矩形 4 条边的距离与旋转矩形的角度。Xue 等[15]则在此基础上还分割了文本的边界区域以助于区分文本实例。

上述基于回归的文本检测方法或是通过调整预设框得到文本的外接四边形,或是在每个点处直接预测文本的外接四边形,都是直接预测文本的整个边界框。这些方法能检测到目标尺寸与网络的感受野大小成正相关,当检测更长的文本目标时,网络需要更大的感受野。在面对长文本时,由于感受野大小有限,并且相应感受野内背景干扰可能更多,这些直接预测整个文本边界的方法难以得到理想的结果。

针对直接预测整个文本边界的方法在检测长文本时的缺陷,本文设计了一种短边顶点回归网络,该网络不再直接预测文本区域的所有边界。具体来说,本文方法在直接回归方法的基础上,分割出文本的两条短边附近的区域以及中间区域。不同于其他直接回归方法中文本区域中的点需要预测文本整个边界框,本方法中,文本边界框顶点由短边附近区域中的点来预测,并且一条短边区域内的点只预测其附近短边的两个顶点,而不需预测另外一条更远短边的顶点。在检测长文本时,相应感受野内背景干扰相对更少,同时对感受野大小的要求更低,所以检测结果更为准确。为了将预测的两组短边顶点结合,本文设计了一种新的后处理方法,利用中间区域与两短边区域相邻或两短边区域直接相邻的特点将文本的两个短边区域组合,两组预测的短边顶点随之结合,便能得到精确完整的文本检测结果。本文所提方法在多个自然场景文本检测数据集上均取得了不错的效果,结果超过了目前绝大部分方法,并且本方法更快速高效。

1 短边顶点回归网络的文本检测方法

图 1 为本文方法的原理流程图,本方法采用了全卷积与多层特征融合的网络架构。图像输入网络后,网络输出 3 种像素级分类结果,即文本中间区域像素、文本左短边区域像素、文本右短边区域像素。文本短边区域是指文本短边边界附近的区域。如图 1 的区域分割结果所示,其中蓝色、绿色、红色区域分别为文本中间区域、左短边区域与右短边区域。在分类短边区域像素的同时,网络还在该点处预测附近一条短边两个顶点的坐标。最后通过后处理,本方法将左短边顶点与右短边顶点的预测结果结合起来,得到最终的检测结果。

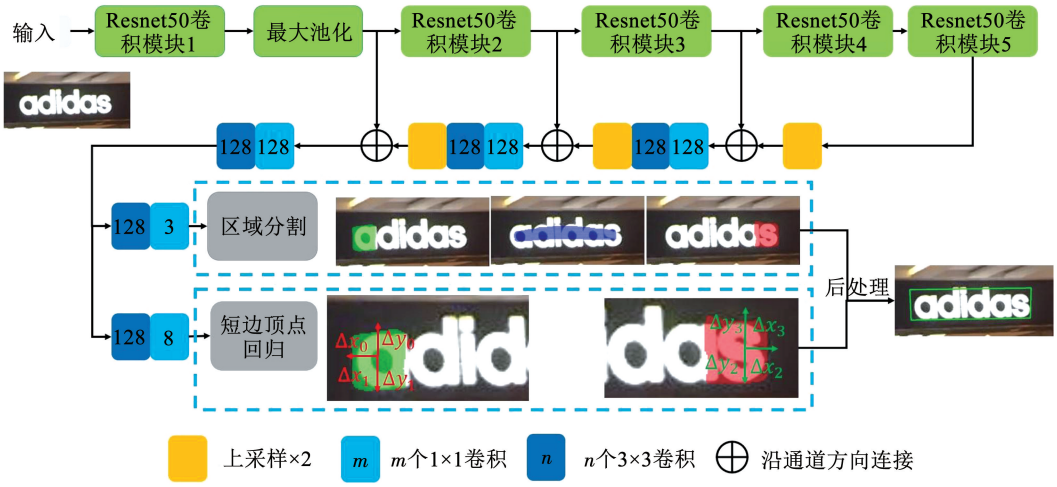


图 1 基于短边顶点回归网络的文本检测方法流程

Fig. 1 Flowchart of scene text detection with short edge vertices regression network

1.1　网络结构

本方法的网络结构可以大致分为 3 个部分：特征提取、特征融合以及分类回归。特征提取部分使用 Resnet 50[16] 的框架，去除 Resnet 50 后面的全连接层，图 1 中绿色模块为 Resnet 50 特征提取部分。相比于经典的 VGG16/19[17]，Resnet 50 参数量更少，计算存储花销更小，而且 Resnet 50 使用了残差结构，能够有效缓解训练时发生梯度消失的情况。

自然场景图像中文本具有尺度变化剧烈的特点，特征提取网络越深，提取到的特征语义范围越广，越有利于大尺度文本的检测，而检测小的文本需要靠浅层局部的特征。为了能够检测不同大小的文本，本方法参考 U-Net[18] 的架构将 Resnet 50 提取到的多层特征进行融合。具体来说，高层的特征首先进行上采样，与低一层的特征的长宽维度保持一致，然后沿通道方向将上采样特征与低一层特征进行连接，最后使用一个 1×1 与一个 3×3 的卷积操作将特征进行融合。融合后的特征继续融合更低层特征，直至融合的特征长宽为原图像的 1/4。

