

引用格式:刘春,田倬韬,刘绍辉,等. 一种改进的图像中文本检测模型 [J]. 微电子学与计算机,2020,37(6):83-88. [LIU C, TIAN Z T, LIU S H, et al. An improved text detection model in image [J]. Microelectronics & Computer, 2020, 37(6): 83-88.]

一种改进的图像中的文本检测模型

刘 春¹, 田倬韬², 刘绍辉², 安 源¹

(1 大庆师范学院, 黑龙江 大庆 163712; 2 哈尔滨工业大学, 黑龙江 哈尔滨 163712)

摘 要: 本文提出了一种基于深度神经网络的高效自然场景文本检测模型 IEAST (Improved EAST). 该模型基于 EAST (An Efficient and Accurate Scene Text Detector) 进行了一系列改进, 同时结合了物体检测以及实例分割的思想, 通过用 Inception Module 替换原本特征融合的方式、加入 Path Aggregation 的思想以及采用 Fuzzy Mask 来提升模型的整体性能. IEAST 是一段式模型, 其不仅检测准度上不落后于绝大多数二段式 State-of-the-art 的文本检测模型, 还能保证极快的检测速度.

关键词: 深度神经网络; 文本检测; 实例分割

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2020)06-0083-06

An improved text detection model in image

LIU Chun¹, TIAN Zhuo-tao², LIU Shao-hui², AN Yuan¹

(1 Daqing Normal University, Daqing 163712, China; 2 Harbin Institute of Technology, Harbin 150001, China)

Abstract: Image Text Detection is the process of accurately locating text area in given images, which is the important prerequisite of OCR (Optical Character Recognition). This paper introduces Improved IEAST (Improved EAST) which is an efficient and accurate scene text detection model based on deep neural network. This model which combines the knowledge of both object detection and instance segmentation is a variant of EAST (An Efficient and Accurate Scene Text Detector) with several improvements. By replacing conv3x3 feature merging with inception module, adding the idea of Path Aggregation and adopting Fuzzy Mask during training, IEAST gains improvements in performance with little compensation. Though IEAST is a one-stage model, it outperforms most of state-of-the-art two-stage text detection models by not only maintaining comparable accuracy but also inferencing faster. IEAST shows good result on ICDAR (International Conference on Document Analysis and Recognition) 2015 dataset

Key words: deep Learning; text detection; image processing

1 引言

自然场景下的文本检测是计算机视觉乃至人工智能领域的一个研究热点, 在汽车自动驾驶导航、图文匹配、图像及视频的语义提取和目标定位等方面有着广泛的应用前景.

自然场景下的文本检测不同于规范化文档的文

本检测(如文档、名片以及身份证等证件), 规范化文档的文本排列整齐、且有规律可循, 一张图片当中的文本字体及风格都十分相近; 而自然场景下的文本更加多变, 角度, 长短, 噪声, 对比度各不相同, 这诸多因素给自然场景下的文本检测带来了极大的困难. 因此自然场景下文本检测模型的研究具有重要的理论意义和实际应用价值.

收稿日期: 2019-10-15; **修回日期:** 2019-11-08

基金项目: 国家重点研发项目(YFC0806802, 2018YFC0832105 资助; 教育部科技发展中心高校产学研创新基金项目(2018A01010); 大庆市指导性科技计划项目(zd-2019-67); 黑龙江省教育科学“十三五”规划项目(GJB1319002)

文本检测可以分为传统文本检测方法和基于神经网络的方法两大类. 在神经网络以及深度网络流行之前, 研究人员们通过分析图片中文本区域的局部特征与周围特征的区别来提取出特定的文本区域, 如基于 Maximal Stable Extremal Regions (MS-ER) 的方法^[1] 以及 Stroke Width Transform (SWT)^[2]. MSER 是基于联通区域进行文本区域的筛选, SWT 是基于图片当中文本边缘特征的分析检测, 通过分析文本区域的边缘特征加以后处理来提取出特定的文本区域. 但基于 MSER 以及 SWT 的方法都是严重依赖手工设计的特征来进行文本检测, 这样的方法能适用于简单的场景, 但是无法应用于自然场景下的复杂情况. 而基于神经网络的方法, 则将检测过程视为一个黑盒, 通过网络来对其映射进行模拟, 经过充分训练之后就能够表达出极其复杂的映射关系, 因而神经网络具有推广、适应、学习以及概括能力, 适当的学习能够推广神经网络, 同时能够增强神经网络的概括能力.

