DOI:10.3969/j.issn.1671-024x.2024.04.009

基于链接关系预测的弯曲密集型商品文本检测

耿 磊[1,2],李嘉琛[2,3],刘彦北[1,2],李月龙[4],李晓捷[1]
(1.天津工业大学 生命科学学院,天津 300387;2.天津工业大学 天津市光电探测技术与系统重点实验室,天津 300387;3.天津工业大学 电子与信息工程学院,天津 300387;4.天津工业大学 计算机科学与技术学院,天津 300387)

摘 要: 针对商品包装文本检测任务中弯曲密集型文本导致的错检、漏检问题,提出了一种由2个子网络组成的基于链接关系预测的文本检测框架(text detection network based on relational prediction, RPTNet)。在文本组件检测网络中,下采样采用卷积神经网络和自注意力并行的双分支结构提取局部和全局特征,并加入空洞特征增强模块(DFM)减少深层特征图在降维过程中信息的丢失;上采样采用特征金字塔与多级注意力融合模块(MAFM)相结合的方式进行多级特征融合以增强文本特征间的潜在联系,通过文本检测器从上采样输出的特征图中检测文本组件;在链接关系预测网络中,采用基于图卷积网络的关系推理框架预测文本组件间的深层相似度,采用双向长短时记忆网络将文本组件聚合为文本实例。为验证RRNet的检测性能,构建了一个由商品包装图片组成的文本检测数据集(text detection dataset composed of commodity packaging, CPTD1500)。实验结果表明:RPTNet不仅在公开文本数据集CTW-1500和Total-Text上取得了优异的性能,而且在CPTD1500数据集上的召回率和*F*值分别达到了85.4%和87.5%,均优于当前主流算法。

关键词: 文本检测;卷积神经网络;自注意力;特征融合;图卷积网络;双向长短时记忆网络
中图分类号: TP183 **文献标志码:** A **文章编号:** 1671-024X(2024)04-0050-11

Text detection of curved and dense products based on link relationship prediction

GENG Lei[1,2], LI Jiachen[2,3], LIU Yanbei[1,2], LI Yuelong[4], LI Xiaojie[1]
(1. School of Life Sciences, Tiangong University, Tianjin 300387, China; 2. Tianjin Key Laboratory of Optoelectronic Detection Technology, Tiangong University, Tianjin 300387, China; 3. School of Electronics and Information Engineering, Tiangong University, Tianjin 300387, China; 4. School of Computer Science and Technology, Tiangong University, Tianjin 300387, China)

Abstract: A detection framework consisting of two sub-networks, text detection network based on relational prediction (RPTNet) is proposed to solve the problem of error detection caused by curved and dense texts in the text detection task of commodity packaging images. In the text component detection network, local and global features are extracted using a parallel downsampling structure of convolutional neural network and self-attention. A dilated feature enhancement module (DFM) is added to the downsampling structure to reduce the information loss of the deep feature maps. The feature pyramid network is combined with the multi-level attention fusion module (MAFM) in upsampling structure to enhance the connections between different features and the text detector detects the text components from the upsampled feature maps. In the link relational prediction network, a relational reasoning framework based on graph convolutional network is used to predict the deep similarity between the text component and its neighbors, and a bi-directional long short-term memory network is used to aggregate the text components into text instances. In order to verify the detection performance of RPTNet, a text detection dataset CPTD1500 composed of commodity packaging images is constructed. The test results show that the effectiveness of the proposed RPTNet is verified by two publicly available text datasets, CTW-1500 and Total-Text. And the recall and *F* value of RPTNet on CPTD1500 are 85.4% and 87.5%, respectively, which are superior to current

收稿日期: 2022-12-01
基金项目: 国家自然科学基金资助项目(61771340);天津市科技计划资助项目(20YDTPJC00110)
第一作者: 耿 磊(1982—),男,博士,教授,主要研究方向为计算机视觉、机器学习等。E-mail:genglei@tiangong.edu.cn
通信作者: 刘彦北(1986—),男,博士,副教授,主要研究方向为机器学习、数据挖掘等。E-mail:liuyanbei@tiangong.edu.cn

mainstream algorithms.

Key words: text detection; convolutional neural network; self-attention; feature fusion; graph convolutional network; bi-directional long short-term memory network

由于场景文本检测具有较高的应用价值和广阔的研究前景,近年来人们对其关注度越来越高。随着深度学习的快速发展,人们对于具有线性、低密度的文本实例检测已经实现了优异的检测效果[1-4]。但自然场景下的文本存在尺寸、形状、密度、字体、透视等方面的多样性,这导致在处理不规则文本实例时,传统的检测算法很难对其几何属性做出精确的判断,无法达到预期的检测效果。近年来,尝试解决这类问题的方法大致可以分为基于回归的方法和基于分割的方法。

基于回归的文本检测方法通常依赖于一般的物体检测框架,如 Faster R-CNN[5]和 SSD[6]等。根据不同文字区域各自的特点,研究者在普通物体检测方法的基础上做了相应的修改,以此解决文本检测中出现的问题。TextBoxes++[7]通过对 TextBoxes[1]做出改进,即通过回归四边形而不是水平边界框来实现多方向文本的检测。Raisi 等[8]用旋转文本表征的方法优化了 DETR[9]的架构,可以更好的表示多方向文本区域。总体而言,上述方法对于倾斜角度较小的多方向文本检测效果优异,但由于矩形或四边形边界框不能足够紧密的包围弯曲文本,故这些方法不能很好地检测弯曲文本。为了更好地适应任意形状文本的检测任务,LOMO[10]利用 Mask-RCNN 作为其基础框架,并引入迭代细化和形状表达模块来细化不规则文本区域的边界框,从而发挥了基于分割和回归的架构优势。MOST[11]用文本特征对齐模块(TFAM)完善了 LOMO 的架构,通过可变形卷积算子进行定位细化,实现了更高的精确率。FCENet[12]首先预测文本实例的紧凑傅里叶特征,然后采用反傅里叶变换(IFT)和非最大抑制(NMS)来重建任意形状文本实例轮廓。