对于最小外接矩形长宽比接近于 1 的文本区域，它的中间区域、左短边区域、右短边区域会有部分重叠。所以在网络输出的结果中，同一个像素可以同时属于中间区域、左短边区域、右短边区域。在最后分类时，中间区域、左短边区域、右短边区域均与背景进行二分类，这样能够避免类间的竞争。具体来说，在最后融合的特征上分别使用一个 1×1 的卷积操作与一个 sigmoid 非线性函数来预测每个像素点属于中间区域、左短边区域、右短边区域的概率。

1.2　训练样本标签生成

短边顶点回归网络的目标之一在于分割文本的中间区域、左短边区域及右短边区域。文本短边区域是文本短边边界附近的部分区域，在确定文本左、右短边区域前，首先需要规定文本的左短边及右短边。由于文本目标大多使用四边形进行标注，而四边形的两条对边不一定为最短的两条边，所以在此通过寻找四边形最小包围矩形的左短边及右短边来确定文本边界四边形的对应短边。

如图 2(a) 所示，黄色四边形是文本区域原始的标注，红色的矩形是该四边形的最小包围矩形 R。矩形 R 顺时针旋转直至长边与水平轴平行，假设此时底部长边为 w，转过的角度为矩形 R 的倾角 θ。当$\theta\leqslant45°$时，R 旋转后，位于 w 右侧的短边对应的是 R 的右短边，位于 w 左侧的短边为 R 的左短边，例如图 2(b) 中的 R1、R2、R3；当 $\theta>45°$时，R 旋转后，位于 w 左侧的短边对应的是 R 的右短边，位于 w 右侧的短边为 R 的左短边，例如图 2(b) 中的 R4、R5、R6、R7、R8。图 2(b) 中矩形 R 绿色短边为左短边，红色短边为右短边。文本四边形的左、右短边与其最小包围矩形 R 的左、右短边一一对应。

在确定文本的左短边与右短边后，再精确定义文本的中间区域、左短边区域、右短边区域。首先定义四边形 $Q=\{q_0,q_1,q_2,q_3\}$，其顶点 $q_i(i=0,1,2,3)$ 的参考长度为

$$l_i=\min(h_{i,(i+1)\bmod4},h_{(i+3)\bmod4,i}) \tag{1}$$

式中：$h_{i,j}$ 为边 q_iq_j 的长度，mod 为求余数。将四边形 Q 的顶点 q_i 沿长边方向、短边方向(指向四边形内部)各缩进 $0.3l_i$ 后，4 个新顶点包围的区域就是文本的中间区域，如图 2(c) 的蓝色区域所示。中间区域的主要作用是确定文本实例，所以不需要将文本区域中所有像素都分割出来，使用缩小的中间区域有助于分离出文本实例，尤其是在检测密集的文本目标时。左短边顶点 q_i、q_j 沿长边方向向内分别移动 l_i、l_j，新的两个顶点与原短边顶点 q_i、q_j 包围的区域便是左短边区域。如图 2(d) 所示，q_0 沿$\overrightarrow{\boldsymbol{q}_0\boldsymbol{q}_1}$方向移动 l_0 至 q_4，q_3 沿$\overrightarrow{\boldsymbol{q}_3\boldsymbol{q}_2}$方向移动 l_3 至 q_5，四边形 $q_0q_4q_5q_3$ 包围的绿色区域即为左短边区域，同理四边形 $q_1q_2q_7q_6$ 包围的红色区域是右短边区域。

1.3　损失函数

本方法是基于直接回归的文本检测方法，所以设计损失函数时参考了其他同类的方法[13-15]。本方法损失函数为：

$$L=L_{\text{cls}}+L_{\text{reg}} \tag{2}$$

$$L_{\text{cls}}=\alpha_1L_{\text{m\_cls}}+L_{\text{l\_cls}}+L_{\text{r\_cls}} \tag{3}$$

$$L_{\text{reg}}=L_{\text{l\_reg}}+L_{\text{r\_reg}} \tag{4}$$

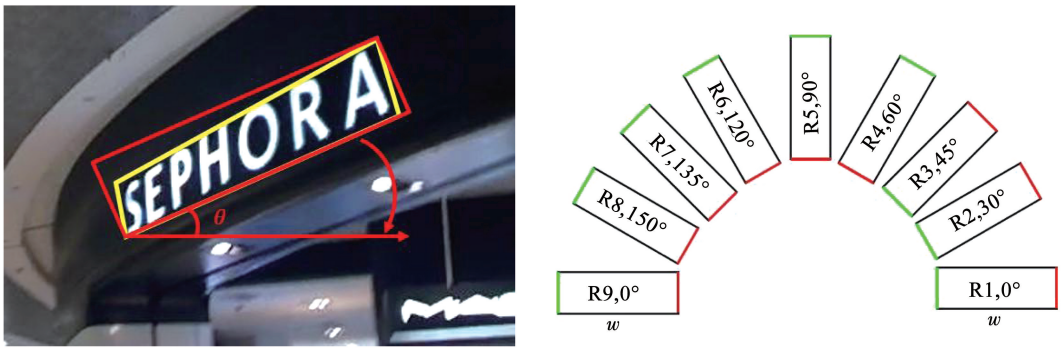
式中：L 为最后总损失；L_{cls} 为 3 类区域的分类损失和；L_{reg} 为短边区域顶点回归损失和；$L_{\text{m\_cls}}$、$L_{\text{l\_cls}}$、$L_{\text{r\_cls}}$ 分别为中间区域、左短边区域及右短边区域的像素分类损失；$L_{\text{l\_reg}}$ 为左短边顶点回归损失；$L_{\text{r\_reg}}$ 为右短边顶点回归损失。由于中间区域起着确定文本实例的作用，相对更加重要，实验中其分类损失权重 α_1 设置为 4。