按照一般的基于深度学习的物体检测网络的结构, 先提取候选区域, 然后对候选区域通过深度网络来进行特征提取并最后分类. 因此, 很自然的, 可以使用传统的 SWT 或 MSER 提取候选区域, 然后再采用深度学习网络进行特征提取与分类. 但显然, 这种滑动窗口的方法由于待检测区域的尺寸变化, 需要穷尽所有可能窗口, 导致计算量巨大. 因此, 替代对每一个候选窗口处理, 从整图进行特征提取, 然后再进行处理. 这种方法具体可以分为 anchor based 以及 anchor free 两类.

Anchor based 方法^[3-5] 主要受到来自物体检测 RCNN, Fast RCNN, Faster RCNN, Mask RCNN 系列 two-stage 以及 SSD 系列 one-stage 方法的影响, 通过设计不同尺寸以及长宽比的 anchor 来初步给定大量待筛选的 proposal, 然后再通过对 proposal 的筛选来得到需要的目标检测框. 而一阶段法以及两阶段方法的区别就在于通过 RPN (Region Proposal Network) 产生的 proposal 是否经过一个过滤网络进行筛选.

尽管上述基于 anchor 方法在物体检测领域已经取得巨大进展, 但是在文本领域效果并不是最理想. 主要原因在于物体检测的界限框一般都是水平检测框, 而文字不仅仅是水平问题, 此时导致 anchor 的产生非常复杂, 并导致过量的候选区域. 因此无 Anchor 方法成为首选, 其典型代表就是 EAST^[6], EAST 利用类似一个 FPN^[7] (Feature

Pyramid Network) 的一段式结构来融合各个层次的特征, 因而得到不同尺度的信息, 最终得到像素级的预测, 预测包括每个点的 text/non-text 评分以及边框距离和角度, 最后再通过简单的算法构建出检测框, 其特点就是在保证一定精准度的同时也保持其检测速度. 同样, 无 anchor 方法中也有基于 two-stage 的方法, 如 AF-RPN^[8], 但该模型并不容易进行训练, 需要进行大量的调参工作. 本文在 EAST 的基础上, 提出了改进的 EAST 检测模型, 后续简称为 IEAST.

2 改进的文本检测模型

由于 PANet 的结构能够缩短信息传递距离, 且能有效地通过 Aggregation 整合数据, 故 IEAST 模型在 EAST 的基础上综合考虑了 PANet 的思想, 将 PANet 中 Augmented Bottom-up Structure 部分改进并结合到 EAST 中, 但 IEAST 最终没有采用 Adaptive Feature Pooling, 主要考虑是 IEAST 没有采用 anchor 机制, 而是直接同 EAST 一样使用的像素级文本/非文本分类以及直接回归预测检测框来完成最终的检测任务. 下面首先分别介绍 EAST 和 PANet 模型, 然后详细介绍所提出的 IEAST 模型.

2.1 EAST 模型和 PANet 模型

EAST 模型结构如图 1 所示. EAST 由三个部分组成: 特征提取主干 (Feature extractor stem)、特征合并分支 (Feature-merging branch) 以及输出层 (Output layer). 特征提取主干和特征合并分支的结构类似于 FPN, 将主干各个层级的特征拿出来做一次合成, 使得最终特征能够包含所有层级的特征. 具体描述为, 从 conv stage 1 到 conv stage 4 的特征图 (feature map) 的尺寸越来越小, 尽管 conv stage 4 的特征图是 conv stage 1 的 $1/8$, 但是 conv stage 4 特征图中的每一个元素在输入图片上的感受野是 conv stage 1 的 8 倍, 所以 conv stage 4 特征图当中的信息更具有宏观信息, 而 conv stage 1 特征图中像素的信息更具有细粒度的信息, conv stage 2 和 conv stage 3 也有类似的关系. 所以如果仅利用 conv stage 4 的特征来做最后的预测, 最后的结果则仅依赖于宏观信息而丢失了细节信息, 因而出现漏检小目标以及边界模糊的问题. EAST 借助 FPN 的思想, 通过 unpool (反池化) 和在最后一个通道 (channel) 的 concat (拼接) 将不同层级信息融合最终在特征融合分支的顶端得到一个原图 $1/4$ 的特征图, 最后进行 3 个简单的 1×1 卷积得到我们所需要的文本/非文本分类以及检测框回归的输

出. 假设 RBOX 代表矩形, 所以需要输出每个点到四条边的距离 ($1 \times 1, 4$) 以及一个角度 ($1 \times 1, 1$), 而 QUAD 表示通用的四边形, 则直接预测包含这一个点的一个四边形的坐标 $x_0, y_0, x_1, y_1, x_2, y_2, x_3, y_3$ ($1 \times 1, 8$).