基于分割的文本检测方法通常首先检测文本组件,然后再将这些文本组件组合成文本实例。近些年来,基于分割的方法在处理任意文本检测问题中被越来越多的研究者采用,根据单元表征的不同,此类方法可分为像素级方法和片段级方法。其中像素级方法通常将文本检测问题作为语义分割或实例分割问题,以全卷积神经网络(FCN)[13]作为框架来预测图片的像素级别的分类图,然后用不同的方法将这些像素组合成文本区域。Zhang 等[14]采用 FCN 预测文本块,然后通过 MSER 提取候选字符,最后使用分组策略来达到多方向文本检测的目的。TextField[15]可以学习到一个深度方向场,此方向场与相邻像素相连接,生成候选文本部分,学习到的方向信息将文本部分分组为文本实例。片段级方法首先检测包含一部分单词或者字符的文本片段,然后将同属于一个文本区域的文本片段组合在一起。PSENet[16]用核去检测每个文本实例,并通过渐进尺度扩展算法去逐渐扩展预定义的核,从而获得最终的检测结果。在 CRAFT[17]中,用亲和力判断相邻的字符之间是否属于同一个文本实例,通过估计字符和字符间的亲和力来检测任意形状文本。Seglink++[18]可以学习文本组件之间的吸引力和排斥力联系,对最小生成树算法改进后,通过实例感知组件,分组检测任意形状文本。DB[19]在分割网络中进行了自适应二值化处理,简化了后处理并提高了检测性能。然而,上述方法往往无法精确分离图像中密集相邻的文本实例,而且检测到的文本轮廓通常包含缺陷和噪声。这是因为现有的基于分割方法的性能在很大程度上依赖于轮廓检测框架的准确性,而忽略了轮廓的自适应调整。

针对上述问题,本文提出了基于链接关系预测的文本检测框架 RPTNet。首先通过文本特征并行采样与多尺度特征融合相结合的方式,解决密集型文本实例间因特征信息提取不充分导致的粘连问题,同时受到 Wang 等[20]在人脸图像聚类工作的启发,通过图来表示非欧几里得数据,使用图卷积网络(Graph Convolutional Network, GCN)推理文本组件间的深度链接关系。通过双向长短时记忆网络(Bi-directional Long Short-Term Memory Network, BiLSTM)[21],根据推理结果将文本组件自适应聚合为文本实例,从而实现了商品包装图像中弯曲密集型文本的精准检测。为了证明 RPTNet 在检测弯曲密集型商品外包装文本实例的有效性,建立了一个由商品包装图片组成的包含大量弯曲密集型文本的文本检测数据集 CPTD1500。实验证明,RPTNet 在 CPTD1500 数据集和公开的曲面文本检测数据集 CTW-1500[22]及 Total-Text[23]上取得了优异的检测效果。

1　研究方法

1.1　整体网络架构

RPTNet 的整体结构如图 1 所示。

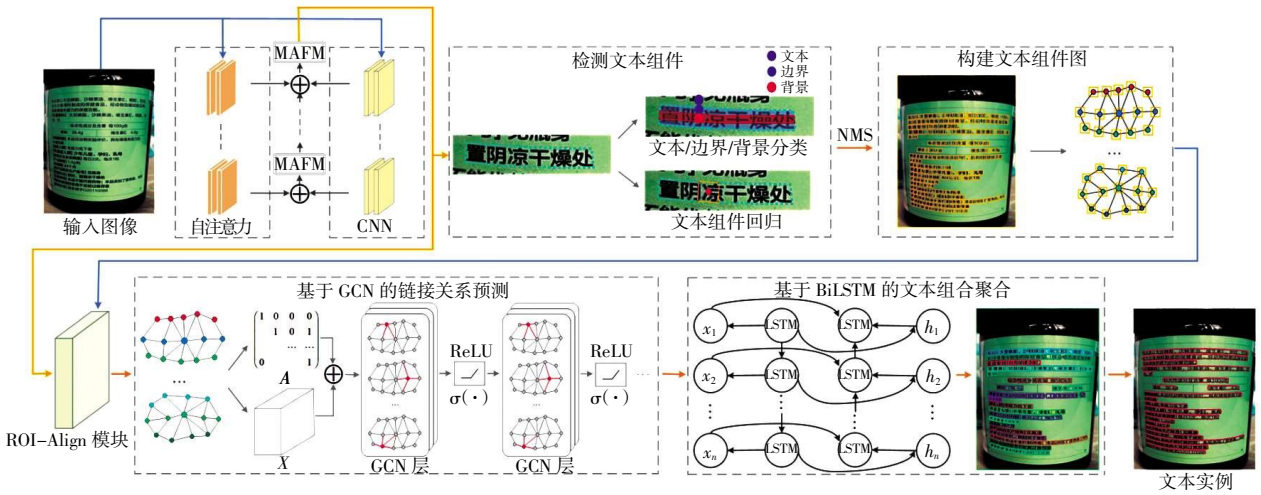


图 1　RPTNet 总体结构

Fig.1　Overall architecture of RPTNet

文本组件检测网络可以细分为特征提取和文本组件检测两部分。链接关系预测网络可以细分为基于 GCN 的链接关系推理和基于 BiLSTM 的文本组件聚合两部分。此外，本文还构建了文本组件图来连接文本组件检测网络和链接关系预测网络，实现 RPTNet 的端到端训练。

1.2　特征提取

CNN 擅长捕捉卷积感受野范围内的局部特征，但对文本实例进行建模时，想要通过 CNN 捕捉全局依赖关系必须增加卷积层深度。理论上，ResNet[24]通过堆叠 residual block 可以实现感受野对文本实例的覆盖，但目前的研究表明其感受野远小于理论值，这对捕捉文本实例中的全局信息造成阻碍。同时，堆叠过深的卷积层也会增加模型参数量，进而引发模型过拟合问题。与 CNN 不同，自注意力擅长提取序列中长距离的全局信息，而 Liu 等提出的 Swin Transformer[25]不仅具备关注全局信息建模的能力，而且可以通过滑动窗口做到跨窗口连接，使特征进行跨窗口交互，解决了卷积结构在捕捉文本实例特征时感受野不足的问题。但是自注意力结构缺少对于局部信息的关注，不能精确地提取文本实例中密集的细节特征。

对于弯曲密集型文本，对其进行特征提取时要求网络同时具备 2 种能力：首先需要能捕捉到文本实例的轮廓特征，这也就要求网络必须具有足够大的感受野；其次还需要关注到弯曲密集型文本的细节信息，这要求网络同时具备对于序列中局部特征的提取能力。通过上述分析，本文构建了一种将 CNN 与自注意力并联的特征提取网络，以学习到文本实例中的多尺度信息，结构设计如图 2 所示。