在自然场景图像中，文本区域往往只占很小一部分，如果分类损失函数使用交叉熵类型的损失函数，很可能由于正负样本不平衡，导致最后分类结果倾向于背景。本方法采用 D(dice coefficient) 函数[19] 作为分类的损失函数，分类损失为：

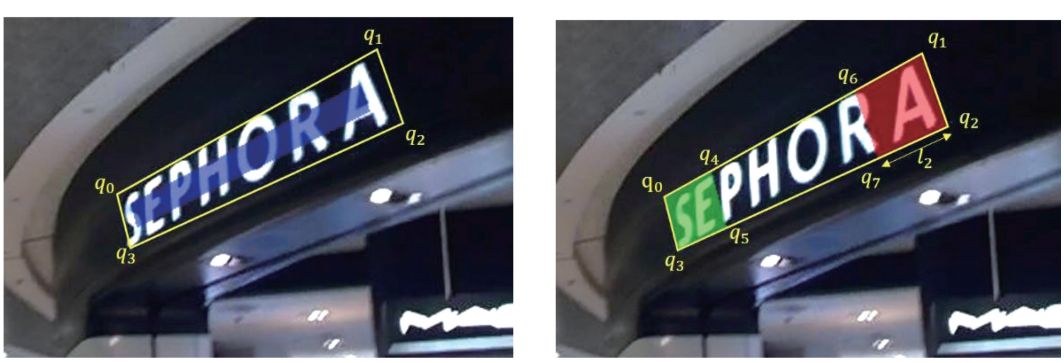
$$L_{\text{cls}}=\alpha_1D(\boldsymbol{P}_{\text{m}},\boldsymbol{G}_{\text{m}})+D(\boldsymbol{P}_{\text{l}},\boldsymbol{G}_{\text{l}})+D(\boldsymbol{P}_{\text{r}},\boldsymbol{G}_{\text{r}}) \tag{5}$$

$$D(\boldsymbol{P},\boldsymbol{G})=1-\frac{2\sum_{x,y}P_{x,y}\times G_{x,y}}{\sum_{x,y}P_{x,y}+\sum_{x,y}G_{x,y}} \tag{6}$$

式中：$\boldsymbol{P}_{\text{m}}$、$\boldsymbol{P}_{\text{l}}$、$\boldsymbol{P}_{\text{r}}$ 分别为中间区域、左短边区域、右短



(a) 文本原始标注框（黄）及其最小包围矩形（红） (b) 常见多方向矩形的左短边（绿）及右短边（红）



(c) 文本中间区域（蓝） (d) 文本左短边区域（绿）与右短边区域（红）

图 2 左短边、中间、右短边区域标签生成

Fig. 2 Generation of labels of region near left short edge, middle region, and region near right short edge

边区域的分类得分预测值；$\boldsymbol{G}_{\mathrm{m}}$、$\boldsymbol{G}_{\mathrm{l}}$、$\boldsymbol{G}_{\mathrm{r}}$为分类得分真实值；$P_{x,y}$、$G_{x,y}$分别为点$(x,y)$分类得分的预测值与真实值。

假设 $V=\{x_{l_1^*},y_{l_1^*},x_{l_2^*},y_{l_2^*},x_{r_1^*},y_{r_1^*},x_{r_2^*},y_{r_2^*}\}$ 为文本区域边界四边形的顶点坐标值，$C_i^k=\{\Delta x_i^k,\Delta y_i^k\}$ 为短边区域点 (x_k,y_k) 与对应短边一个顶点 (x_i,y_i) 坐标差的真实值，Δx_i^k、Δy_i^k 可以使用下式计算：

$$\Delta x_i^k=x_k-x_i \tag{7}$$

$$\Delta y_i^k=y_k-y_i \tag{8}$$

式中：$i\in\{l_1^*,l_2^*,r_1^*,r_2^*\}$，$l_1^*$、$l_2^*$ 为四边形 Q 中左短边顶点的索引；r_1^*、r_2^* 为右短边顶点的索引。左、右短边顶点回归损失函数为：

$$L_{\mathrm{l\_reg}}=\sum_{k=1}^{N_{\mathrm{l}}}\sum_{i\in\{l_1^*,l_2^*\}}\sum_{c\in C_i^k}\frac{f(c-\hat{c})}{4\times l_i\times N_{\mathrm{l}}} \tag{9}$$

$$L_{\mathrm{r\_reg}}=\sum_{k=1}^{N_{\mathrm{r}}}\sum_{i\in\{r_1^*,r_2^*\}}\sum_{c\in C_i^k}\frac{f(c-\hat{c})}{4\times l_i\times N_{\mathrm{r}}} \tag{10}$$

$$f(x)=\begin{cases}0.5x^2,&|x|<1\\|x|-0.5,&|x|\geqslant 1\end{cases} \tag{11}$$

式中：$\hat{c}$ 为网络预测的坐标差值；c 为坐标差的真实值；i 为四边形顶点索引；l_i 作为一个归一化项；l_i 为式(1)计算得到的顶点 q_i 的参考长度，加入归一化项能使模型对目标的尺度不敏感；N_{l}、N_{r} 分别为左短边区域、右短边区域所有点的数量；f 为 smooth L1[2] 损失函数。