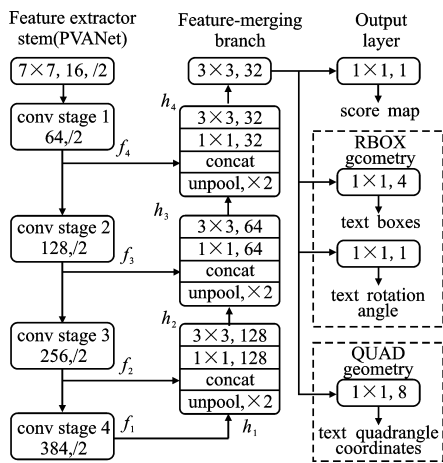


图1 EAST 的结构

PANet 的设计核心就是缩短不同层级特征之间的距离, 以达到信息有效传递的目的. PANet 的结构如图 2 所示, PANet(a) 部分还是基于 FPN 设计, 通过提取不同层次的主干网络特征来进一步融合最后作出预测, 这样能够提取到各个维度层级的特征, 从而能够适应大小多变的物体分割和检测任务. FPN 的一个不足之处在于信息传递的效率不够高, 信息从输入图片之后的第一个特征图传递特征到 P_5 特征图 (整个结构的最顶层) 的过程可以表示为红色虚线, 在以前的 FPN 当中信息需要整体经过一个主干特征提取网络才能到达 P_5 , 如果主干特征提取网络为 Resnet-101, 那么这一趟信息传递至少需要经过 100 层才能到达, 无论是前向还是反向传递的过程中, 信息都会有相对较多的损失. 而 PANet 则通过加上 Augmented Bottom-up 结构给信息传递添加了一个捷径 (shortcut), 使得上述信息传递过程可以由绿色虚线表示, 输入图片后的第一层特征图的信息则可以直接跳过复杂的主干特征提取网络到达 P_2 , 再在 (b) 中经过几次简单的卷积操作就能够到达 N_5 最后作出预测, 这整个过程在 PANet 当中只需要经过不到 10 层, 同时也能保证信息准确. 所以有效的信息传递方式是 PANet 成为 state-of-the-art 模型的关键要素之一.

另外, PANet 表明尽管小的 proposal 被分配到层次较低 feature map 上, 但是层次较高的 feature map 同样能够提供全局特征, 为 proposal 的分割以

及检测提供额外的信息. 实验进一步表明将各个层级特征融合有助于最后的预测.

因此, 我们提出的模型采用 PANet 与 EAST 结合来改进文字检测性能.

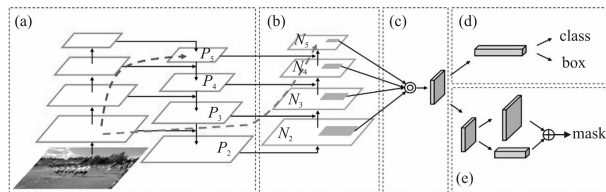


图2 PANet 的结构

2.2 IEAST 模型

IEAST 结构如图 3 所示, 可以看作是 EAST 与 PANet 的结合, 整体是基于 FPN 的思想构建. 首先图片输入之后进入主干特征提取网络 (采用的主干特征提取网络为 resnet-50 v1), 经过一个 7×7 卷积的之后, 再依次经过四层的特征提取, 其中每一层特征图是上一层的二分之一, 主干特征提取网络的顶层特征图则等于输入图片的 $1/32$ ($1/4 * 1/2 * 1/2 * 1/2 = 1/32$). 然后经过反复地上采样 (Unpool) 以及拼接 (Concat) 得到初步特征融合分支 (图 3 左起第二个部分), 再通过卷积降采样以及相加的过程, 将初步特征融合分支的各部分融合得到综合特征融合分支 (图 3 左起第三个部分), 最后将综合特征融合分支经过上采样相加后得到预测结果. 下面介绍各个部分的具体实现.

主干特征提取网络 (Feature Extractor Network, 后称为 stem) 是模型的重要组成部分, 模型后续所应用到的特征全部都来源于 stem, 因而一个性能较佳的 stem 是构建一个检测、识别或者分割模型的关键. 检测、识别以及分割领域今年来的提升离不开 stem 的发展, 从 VGG 到 inception、resnet、mobilenet 等结构的提出, 表明 stem 的巨大潜力.