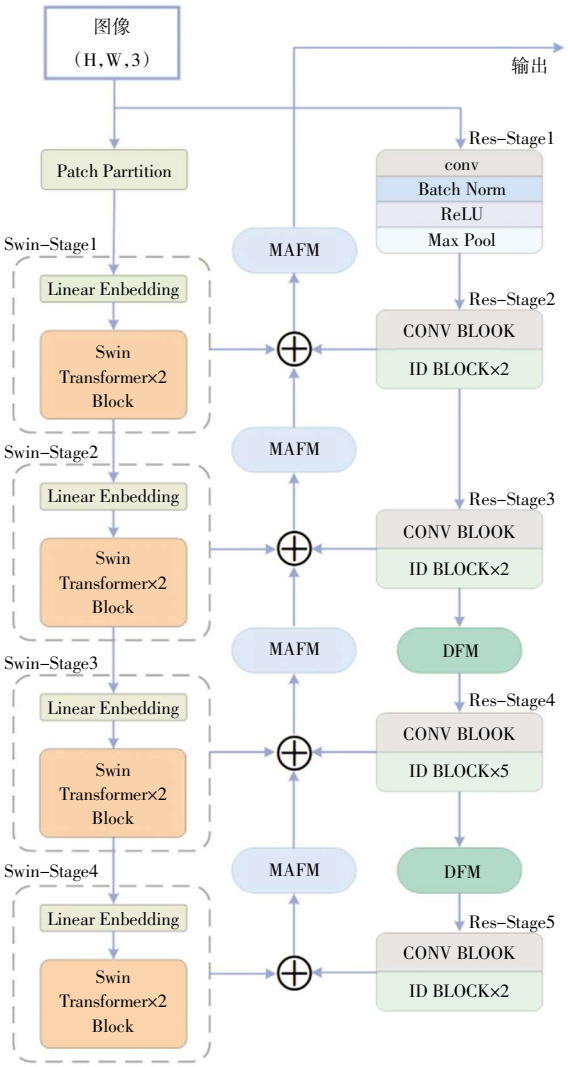
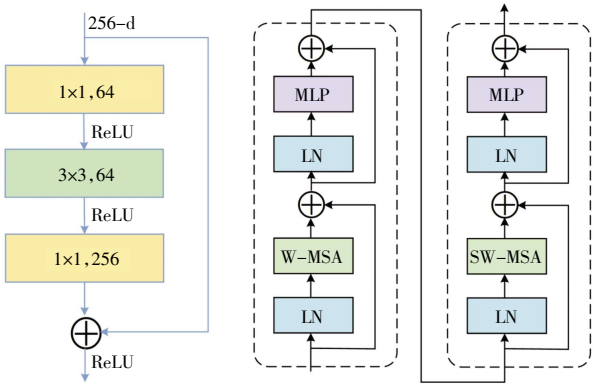


图 2　特征提取框架

Fig.2　Framework for feature extraction

在下采样过程中，输入图像通过并行的 ResNet-

50 和 Swin Transformer 来提取文本的局部和全局特征,ResNet–50 和 Swin Transformer 的 Block 如图 3 所示。同时,在 ResNet–50 的 Res–Stage3、Res–Stage4 之间及 Res–Stage4、Res–Stage5 之间加入空洞特征增强模块(DFM),起到增大特征图感受野、增强文本区域之间关联性的作用。



(a) ResNet–50 的 Block　　(b) Swin Transformer Block

图 3　ResNet–50 和 Swin Transformer 的 Block

Fig.3　Blocks of ResNet–50 and Swin Transformer

在上采样过程中，将 ResNet–50 的 2–5 层结构(Res–Stage2 至 Res–Stage5 层）与 Swin Transformer 的 1–4 层结构(Swin–Stage1 至 Swin–Stage4 层)进行多级特征融合,如图 2 所示。具体来讲,Res–Stage 与 Swin–Stage 相同层之间的特征图维度统一后,经过特征融合模块，输出结果依次为 G4 层、G3 层、G2 层和 G1 层，输出维度大小依次为 16×16×2 048、32×32×1 024、和 128×128×256。

1.2.1　空洞特征增强模块

对于 ResNet–50 网络的 Res–Stage1 至 Res–Stage5 层:浅层特征图尺度大,包含的空间信息较多,但包含的语义信息较少;深层的特征图尺度小,包含的语义信息丰富,但包含的空间信息较少。为了增大特征图的感受野,增加文本区域之间的关联度,本文将空洞特征增强模块(DFM)引入 ResNet–50 中。空洞特征增强模块整体结构借鉴了 Inception[26]的思想。DFM 在 Inception 多分支卷积层结构的基础上,引入了 3 个空洞卷积,从左至右空洞率分别为 1、3、5,从而有效的增加了感受野,如图 4 所示。

1.2.2　基于全局坐标注意力机制的多级特征融合模块

为了最大限度保留文本实例中的全局特征和局部特征,本文提出了一种基于全局坐标注意力机制的多级特征融合模块(MAFM),如图 5 所示,在训练过程中自动融合多级信息以增强网络的表征学习。

通道注意力机制例如 SE、SK 等,虽然能够充分考

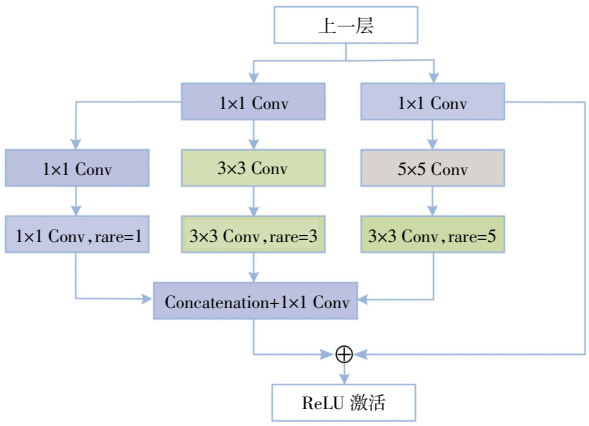


图 4　空洞特征增强模块

Fig.4　Dilated feature enhancement module

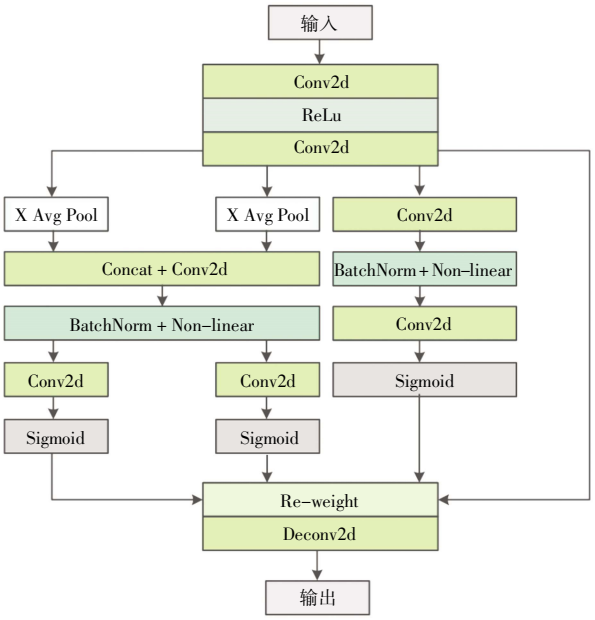


图 5　基于全局坐标注意力机制的多级特征融合模块

Fig.5　Multi–scale feature fusion module based on global coordinate attention mechanism

虑到通道间的联系,实现特征图不同通道间的权重分配,但是没有反映位置间的相关性。基于通道注意力机制的以上不足所提出的坐标注意力机制 CA(coordinate attention）通过引入水平和垂直 2 个方向的注意力,将位置信息嵌入到通道注意力中,以捕捉特征位置间的相关性。但研究发现,坐标注意力机制仅仅考虑到了各个通道特征图中不同位置之间的联系,没有考虑全局信息对于输出特征图的影响。以此为出发点,对坐标注意力机制做出进一步优化,提出了全局坐标注意力机制,其在关注不同位置间依赖关系的同时,还能捕捉到全局信息对于输出特征图的影响。基于全局坐标注意力机制的多级特征融合模块由 2 部