1.4 后处理

从网络的输出中不能直接得到文本区域的检测结果，还需要进行后处理才能得到完整的结果。对于点(x,y)，用 S_{m}、S_{l}、S_{r} 分别表示该点属于文本中间区域、左短边区域、右短边区域的分类得分。T_{m}、T_{l}、T_{r} 分别表示中间区域、左短边区域、右短边区域的分类阈值。当 $S_{\mathrm{m}}>T_{\mathrm{m}}$，$S_{\mathrm{l}}<T_{\mathrm{l}}$，$S_{\mathrm{r}}<T_{\mathrm{r}}$ 时，称点(x,y)为有效中间区域点，这类点连接形成的区域称为有效中间区域，当 $S_{\mathrm{m}}>T_{\mathrm{m}}$，$S_{\mathrm{l}}\geqslant T_{\mathrm{l}}$ 时，称该点为有效左短边区域点，这类点连接形成的区域称为有效左短边区域，当 $S_{\mathrm{m}}>T_{\mathrm{m}}$，$S_{\mathrm{r}}\geqslant T_{\mathrm{r}}$ 时，称该点为有效右短边区域点，这类点形成的区域称为有效右短边区域。舍弃其他短边区域点的原因在于密集的文本实例的短边区域可能存在误连接的情况，容易导致文本边界顶点预测不准确。

整个后处理的流程如图 3(a)所示，具体如下：

1)寻找有效区域。遍历所有的像素点，找到所有的有效中间区域、有效左短边区域及有效右短边区域，分别如图 3(c)中蓝色、绿色、红色区域所示。同时记录下各有效区域的相邻区域。当某个有效区域的点与其他有效区域的点相邻或重叠时，则这两个有效区域相邻，如图 3(d)所示。

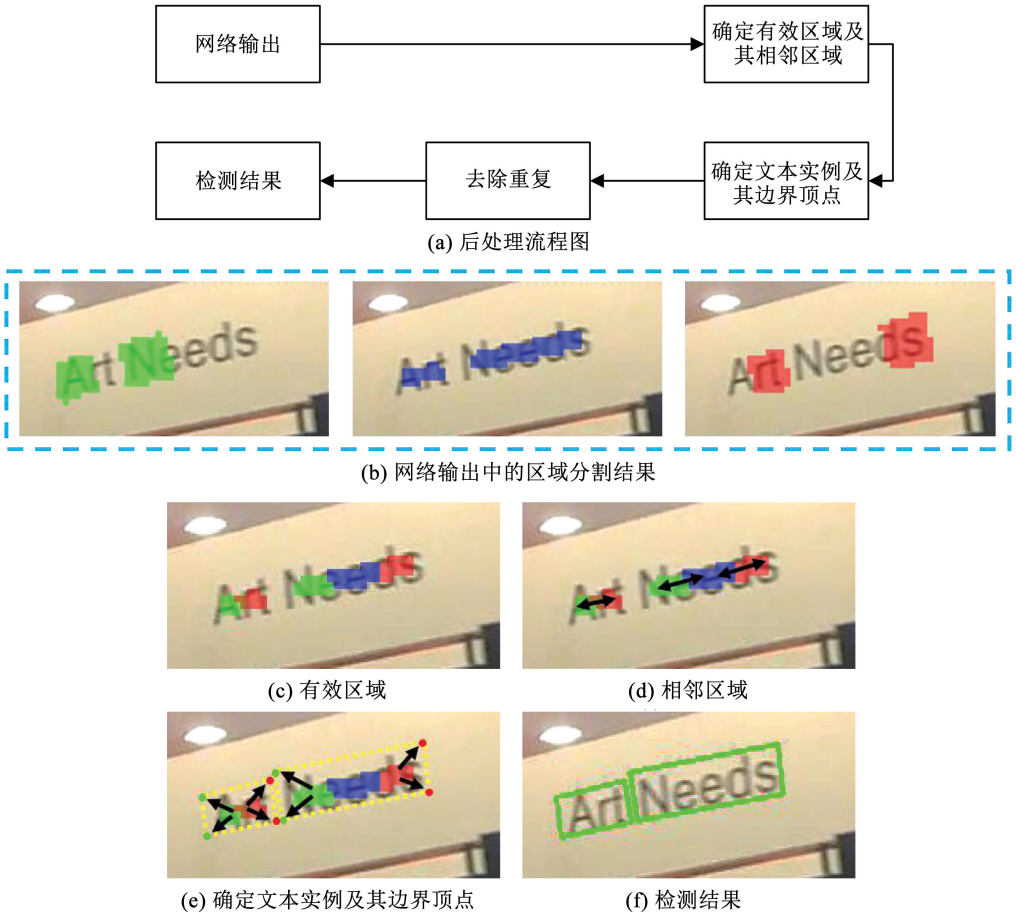
2)确定文本实例及其边界顶点。遍历所有的有效中间区域，当该有效中间区域相邻的有效左短边区域、有效右短边区域数目均不小于 1 时，则 3 种区域共同构成一个文本实例。若相邻的有效短边区域数大于 1，只选最大的有效短边区域。遍历所有的有效左短边区域，当该有效左短边区域相邻的有效

中间区域数为 0，相邻的有效右短边区域数大于 0 时，则两种短边区域同样构成一个文本实例。在确定文本实例后，综合计算左、右短边顶点坐标，计算方式为

$$x = \frac{\sum_{i=1}^{n} s_i \times x_i}{\sum_{i=1}^{n} s_i} \tag{12}$$

式中：x_i 为由单个有效短边区域点预测的短边顶点坐标；s_i 为该有效短边区域点的短边区域分类得分；n 为该短边区域有效点数；x 为最后综合计算的短边顶点坐标结果。