IEAST 的基本模型基于 Tensorflow Slim 构建, 采用 slim resnet 50 作为其 stem, 在 resnet50 stem 中提取了 conv2_x、conv3_x、conv4_x 以及 conv5_x 作为初步特征融合分支的输入, 所以初步特征融合分支的末端特征图的大小与 conv2_x 相当 (为输入图片的四分之一).

第二个部分是初步特征融合网络, 从主干特征提取网络得到各个层级的特征之后通过一定方式将其融合起来向更上一层传递. 在 EAST 当中, 该部分仅仅是 1×1 卷积降低通道数量后接上一个 3×3 的卷积进行特征提取, 这样做尽管效率很快, 但是其

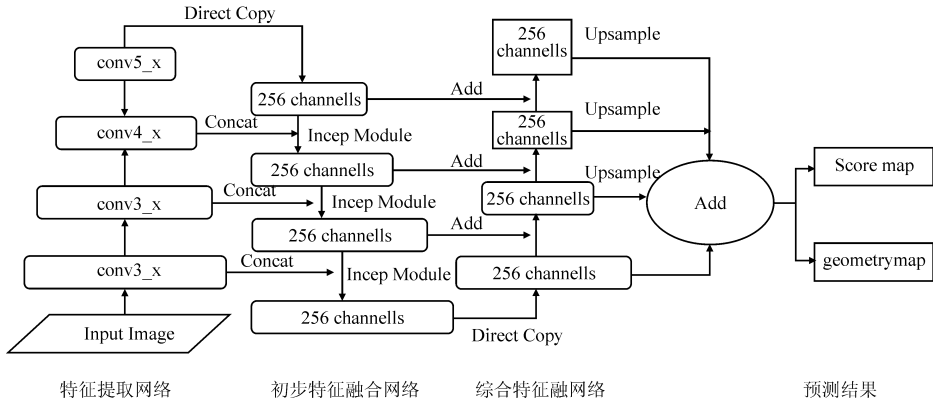


图 3 提出的 IEAST 模型的结构

提取的特征较为粗糙。

受到文献 [9] 中提到 Inception Module 的影响, Inception Module 通过拼接不同卷积操作的结果来提取特征图当中不同形式的特征, 最后进行综合得到提升的结果。通过实验, 在初步特征融合网络的这一部分, IEAST 的特征提取采用文献 [10] 中的 Inception 模块。

在初步特征融合网络融合主干特征提取网络送过来的不同层级的特征图时, 因为 Resnet 50 不同层级的特征图通道数各不相同, 而又不希望通过一个 1×1 卷积降低通道数损失最初提取到的信息, 所以在此使用拼接 (Concat) 的操作将特征融合起来。而在融合之后进入 Inception Module 提取, 我们根据文献 [9] 的设置, 将所有融合特征之后的通道数都设定为 256。

对于 Inception 的位置考虑上, 将原本的 1×1 和 3×3 卷积替换为 Inception Module, 实验结果表明该方案较其余方案性能更优。

IEAST 的损失函数主要采用的 EAST 的损失函数, 损失函数 f 分为两个部分: 1. score map 的损失, 2. Geometry map (像素到检测框四周距离以及角度) 的损失。损失函数 f 可以用下面这个式子表示:

$$L = L_s + numda_g * L_g \quad (1)$$

式中, L_s 表示 score map 的损失; L_g 表示 geometry 的损失; $numda_g$ 表示一个均衡 score map 以及 geometry 二者损失之间的一个参数, IEAST 在这里将 $numda_g$ 设置为了 1, 表明 score map 和 geometry 同样重要。针对 score map 的损失函数, IEAST 采用的是 dice loss, dice loss 是一种广泛应用于物体分割当中的 loss 函数, 在 V-net^[9] 当中被 Fausto Milletari 等人第一次提出来。而 Geometry 的损失函

数包括两个部分, Geometry 的损失可以用下面这个式子来表示, 其中 L_{AABB} 是对四周检测框距离回归的损失, L_{θ} 是对角度预测的损失, $numda$ 表示一个权重系数, 在 IEAST 实现当中 $numda$ 为 20。 $L_g = L_{AABB} + numda * L_{\theta}$ (2) L_{AABB} 利用根据预测出来的 geometry map (点到检测框四边距离) 得到的检测框 R 和 Ground Truth 的检测框 R^* 之间的 IoU 来计算, 如下所示。

$$\begin{aligned} L_{AABB} &= -\log IU(R, R^*) \\ &= -\log \frac{|R * R^*|}{|R + R^*|} \end{aligned} \quad (3)$$