分组成：

（1）第 1 部分和坐标注意力机制相同，在水平与垂直 2 个方向集成特征，生成方向相关特征图。具体来讲，输入特征维度为 $H\times W\times C$，首先在空间维度上分解成 2 个张量 $\boldsymbol{f}^h\in\boldsymbol{R}^{R/r\times H}$ 和 $\boldsymbol{f}^w\in\boldsymbol{R}^{C/r\times W}$，通过 2 个 1×1 卷积操作 F_h 和 F_w，让 $\boldsymbol{f}^h$ 和 $\boldsymbol{f}^w$ 2 个张量的通道数变为一致，如式(1)和式(2)所示：

$$g^h=\sigma(F_h(\boldsymbol{f}^h)) \tag{1}$$

$$g^w=\sigma(F_w(\boldsymbol{f}^w)) \tag{2}$$

式中：σ 表示 Sigmoid 激活函数。在上述工作的基础上，将 g^h 和 g^w 分别作为注意力权重进行分配，得到坐标注意力机制的输出 z_c，如式(3)所示：

$$z_c(i,j)=x_c(i,j)\times g^h{}_c(i)\times g^w{}_c(j) \tag{3}$$

式中：$x_c(i,j)$表示第 c 个通道的高度坐标 i 与宽度坐标 j 位置特征图的数值。

第二部分考虑了特征图自身对于输出的影响。将输入特征图送入到共享 1×1 卷积，之后进行标准化操作，最后采用额外的 1×1 卷积和 Sigmoid激活，输出与输入相同维度的张量 z_n，如式(4)所示：

$$z_n=\sigma(F_1(\sigma(F_1(x_c(i,j))))) \tag{4}$$

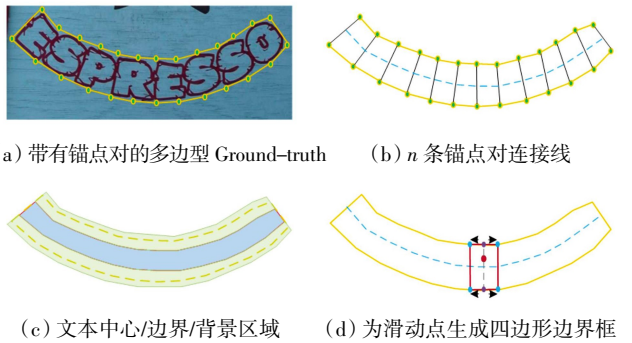
最终输出得到的张量与输入维度相同，如式(5)所示：

$$Z=z_c+z_n \tag{5}$$

通过上述方法，该模块不仅可以关注到位置信息间的相关性，还加权了输入本身对于输出的影响，在提高特征提取效率的同时，也加强了不同特征间的融合。

1.3　文本组件的检测

由于 DenseBox[27]在检测小尺寸、遮挡严重、不规则的物体上具有突出的优势，因此本文通过借鉴 Dense-Box 的思想，从上采样输出的特征图中检测文本组件，如图 6 所示。



（a）带有锚点对的多边型 Ground–truth　（b）n 条锚点对连接线

（c）文本中心/边界/背景区域　（d）为滑动点生成四边形边界框

图 6　文本组件的 Ground–truth 边界框生成

Fig.6　Generation of ground–truth bounding box of text component

首先用一个 Ground–truth 多边形来表示任意形状文本实例的边界，在它的上和下 2 个长边上有组上下对称的锚点对(图 6(a))；然后用条线来连接组上下对称的锚点对(图 6(b))，求出这条线长度的平均值，作为此文本实例的尺度。

在训练过程中，多边形被分配到上采样输出的特征图中。然后，使用特定比例的检测模块从特征图中检测文本实例所包含的文本组件。对原始文本图像中的每一个 Ground–truth 多边形，将其尺度按 0.5 的比例缩小，得到图 6(c)中的蓝色区域，即文本中心区域；同时将 Ground–truth 多边形的尺度按 1.2 的比例扩大，将大于 0.5 且小于 1.2 缩放比例的区域定义为文本边界区域，如图 6(c)中的绿色区域；最后将缩放比例大于 1.2 的区域定义为背景区域。

在本文中，特征图中的每一个像素都与原始文本图像中的一个滑动点相映射。对于任何一个像素来讲，如果与其相映射的那个滑动点位于 Ground–truth 多边形的文本中心区域或边界区域或背景区域，则该像素被贴上“文本”或“边界”或“背景”的标签。对于每一个被贴上“文本”标签的像素，本文使用图 6(d)的方法来生成与其相对应的文本组件 Ground–truth 边界框。具体来讲，用 p 表示 1 个滑动点，l 表示垂直于文本中心线且通过点 p 的线。与 Ground–truth 多边形的 2 个长边分别相交于 p_1 和 p_2 点。然后，本文将 p_1 和 p_2 分别沿着 2 个长边向前和向后移动 d 个像素，最终得到文本组件中的 4 个顶点，由 4 个顶点所围成的区域就是文本组件 Ground–truth 边界框内的区域。在训练过程中，取 $d=2$。检测模块用 1 个 3×3 卷积和 2 个 1×1 卷积来表示，分别用于文本/边界/非文本分类和 Ground–truth 边界框的回归。同时，为了减少计算量和减小误差，只保留得分高于预定阈值的标签为“文本”的像素，本文将阈值设定为 0.85。最后，在特征图中使用标准的 NMS 算法，以 0.6 的交并比(IoU)阈值来删除多余的文本组件。

1.3.1　损失函数

文本组件检测损失由 2 部分组成，分别是文本/边界/非文本的分类损失和 Ground–truth 边界框的回归损失。其中文本/边界/非文本损失使用二元交叉熵计算取值像素的预测和 Ground–truth 标签的损失并取其平均值，Ground–truth 边界框的回归损失使用 计算取样像素中正像素的预测值和 Ground–truth 值 8–d 归一化坐标偏移的损失并取其平均值，方法如式(6)所示：

$$L_{\text{det}}=\frac{1}{N}\sum_k L_{\text{cls}}(c_k,c^*_k)+\frac{1}{N_{\text{ps}}}\sum_l L_{\text{reg}}(t_l,t^*_l) \tag{6}$$

式中：N 为取样像素的数量；c_k 和 c^*_k 分别为第 k 个取

样像素的预测和 Ground-truth 标签;$L_{cls}(c_k,c^*_k)$为二元交叉熵的分类损失;N_{ps} 为取样像素中正像素的数量,其中,$N_{ps}\in N$,t_l 和 t^*_l 分别为第 l 个正取样像素的预测值和 Ground-truth 值 8-d 归一化坐标偏移;$L_{reg}(t_l,t^*_l)$为 Smooth-L_1 的回归损失。