3）去除重复。当同一个连通区域内有多个重叠的检测结果，去除面积较小的。



(a) 后处理流程图

(b) 网络输出中的区域分割结果

(c) 有效区域　(d) 相邻区域

(e) 确定文本实例及其边界顶点　(f) 检测结果

图 3　后处理流程图及中间结果示例

Fig. 3　Flowchart of post process and illustration of intermediate results

2　实验比较与分析

为了验证本文方法的效果，本文将在常用的 3 个公开的自然场景文本检测数据集及一个长文本数据集上进行测试比较。

2.1　数据集与评价指标

1）长文本数据集。根据主观经验，当一个文本实例长边与短边之比大于 7 时，认为该文本实例为长文本。从 MLT 数据集[20]中选取 742 张含有长文本实例的图片作为长文本数据集。该数据集的文本实例均为英文。该数据集均为测试集。

2）MSRA-TD 500[21]。MSRA-TD 500 包含 500 张图片，其中训练集有 300 张，测试集有 200 张。该数据集包含中文与英文两种类型文本，标注的目标为文本行，标注的类型为旋转矩形。该数据集中的文本具有大小变化剧烈、长宽比变化剧烈、多方向的特点。该数据集中含长文本的图像占 40%，长文本实例占总文本实例的 27.3%。

3）ICDAR 2015[22]。该数据集来自于 ICDAR 2015 鲁棒阅读竞赛。该数据集包含 1 500 张图片，训练集有 1 000 张，剩余的 500 张为测试集。该数据集包含的文本为英文，文本实例标注是英文单词的边界四边形。与 MSRA-TD 500 相比，该数据集的文本同样具有多方向的特点，但大小、长宽比变化相对较小。该数据集中长文本实例只占 1.5%。

4）ICDAR 2013[23]。该数据集一共有 462 张图片，训练集有 229 张，测试集有 233 张。该数据集的文本为英文，对每一个词进行标注，标注类型为轴向矩形。其中长文本实例占 6%。

当一个文本实例的检测结果与真实目标交占比

大于 0.5 时,该检测结果被认为是正确的检测结果,否则为一个错误的检测结果。文本检测的评价指标有 3 个,召回率(r,recall),准确率(p,precision),综合得分(f,f-score),其计算方式为:

$$r = \frac{|\mathrm{TP}|}{|\mathrm{GT}|} \tag{13}$$

$$p = \frac{|\mathrm{TP}|}{|\mathrm{DT}|} \tag{14}$$

$$f = \frac{2 \times r \times p}{p + r} \tag{15}$$

式中:|TP|为正确的检测结果数目;|GT|为真实的文本实例数目;|DT|为检测结果数。

2.2 实验实施细节

由于各个数据集训练集规模都较小,本方法参考了多种文本检测方法[6,8-9,13-14,24-28]通过加入其他数据与仿射变换的方式增大训练数据量,提高模型的泛化能力。HUST-TR 400 数据集是使用文本行标注,与方法[6,13-14,25-27]一样,将该数据集加入到 MSRA-TD 500 训练集中。参考方法[8-9,13-14,24-28],在 ICDAR 2013 训练集中加入其他训练样本,将部分 MLT 数据集加入到 ICDAR 2013 训练集中。本方法使用缩放、旋转、随机截取 3 种方式进行数据扩充。对于 ICDAR 2013 与 ICDAR 2015 训练集,在保持长宽比不变的条件下,图片长边被随机缩放到[640,2 560]之间。然后在[-10°,10°]之间随机旋转图像。最后随机截取 512 ×512 大小的图像块作为训练样本。对于 MSRA-TD 500 数据集,图片长边被随机缩放到为原始长度的[0.5,2.0]倍,最后随机截取1 024 ×512 的图像块作为训练样本。

本方法使用 Adam[29]作为网络训练优化器,学习率设置为 0.000 1。使用多步调整为学习率调整策略,每经过 10 000 次迭代,学习率衰减为原来的 0.94倍。使用在 ImageNet[30]上预训练的 Resnet 50 模型初始化本网络中特征提取部分的模型参数,其余新加入层的参数使用符合均值为 0,方差为 0.01 的高斯分布的随机数进行初始化。

在测试时,将 3 类区域的分类阈值均设置为 0.9。实验的硬件环境是 Intel Core 7700 CPU,16 GB RAM,Nvidia GTX 1080 显卡,操作系统为 Ubuntu 16.04。

2.3 结果及分析

表 1 为各方法在长文本数据集上测试结果。表中各方法均是在 ICDAR 2015 训练集上进行训练,在长文本数据集上进行测试。由于训练集与测试集存在一定差异,所以总体指标数据均不高。但是本方法在准确率与召回率均高于其他方法,综合得分至少高于其他方法 5%。这充分表明了本方法在长文本检测方面的优势。

表 1 各方法在长文本数据集中测试结果比较

Tab. 1 Comparison of performances of different methods on long text dataset %

方法	准确率	召回率	综合得分
EAST[14]	41.56	66.28	51.25
TextBoxes + +[8]	39.33	47.20	42.91
PixelLink[24]	37.22	65.61	47.49
Proposed	**47.98**	**69.24**	**56.68**