IEAST 输出的仅是 geometry map, 里面包含的信息仅仅是点到四边的距离, R 的计算得利用下面的计算量, d_1 、 d_2 、 d_3 以及 d_4 分别代表一个像素到达该像素所在检测框的上边界、右边界、下边界以及左边界的距离, 所以根据这四个量, 可以得到 R 和 R^* 的交集区域 $|R * R^*|$, w_i 和 h_i 分别为该像素所在检测框的宽度和高度, 其中 i 表示像素坐标。

$$w_i = \min(d_2, d_2^*) + \min(d_4, d_4^*) \quad (4)$$

$$h_i = \min(d_1, d_1^*) + \min(d_3, d_3^*) \quad (5)$$

$|R + R^*|$ 的即可通过 R 和 R^* 以及根据 d 得到的 $|R * R^*|$ 计算得到。

$$|R + R^*| = |R| + |R^*| - |R * R^*| \quad (6)$$

至此, L_{AABB} 即可求出。

L_{θ} 根据下式即可计算得到, 其中 θ 为预测的角度, θ^* 为 Ground Truth 的角度。

$$L_{\theta}(\theta, \theta^*) = 1 - \cos(\theta, -\theta^*) \quad (7)$$

2.3 IEAST 模型的训练

IEAST 基于 Adam Optimizer 进行训练, 因为 IEAST 是 one stage (一段式) 模型, 从而可以进行端到端训练, 中间无需额外的处理过程。

IEAST 采用了 EAST 的训练策略, 从原图当中

随机裁剪 512 * 512 大小的图片作为输入来加速训练,得益于轻量的网络结构,训练过程当中 batch size 为 16,GPU 数量为 2,故有效 batch size 为 32,相较于 Faster RCNN 等模型 batch size 为 2 的训练要更高效.

IEAST 初始学习率设置为 1e-3,采用 exponential decay 的学习衰减策略,decay rate 设置为 0.94,decay step 为 5000(双卡 5000,单卡 10000),最佳结果是通过双卡训练 50000 次得到(取 40000 次左右的储存点). IEAST 在训练过程当中时刻检测 loss 的变化,当 loss 发散到一个较大数值未收敛的时候就会终止训练.

3 实验结果及分析

3.1 数据集

IEAST 的评价主要基于 ICDAR 2015 进行. ICDAR 2015 数据集指的是 ICDAR2015 Robust Reading Competition 的 Challenge 4 中所使用的数据. 该数据集主要包含 1500 张图片,其中 1000 张图用来作为训练数据,另外 500 张图用作测试数据. 在 IEAST 训练过程当中,我们取 1000 张训练数据中的 100 张作为 Validation Set,另外 900 张作为真正的训练数据. ICDAR 2015 数据集的标注具体形式即 $x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4, text$. 这对应于 EAST 当中提到的 QUAD 形式.

3.2 实验结果对比

IEAST 的实验结果将与基于 AF-RPN 的模型、EAST 中介绍的 Corner Based 模型(CVPR 2018)、PixelLink 以及 TextBoxes++ 等相关方法于 ICDAR2015 数据集上做对比,结果如表 1 所示.

表 1 主流方法结果对比(训练数据集为 ICDAR15)					
方法	召回率	精度	F 分数	FPS (720p)	
TextBoxes++	0.767	0.872	0.817	11.6	
EAST PVANET2x	0.735	0.836	0.782	13.2	
CTPNVGG16	0.516	0.742	0.609	7.1	
SeglinkVGG16	0.768	0.731	0.750	—	
Method in	0.707	0.941	0.807	3.6	
PixelLink VGG164s (conv3_3)	0.817	0.829	0.823	7.3	
PixelLink VGG162s (conv2_2)	0.820	0.855	0.837	3.0	
R2CNN	0.797	0.856	0.825	—	
RRD	0.790	0.865	0.822	6.5	
IncepText (Alibaba-PAI)	0.806	0.905	0.853	—	
IEAST(无预训练)	0.805	0.845	0.825	12.5	
IEAST(有预训练)	0.804	0.888	0.844	12.5	

PVANET2x 表示在原 PVANET 上面将通道数变成了原来的两倍. PixelLink 的 2s 以及 4s 分别表示预测结果是从 conv2_2 以及 conv3_3 得到.