1.4　文本组件图的构建

为了通过图卷积网络预测文本组件的深度相似性,将每一个文本组件用 1 个节点来表示。将所有的节点和节点间的连接线用 1 个集合来表示,记为 $A=\{V,L\}$。其中,$V=\{V_1,V_1,\cdots,V_i,\cdots,V_M\}$为所有节点(文本组件)的集合,$V_i$ 为第 i 个节点。$L=\{l_{i-j}=(V_i,V_j)|V_i,V_j\in V\}$为连接线的集合,$l_{i-j}$ 表示从节点 V_i 指向节点 V_j 的连接线。但是,如果本文考虑所有节点之间的连接线,那么计算量会十分巨大。受 Wang 等[11]工作的启发,只需要建立与每一个节点最相邻近的 k 个节点的连接关系即可。在训练过程中,设置 $k=8$。本文将 2 个节点之间的欧氏距离作为测量距离,以此来衡量 2 个节点间的邻近关系。给定 2 个节点 V_i、V_j,本文通过式(7)来判断 V_i 是否有一条指向 V_j 的连接线。

$$l_{i-j}=\begin{cases}1, & V_i\in \mathrm{KNN}(V_j)\\ 0, & \text{其他}\end{cases} \tag{7}$$

式中:$\mathrm{KNN}(V_j)$表示与 V_j 最相邻的 k 个节点。如果 V_i 属于与 V_j 最相邻的 k 个节点,则 $l_{i-j}=1$,会有连接线从 V_i 指向 V_j;如果 V_i 不属于与 V_j 最相邻的 k 个节点,则 $l_{i-j}=0$,不会有连接线从 V_i 指向 V_j。

通过上述方法,将 1 个文本实例划分为多个文本组件图。每一个文本组件图都由 1 个枢轴节点和 k 个邻居节点组成。首先,本文将 V 中的每一个节点都作为枢轴节点构建文本组件图,这样本文共构建了 M 个文本组件图。但是,为了避免在训练过程中因出现很多相似文本组件图而造成的梯度累积现象,本文以 ξ 为交并比(IoU)阈值来删除多余的文本组件图,如式(8)所示:

$$G_{iou}=\frac{G_m\cap G_n}{G_m\cup G_n}<\xi \tag{8}$$

式中:G_m 和 G_n 为 2 个文本组件图,同属于 1 个文本实例;$G_m\cap G_n$ 为 G_m 和 G_m 各自 k 个邻居节点的交集;$G_m\cup G_n$ 为 G_m 和 G_n 各自 k 个邻居节点的并集。在本文实验中,ξ 设置为 0.8。通过这种方法,本文减少了相似文本组件图的数量,达到了样本平衡的目的。

1.5　基于 GCN 的链接关系预测

为了预测节点间链接的更多可能性,基于图卷积网络,在文本组件图的基础上进一步推理节点间的链接关系。图通常表示为 $g(\boldsymbol{X},\boldsymbol{A})$,图卷积网络的输入包括 2 部分,即特征矩阵 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}$。

为了获得节点特征,本文使用 RoI-Align 提取文本组件的特征。首先,将文本组件图与上采样后输出的特征图一起送入 RoI-Align 层,RoI-Align 层的输出 F_r 被作为节点特征;然后,对节点特征进行归一化的操作。对于任意一个文本组件图 G_p,V_p 为 G_p 中的枢轴节点,x_p 为枢轴节点 V_p 的特征;对文本组件图 G_p 中的每一个节点的节点特征执行减去 x_p 的操作,目的是将中枢节点的特征编码到文本组件图中,可以使链接关系预测网络更加充分地了解到枢轴节点与邻居节点间的连接关系。通过式(9)计算得到 $\boldsymbol{F}_p$,令 $\boldsymbol{X}=\boldsymbol{F}_p$,从而完成对文本组件图中节点特征的归一化。

$$\boldsymbol{F}_p=[\cdots,x_q-x_p,\cdots]^T,x_q|V_q\in G_p \tag{9}$$

式中:x_q 为文本组件图 G_p 中节点 V_q 的节点特征。

使用邻接矩阵 $\boldsymbol{A}_p\in R^{N\times N}$ 表示文本组件图的拓扑结构,N 为节点个数。在获得特征矩阵 $\boldsymbol{X}$ 和邻接矩阵 $\boldsymbol{A}$ 后,本文使用图卷积网络来推理节点之间的链接关系。图卷积层可以表示为:

$$Y=\sigma([\boldsymbol{X}\|\boldsymbol{G}\boldsymbol{X}]\boldsymbol{W}) \tag{10}$$

$$\boldsymbol{G}=\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Lambda}^{\frac{1}{2}} \tag{11}$$

式中:$\boldsymbol{X}\in\boldsymbol{R}^{N\times d_{in}}$;$\boldsymbol{Y}\in\boldsymbol{R}^{N\times d_{out}}$;$d_i$、$d_o$ 分别为输入、输出节点特征的维度;N 为节点个数;$\boldsymbol{G}$ 为大小为 $N\times N$ 的聚合矩阵,每一行的和为 1;运算符‖表示沿着特征维度拼接矩阵;$\boldsymbol{W}$ 为图卷积网络的权重矩阵;$\sigma(\cdot)$为 ReLU 激活函数;$\boldsymbol{A}$ 为文本组件图的邻接矩阵;$\boldsymbol{\Lambda}$ 为对角矩阵,其中 $\Lambda_{ii}=\Sigma_j\Lambda_{ij}$。

1.6　文本组件聚合

所有节点经过 4 个图卷积层的推理预测后,通过双向长短时记忆网络(BiLSTM)动态地对各节点的特征信息进行聚合。图卷积层的输出表示为 $H=[\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_i,\cdots,\boldsymbol{h}_n]$,其中 $\boldsymbol{h}_i$ 表示第 i 个节点经过图卷积层后输出的隐藏向量。在本文中,BiLSTM 中细胞单元的输入为节点的隐藏向量,第 j 个细胞单元的运算过程如下:

$$c_j=\tanh(W_c[a_{j-1},\boldsymbol{h}_j]+b_c) \tag{12}$$

$$u_j=\sigma(W_u[a_{j-1},\boldsymbol{h}_j]+b_u) \tag{13}$$

$$f_j=\sigma(W_f[a_{j-1},\boldsymbol{h}_j]+b_f) \tag{14}$$

$$o_j=\sigma(W_o[a_{j-1},\boldsymbol{h}_j]+b_o) \tag{15}$$

$$c_j=u_j\odot\tilde{c}_j+f_j\odot\tilde{c}_{j-1} \tag{16}$$

$$a_j=o_j\odot\tanh(c_j) \tag{17}$$

式中:a_j 和 c_j 分别为第 j 个细胞单元的隐藏状态和细胞状态;W 为向量权重;b 为对应的偏置;tanh 函数为双曲正切激活函数;σ 为 Sigmpid 激活函数;⊙表示哈达玛乘积;$\tilde{c}_j$、u_j、f_j、o_j 分别为细胞门、更换门、遗忘门