图4(a)、(b)分别是一种间接回归方法 TextBox + +[8]与一种直接回归方法 EAST[14]检测一个较长文本的效果示例。图 4(a)中品红色的虚线框为预设框,黄色框为最后检测结果,TextBox + +[8]只能检测到长文本的一部分。图 4(b)中品红色的四边形是 EAST[14]在文本区域中右侧某像素点处预测的检测结果。该点距离文本区域的左侧边界较远,由于该点处的感受野不足导致其预测结果中左侧两个顶点的定位精度非常差,而该点距离右侧边界较近,右侧边界定位较为准确。

图 4(c)为本文方法检测长文本结果,其中黄色框为最后检测结果,绿色、红色及蓝色区域分别为文本左短边区域、右短边区域及文本中间区域,左短边区域内像素点只预测文本左短边的两个顶点,右短边区域内的点只预测右短边的两个顶点。与其他两种方法比较,在预测文本边界框顶点时,本文方法只需要关注文本短边附近一小块区域,而不用关注整个文本区域,对网络的感受野要求较低。所以在检测长文本时,本文方法检测精度要明显优于预测整个文本边界的方法。

(a) TextBox++[8]检测长文本效果

(b) EAST[14]检测长文本效果

(c) 本文方法检测长文本效果

图 4 长文本测试结果比较示例

Fig. 4 Illustration of comparison of long text test results

表 2 为各方法在 MSRA-TD 500 数据集上测试结果，其中其他方法的结果来自各自的文献。本方法在 MSRA-TD 500 测试集上分别使用了单尺度与多尺度图像进行测试，单尺度图像长宽被缩放为原图像的 0.6 倍，多尺度图像分别被缩放为原来的 0.25、0.50、1.00 倍。表 2 中一些方法的准确率高于本方法的原因在于它们牺牲了一定的召回率。本方法最高综合得分为 82.66%，高于文献[27]中的 1%。MSRA-TD 500 数据集检测目标是文本行，其中含有许多长文本。表 2 的结果再次表明了本方法在长文本检测方面的有效性。

表 2　各方法在 MSRA-TD 500 数据集中测试结果比较

Tab. 2　Comparison of performances of different methods on MSRA-TD 500　%

方法	准确率	召回率	综合得分
RRPN[6]	82.00	68.00	74.00
DDR[13]	77.00	70.00	74.00
EAST[14]	87.30	67.40	76.10
Seglink[25]	86.00	70.00	77.00
PixelLink[24]	83.00	73.20	77.80
RRD[26]	87.00	73.00	79.00
文献[27]*	**87.60**	76.20	81.50
Proposed	86.30	74.60	80.00
Proposed*	86.63	**79.04**	**82.66**

注：* 为多尺度测试结果。

表 3 所示为各方法在 ICDAR 2015 数据集上的测试结果比较，其中其他方法的结果来自各自的文献。本方法测试图像大小为 1 728 ×972。从表 3 数据可以看到，虽然文献[27]准确率高于本方法，但其召回率较低，所以综合性能落后于本方法。与综合得分为第 2 的方法 RRD[26] 相比，本方法的综合得分为 85.44%，高于其 1.6%。

表 3　各方法在 ICDAR 2015 数据集中测试结果比较

Tab. 3　Comparison of performances of different methods on ICDAR 2015　%

方法	准确率	召回率	综合得分
CTPN[28]	74.00	52.00	61.00
RRPN[6]	84.00	77.00	80.00
R2CNN[9]*	85.62	79.68	82.54
TextBoxes + +[8]*	87.80	78.50	82.90
Seglink[25]	73.10	76.80	75.00
DDR[13]	82.00	80.00	81.00
EAST[14]*	83.27	78.33	80.72
PixelLink[24]	85.50	**82.00**	83.70
RRD[26]	88.00	80.00	83.80
文献[27]	**94.10**	70.70	80.70
Proposed	89.42	81.80	**85.44**

注：* 为多尺度测试结果。

表 4 所示为多种方法在 ICDAR 2013 数据集上的测试结果，其中其他方法的结果来自各自的文献。本方法在测试之前，将一些过大的图像缩小为原来的 0.5 倍。本方法单尺度测试图像大小基本为原图像大小。而多尺度测试时，对于较小的图像，所使用的尺度为 0.5、1.0、2.0，对于较大的图像，所使用的尺度为 0.25、0.50、1.00。不同于表 4 中一些方法，本方法能在获得较高准确率的同时，获得高召回率，所以本方法综合得分能达到 90.1%，超过了表 4 中其他所有方法。

表 4　各方法在 ICDAR 2013 数据集中测试结果比较

Tab. 4　Comparison of performances of different methods on ICDAR 2013　%

方法	准确率	召回率	综合得分
CTPN[28]	93.00	83.00	88.00
R2CNN[9]*	**93.55**	82.59	87.73
TextBoxes + +[8]	88.00	74.00	81.00
TextBoxes + +[8]*	91.20	84.40	87.60
Seglink[25]	87.70	83.00	85.30
DDR[13]	92.00	81.00	86.00
EAST[14]*	92.98	82.98	87.69
PixelLink[24]	86.40	83.60	84.50
PixelLink[24]*	88.60	87.50	88.10
RRD[26]	88.00	75.00	81.00
RRD[26]*	92.00	86.00	89.00
文献[27]	93.30	79.70	85.80
文献[27]*	92.00	84.40	88.00
Proposed	89.47	85.49	87.38
Proposed*	90.43	**89.77**	**90.10**