IEAST 表示提出的方法,有预训练版本是经过 COCO text 2014 的 28562 张图预训练(单卡 120k iterations,双卡 60k iterations)后再由 ICDAR 2015 的 1000 张训练图片训练而来(单卡 100k iterations,双卡 50k iterations). 表中 IEAST 的 FPS 是基于 NVIDIA Tesla P40 测量得到. 部分预测结果如图 4 所示.

可以发现,IEAST 在保持相对较佳的识别精度的还能保持不错的识别速率,是一个快速且精确的文本检测模型.

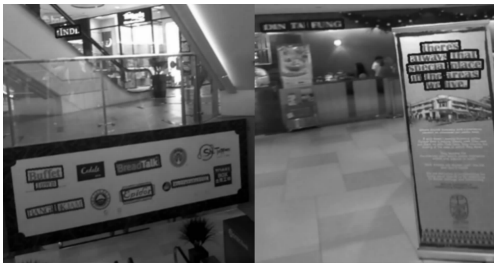


图 4 IEAST 的部分检测结果

4 结束语

IEAST 模型通过能够直接预测出文本\非文本区域并且通过直接回归得到相应的文本检测框,通过用 Inception Module 替换原本特征融合的方式、加入 Path Aggregation 的思想以及采用 Fuzzy Mask 来提升模型的整体性能. 同时加以对目前业界领先的结构的讨论,最终得到的模型在能够保证较优秀的检测准度的同时还能保证极快的检测速度,亦达到业界领先的水准.

此外,该模型普适性以及移植性较高,没有针对文本检测进行过多的修改(只有 Fuzzy Mask 是针对文本进行的操作),所以经过简单修改亦可用于其他物体检测,不过适用于体积相对较小物体,例如行人检测或者自然场景下的大规模人脸检测等. 若实际需求对检测精度的要求较高,IEAST 亦可作为两段式模型的一个 anchor free RPN 提供一个精准的 proposal 预测以用来后期处理.

尽管 IEAST 是一个精准且高效的文本检测模型,但在针对特定场景的应用中,仍有改进的空间,在后面的工作中我们计划对其进行进一步的优化和

性能提升.

参考文献:

[1] NEUMANN L, MATAS J. A method for text localization and recognition in real-world images [C]. In Proc. of ACCV, German: Springer, 2010: 770-783.

[2] VENKATESWARLU K, SREERAMA M V. Text detection on scene images using msr [J]. International Journal of Research in Computer And Communication Technology, 2015,4(7):452-456.

[3] ZHONG Z Y,JIN L W, HUANG S P. DeepText: A new approach for proposal generation and text detection in natural images[C]. In ICASSP, LA: IEEE, 2017:1208-1212.

[4] LIAO M H, SHI B G, BAI X et al. TextBoxes: A fast text detector with a single deep neural network[C]. In AAAI, California: AAAI Press, 2016:4164-4167.

[5] MA J Q, SHAO W Y, YE H et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Trans. Multimedia, 2017, 20(11):3111-3122.

[6] ZHOU X Y, YAO C, WEN H et al. EAST: an efficient and accurate scene text detector[C]. In Proc. of CVPR, Honolulu: IEEE, 2017: 2642-2651.

[7] LIN T Y, DOLLAR P, GIRSHIK R B et al. feature

pyramid networks for object detection[C]. In Proc. Of CVPR, Honolulu: IEEE, 2017:936-944.

[8] ZHONG Z Y, SUN L, HUO Q et al. An anchor-free region proposal network for faster r-cnn based text detection approaches[J]. IJDAR 2019, 22(3):315-327.

[9] MILLETARI F, NAVAB N, AHMADI S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C]. In Proc. of 3DV, USA CA: IEEE, 2016: 565-571.

[10] LIU S, QI L, QIN H F et al. Path aggregation network for instance segmentation [C]. In Proc. of CVPR, USA UT: IEEE, 2018:8759-8768.

作者简介:

刘 春 女,(1970-),硕士,教授. 研究方向为数据库、云计算、软件工程.

田倬韬 男,(1996-),硕士研究生. 研究方向为人工智能,深度学习,计算机视觉.

刘绍辉(通讯作者) 男,(1977-),博士,副教授,硕士生导师. 研究方向为计算机视觉、图像处理、目标跟踪.

E-mail:shliu@hit.edu.cn.

安 源 男,(1982-),硕士,讲师. 研究方向为图像处理,软件工程.