和输出门。

在 BiLSTM 中，第 j 个细胞单元的向量输出需要考虑到双向信息，可以表示为：

$$y_j = \sigma(W_y[\overrightarrow{a_j}, \overleftarrow{a_j}] + b_y) \tag{18}$$

式中：$\overrightarrow{a_j}$、$\overleftarrow{a_j}$分别为第 j 个细胞单元的前向隐藏状态和后向隐藏状态。最后，本文使用 MinPath 算法，根据聚类结果对每组文本组件进行排序，从而生成文本实例的边界，得到一个完整的文本实例。

2　实验结果与分析

本文在 Ubuntu 16.04 操作系统下，通过 PyTorch 1.2.0 框架实现了 RPTNet，并在 2 块 NVIDIA GeForce GTX 1080Ti 的 GPU 上进行了实验。

2.1　实验数据

为了评估本文提出的 RPTNet 的性能，本文建立了一个由商品外包装图像组成的文本检测数据集 CPTD1500。其中，CPTD1500 数据集的标注方式与基准数据集 CTW-1500[22]相似。与 CTW-1500 数据集不同的是，CPTD1500 数据集采用弯曲或密集型的商品包装上的文本实例作为训练集与测试集。本文建立此数据集的目的在于评估 RPTNet 在检测弯曲、密集型的商品外包装文本实例时的性能。为了更好地评估 RPTNet 的性能，本文分别在 CPTD1500 数据集和 2 个场景文本检测基准数据集 CTW-1500 和 Total-Text[23]上进行消融实验和对比实验。

CPTD1500 数据集由 1 000 张训练图像和 500 张测试图像组成，均为商品外包装文本图像。图像中的文本实例以中文和英文为主。同时还包含少量日文和韩文，以验证网络检测不同语言的泛化能力。在数据集统计过程中，将实物中文本实例弯曲弧度大于 5 度的归类为弯曲文本，否则归类为四边形文本；将实物中文本实例间的间隙大于 1 mm 的 2 个文本实例归类为密集文本，否则归类为稀疏文本。对 CPTD1500 数据集中的文本实例类型进行统计，统计结果如表 1 所示。每个文本实例均通过 14 点多边形进行标注。标注示例如图 7 所示。数据集采用基于 PASCAL VOC 文本评测准则。

CTW-1500 数据集由 1 000 张训练图像和 500 张测试图像组成。每张图像至少有一个弯曲的文本实例。该数据集存在很多的艺术体、模糊小文本和类似文本等干扰因素。图像中的文本实例以英文为主，包含少数中文。每个文本实例均采用 14 点多边形进行标注。

Total-Text 数据集由 1 255 张训练图像和 300 张测试图像组成。该数据集包含许多曲线和多方向文本实例。每一个文本实例用多边形标注框标注在字符级别上。

表 1　CPTD1500 数据集训练集和测试集文本实例数量

Tab.1　Number of text instances in training set and test set of CPTD1500

文本实例类型	文本实例数量		
	训练集	测试集	合计
四边形密集文本	2 492	1 124	3 616
四边形稀疏文本	1 222	572	1 794
弯曲密集文本	10 362	4 932	15 294
弯曲稀疏文本	4 684	2 217	6 901

图 7　CPTD1500 标注示例

Fig.7　Annotation examples of CPTD1500

在对比实验中，模型在 SynthText 数据集上预训练 2 个 epoch，消融实验部分没有设置预训练步骤。

在消融实验和对比实验中，分别在本文构建的数据集和基准数据集上做 700 个 epoch 的微调训练，批次设置为 4。同时，使用带动量的 SGD 优化器来训练模型，动量设置为 0.9，权重衰减设置为 0.000 5，初始学习率设置为 0.01，学习率衰减使用 Poly 策略。

为了提高训练后模型的泛化能力，本文也采用随机旋转[-10°~10°]、随机裁剪、随机翻转来对训练图像做数据扩充，最后将图像调整成 1 024 pixel×1 024 pixel 大小送入网络训练。

2.2　评价指标

本文算法的性能由精确率 P、召回率 R、F 值和检测速率 v 共 4 个指标来衡量。其中精确率 P、召回率 R 和 F 值的计算过程分别如式(19)—式(21)所示：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

$$F = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{21}$$

式中：TP、FP 和 FN 分别为真阳性、假阳性和假阴性文本实例的数量；精确率 P 和召回率 R 分别反映了模型识别负样本和正样本的能力；F 值为由精确率和召回率的平均值计算出的总体评价分数。

2.3　消融实验

为了验证文本组件检测网络和链接关系预测网络的有效性，在 CPTD1500 数据集和 Total–Text 数据集上分别进行了消融实验。实验结果如表 2 所示。

表 2　CPTD1500 数据集和 Total–Text 数据集上的消融实验结果

Tab.2　Results of ablation experiments on CPTD1500 and Total–Text

骨干网络	DFM	MAFM	GCN+BiLSTM	CPTD1500 数据集			Total–Text 数据集		
				P/%	*R*/%	*F*/%	*P*/%	*R*/%	*F*/%
ResNet	√	√	√	87.05	84.04	85.52	87.85	83.88	85.82
Swin	√	√	√	87.14	83.49	85.28	87.71	83.74	85.68
ResNet+Swin	×	×	×	83.77	82.91	83.34	85.24	80.71	82.91
ResNet+Swin	√	×	×	84.03	83.32	83.67	85.87	81.19	83.46
ResNet+Swin	√	√	×	84.71	82.96	83.83	86.28	81.04	83.58
ResNet+Swin	×	×	√	87.69	83.47	85.53	88.11	84.23	86.13
ResNet+Swin	×	√	√	88.45	84.13	86.41	88.87	84.79	86.78
ResNet+Swin	√	×	√	88.13	**84.56**	86.31	88.48	85.03	86.72
ResNet+Swin(ours)	√	√	√	**88.94**	84.53	**86.67**	89.05	**85.14**	**87.05**

由表 2 可以看出，在同样加入 DFM 模块、MAFM 模块、GCN 链接关系预测网络和 BiLSTM 聚合模块的前提下，选取 ResNet–50 和 Swin Transformer 的多级特征融合网络作为骨干网络比单独选取 ResNet–50 网络作为骨干网络时拥有更高的精确率、召回率和 *F* 值，三者在 CPTD1500 数据集上分别有 1.89%、0.49%、1.15%的提升，在 Total–Text 数据集上分别有 1.20%、1.26%、1.23%的提升。然后，在选取 ResNet–50 和 Swin Transformer 的多级特征融合网络作为骨干网络的前提下，分别加入 DFM 和 MAFM 模块后，精确率和 *F* 值较之前分别有不同程度的提升。最后，在骨干网络和各个模块不变的前提下，添加 GCN 关系预测网络和 BiLSTM 聚合模块后，相较于添加前，在精确率、召回率和 *F* 值上均有较大幅度的提升。由此说明，在 CPTD1500 数据集和 Total–Text 数据集上的测试结果证实了上述各模块的重要性。CPTD1500 数据集下不同模型的损失曲线如图 8 所示。