注：* 为多尺度测试结果。

ICDAR 2015 数据集、ICDAR 2013 数据集的检测目标为词，长文本实例数目不多。相比于文本行，词相对较短，而本方法在这两个数据集上的效果依然超过了目前绝大部分方法。原因在于：1）词通常是以多个密集出现，短边区域能够将密集的文本实例分离开，缩小的中间区域能防止相邻的文本实例误连接；2）不再直接预测整个文本边界，短边区域内的像素点只预测与之邻近的短边的顶点，这样的任务相对更简单，所以能更精准地预测文本边界顶点。

图 5 为本方法在各个数据集上的单尺度测试的一些结果样例。1 ~ 4 行分别为长文本数据集、MSRA-TD 500 数据集、ICDAR 2015、ICDAR 2013 数据集测试样例结果。

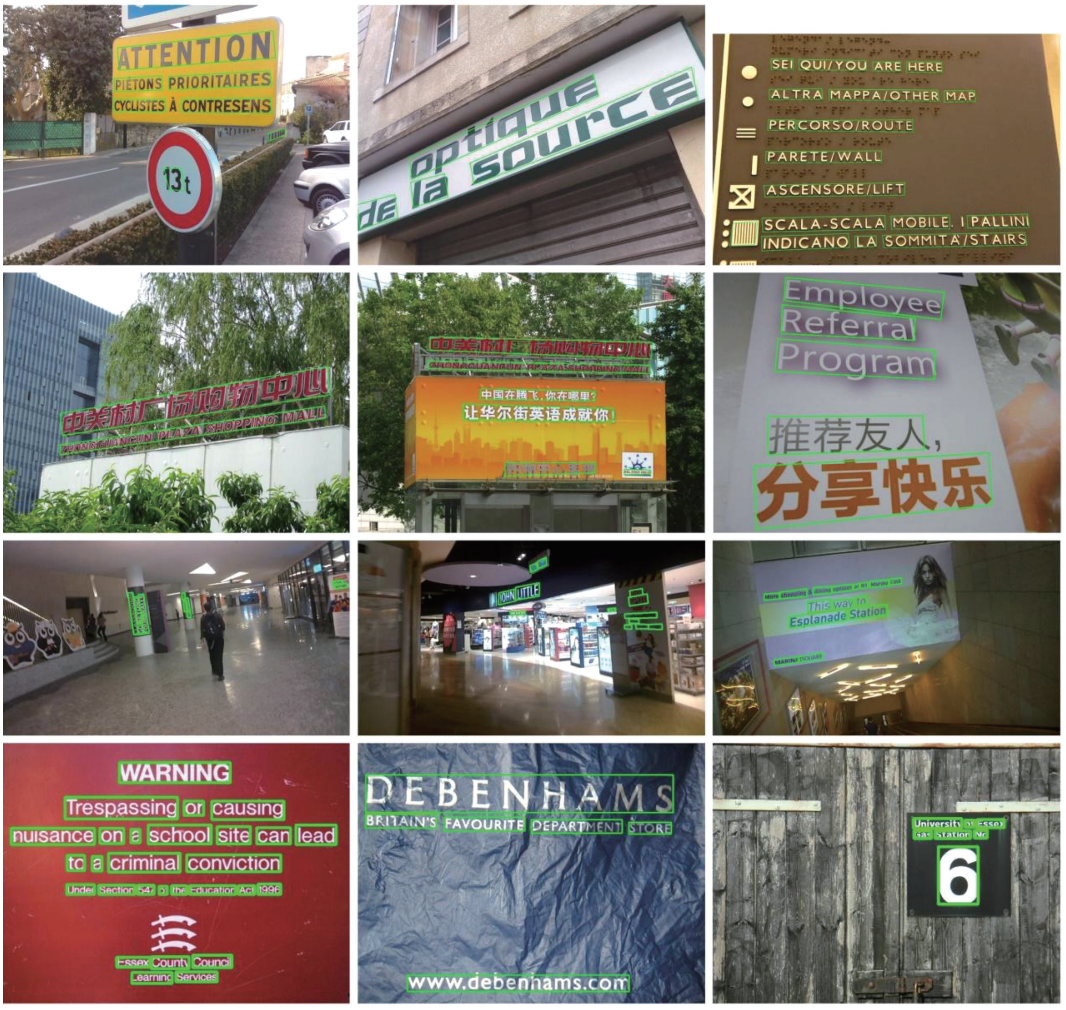


图 5 本方法的测试结果样例

Fig. 5 Samples of test results of the proposed method

2.4 速度比较

表 5 所示为各方法运行速度测试结果。基本上所有基于深度学习的检测方法测试过程都可分为两阶段,网络前向推理阶段与后处理阶段,其中网络前向推理阶段占大部分时间开销,测试图像的大小对速度有直接的影响。各方法测试时,图像大小与实验设备平台不一样。

表 5 各方法速度比较

Tab. 5 Speed comparison of different methods

方法	测试图像大小	GPU 设备	帧率/(帧·s^{-1})
R2CNN[9]	1 280×720	K80	2.2
TextBoxes + +[8]	1 024×1 024	Titan XP	11.6
Seglink[25]	768×768	Titan XP	8.9
EAST[14]	1 280×720	Titan X	13.2
文献[27]	1 280×768	Titan XP	3.6
PixelLink[24]	1 280×768	Titan X	7.3
RRD[26]	1 024×1 024	Titan XP	6.5
Proposed	1 280×704	GTX 1080	13.0

表 5 列出了每种方法测试的图像大小与使用的 GPU。在测试图像大小相近的条件下,EAST[14] 只比本方法稍快一点,然而其测试所用 GPU 设备性能要大大强于本方法。本方法能够如此快速,原因在于:1)本方法网络为单阶段的全卷积网络;2)网络输出结果边长为原图的 1/4,这不仅减少了特征融合部分的卷积运算量,还降低了后处理的运算量。

3 结 论

1)针对长文本难以有效检测的问题,本文提出了一种全新的短边顶点回归网络。本方法分割出文本的中间区域、左短边区域、右短边区域,左、右短边区域的点预测各自短边的顶点,再利用区域的相邻关系将两种短边区域连接组合起来,便可得到精确完整的文本检测结果。

2)在长文本数据集,MSRA-TD 500,ICDAR 2015 及 ICDAR 2013 文本检测数据集上的实验测试结果表明本方法高速有效。

3)本方法目前主要适用于直线文本,在未来的

工作中，将研究如何改善本方法使其具有更强的泛化能力。

参考文献

[1] LIU Li, OUYANG Wanli, WANG Xiaogang, et al. Deep learning for generic object detection: A survey [J]. International Journal of Computer Vision, 2020, 128: 261. DOI: 10.1007/s11263-019-01247-4

[2] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137. DOI:10.1109/TPAMI.2016.2577031

[3] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21. DOI:10.1007/978-3-319-46448-0_2

[4] LIAO Minghui, SHI Baoguang, BAI Xiang, et al. Textboxes: A fast text detector with a single deep neural network[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 4161

[5] ZHONG Zhuoyao, JIN Lianwen, ZHANG Shuangping, et al. Deeptext: A unified framework for text proposal generation and text detection in natural images[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. DOI: 10.1109/ICASSP.2017.7952348

[6] MA Jianqi, SHAO Weiyuan, YE Hao, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111. DOI:10.1109/TMM.2018.2818020

[7] LIU Yuliang, JIN Lianwen. Deep matching prior network: Toward tighter multi-oriented text detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 1962. DOI:10.1109/CVPR.2017.368

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

[9] JIANG Yingying, ZHU Xiangyu, WANG Xiaobing, et al. R2CNN: Rotational region CNN for orientation robust scene text detection [EB/OL]. (2015-06-30). https://arxiv.org/abs/1706.09579

[10] ZHU Yixing, DU Jun. Sliding line point regression for shape robust scene text detection[C]//Proceedings of the 24th International Conference on Pattern Recognition. Beijing: IEEE, 2018: 3735. DOI: 10.1109/ICPR.2018.8545067

[11] HUANG Lichao, YANG Yi, DENG Yafeng, et al. Densebox: Unifying landmark localization with end to end object detection [EB/OL]. (2015-09-19). https://arxiv.org/abs/1509.04874

[12] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4): 640. DOI:10.1109/TPAMI.2016.2572683

[13] HE Wenhao, ZHANG Xuyao, YIN Fei, et al. Deep direct regression for multi-oriented scene text detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 745. DOI:10.1109/ICCV.2017.87

[14] ZHOU Xinyu, YAO Cong, WEN He, et al. EAST: An efficient and accurate scene text detector[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2642. DOI:10.1109/CVPR.2017.283f

[15] XUE Chuhui, LU Shijian, Zhan Fangneng. Accurate scene text detection through border semantics awareness and bootstrapping [C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2018: 355. DOI: 10.1007/978-3-030-01270-0_22

[16] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770. DOI:10.1109/CVPR.2016.90

[17] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2015-04-10). https://arxiv.org/abs/1409.1556

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234. DOI: 10.1007/978-3-662-54345-0_3

[19] MILLETARI F, NAVAB N, AHMADI S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//Proceedings of the 4th International Conference on 3D Vision. Stanford: IEEE, 2016: 565. DOI: 10.1109/3DV.2016.79

[20] NAYEF N, YIN Fei, BIZID I, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT [C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto: IEEE, 2017: 1454. DOI:10.1109/ICDAR.2017.237

[21] YAO Cong, BAI Xiang, LIU Wenyu, et al. Detecting texts of arbitrary orientations in natural images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE, 2012: 1083. DOI: 10.1109/CVPR.2012.6247787

[22] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on robust reading[C]//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis: IEEE, 2015: 1156. DOI:10.1109/ICDAR.2015.7333942

[23] PRATIKAKIS I, GATOS B, NTIROGIANNIS K. ICDAR 2013 document image binarization contest (DIBCO 2013) [C]// Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington DC: IEEE, 2013: 1471. DOI: 10.1109/ICDAR.2013.219

[24] DENG Dan, LIU Haifeng, LI Xuelong, et al. Pixellink: Detecting scene text via instance segmentation[C]//Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI, 2018: 6773

[25] SHI Baoguang, BAI Xiang, 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: 3482. DOI:10.1109/CVPR.2017.371

[26] LIAO Minghui, ZHU Zhen, SHI Baoguang, et al. Rotation-sensitive regression for oriented scene text detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5909. DOI:10.1109/CVPR.2018.00619

[27] LYU P, YAO Cong, WU Wenao, et al. Multi-oriented scene text detection via corner localization and region segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7553. DOI:10.1109/CVPR.2018.00788

[28] TIAN Zhi, HUANG Weilin, HE Tong, et al. Detecting text in natural image with connectionist text proposal network[C]// European Conference on Computer Vision. Cham: Springer, 2016: 56. DOI:10.1007/978-3-319-46484-8_4

[29] KINGMA D P, BA J L. Adam: A method for stochastic optimization [EB/OL]. (2014-12-22). https://arxiv.org/abs/1412.6980

[30] DENG Jia, DONG Wei, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248. DOI:10.1109/CVPR.2009.5206848

（编辑　张　红）