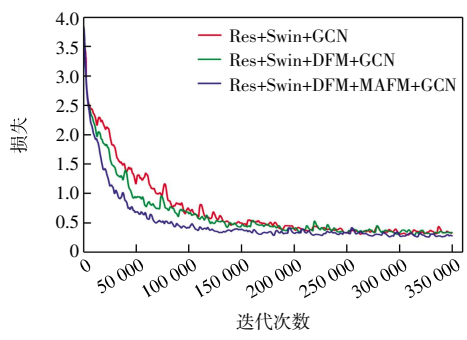


图 8　CPTD1500 数据集下不同模型的损失曲线

Fig.8　Loss curves for different models on CPTD1500

由图 8 可知，在 CPTD1500 数据集中，3 个模型在经过 150 000 次迭代之后，损失值逐渐变得平稳。通过比较 3 个模型的损失曲线，在分别加入 DFM 模块和 MAFM 模块后，模型具有更快的收敛速度，同时在收敛之后损失值更加稳定，曲线更加平滑。由此可以得出，DFM 模块和 MAFM 模块的加入对于 RPTNet 具有重要作用。

2.4　对比实验

CTW–1500 和 Total–Text 是 2 个典型的弯曲和多方向的文本数据集，分别基于行级和字符级进行了注释。在测试过程中，由于图像的长宽比变化较大，首先把图像的短边统一为 512，然后根据短边的变化比例，对长边做出相同比例的调整，以保证图像的长宽比不变。RPTNet 在 CTW–1500 数据集和 CPTD1500 数据集上的可视化检测结果如图 9 和图 10 所示。

图 9　RPTNet 在 CTW–1500 数据集中的可视化结果

Fig.9　Visualisation results of RPTNet on CTW–1500

图 10　RPTNet 在 CPTD1500 数据集中的可视化结果

Fig.10　Visualisation results of RPTNet on CPTD1500

由图 9 可以看出,RPTNet 在检测弯曲度很大的文本时具有较好的鲁棒性,即使图中有个别的文本实例存在多个弯曲方向,本文所提出的模型也能够对其完成精准的检测;CTW-1500 数据集中包含大量的复杂场景图像,RPTNet 能够准确区分出图像中的背景和文本,取得了较好的检测效果。

由图 10 可以看出,RPTNet 可以有效的处理任意形状的密集型文本,虽然在商品说明区域中的文本实例存在弯曲、密集、字体小等检测难点,但通过二值化分类图可以看出,本文所提出的模型可以准确区分出相邻文本实例,并未出现文本粘连的问题,由此可以说明,RPTNet 在弯曲密集型小文本实例的检测中有较好的鲁棒性;CPTD1500 数据集中包括多种语言、符号及数字表示,RPTNet 均能够正确的提取出相应的文本,有着较好的泛化性。

不同数据集上的对比实验结果如表 3 所示。由表 3 可知,对于 CTW-1500 数据集而言,RPTNet 的精确率、召回率、F 值和检测速率分别达到 87.9%、84.1%、86.0% 和 12.7 fps,均优于最新的方法。在 Total-Text 数据集中,RPTNet 的召回率和 F 值分别达到 86.1%和 88.1%,取得了最优的结果,同时 RPTNet 在精确率比 ABPNet 低 0.6%的前提下,召回率和 F 值分别高出 ABPNet 0.9%和 0.2%,有着比 ABPNet 更均衡的综合性能指标。在CTW-1500 数据集和 Total-Text 数据集上的检测结果验证了 RPTNet 在处理行级和字符级弯曲文本时的优势。

表 3　不同数据集的对比实验结果

Tab.3　Results of comparison experiments on different datasets

方法	CTW-1500				Total-Text				PTD1500			
	P/%	R/%	F/%	v/fps	P/%	R/%	F/%	v/fps	P/%	R/%	F/%	v/fps
LOMO[10]	85.7	76.5	80.8	4.4	87.6	79.3	83.3	4.4	88.8	78.1	83.1	4.4
SegLink++[18]	82.8	79.8	81.3	7.1	82.1	80.9	81.5	—	85.1	81.7	83.4	6.9
TextField[15]	83.0	79.8	81.4	6.0	81.2	79.9	80.6	5.9	83.2	81.7	82.4	4.7
MSR[28]	85.0	78.3	81.5	4.3	73.0	85.2	78.6	4.3	85.3	79.8	82.5	4.3
PSENet[16]	84.1	79.0	82.2	3.9	84.0	78.0	80.9	3.9	85.8	80.7	83.2	4.2
CRAFT[17]	86.0	81.1	83.5	—	87.6	79.9	83.6	—	88.1	82.1	85.0	—
ABPNet[29]	87.7	80.6	84.0	12.0	90.7	85.2	87.9	10.5	88.7	81.9	85.2	10.9
ABCNet v2[30]	85.6	83.8	84.7	10.0	90.2	84.1	87.0	10.0	86.8	84.7	85.7	10.0
PCR[31]	87.2	82.3	84.7	11.8	88.5	82.0	85.2	13.4	89.2	83.2	86.1	11.2
FCE[12]	87.6	83.4	85.5	—	89.3	82.5	85.8	—	90.1	84.1	87.0	—
本文方法	87.9	84.1	86.0	12.7	90.1	86.1	88.1	12.1	89.8	85.4	87.5	11.2

在 CPTD1500 数据集上的测试结果验证了RPTNet 在检测弯曲密集型商品外包装文本的有效性。该方法在召回率、F 值和检测速率上均取得了最优的结果。其中 F 值高达 87.5%,相比于针对曲面密集型文本检测任务的 SegLink++提高了 4.1%,相比于精确率最高的 FCE 提高了 0.5%, 从而验证了本文所提出的 RPTNet 相较于其他最新方法,在弯曲密集型商品外包装文本检测任务中有着较大的竞争优势。

3　结　论

本文提出一种片段级文本检测方法(RPTNet)来检测弯曲密集型商品包装文本。通过 MAFM 模块和DFM 的结合进行局部特征和全局特征的融合,以更好地检测文本组件。基于 GCN 和 BiLSTM 的链接关系预测网络可以有效推理文本组件间链接的更多可能性。在 2

个公开数据集和本文构建的 CPTD1500 数据集上的测试结果表明,RPTNet 召回率为 85.4%和 F 值为87.5%,与最新的方法相比都有一定的提升。未来的研究中,将进一步优化文本检测算法,对模型的轻量化展开深入研究。同时,希望将 RPTNet 与文本识别算法相结合,设计一种端到端的针对任意形状文本的文本识别网络。

参考文献:

[1] LIAO M H, SHI B G, BAI X, 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, USA:ACM, 2017: 4161-4167.

[2] ZHOU X Y, YAO C, WEN H, et al. EAST: An efficient and accurate scene text detector[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA:IEEE, 2017: 2642-2651.

[3] MA J Q, SHAO W Y, YE H, et al. Arbitrary-oriented scene text detection via rotation proposals[J]. IEEE Transactions on Multimedia, 2018, 20(11): 3111-3122.

[4] HE W H, ZHANG X Y, YIN F, et al. Deep direct regression for multi-oriented scene text detection[C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy:IEEE, 2017: 745-753.

[5] REN S Q, HE K M, 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): 1137-1149.

[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector[C]//European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

[7] LIAO M H, SHI B G, BAI X. TextBoxes++: A single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2018, 27(8): 3676-3690.

[8] RAISI Z, NAIEL M A, YOUNES G, et al. Transformer-based text detection in the wild[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Nashville, USA:IEEE, 2021: 3156-3165.

[9] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Computer Vision – ECCV 2020: 16th European Conference. Glasgow, UK:ACM, 2020: 213-229.

[10] ZHANG C Q, LIANG B R, HUANG Z M, et al. Look more than once: An accurate detector for text of arbitrary shapes[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA:IEEE, 2019: 10544-10553.

[11] HE M H, LIAO M H, YANG Z B, et al. MOST: A multi-oriented scene text detector with localization refinement[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA:IEEE, 2021: 8809-8818.

[12] ZHU Y Q, CHEN J Y, LIANG L Y, et al. Fourier contour embedding for arbitrary-shaped text detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA:IEEE, 2021: 3122-3130.

[13] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence. [s.n.]:IEEE, 2017: 640-651.

[14] ZHANG Z, ZHANG C Q, SHEN W, et al. Multi-oriented text detection with fully convolutional networks [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA:IEEE, 2016: 4159-4167.

[15] XU Y C, WANG Y K, ZHOU W, et al. TextField: Learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society, 2019, 28(11): 5566-5579.

[16] WANG W H, XIE E Z, LI X, et al. Shape robust text detec-tion with progressive scale expansion network [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA. IEEE, 2019: 9328-9337.

[17] BAEK Y, LEE B, HAN D, et al. Character region awareness for text detection[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA:IEEE, 2019: 9357-9366.

[18] TANG J, YANG Z B, WANG Y P, et al. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping[J]. Pattern Recognition, 2019, 96: 106954.

[19] LIAO M H, WAN Z Y, YAO C, et al. Real-time scene text detection with differentiable binarization[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11474-11481.

[20] WANG Z D, ZHENG L, LI Y L, et al. Linkage based face clustering via graph convolution network[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA:IEEE, 2019: 1117-1125.

[21] KIPERWASSER E, GOLDBERG Y. Simple and accurate dependency parsing using bidirectional LSTM feature representations[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 313-327.

[22] LIU Y L, JIN L W, ZHANG S T, et al. Curved scene text detection via transverse and longitudinal sequence connection[J]. Pattern Recognition, 2019, 90(C): 337-345.

[23] CH′NG C K, CHAN C S. Total-text: A comprehensive da-

(下转第 74 页)

2007, 36(6): 1008–1012(in Chinese).

[16] SHI J, YANG F, XU W, et al. High–resolution temperature sensor based on intracavity sensing of fiber ring laser [J]. Journal of Lightwave Technology, 2020, 38(7): 2010–2014.

[17] 吴映. 光纤法–珀传感器解调算法研究与实现[D]. 武汉: 华中科技大学, 2021.
WU Y. Research and implementation of the demodulation algorithm for fiber Fabry–Perot sensors[D]. Wuhan: Huazhong University of Science and Technology, 2021(in Chinese).

[18] 孟建兴. 带式输送机液压张紧装置设计和控制研究[J]. 机械管理开发, 2021, 36(3): 54–56.
MENG J X. Research on design and control of hydraulic tensioning mechanism of belt conveyor[J]. Mechanical Management and Development, 2021, 36(3): 54–56(in Chinese).

[19] 周宁. 强度调制型光纤压力传感器设计与研究[D]. 太原: 中北大学, 2020.
ZHOU N. Design and research of intensity modulation fiber optic pressure sensor[D]. Taiyuan: North University of China, 2020(in Chinese).

[20] 白浪, 郑刚, 张雄星, 等. 调频连续波光纤压力传感器及其测量特性分析[J]. 光学学报, 2021, 41(3): 0328002.
BAI L, ZHENG G, ZHANG X X, et al. Optical fiber pressure sensor based on frequency–modulated continuous–wave and analysis of its measurement characteristic[J]. Acta Optica Sinica, 2021, 41(3): 0328002(in Chinese).

本文引文格式:

苗长云, 张豫飞. 基于光纤传感的带式输送机张力检测技术研究[J]. 天津工业大学学报,2024, 43(4): 67–74.
MIAO C Y, ZHANG Y F. Research on tension detection of belt conveyor based on fiber optic sensing [J]. Journal of Tiangong University, 2024, 43(4): 67–74(in Chinese).

(上接第 59 页)

taset for scene text detection and recognition[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto, Japan:IEEE, 2017: 935–942.

[24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, USA:IEEE, 2016: 770–778.

[25] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical vision transformer using shifted windows[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada:IEEE, 2021: 9992–10002.

[26] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA:IEEE, 2015: 1–9.

[27] HUANG L C, YANG Y, DENG Y F, et al. DenseBox: Unifying landmark localization with end to end object detection [EB/OL]. [2015–09–15]. http://arxiv.org/abs/1509. 04874

[28] XUE C H, LU S J, ZHANG W. MSR: Multi–scale shape regression for scene text detection[EB/OL]. [2019–01–19]. http://arxiv.org/abs/1901.02596

[29] ZHANG S X, ZHU X B, YANG C, et al. Adaptive boundary proposal network for arbitrary shape text detection [C]//2021 IEEE/CVF International Conference on Com–puter Vision (ICCV). Montreal, Canada:IEEE, 2021: 1285–1294.

[30] LIU Y L, SHEN C H, JIN L W, et al. ABCNet v2: Adaptive bezier–curve network for real–time end–to–end text spotting[EB/OL]. [2021–05–21]. http://arxiv.org/abs/2105.03620

[31] DAI P W, ZHANG S Y, ZHANG H, et al. Progressive contour regression for arbitrary–shape scene text detection[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA:IEEE, 2021: 7389–7398.

本文引文格式:

耿磊,李嘉琛,刘彦北,等. 基于链接关系预测的弯曲密集型商品文本检测[J]. 天津工业大学学报,2024, 43(4): 50–59,74.
GENG L, LI J H, LIU Y B, et al. Text detection of curved and dense products based on link relationship prediction[J]. Journal of Tiangong University, 2024, 43(4): 50–60(in Chinese).